International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 3, March-April 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Survey on Key Phrase Extraction using Machine Learning Approaches

Preeti Sondhi[1], Aakib Jabbar[2]

[1]Assistant Professor, [2]M.Tech Scholar,
[1,2]Universal Group of Institutions, Lalru, Punjab, India

ABSTRACT

The automated keyword extraction task is to define a collection of representative terms for the text. Extracting keywords defines a small collection of terms, key phrases and keywords that define the document's context. Keyword search allows large document collections to be searched effectively. To allocate suitable key-phrases to new documents, text categorization techniques can be applied. A predefined collection of key-phrases from which all key-phrases for new documents are selected is given in the training documents. The training data for each key-phrase describes a collection of documents associated with it. Standard machine learning techniques are used for each key-phrase to construct a "classifier" from the training materials, using those relevant to it as positive examples and the rest as negative examples. Provided a new text, it is processed by the classifier of each key-phrase.

KEYWORDS*: Text mining, Keyphrases, clustering, supervised learning, unsupervised learning etc*

How to cite this paper*:* Preeti Sondhi | Aakib Jabbar "Survey on Key Phrase Extraction using Machine Learning Approaches" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-3, April 2021, pp.485-489, URL: www.ijtsrd.com/papers/ijtsrd39890.pdf

IJTSRD39890

INTRODUCTION

The pervasion of enormous quantities of data through the World Wide Web (WWW) has created an increasing imperative for the advancement of information discovery, access, and sharing techniques. The key phrases help readers understand, organise, view, and distribute a document's data easily. Phrases consisting of one or more important words are main phrases. To promote knowledge search on the internet, key phrases can be inserted into the search results as topic metadata [1]. An indicative description or document metadata may serve as a list of key phrases associated with a document, which helps readers search for relevant information. Key phrases are intended to serve different purposes. For instance, [1] when they are printed on the first page of a journal article, the purpose is to summarise. They help the reader to quickly decide if the article in question is worth reading in detail. [2] Once they are added to a journal 's cumulative index, indexing is the purpose.

Automatic Key phrase Extraction is used to extract key terms or phases from text in order to better identify the main content of the document without any human interference (could be of any sort, i.e. newspaper clipping, site, conversational). Manual extraction is time-consuming, repetitive and costly, as human resources are needed[2]. Focus on 4 intrinsic properties: coverage of the subject, significance of the subject covered, phrases, facts.

In many fields such as education, biomedical, science, and many more where data in the form of text is present, key phrase extraction is helpful.

We need to quickly go through vast quantities of textual material nowadays to find documents relevant to our interests and this paper space is growing at an overwhelming rate on a daily basis. Nowadays, storing several million web pages and hundreds of thousands of text files is popular. It can be made easier to analyze such large amounts of data if we can have a subset of terms (keywords) that can provide us with the document's main features, concept, theme, etc. Appropriate keywords can serve as a very succinct overview of a document and assist us to organize and retrieve documents quickly based on their material. In scholarly journals, keywords are used to give the reader an idea about the content of the paper. They are helpful for readers to recognize and maintain the key points about a specific section of their minds in a textbook[3][4]. As keywords represent a text's main theme, they can be used for text clustering as a measure of similarity.

Keyword is a phrase that defines the subject, or an aspect of the subject, discussed in a paper succinctly and accurately. Key terms can refer to both single words (keywords) and phrases (key phrases). In their book Foundations of Mathematical Natural Language Processing, Manning and Schutze have the following to say about phrases: Words do not appear in just any old order. Languages have word order constraints. But it is also the case that, like beads on a necklace, the words in a sentence are not only strung together as a series of parts of expression. Words, instead, are grouped into phrases, groupings of words that are clumped together as a unit. One simple principle is that some groupings of words[5] serve as constituents.

Humans tend to choose key sentences over keywords. The size of the main phrase depends on the application intended for it. It is possible to allocate keywords either manually or automatically, but the former technique is very time-consuming and costly. Thus, an automated process that extracts keywords from documents is required.

How does keyword extraction work

The task of identifying appropriate terms and phrases in unstructured text is simplified. This involves emails, social media messages, chat conversations, and all other forms of data that are not structured in any predefined way.

Keyword extraction will help you simplify some of your workflows, such as marking incoming survey responses or answering urgent customer questions, saving you a great deal of time. It also provides you with actionable information that can be used to make smart decisions about companies. But the best thing about keyword extraction models is that they're simple to set up and implement.

To extract automatic keywords, there are numerous techniques you can use. From simple statistical approaches that detect keywords through word frequency counting, to more advanced machine learning approaches that allow you to construct more complex models that can learn from previous examples.

APPROACHES FOR KEYWORD EXTRACTION

Broadly speaking there can be different approaches for automatic keyword/key phrase extraction, each having its own pros and cons, but there are four major methods.

Rule Based Linguistic approaches:

In general, these methods are rule-based and derive from linguistic knowledge / features. These methods can be more specific, but they are computer-intensive and, in addition to language expertise, require domain knowledge. These techniques use the linguistic features of the expressions, especially sentences and articles. Lexical analysis, syntactic discourse analysis analysis and so on are part of the linguistic approach.

Statistical approaches: In general, these approaches are based on a linguistic corpus and a corpus-derived statistical function. The most important advantage of them is that they are independent of the language in which they are implemented and can thus be used in many languages using the same technique[6]. Compared to linguistic ones, these approaches do not provide reliable results, but the availability of a large number of datasets has provided good results.

Machine Learning approaches: Machine Learning techniques typically use methods of supervised learning. Keywords are extracted from training documents in these methods to learn a model, the model is tested via a testing module. It is used to find keywords from new files after a satisfactory model is built. Naïve Bayes, Support Vector Machine, etc. include this approach. Supervised learning methods, however, require a tagged corpus of documents that is hard to create. In the absence of such an entity, unsupervised and semi-supervised methods of learning are used.

Domain specific approaches: Different methods can be applied to a particular domain corpus, using the domain-related backend information (such as ontology) and the inherent structure of that specific corpus to define and extract keywords.

MODES OF KEYWORD AND KEYPHRASE GENERATION

There are two fundamental approaches for automatic key phrase generation:

Keyphrase/keyword assignment: The set of potential key phrases is constrained by a predefined vocabulary of terms in this method. The aim is to find a small collection of words, independent of the domain to which it belongs, that describes an individual text.

Simplicity and continuity are the benefits. The same key phrases can describe similar documents and the use of a controlled vocabulary ensures the appropriate scope of document coverage.

The disadvantages of this approach are: the development and preservation of controlled vocabulary is costly and therefore not always usable. If they are not in the vocabulary, possible main terms that appear in the text are overlooked.

Keyword/key phrase extraction:

The most important words present in the document are chosen by this method and the collection is not based on any vocabulary and extracted words are present in the document itself[7][8].

The benefits are: there is no need to develop and retain vocabulary, and it is possible to pick relevant keywords and key phrases that appear in the text.

The disadvantages of this approach are: lack of consistency; since various key phrase can represent similar documents and it is difficult to choose the most appropriate key phrases; i.e. the required scope of coverage of the document is not ensured.

APPLICATION AREAS

- ➢ Text highlighting
- ➢ Text summarization
- ➢ Information search
- ➢ Text categorization
- ➢ Text clustering
- ➢ Automatic indexing
- ➢ Ontology learning

LITERATURE REVIEW

DhruvaSahrawat**(2019)** In this paper, we formulate the extraction of key phrases from scholarly papers as a sequence labelling task solved using a BiLSTM-CRF where the terms are represented using deep contextualised embedding in the input text. The proposed architecture is tested on three separate benchmark datasets (Inspec, SemEval 2010, SemEval 2017) using both contextualised and fixed word embedding models and contrasted with current common unsupervised and supervised techniques.

Lee Xiong et al. (2019) In real-world scenarios where documents are from different domains and have variant content quality, this paper studies key phrase extraction. OpenKP, a large-scale open domain key phrase extraction dataset with almost one hundred thousand web documents and expert key phrase annotations, is being curated and published.

EiriniPapagiannopoulou (2019) Key phrase extraction is a processing task of textual information concerned with the automated extraction from a document of representative and characteristic phrases that convey all the key aspects of its content. Key phrases constitute a concise conceptual description of a document that is very useful for semantic indexing, faceted search, document clustering and classification in digital information management systems.

KamilBennani-Smires et al. (2018) The process of automatically selecting a small collection of phrases that better represent a given free text document is key phrase extraction. Supervised extraction of key phrases involves vast quantities of labelled training data and generalises very poorly outside of the training data domain. At the same time, unsupervised systems have low accuracy and often do not generalise well, as the input document is expected to belong to a larger corpus that is also given as input.

ImtiazHossain Emu (2017) Key phrases are set of words that reflect the main topic of interest of a document. It plays vital roles in document summarization, text mining, and retrieval of web contents. As it is closely related to a document, it reflects the contents of the document and acts as indices for a given document. Extracting the ideal keyphrases is important to understand the main contents of the document. In this work, we present a keyphrase extraction method that efficiently finds the keywords from English documents.

Allamanis et al. (2016) There is a distinction between the performance of the goal during planning and the metrics during assessment. By integrating new algorithms into preparation, a few studies have attempted to remove this difference. Key phrase offers highly summative data that can be used to recognise, organise and retrieve text information effectively. While several workable solutions for automated key phrase extraction have been introduced in previous studies, the content to be summarised was usually split into several pieces of text, then the most important ones were identified and selected.

Gu et al., (2016) Review proposed to allow our model to locate important parts based on positional details. Therefore, our model can produce key phrases based on an understanding of the text, regardless of the presence or absence of key phrases in the document; at the same time, essential in-text knowledge is not lost.

Aditisharan et al (2015) Present a rundown of the different techniques available in text mining for keyword and key phrase extraction. In addition to being useful for many other uses, keywords and key phrases are very useful for easily and efficiently searching for vast quantities of textual content over the internet. Keywords and key phrases are a selection of a text's representative words that provide high-level content specification for interested readers. These are particularly used in the field of computer science in information retrieval and natural language processing and can be used for index construction, query refinement, text description, author assistance, etc.

RuiMeng et al (2015) Highly summative data is provided by the proposed Key phrase analysis, which can be used effectively to understand, organise and retrieve text material. While several workable solutions for automated key phrase extraction have been introduced in previous studies, the content to be summarised was usually split into several pieces of text, then the most important ones were identified and selected. Such strategies were unable to recognise key phrases that did not appear in the text, nor were they able to capture the real semantic meaning behind the text.

Cho et al. (2014) The suggested key phrase analysis consists of single or multiple word phrases in a document that defines the main points of the document. These key sentences help readers obtain an overview of the document. In this paper, we proposed a system that uses Recurrent Neural Network (RNN) Long Short-Term Memory (LSTM) to automatically detect key phrases from a text. To compare the efficiency of our proposed LSTM solution, we also implemented the Multilayer Perceptron (MLP) network.

Sutskever et al. (2014) recommended analysis to solve problems with the translation. Because it offers a powerful tool for end-to-end modelling of variable-length sequences, it suits many natural language processing tasks and can achieve great results quickly.

Bahdanau et al., (2014) Is a soft alignment approach that helps the model to locate the appropriate input components automatically. Several studies found ways of copying some parts of the material from the source text and pasting them into the target text in order to make use of the important information in the source text.

Abilhoa and de Castro (2014) Propose a method for keyword extraction to identify tweets (micro blogs) as graphs and apply central steps to define keywords in the question. In the pre-processing level, they build a technique called Twitter Keyword Graph where they use the method of eliminating tokenization, stemming, and stop words. Using measurements of graph centrality-closeness and eccentricity, keywords are extracted from the graph by cascade. The algorithm is evaluated on a single text from the literature and is contrasted with the TF-IDF method and the KEA algorithm. Finally, on five sets of increasingly large tweets, the algorithm is tested. A robust proposal to extract keywords from texts, particularly from short texts such as micro blogs, proved to be the calculation time needed to run the algorithms.

Zhou et al. (2013) Investigate weighted, complex network-based keyword extraction that combines network systemic exploration and linguistic awareness. The emphasis is on building lexical network including fair node selection, clear definition of word-to-word relationships, simple weighted network and TF-IDF. Reasonable collection of terms from texts as lexical nodes from a linguistic perspective, proper description of the relationship between words, and development of node attributes attempts to more accurately represent texts as lexical networks. Jaccard coefficient is used in the process of network construction to represent the associations or relationships of two terms, rather than the normal co-occurrence criterion. Importance of each node to become a candidate for the keyword is determined with centrality of closeness. Compound measure is used which takes into account node attributes (word length and IDF). Method is contrasted with three competitive approaches to the baseline: binary network, simple weighted network, and TF-IDF method. Experiments for Chinese suggest that, over the classic TF-IDF process, the lexical network built by this approach achieves superior impact on precision, recall and F-value.

Boudin (2013) Compares various core measures for graph-based key phrase extraction. Experiments on standard

English and French data sets show that simple degree centrality produces results comparable to the widely used Text Rank algorithm, and that the best results on short documents are obtained through centrality of closeness. Undirected and weighted co-occurrence networks are syntactically constructed (only nouns and adjectives) from parsed and lemmatized text using the co-occurrence window. The degree, closeness, between's and own vector centrality are compared to the Page Rank ad proposed by Mihalcea (2004) as a guide. The degree's centrality achieves the same output as the far more complex Text Rank. Closeness centrality outperforms Text Rank (abstracts from scientific papers) on short articles.

Yang et al. (2013) centered on keyword extraction based on the entropy difference between intrinsic and extrinsic modes, which refers to the fact that the author's written purpose represents significantly the relevant words. Their method uses the entropy difference between the intrinsic and extrinsic mode of the Shannon, which refers to words occurring being modulated by the intent of the author, while the irrelevant words are randomly distributed in the text.We suggest that this work's ideas can be extended to any natural language with clearly identified words, without needing any prior knowledge of semantics or syntax.

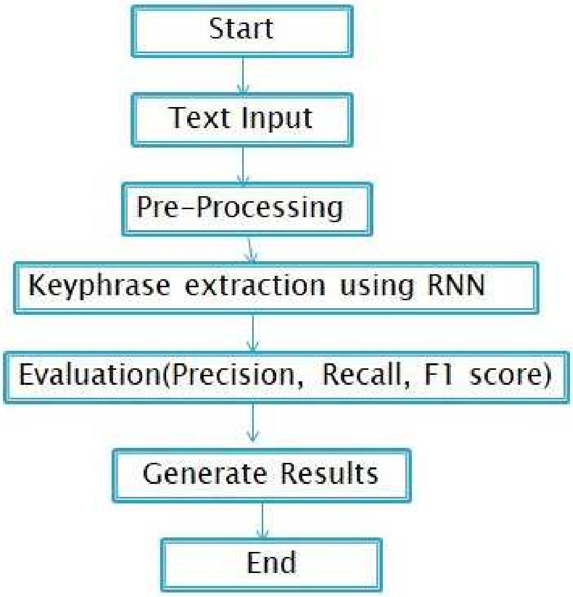
PROPOSED WORK

Extraction of key phrases is based on vocabulary. Error over generation indicates that a sentence containing a keyword is extracted, but not necessarily a key phrase. In frequency error, key phrases fail to predict the occurrence of candidates once or twice in text. Behind the text, current methods don't catch real semantic meaning.

RESEARCH OBJECTIVES

- To study and analyze the recent and existing Keyphrase extraction techniques.
- To propose new framework for keyword extraction system from text using machine learning applications.
- To compare the proposed framework system with the existing state of artbased keyword extraction.
- Validation of extracted keyword by human expert.

RESEARCH METHODOLOGY


Start
Text Input
Pre-Processing
Keyphrase extraction using RNN
Evaluation(Precision, Recall, F1 score)
Generate Results
End

Conclusion

Key phrases offer a convenient way to explain a text, providing some hints about its contents to the reader. In a variety of applications, such as retrieval engines, browsing interfaces, thesaurus creation, text mining, etc., key phrases can be useful. As we address in this article, there are also other activities for which main phrases are helpful. This paper describes an approach to key phrase extraction based on machine learning.

REFERENCES

[1] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", *Research.ijcaonline.org*, 2015. [Online]. Available: https://research.ijcaonline.org/volume109/number2/pxc3900607.pdf. [Accessed: 01- Nov- 2018]

[2] H. M. Lynn, E. Lee, C. Choi, and P. Kim, "Swift Rank: An Unsupervised Statistical Approach of Keyword and Salient Sentence Extraction for Individual Documents," in *Procedia Computer Science*, 2017, vol. 113, pp. 472–477.

[3] Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, "A Graph-based Approach of Automatic Key phrase Extraction," in *Procedia Computer Science*, 2017, vol. 107, pp. 248–255.

[4] F. Xie, X. Wu, and X. Zhu, "Efficient sequential pattern mining with wildcards for key phrase extraction," *Knowledge-Based Syst.*, vol. 115, pp. 27–39, 2017.

[5] Q. Wang, V. S. Sheng, and X. Wu, "Document-specific keyphrase candidate search and ranking," *Expert Syst. Appl.*, vol. 97, pp. 163–176, 2018.

[6] S. Štajner and G. Glavaš, "Leveraging event-based semantics for automated text simplification," *Expert Syst. Appl.*, vol. 82, pp. 383–395, 2017.

[7] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review," *Int. J. Comput. Appl.*, vol. 109, no. 2, pp. 18–23, 2015.

[8] J. Rafiei-Asl and A. Nickabadi, "TSAKE: A topical and structural automatic keyphrase extractor," *Appl. Soft Comput. J.*, vol. 58, pp. 620–630, 2017.

[9] E. Papagiannopoulou and G. Tsoumakas, "Local word vectors guiding keyphrase extraction," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 888–902, 2018.

[10] J. Hu, S. Li, Y. Yao, L. Yu, G. Yang, and J. Hu, "Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification," *Entropy*, vol. 20, no. 2, p. 104, 2018.

[11] Kathait, S. S., Tiwari, S., Varshney, A. and Sharma, A. "Unsupervised Key-phrase Extraction using Noun Phrases", International Journal of Computer Applications, 162(1), 2017.

[12] Gadag, Ashwini I., and B. M. Sagar. "N-gram based paraphrase generator from large text document", In Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on, pp. 91-94. IEEE, 2016.

[13] Shirakawa, Masumi, Takahiro Hara, and ShojiroNishio. "N-gram idf: A global term weighting scheme based on information distance", In Proceedings of the 24th International Conference on World Wide Web, pp. 960- 970. International World Wide Web Conferences Steering Committee, 2015.

[14] Chatterjee, Niladri, and NehaKaushik. “RENT: Regular Expression and NLP-Based Term Extraction Scheme for Agricultural Domain”, In Proceedings of the International Conference on Data Engineering and Communication Technology, pp. 511-522. Springer Singapore, 2017.

[15] Nesi, Paolo, Gianni Pantaleo, and GianmarcoSanesi. “A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction From Web Pages and Documents”, In DMS, pp. 155-161. 2015.

[16] Onan, Aytuğ, SerdarKorukoğlu, and HasanBulut. “Ensemble of keyword extraction methods and classifiers in text classification”, Expert Systems with Applications 57 pp. 232-247, 2016.

[17] Habibi, M. and Popescu-Belis, A. . “Keyword extraction and clustering for document recommendation in conversations”, IEEE/ACM Transactions on audio, speech, and language processing, 23(4), pp. 746-759, 2015.