IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY OF KEYWORD EXTRACTION AND TEXT SUMMARIZATION

Dr. Prashanth CM[1], Sumedha G[2], Riya Singh[2], Urwashi Priya[2], Neha Chauhan[2]

[1] *Associate Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management (DSATM), Bengaluru, Karnataka, India.*

[2] *Student [BE], Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management (DSATM), Bengaluru, Karnataka, India.*

Abstract: The amount of knowledge being generated in numerous domains, like news, social media, education, banking etc. has increased exponentially. Because of the variability of data we are exposed to, there's a desire for an automatic summarizer capable of condensing the textual data within the original document, while keeping the integrity of the information intact. Text summarization has emerged as a necessary research area within the recent past. This paper presents contemporary literature on automatic keyword extraction and text summarization since the text summarization process is highly dependent on keyword extraction. This literature includes a discussion about the various methodologies used for keyword extraction and text summarization. Finally, it briefly discusses the direction these processes may take in the future.

***Keywords*— Keyword Extraction, Natural language processing, Text Summarization**

I. Introduction

In today's rapidly growing world, scientists and scholars are constantly faced with the task of keeping up with knowledge in their field. Various disciplines are inter-connected to each other which forces the scientists and scholars to learn about other disciplines in a very short amount of time. Authors of books need to write short and detailed summary of their work which can take a lot of time and human effort [1]. Automatic Keyword Extraction and Text summarization saves this time and effort. Its applications range from scientific research, e-Newspapers, software bugs reports, journal articles, transcription dialogues etc. The technique of Keyword extraction is used to extract main features in various studies, text categorization, topic detection, information retrieval, document summarization, etc. [2]. Automatic keyword extraction is targeted to apply the power and speed of current computation abilities to the problem of recovery and access, stressing upon information organization without the added costs of human annotators. The technique of text summarization is used to efficiently retrieves the relevant information from documents [3]. It involves reducing a text document into a short set of words or paragraph that conveys the main meaning of the text [4]. Summaries are usually around 17% of the original text and yet contain everything that could have been learned from reading the original article [5]. In the wake of big data analysis, summarization is an efficient and powerful technique to give a glimpse of the whole data. The text summarization is achieved in mainly two ways namely, abstractive summary and extractive summary. The abstractive summary is a topic under tremendous research. However, no standard algorithm has been achieved yet. These summaries are derived from learning what was expressed in the article and then converting it into a form expressed by the computer. It resembles how a human would summarize an article after reading it. Whereas, extractive summary extracts details from the original article itself and presents it to the reader.

II. Keyword Extraction

On the premise of the literature survey of the work done in the field of keyword extraction, we identified the following methods of keyword extraction:

a) TF-IDF

Let D = d1, d2, … dm be a set of documents that belong to each domain. And let Tj = tj1, tj2, … tjn be a set of n terms extracted from a single document 'dj'. T a set of terms extracted from a document set D, is union of T1, T2, … Tm. Weight the terms of T to extract the exact keywords from T, after that sort the terms of T by this weight and extract terms that are weighted highly. Then, with these terms, make a word list named as 'Candidate Keyword List'.

TF-IDF value is composed of two components: TF (Term frequency) and IDF (Inverse Document Frequency) values. The equations are given as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j|}}$$

The numerator is the number of occurrences of the considered term in document dj and the denominator is the number of occurrences of all term in document dj.
The equation for idf term is:

$$idf_i = log\frac{|D|}{|d_j : t_j \epsilon d_j|}$$

The numerator is the number of documents in the corpus and denominator is the number of documents containing the term tj.
Finally, the tf-idf term is given by:

$$tfidf_{i,j} = tf_{i,j} * idf_i$$

b) Rapid Automatic Keyword Extraction

Rake is a keyword extraction algorithm which is domain independent. It partitions the textual data into candidate keywords which are sequence of one or more content word that occur in a text. It extracts candidate keywords by analyzing the frequency of cooccurrence of these content words within a candidate keyword. For keyword extraction, Rake splits the textual data into an array of words. Then, this array of words is split into sequence of contiguous
words separated by phrase delimiters and stop word position. The sequence of contiguous word is called as candidate keyword and each candidate keyword is assigned a same position in the text. A matrix of word co-occurrence is constructed which indicates the frequency of co-occurrence of each content word within a candidate keyword.

Candidate words are also called as sequence of one or more content words (informative words) that occur in a text. After the identification of each candidate keyword, each keyword is assigned a score. The sum of the score of each content word is the total score of each candidate keyword. [6] The process of assigning the score for every keyword is illustrated as follows:

- First, the frequency (freq) of each content word (CW) is calculated in a given textual document represented by freq(CW).
- After computing the frequency, degree of a word is calculated, represented by deg (CW). To compute the degree, total number of words that appear in candidate keywords consisting the content word is counted.
- At last, ratio of degree of content word to frequency of content word is computed, represented by

$$\frac{Degree(CW)}{Freq(CW)}$$

C. Word Co-occurrence

One of the most important criteria for a word to be selected as keyword is its relevance for the text. The tf.idf score of a term is a widely used relevance measure. While easy to compute and giving quite satisfactory results, this measure does not take (semantic) relations between words into account. The main idea is to use cooccurrence of words as the primary way of quantifying semantic relations between words. According to the distributional hypothesis, semantically similar words occur in similar contexts, i.e. they co-occur with the same other words. [7]

Therefore, rather than using the immediate co-occurrence of two terms as a measure for their semantic similarity we will compare the co-occurrences of the terms with all other terms. This intuition is defined as so called co-occurrence distribution of each word which is simply the weighted average of the word distributions of all documents in which the word occurs. The “semantic similarity” of two terms is computed by similarity measure(s) for their cooccurrence distributions. The co-occurrence distribution of a word can also be compared with the word distribution of a text. This gives a measure to determine how typical a word is for a text.

Finally, different keyword extraction algorithms are defined by selecting different relevance measures and the results show that using word co-occurrence information can improve precision and recall over tf.idf.

III. Text Summarization

Based on the literature survey, text summarization can be classified into the following approaches:

A. Extractive Text Summarization

The summarizer evaluates the sentences and words based on statistical and linguistic features to derive the most relevant sentences from the document to form a final summary. Various supervised and unsupervised learning algorithms can be used to obtain the result. Jindal et al. [3] used Fuzzy C-Means clustering to form clusters of sentences based on distances or similarity. Membership value of each sentence is computed and based on it, each sentence is assigned to a cluster with minimum Euclidean distance from the center of the cluster. From each cluster, sentences with high degree of membership are selected. Hierarchical clustering was then applied to derive the most relevant sentences.

Qazvinian et al.[1] developed C-LexRank, a graph based summarization system. It modelled sentences as vertices where edges represent their lexical similarity. It then identified vertex communities (clusters) in this network, and selected sentences from different communities to increase diversity in the summary. It was performed on single scientific articles based on citations, which employed community detection and extracted information-rich sentences.

Derek [10] developed summary for lectures using BERT (Bidirectional Encoder Representations from Transformers), a deep learning NLP model in combination with K-Means clustering. Tokenized sentences were passed to the BERT model to output embeddings, which were then clustered using K-Means. Embeddings that were closest to the centroid were selected as the candidate summary sentences. The core BERT implementation used the pytorch-pretrained-BERT library from the "huggingface" organization.

B. Abstractive Text Summarization

In Abstractive Text Summarization a machine takes in the idea of all the input documents and the formulates a summary. It uses linguistic methods to evaluate and understand the text and finds the new relations and concepts that best describe it. It then generates new shorter text that conveys the most relevant and significant details from the original text document. It can be done using a structure based approach or a semantic based approach. [5] In semantic based technique, linguistics illustration of document(s) is employed to feed into natural language generation (NLG) system. This technique specialize in identifying noun phrases and verb phrases by processing linguistic data. It understands and exploits the relationship between related and co-occurring words. A linguistics model, that captures concepts and relationship among ideas, is made to represent the contents like text and images that are used for multimodal documents. The important ideas are rated using some measures and eventually the chosen concepts are expressed as sentences to create summary. Structured based approach encodes most vital data from the document(s) through psychological feature schemas like templates of existing word structures and extraction rules along with graphs of sentence or word statistics. The documents to be summarized are depicted in terms of classes and listing of aspects. A content choice module selects the most effective candidate among those generated by data extraction rules to answer one or lot of aspects of a category. Finally, generation patterns are used for generation of outline sentences. [8]

IV. Evaluation Methods

The various evaluation methods used for checking the relevance and accuracy of the summary generated are:

A. ROUGE

Recall Oriented Understudy for Gisting Evaluation (ROUGE) usually measures the recall. The recall tells us how much the words(ngrams) from the human generated summaries were present in the machine generated summaries. [8]

1) Precision and Recall: Positive predictive value (precision) can be defined as the fraction of relevant instances by the retrieved instances.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{\{retrieved\ documents\}}$$

Recall (also known as sensitivity) can be defined as the fraction of relevant instances that were retrieved.

2) F-score: The F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the ratio of number of true positive results and the number of all positive results, including those identified incorrectly. The recall is the ratio of number of true positive results and the number of all samples that should have been identified as positive.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

B. Blue

Bilingual Evaluation Study measures the precision. This method tells us how much the words(n- grams) from the machine generated summaries were present in the summaries generated by humans.

$$F1 = 2 * (Bleu * Rouge)/(Bleu + Rouge)$$

Just like Precision and Recall, Rouge and Bleu methods are also complementary to each other. Bleu score will be more if machine generated words appear in human generated summaries more. Rouge score will be more if human generated words appear more in the machine generated summaries.

C. Pyramid Method

Summaries conveying different content can be equally good. This method incorporates the human variation in the summary extraction. An Assumption is made that multiple summaries will be required for the evaluation. By the use of a Set of Contributors(SCU) in the reference summaries a Pyramid is created. The number of contributors in an SCU is given by the frequency with which an SCU is expressed in the pool of model summaries, this frequency is used to weight the importance of the SCU. [9]

V. Challenges and future scope

The constant problem faced in the field of text summarization is the lack of proper datasets. Target value for a corpus is a manually generated summary, good in both factual content and fluency. It is not possible to generate innumerable such summaries manually beforehand for the given dataset for training. This problem can be resolved to some extent by using unsupervised or deep learning methods but still the problem persists. Current algorithms that work on generating topic summaries are focused almost exclusively on extracting relevant, information or feature-rich summaries. Meanwhile, the fluency of the produced summaries has been mostly ignored. In abstractive summarization, there is no generalized framework. Extracting the important sentences and ordering them as in the original source document for producing an efficient summary is an open issue.

• The models can be tweaked to suit semantic differences in different languages and used for datasets consisting of multilingual documents. This can be used for analyzing legal or technical reports that are generally written in native languages, and thus be extended to the lesser researched languages like Dutch, Korean, Polish, Telugu, Tamil etc.

• These keyword extraction techniques can be used to extract features from a specified document set and applied to opinion mining.

• There are also various challenges of in terms of space and time complexity due to usage of layered neural networks, which need to be improved upon.

VI. Conclusion

This paper contains the literature review of recent work in text summarization from the point of view of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices. Text summarization automates the process of getting useful information from large texts in a stipulated or short time. Keyword extraction is used to extract main features from various studies for topic detection, document summarization, etc.

To extract keywords, TF-IDF weights, RAKE, word co-occurrence methods are described and compared. For summarizing documents in a variety of domains, different approaches via extractive method and abstractive method are described. The results of these processes are then evaluated and compared using different methods like ROUGE, Bleu and Pyramid method. Some important research issues in the area of text summarization are also highlighted in the paper.

VII. References

[1] Qazvinian, Vahed & Radev, Dragomir & Mohammad, Saif & Dorr, Bonnie & Zajic, David & Whidby, Michael & Moon, Taesun, “Generating Extractive Summaries of Scientific Paradigms”, in Journal of Artificial Intelligence Research, February 2014.

[2] S Lee, H Kim, “News Keyword Extraction for Topic Tracking,” in 4th International Conference on Networked Computing and Advanced Information Management, Vol. 2, 2008, IEEE.

[3] S. G. Jindal and A. Kaur, “Automatic Keyword and Sentence-Based Text Summarization for Software Bug Reports,” in IEEE Access, vol.8, pp. 65352-65370, 2020, IEEE.

[4] Varun Pandya, “Automatic Text Summarization of Legal Cases: A Hybrid Approach”, ArXiv, vol. abs/1908.09119, 17 August, 2019.

[5] Bharti, Drsantosh & Babu, Korra, “Automatic Keyword Extraction for Text Summarization: A Survey”,8 February 2017.

[6] Yan Du, Hua Huo, “News Text Summarization Based on Multi-feature and Fuzzy Logic”, in IEEE access Vol. 8, pp. 140261 - 140272, 2020, IEEE.

[7] Christian Wartena, Rogier Brussee, Wout Sklakhorst, “Keyword Extraction using Word Co-occurrence”, in 2010 Workshop on Database and Expert Systems Applications

[8] Rada Mihalcea, Paul Tarau, “TextRank: Bringing Order into Text”, in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 404–411, July, 2004.

[9] Qazvinian, Vahed & Radev, Dragomir & Mohammad, Saif & Dorr, Bonnie & Zajic, David & Whidby, Michael & Moon, Taesun, “Generating Extractive Summaries of Scientific Paradigms”, Journal of Artificial Intelligence Research, 2014 [10] Miller, Derek, “Leveraging BERT for Extractive Text Summarization on Lectures”, 2019