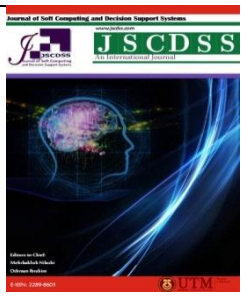
An International Journal
JSCDSS
http://www.jscdss.com
Vol.3 No.3 June 2016: 35-41
Article history:
Accepted 5 April 2016
Published online 5 April 2016

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Sentence Similarity Techniques for Automatic Text Summarization

Yazan Alaya AL-Khassawneh [a,*], Naomie Salim [a], Adekunle Isiaka Obasae [a]
[a] Faculty of Computing, Universiti Technologi Malaysia, Johor, Malaysia

*** Corresponding author email address**: yakhassawneh@yahoo.com

Abstract

The technology of summarizing documents automatically is increasing rapidly and may give an answer for the information overload quandary. These days, document summarization is assumed an imperative part of information retrieval. With expansive amounts of documents, giving the user a short version of every document incredibly encourages the errand of discovering required documents. Text summarization is a procedure for making a packed form of a particular document that gives the users utilizable info, and summarization of multi document is engender summary distributing the meaning of the most info either explicitly or implicitly from a group of documents about main topic. In text summarization, resemblance among several sentences in a text has a major role. As such, development of methods of summarization has taken into consideration the aspect of similarities between several sentences in a text. This paper seeks to investigate different techniques of automatic summarization based on the element of sentence resemblance. Comparison is also developed for functionalities of various techniques with respect to recall, precision and F-measure values.

Keywords: Text summarization, Extractive summarization, Abstractive summarization, Sentence similarity

1. Introduction

When an extracted or generated text carries information that is a vital segment of the primary document, it is deemed as summary for the main text. Moreover, when this occurs mechanically with involvement of a computerized program, it is known as an Automatic Text Summarization (ATS). In brief, a summary ought to sustain the mainstay of the document which paves the way for quick detection of pertinent information. Radev et al. (2002), opined that a summary could be defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". Such definition suggests that summaries, which can be generated from one or more documents, should be reasonably brief and hold significant information derived from the primary text(s). Automatic text summarization is categorized under two procedures in accordance with the input. In a circumstance where the input involves a sole document, the procedure is termed a Single Document Summarization. An input that involves several documents of a similar nature, the procedure is known as a Multiple Document Summarization.

The aim of summarizing text is to display the ultimate essential information using smaller form from the first content while saving its principle substance and assists the user to rapidly see huge amounts of info. Text summarization studies the quandary of choosing the paramount parts from the text and the quandary of creating cohesive and reasonable summaries. The automatic procedure is fundamentally not quite the same as that of summaries generated by human, since human can catch and connect profound implications and subjects of text documents, while automatic programs' ability of such adeptness is very hard.

Programmed text summarization dated back to 1958- (Luhn, 1958). Since this time, efforts are being made by scholars to suggest systems for producing summaries. Several of scholars have suggested strategies for programmed text summarization, which could be classified into two: extractive and abstractive.

Determination of highest scored sentences or paragraphs from the pristine text and put them together to create shorter text while keeping the main meaning of the source text is known as Extractive summary. While in Abstractive summary scheme the linguistic means are used to inspect and explicate the text. Extraction systems are mostly used nowadays, to engender summary.

With organized documents, automatic text summarization works better, for example, scientific papers, reports, news and articles. The initial phase in extractive summarization is the determination of critical components, for example sentence location (Fattah and Fuji Ren, 2008), sentence length, number of numerical data (Lin, 1999), term frequency (Salton, 1989), number of opportune entities (Kupiec et al., 1995) and number of words occurring in denomination (Salton and Buckley, 1997).

The sentence similarity within the document is essential during the summarization process. This aspect has formed the basis of different summarization techniques. This paper surveys different automated methods of text summarization based on the basis of sentence resemblance aspect. The other sections of the paper, Section 2 covers the related work of text summarization. Section 3 discusses six summarization techniques based on sentence similarity. Section 4 compares the techniques in the previous section with respect to precision, recall and F-measure and conclusion is given in Section 5.

2. Related Work

Extraction-based text summarization and abstraction-based text summarization are the procedures utilized for automatic summarization.

The central goal of an extractive summary is to identify the most significant areas in the context of words, sentences and paragraphs, among others, in the input sourced from one or more documents. Summaries derived from the extraction procedure hold some concatenated sentences expressed precisely as they occur in the documents targeted for summarization.

In the extractive summarization procedure, a ruling is arrived at on whether or not a specific sentence ought to be extracted for inclusion in the summary. Search engines, for instance, employ extractive summary generation to realize summaries from web pages.

A variety of approaches utilizing rational and arithmetical principals is available for grading the areas and identifying those with a superior count from the text to be added to the summary. Extraction of sentence can be described as a text sieve permitting the passage of only applicable sentences. For the most part, studies on summarization train their sights on extractive summarization due to its less complicated implementation process (Luhn, 1958; Edmondson, 1969; Barzilay et al., 1999); Marcu and Daniel, 1999; Hovy et al., 1998; Kan, 1999; Chen, 2000; Copeck et al., 2002; McKeown et al., 2002; Farzindar et al., 2005; AL-Khassawneh et al., 2014; Kågebäck et al., 2014; Mogren et al., 2015) are few examples for systems which use Extractive based methods to generate summary.

However, it should be noted that people record summaries in abstract form. Subsequent to digesting a text, an individual can comprehend the subject matter and jot down an exceptionally brief summary without leaving out significant facts. Machines do not have this capability. The generation of abstractive summaries can turn out to be a tall order. Thus, it can be construed that the objective of abstraction based summarization is to produce (as humans do) a summary comprising grammatically flawless fresh sentences through the utilization of sophisticated natural language generation processes.

The realization of this objective necessitates a good comprehension of the subject matter in the primary text. Abstractive summary generation is comparatively more demanding as it calls for semantic knowledge of the text to be submitted into the National Language Generation system. A significant stumbling block in this area is the lack of development in the realm of sentence synthesis that can lead to an illogical generated summary (Barzilay et al., 1999; Jing et al., 2000; Saggion et al., 2002; Moawad and Aref 2012; Thomas et al., 2015; Li, 2015) are some examples of systems that generate abstractive summaries.

Summarization methods could be categorized into two sets: supervised and unsupervised methods. Supervised methods deal with the summarization mission as a double class grouping issue at the sentence level. The sentences contained in the summary are positive, while the rest sentences are negative. After representing every sentence with a set of features, the grouping function could be prepared in two dissimilar methods (Mihalcea and Ceylan, 2007). One is a discriminative manner with understood algorithms, for example, Support Vector Machine (SVM) (Yeh et al., 2005).

Numerous unsupervised techniques were created for summarizing documents by utilizing distinctive features and relations from the sentences (Radev et al., 2004; Erkan and Radev, 2004; Alguliev et al., 2005; Alguliev and Aliguliyev, 2005; Aliguliyev, 2006; Aliguliyev, 2007; Alguliev and Alyguliev, 2007), are some examples of these techniques.

Also, summarization can be classified based on the task, either query or generic based. Query based summary introduces the most pertinent information to the offered queries. (Fisher and Roark, 2006; Li et al., 2007; Dunlavy et al., 2007; Wan, 2008).

While generic based summary provides a general meaning of the document's content. In another meaning, generic summaries attempt to identify salient information in text without the context of a query. (Salton et al., 1997; Gong and Liu, 2001; Alguliev et al., 2005; Alguliev and Aliguliyev, 2005; Aliguliyev 2006; Aliguliyev 2007; Alguliev and Alyguliev, 2007; Li et al., 2007; Dunlavy et al., 2007; Jones, 2007; Wan, 2008).

Dunlavy et al. (2007) proposed a new system called Query, Cluster, and Summarize (QCS). This system implements some tasks to meet query requirements: recovery of related documents; grouping the recovered documents based on topics and generating summary for every group. QCS is an instrument for retrieving documents, which allows the user to determine the desired documents. McDonald and Chen (2006) improved query based, generic, and hybrid summarizer, with varying measures of document content for each.

Query based summarizer utilize query term info only. While generic summarizer utilize a combination of discourse info and info acquired through conventional analysis for the surface level, and the hybrid summarizer utilized query term info combined with discourse info. Fung and Ngai (2006) show a multilingual, theme based, multi document summarization system based on the text modelling with respect to displaying content coherence.

Automatic text summarization is an immensely multidisciplinary research area associated with multimedia, computer science, cognitive psychology and statistics.

An International Journal JSCDSS E-ISSN: 2289-8603

Event Indexing / Summarization (EIS) intelligent system was proposed by (Guo and Stylios, 2005), for automatic text summarization. This system depends on a perceptive psychology paradigm, the roles, sentences relevance, and their structure in document comprehension. The EIS includes grammatical analyzation of sentences, indexing and grouping sentences by five indicators from the event indexing paradigm, and finally extracting the maximum notable content, by lexically analytical at phrase / clause levels.

Resemblance measures have growing significant role in Information Retrieval (IR) and Natural Language Processing (NLP). Resemblance measures were used in the research area of text and its applications, as information retrieving, text summarization, text clustering, and text mining. From those applications, it was shown that the calculation of sentence resemblance had become a general factor in knowledge representation and discovery era. There is a large-scale studies focus on computing the resemblance among documents, but very little studies are on the computing of resemblance among sentences and short texts (Liu et al., 2007; Li et al., 2008).

Li et al. (2006) suggested a method to find the similarity among sentences, on the base of word order and semantic info. First, semantic verbal resemblance is derivative from a corpus and a lexical base. Secondly, the suggested approach considers how the order of words can affect the sentence meaning. The integration of semantic similarity and word order similarity is defined as the whole sentence resemblance.

Liu et al. (2007) raise a fresh technique to compute resemblance among sentences by utilizing Dynamic Time Warping method and analysing parts of speech.

Wan (2007) suggested a new measure on the base of Earth Mover's Distance (EMD) to calculate document resemblance by permitting many to many conformity among subtopics. Firstly, every document is disintegrated into a group of subtopics, and after that, the EMD is used to calculate the similarity among two groups of subtopics in two different documents dissolve the problem of transportation. The suggested measurement is an enhancement of the prior Optimal Matching (OM) measure, which permits only one-to-one matching among subtopics.

Bollegala et al. (2007) suggested a technique that combines both scraps and page counts to measure semantic resemblance among specified couple of words. Through this work, improved four standard measures; Dice, Jaccard, Point-wise Mutual Information (PMI) and Overlap (Simpson), were used to calculate semantic similarity through page counts.

3. Methods of text summarization based on sentence resemblance

Some approaches have been designed for summarization of texts. This paper aims at exploring six automated text summarization approaches based on sentence similarity. Additionally, a comparison of their performances with regard to F-score, recall and accuracy has been initiated.

3.1 Query-based summarizer based on similarity of sentences

Kumar et al. (2011), suggest a query-based summarizer that revolves around word frequency and sentence similarity. In this technique, a summarizer utilises Vector Space Model (VSM) in identifying sentences that have a similarity to the question as well as sum focus for identification of word recurrence. In the current work, the authors suggested a query based summarizer that revolves around classification of similar sentences. In the suggested structure, the query is pre-processed, and then the summarizer obtains relevant documents and produces a final summary.

After pre-processing, summary creation entails several steps that include:

i) Calculation of sentence similarity in documents having user query.

ii) After calculation of similarity, sentences are classified basing on similarity values.

iii) Calculation of sentence score using sentence identification feature and word frequency.

iv) Selection of sentences with highest scores from each category and summarizing it.

v) Compression of the summary to exactly 100 words.

3.2 Extractive Multi-Document Summarizer (EMDS) algorithm

The extractive multi-document summarizer was proposed by (Amit and Aarati, 2014). This method is a graph-based multi-document summarizer that follows the following steps. A set of related texts forms the input. In the first phase, pre-processing of documents is done.

The undirected acyclic graph is generated for every text with sentences representing nodes and similarities representing edges. Then, the weighted ranking algorithm is executed to allow generation of significant score for every sentence in the text. Sentences ranking is done based on their respective significant scores. The highest ranking sentences are selected from the summary of every text. In the second phase, all the single summary of every text is assembled to form a single text. In the last phase, the above process is employed in combining document, thus forming the last extractive summary.

An International Journal JSCDSS E-ISSN: 2289-8603

3.3 Language Independent Sentence Extraction Based Text Summarization

Krish and Bidyut (2011) suggested a text summarization technique based on extraction of language-independent sentences and one that adopts a structural-feature based sentence scores and a PageRank-oriented sentence ranking.

The efficiency of the suggested method had been ascertained in both Tamil and English documents through application of ROUGE assessment.

The technique was undertaken in four separate stages that include

i) Pre-processing, where stemming and removal of stop words were conducted for preparation of the information sources for production of summary.

ii) Scoring, where the scores are awarded to sentences using their TF-IDF feature, topic similarity, length and position such that long sentences having sentences that relate to the document title and existing at the document beginning are scoring highly.

iii) Ranking, where ranking of the sentences is based on Google's "PageRank formula" and lastly.

iv) Summary generation, where the finalised summary contains highly ranked sentences that are displayed in a manner similar to that of the source document text.

The quantity of highly rated sentences chosen for summarization could be user-defined according to compression ratio or sentence quantity with regard to the breadth of the source document text.

3.4 Fuzzy logic based method for text summarization

A method for text summarization, based on fuzzy logic, was suggested by Suanmali et al. (2009). Regarding this method, segmentation of sentence, tokenization, removal of stop words and stemming of words, were performed in the pre-processing step. Thereafter, critical features of every sentence are extracted. The features may entail title, length of sentence, weight of term, position of sentence, sentence resemblance, proper noun, numerical data and thematic word. The numerical vectors, which correspond to the elements, are computed to derive the sentence score base, which will be utilized on the fuzzy logic technique. A group of sentences with top scores is derived from text summary based on compression ratio. Four elements including fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base make up the fuzzy logic system. The fuzzifier translates input texts into linguistic values via the use of a membership function and the outcome is applied as the input variables. Following fuzzyfication (fuzzifier step), the inference engine implies the rule base, which contains IF-THEN rules. Regarding the final phase, there is conversion of the output variables to the last crisp values via the defuzzifier through the use of a membership function to represent the score of the last sentence. The input membership operation for every element is separated into five fuzzy groups, which are made up of High (H), Very High (VH), Low (L), Very Low (VL) and Median (M).

3.5 Text summarization using clustering technique

Anjali and Lobo (2013) suggested this novel approach to multi-document summarization that warrants excellent coverage and avoidance of generation of redundant sentences. The input to the query is the group of texts and query. These authors have retained a map list, where every expression with its frequency from the text group is stowed in a map. The technique of query modification is applied as follows: query splitting into tokens and finding the synonym for every token, and if the synonym or tokens in a text collection then the highest frequent add the synonym of the query with highest frequent to query. The terms occurring most frequently among corpus are chosen and added to the query for strengthening of the query. The elements are employed in the computation of sentence score which include title feature, numerical data, cue phrase, sentence length, sentence centrality, sentence position, upper case word, term frequency, sentence similarity, and inverse document frequency. Clustering of the text is done via use of cosine similarity as a means of generating the necessary documents clusters. Thereafter, from each text cluster, clustering of sentences is based on the values of resemblance. Each group`s score is then calculated. Sorting of the sentence clusters is then done in the reverse order of the group score. Lastly, for every cluster, the best sentence score is selected and added to the final summary.

3.6 Information content based Sentence Extraction

Daniel et al., (2004) suggested a FULL-COVERAGE algorithm, which extracts sentences covering the entire document's concept space through iterative measurement of all sentence similarities in the entire document and elimination of already featured words. It revolves around the notion that the sentence relevance lies proportionally towards its similarity within the entire document.

The initial stage of the algorithm entails parsing a document into sentences. Furthermore, this stage is characterized by the application of Porter stemming algorithm, elimination of stop-words and tokenization. The second algorithm phase entails FC calculation, the sentences subset covering the whole document's concept space. The technique of calculating FC entails treating every single sentence S_i $(i = 1,2,3, \dots, N)$ of D as a document in the final group of D. Next phase entails using the whole document as a question against all individual sentences and incorporating the highly rated sentence into the FC set. After the determination of highly rated FC set,

An International Journal JSCDSS E-ISSN: 2289-8603

the third phase that involves actual generation and return of a summary is launched. Given that p is the percentage, a $CR\ (p)$ summary encompasses the initial n sentences from full-coverage whereby $n \leq p * |FC|$.

4. Comparative Study Of Sentence Similarity Based Summarization Techniques

The query based summarizer (Kumar et al., 2011), summarized texts based on the sentence resemblance and word frequency. The Information-Retrieval Text Research Collections (AQUAINT-2) was applied as the corpus and the generated summary sentences were assessed via use of ROUGE metrics. Expensive linguistic data is not used in this summarizer. The summarizer employs the Vector Space Model (VSM) to determine sentences resembling the query and sum focus for determination of word frequency.

The accuracy obtained via the use of the suggested method was compared to the Text Analysis Conference (TAC) system. For evaluation of the suggested system, datasets of TAC2009 were employed as recommended by NIST for updated summarization. The text was composed of 48 topics, where every topic had 20 texts categorized into "A" and "B" clusters in terms of their sequential scope of topic. Cluster "A" texts of datasets of TAC2009 were used for system evaluation. The precision value was 0.29034, recall value was 0.30127 and F-measure value was 0.29961.

The Extractive Multi-Document Summarizer (EMDS) algorithm (Amit and Aarati, 2014) suggests a method where extractive summary for many appropriate documents is created using different sentence features that include sentence similarity, sentence length and word class.

For assessment of Extractive Multi-Document summarization, a comparison of two summarizers has been undertaken, they include LEAD and Random. The resultant summaries are evaluated through the automated metric ROUGE as well as manual assessment. The F-measure, recall and precision values were 0.56062, 0.57328 and 0.54850.

The Language Independent Sentence Extraction Based technique (Krish and Bidyut, 2011), entails application of structural feature-oriented sentence scoring alongside a PageRank-based sentence ranking. The suggested technique's efficiency has been ascertained using Tamil and English documents through ROUGE assessment. The outcomes for English were arranged through DUC 2002 data on single-document summarization, whereas the ones for Tamil were arranged through 100 sets of human created summaries. The suggested system was contrasted to the baseline summarizer. The Tamil and English recall values were 0.4877 and 0.5200.

The fuzzy logic technique (Suanmali et al., 2009) obtained eight critical aspects and computed their scores for all sentences. The technique suggests text summarization according to fuzzy logic for the improvement of the summary quality created through the general statistical technique.

The system testing was achieved through DUC 2002 data set. They contrasted their outcomes using Microsoft Word 2007 and baseline summarizers. ROUGE-I assessment tool indicates an F-measure of 0.47181, recall of 0.45706 and a precision of 0.49769.

The method based on clustering (Anjali and Lobo, 2013) offered an extractive method of text summarization by combined technique between sentence and text clustering. The clustering based technique, which assemblies similar texts into clusters first, then clustering of sentences from every text cluster and lastly sentence scoring is done and top score sentences are picked in to the last summary version.

The comparison was developed between this method and methods based on statistical features as well as methods based on clustering of texts only. The values for precision, recall and F-measure were 0.57, 0.48 and 0.52 respectively.

The method based on content of information (Daniel et al., 2004), suggests the FULL COVERAGE algorithm based on the aspect that the significance of a sentence is proportional to its resemblance to the entire text. The system is examined with information from SMARTs TIME Magazine Collection and the TREC documents utilized for the 2002 DUC edition.

A summary text of the TIME Magazine collection forms 40% of the size of the initial text. Regarding DUC, the algorithm generates 22% of the size of the initial document. For comparing with the suggested method, additional two baseline methods, random and lead-based summarizer, were employed. The system`s precision was 0.346 via use of ROGUE-1 assessment.

A comparison of precision, recall and F-measure values of different techniques are given in Table 1, which shows that the Extractive Multi-Document Summarizer (EMDS) algorithm gives the best results in summarization. Fig. 1 shows those these results.

Table 1
Comparison of sentence similarity based summarization techniques

Summarizer	Precision	Recall	F- measure
FUZZY	0.49769	0.45706	0.47181
QUERY	0.29034	0.30127	0.29961
CLUSTERING	0.57	0.48	0.52
EMDS	0.5485	0.57328	0.56062

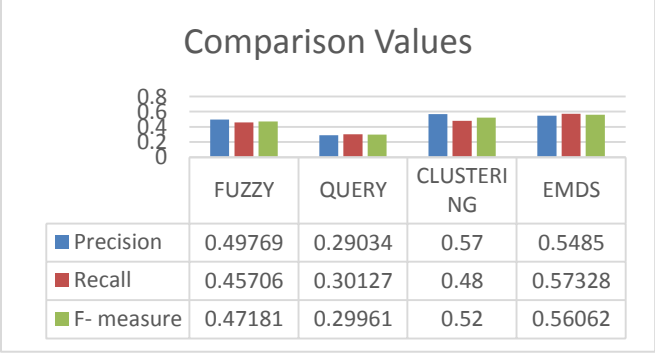


Fig. 1. Comparison of sentence similarity based summarization techniques

An International Journal JSCDSS E-ISSN: 2289-8603

5. Conclusion

Automated generation of text summaries has been an area that has attracted much attention for many years. Similarities between sentences in a text have a crucial role to play concerning summarization. This paper examines methods of text summarization, which are based on the feature of sentence similarity. Comparison is developed between six different techniques, with regard to their precision, recall, F-score and test corpus among other measures. The ROUGE toolkit was applied by approximately all techniques for assessing the summaries produced. Comparative analysis of the various methods was then carried out. Best summarization outcomes with regard to the values of precision and recall were found to be generated by EMDS (Extractive Multi-Document Summarizer).

References

Alguliev, R. M., & Alyguliev, R. M. (2007). Summarization of text-based documents with a determination of latent topical sections and information-rich sentences. Automatic Control and Computer Sciences, 41, 132–140.

Alguliev, R. M., & Aliguliyev, R. M. (2005). Effective summarization method of text documents. In Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence (WI'05), 19–22 September (pp. 264–271), France.

Alguliev, R. M., Aliguliyev, R. M., & Bagirov, A. M. (2005). Global optimization in the summarization of text documents. Automatic Control and Computer Sciences, 39, 42–47

Aliguliyev, R. M. (2006). A novel partitioning-based clustering method and generic document summarization. In Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI–IAT 2006 Workshops) (WI–IATW'06), 18–22 December (pp. 626–629), Hong Kong, China.

Aliguliyev, R. M. (2007). Automatic document summarization by sentence extraction. Journal of Computational Technologies, 12, 5–15.

AL-Khassawneh, Y. A., Salim, N., & Isiaka, O. A. (2014). Extractive Text Summarisation using Graph Triangle Counting Approach: Proposed Method. In 1 st International Conference of Recent Trends in Information and Communication Technologies in Universiti Teknologi Malaysia, Johor, Malaysia (pp. 300-311)

Amit S. Zore, Aarati Deshpande Extractive Multi-Document summarizer algorithm International Journal of Computer Science and Information Technologies, Vol. 5, 5245-5248, 2014.

Anjali R Deshpande, Lobo L M R J Text summarization using Clustering technique, International Journal of Engineering Trends and Technology, Volume 4, Issue 8 (August 2013).

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. Advances in automatic text summarization, 111-121.

Barzilay, R., McKeown, K. R., & Elhadad, M. (1999, June). Information fusion in the context of multi-document summarization. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 550-557). Association for Computational Linguistics.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In Proceedings of 16th world wide web conference (WWW16), May 8–12 (pp. 757–766) Banff, Alberta, Canada.

Chen, H. H., & Lin, C. J. (2000, July). A multilingual news summarizer. InProceedings of the 18th conference on Computational linguistics-Volume 1(pp. 159-165). Association for Computational Linguistics.

Copeck, T., Szpakowicz, S., & Japkowicz, N. (2002, July). Learning How Best to Summarize. In Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics, Philadelphia, PA.

Dunlavy, D. M., O'Leary, D. P., Conroy, J. M., & Schlesinger, J. D. (2007). QCS: A system for querying, clustering and summarizing documents. Information Processing and Management, 43, 1588–1605.

Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264-285.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 457-479.

Farzindar, A., Rozon, F., & Lapalme, G. (2005, October). CATS a topic-oriented multi-document summarization system at DUC 2005. In Proc. of the 2005 Document Understanding Workshop (DUC2005).

Fattah, M. A., & Ren, F. (2008). Automatic text summarization. World Academy of Science, Engineering and Technology, 37, 2008.

Fisher, S., & Roark, B. (2006). Query-focused summarization by supervised sentence ranking and skewed word distributions. In Proceedings of the document understanding workshop (DUC 2006), 8–9 June (pp. 8) New York, USA.

Fung, P., & Ngai, G. (2006). One story, one flow: Hidden Markov story models for multilingual multidocument summarization. ACM Transaction on Speech and Language Processing, 3, 1–16.

Gong, Y., & Liu, X. (2001). Creating generic text summaries. In Proceedings of the 6th international conference on document analysis and recognition (ICDAR'01), 10–13 September (pp. 903–907) Seattle, USA.

Guo, Y., & Stylios, G. (2005). An intelligent summarization system based on cognitive psychology. Information Sciences, 174, 1–36.

Hovy, E., & Lin, C. Y. (1998, October). Automated text summarization and the SUMMARIST system. In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998 (pp. 197-214). Association for Computational Linguistics.

Jing, H., & McKeown, K. R. (2000, April). Cut and paste based text summarization. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 178-185). Association for Computational Linguistics.

Jones, K. S. (2007). Automatic summarizing: The state of the art. Information Processing and Management, 43, 1449–1481.

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2014, April). Extractive summarization using continuous vector space models. InProceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL (pp. 31-39).

Kan, M. Y., & McKeown, K. (1999). Information extraction and summarization: Domain independence through focus types.

Kumar, A. S., Premch, P., & Govardhan, A. (2011). Query-based summarizer based on similarity of sentences and word

An International Journal JSCDSS E-ISSN: 2289-8603

frequency. International Journal of Data Mining and Knowledge Management Process, vol.1, no.3.

Kupiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73). ACM.

Li, J., Sun, L., Kit, C., & Webster, J. (2007). A query-focused multi-document summarizer based on lexical chains. In Proceedings of the document understanding conference 2007 (DUC 2007), 26–27 April (p. 4.) New York, USA.

Li, W. (2015) Abstractive Multi-document Summarization with Semantic Information Extraction. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1908–1913, Lisbon, Portugal.

Li, Y., Luo, C., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. IEEE Transactions on Knowledge and Data Engineering, 20, 641–652.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18, 1138–1150.

Liu, X., Zhou, & Y., Zheng, R. (2007). Sentence similarity based on dynamic time warping. In Proceedings of the first international conference on semantic computing (ICSC 2007), 17–19 September (pp. 250–256) Irvine, USA.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165..

Mallett, D., Elding, J., & Nascimento, M. A. (2004, April). Information-content based sentence extraction for text summarization. In Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on (Vol. 2, pp. 214-218). IEEE.

Marcu, D. (1999). Discourse trees are good indicators of importance in text.Advances in automatic text summarization, 123-136.

McDonald, D. M., & Chen, H. (2006). Summary in context: Searching versus browsing. ACM Transactions on Information Systems, 24, 111–141.

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., ... & Sigelman, S. (2002, March). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. InProceedings of the second international conference on Human Language Technology Research (pp. 280-285). Morgan Kaufmann Publishers Inc.

Mihalcea, R., & Ceylan, H. (2007). Explorations in automatic book summarization. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), 28– 30 June (pp. 380–389) Prague, Czech Republic.

Moawad, I. F., & Aref, M. (2012, November). Semantic graph reduction approach for abstractive Text Summarization. In Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on (pp. 132-138). IEEE.

Mogren, O., Kågebäck, M., & Dubhashi, D. (2015) Extractive Summarization by Aggregating Multiple Similarities. In Proceedings of Recent Advances in Natural Language Processing, pages 451–457, Hissar, Bulgaria .

Perumal, K., & Chaudhuri, B. B. (2011). Language independent sentence extraction based text summarization. In Proceedings of ICON-2011: 9th International Conference on Natural Language Processing.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational linguistics, 28(4), 399-408.

Radev, D. R., Jing, H., Stys, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing and Management, 40, 919–938.

Saggion, H., & Lapalme, G. (2002). Generating indicative-informative summaries with sumUM. Computational linguistics, 28(4), 497-526.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. Information Processing and Management, 33, 193–207.

Suanmali, L., Salim, N., & Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

Thomas, S., Beutenmüller, C., de la Puente, X., Remus, R., & Bordag, S. (2015, September). ExB Text Summarizer. In 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (p. 260).

Wan, X. (2007). A novel document similarity measure based on earth mover's distance. Information Sciences, 177, 3718–3730.

Wan, X. (2008). Using only cross-document relationships for both generic and topic-focused multi-document summarizations. Information Retrieval, 11, 25–49.

Yeh, J-Y., Ke, H-R., Yang, W-P., & Meng, I-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information Processing and Management, 41, 75–95.

An International Journal JSCDSS E-ISSN: 2289-8603