International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4577 - 4581

Abstract Creation of Research Paper Using Feature Specific Sentence Extraction based Summarization

Ms. Jagtap Jayanti
Dept. Of Computer Engineering
Dnyanganga College Of Engineering And Research
Pune, India
Email:jagtapjayanti@gmail.com

Prof. Patel H.H
Dept. Of Computer Engineering
Dnyanganga College Of Engineering And Research
Pune, India
Email: helly.patel@zealeducation.com

Abstract— Several techniques for identifying essential content for text summarization have been created to date. Subject representation techniques is primary infer a midway reflection of the content that that grabs the styles discussed in the data. Considering these representations of topics, phrases in the details records are obtained for each and every relevance. In our suggested system sentence relevance detection is applied determines a score for each sentence based on its significance. Then an overview is produced by selecting most calculated sentences. The produced overview is use for producing subjective by Enhanced summation technique, choosing the sentences from the overview one by one and create word chart. In our system enhance edge weighting strategy is applied for high connection throughout words of produced chart. For discovering few shortest path sentences suggested method use dijkstras algorithm. Before choosing the best quickest path sentences, system examine framework of phrase grammatically. Outcomes demonstrate that extractive and abstractive-oriented overviews produced by Improve COPMENDIUM outshine current system of summation system. We used feature specific sentence extraction techniques which enhance the effectiveness of the summarization strategy.

Keywords- *Sentence Extraction; Dijkstras; Feature Profile; Word Graph; Abstractive Summary; Extractive summary ; Text Summarization;*

*****

I. INTRODUCTION

Today, The large amount of data on World Wide Web is growing at an exponential pace. Nowadays, people use the on World Wide Web to find information or data through information retrieval tools such as Google, Yahoo, Gmail, Bing and vices-versa. So, the exponential growth of information available on the internet, summary of the retrieved results has been necessary for peoples. The concept of summary is commonly used in everyday language. The words summary and abstract are used as synonyms. These definitions use the word abstract instead of summary because they are proposed for the field of human summarization where extracts are produced. Summary means as exact, highly structured, concise, abbreviated, accurate representation of the content of a document. The abstract is a time saving tool that can be used to find a main content of the article. Text summarization has become an important and timely tool for peoples to quickly for assisting and interpreting and understand the large amount of information. The goal of automatic text summarization is to condense the documents into shorter version and preserve important contents of document.

The main motive of this paper is that text summarization tool has used to create abstract of research paper using features extraction. Abstract generation is very challenging task in text summarization. These abstracts are very important because, reader first decide whether or not read complete paper after going though summarie.That means reader first look abstract of research papers before reading a complete papers. In abstract creation, first the most salient sentences has to be identified and then developed small paragraph using the most relevant sentences in their whole document.

II. RELATED WORK

In [1] paper, generating abstracts of biomedical papers using text summarization system. COMPENDIUM text summarization system are have two approaches such that $COMPENDIUM_E$ is generating extractive summaries that only selected and extracted most relevant sentences, clause form document and produced summary without changing text in original document and $COMPENDIUM_{E-A}$ is generates the abstractive oriented summaries in which express idea of original document using different words. It generated new information in form of sentences compression using natural language processing (NLP).

In paper [2], explains the performance of MEAD. It is an extensive, public sector, free, multi document multilingual summarization environment. It is base on sentences extraction. For each sentence in cluster of related document MEAD computed three features such as centroid score, position of sentences in document and overlap with first sentence. These features are used to find salient sentences in document

In this paper [3], we examine a strategy for developing an extensive textual summary of a topic consisting of details attracted from the World Wide Web. We use the high-level framework of human-authored text messages to instantly generate a segment particular design for the subject framework of a new summary. It used a technique to learn topic particular extractors for content choice together for the whole template.

The standard perceptron algorithm with a worldwide integer linear programming optimized both local fit of data into each topic and global coherence across the whole summary. The outcomes of our assessment validate the advantages of integrating architectural details into content selection procedure.

In paper [4], QCS is tool for querying, clustering, and summarizing document sets. It is used for document tretrival.This methods used for retrieving a set of documents that best match a query, clustering a set of documents by topic, and creating a summary of single or multiple documents

In paper [5], MUSE is MUltilingual Sentence Extractor which is a new approach for multilingual single-document extractive summarization. This summarization is considered as an optimization or a search problem. A Genetic Algorithm is used to find an optimal weighted linear combination of statistical sentence scoring methods which are all language-independent and are based on either a vector or a graph representation of a document, where based on a word segmentation. MUSE have significantly out-performed TextRank, the best known language-independent approach to generated summary in both Hebrew and English languages.

In [6],propose a system retrieval of similar clinical cases, based on mapping the content onto UMLS ideas and present the patient records as semantic graph. But is it not consider geographical location, age and gender into consideration when ranking then results for any particular patient.

In this paper [7], we recommend a summarization strategy, AZOM that brings together mathematical and conceptual property of text and in regards of documents framework, ingredients the conclusion of text. AZOM is also capable of outlining unstructured records. Suggested strategy is nearby for Persian language but effortlessly can implement to other languages. The scientific outcomes show relatively excellent outcomes than frequent organized text summarizers, also than current Persian text summarizers.

In [8], proposed CBSEAS which generate extractive sentiment based summaries. In this method sentence selection is directly based on redundancy location. Also, redundancy elimination is crucial in multi-document summarization, takes place in the same step as sentence selection. There is need to improve performance for high-traffic applications that use large grammars, the web service could cache responses.

III. PROPOSED MODEL

The following figure.1 represents the proposed system architecture. This architecture model consists of the following stages:

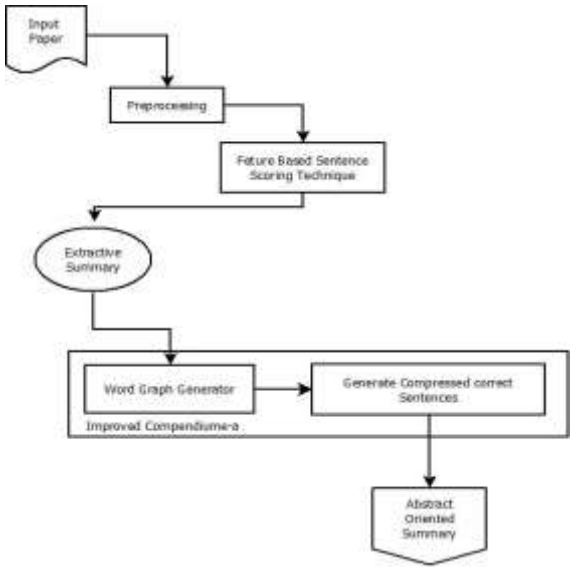
A. PREPROCESSING

The pre- processing is a primary step to upload the text into the system, and to obtain a structured representation of the original text that improve the accuracy of the system. It is carried out so that the text can be organized for the further processing. It involves 4 steps

- Sentence Segmentation and Tokenization
- POS tagging
- Removing Stop Word
- Word Stemming

B. FEATURE BASED SENTENCE SCORING TECHNIQUES

After processing of text, each sentence of the document is represented by an attribute vector of features. The information



Figuter.1: System Architecture

of text is described by a large number of terms or words that occurs in text and all of them are not useful and irrelevant. So that feature extraction is needed. In feature extraction, each sentence of document obtains a feature score based on its importance. The proposed method used six features for each sentence. It combines features called term feature [9] with following features [10] like sentence position, sentence length, sentence centrality, number of proper nouns in the sentence which are obtain feature profile. Each feature is given a value between 0 and 1.There are six features as follows:

1. Term feature:

The Frequency of terms or words which has maximum possible time occurred in text document can be considered as main terms or words which has been used for calculating the importance of sentence. The sum of the score of words in the sentence can be used for calculated score of each sentence. Term Feature (T_F) is defined as

$$TF\ (S_{i,k}) = \sum Term\ Weight\ (t).f\ (t, S_{i,k})$$

Where f (t,$S_{i,k}$) is the frequency of each term t in sentence $S_{i,k}$.

2. Proper noun:

The sentence that contains more proper nouns is most important which are probably included in the document summary. Proper noun feature is the ratio of the number of proper nouns that occur in sentence over the sentence length. Proper Nouns Feature (PN_F) in the sentence is defined as,

$$PN\_F\ (S_{i\,k}) = No.\ Proper\ nouns\ in\ Sentence\ /\ Sentence\ Length$$

3. Position feature:

Assume that the first sentences of a document are the most important. Therefore, we rank a sentence of document according to their position and we consider maximum positions of 3. For example, the first sentence in a document has a score value of 3/3, the second sentence has a score 2/3 and so on.

4. Sentence length

The short and very long sentences are not necessary to belong to the summary. The Sentence Length Feature (SL_F) is defined as,

$$SL\_F\ (S_{i\,k}) = (N * length\ (S_{i\,k})) / length\ (d_k)$$

Where N is the number of sentences in the articals.

5. Sentence centrality

This feature is finding a similarity between sentences. For each sentence *S*, sentence centrality feature is identify vocabulary overlap between sentence words with each other sentences words in document are also more essential in sentence scouring. The Sentence Centrality Feature (C_F) is defined as,

C_F (Si, k) = (keywords in sentence ∩ keywords in other sentence) / (keywords in sentence ∩ keywords in other sentence)

6. Cue phrase

The hypothesis of this feature is that the relevance of a sentence is computed by the presence or absences of certain cue words in the cue dictionary will be compute the significance of a sentence. Sentences containing cue phrase such as “This letter”, “this paper” ,”The proposed work”, “this report” , “develop”, “describes”, “significantly” etc are candidate sentences to be included in the summary.

C. SENTENCE SCORING:

The score of each sentence is calculating by calculated all featured weight.. The score of the sentence is the sum of all the scores of every feature. The score of the each sentence is called the rank of the sentence.

D. SENTENCE SELECTION

Finally sentences are ranked using their relevance and the highest ones are selected .This sentences are put into the summary in the order of their positions in the original document and then generated extractive summary.

E. WORD GRAPH GENERATION WITH IMPROVED EDGE WEIGHTING TECHNIQUE

Generate a weighted directed word graph from extracted summary for finding compressed sentence. A weighted directed word graph is built taking as input the generated extract, where the words represent the nodes of the graph, and the edges are adjacency relationships between two words [1].

- Weighting function[12]:

The weight of each edge is calculated using weighting function. The objective of this function is (i) To produced a grammatical compression, it favors strong links i.e. links or edges between words which are strongly associated with each other. Strong links indicates how strong the association between two words is; (ii) to generate an informative compression, and paths passing through relevant nodes.

- Dijkstra’s algorithm:

Dijkstar’s algorithm are use to find L shortest paths sentences form start to end using weighting function.

F. REDUCED INCORRECT SENT WITH ADDITIONAL RULE:

Shortest path are generated compressed sentence which are not grammatically correct so we have use some additional rules for better accuracy of sentences. These rules are:

1. The minimal length for a sentence must be 3 words, since we assume that three words (i.e., subject + verb + object) is the minimum length for a complete sentence [1].
2. Every sentence must contain a verb [1].
3. The sentence should not end in an article (E.g. a, the), a preposition (e.g. of), an interrogative word (e.g. who), nor a conjunction (e.g. and) [1].
4. Subject-Verb Agreement:

To generate a sentence to be grammatically correct, the subject and verb must both be singular or plural. That is the subject and verb must be agree with one another in their tense.

5. Run-on Sentences:

A run-on sentence is one which really includes two or more complete sentence without the proper punctuation to make separate sentence.

6. Parallel Structure:

When one or more phrase or description is used in a sentence, those phrases or descriptions should be consistent with one

ISSN: 2321-8169
4577 - 4581

another in their form and wording. Parallel structure is important because it known the ease with which the reader can follow the writer's idea.

G. INFORMATION MIXTURE

In the last step decide which of the new sentences are more appropriate. Those sentences are to be included in the final summary. Here cosine similarity measure is used to compute the similarity between two sentences. Finally, a sentence in the extract summary has an equivalent in the set of new generated sentences we take sentences form new generated sentences otherwise; we take the sentence in the extract.

IV. EXPERIMENTAL RESULTS

Performers our system is better for most of the metrics. This is due to the fact that for building the abstract oriented summaries, we rely on the sentences scored as important in the sentence scoring stage, and we compress or merge some information within them for next process. Therefore obviously the resulting summaries are shorter than the extracted summary, and since no extra information is added. One possible solution to address this issue would be to depends on the source document and generate the new sentences from it instead of the extracted summary sentences. Table 1 shows system performance based on precision value.

TABLE I. PRECISION OF SYSTEMS

Research Paper	Compendium Precision value(%)	Feature base Precisionvalue(%)
P1	61.53	69.23
P2	55.55	75
P3	69.23	83.33

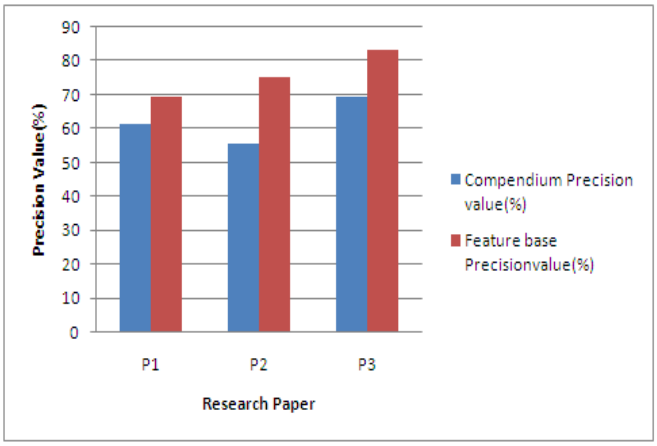


Fig.3: shows Precision base comparison between Existing and proposed system

V. CONCLUSION

In this paper, we proposed a new framework to Abstract Creation of Research Paper using Feature Specific Sentence Extraction method. In order to reduce workload of author to find out the main topics addressed in their research to generate a small paragraph Here we proposed the feature specific sentence extraction method for improving the performance of the summary generation. By using the feature specific sentence extraction method, the performance of the system is improved. Also in the proposed method both abstractive and extractive approaches have been attempted.

REFERENCES

[1] Elena Lloret, Mara Teresa Rom-Ferri, Manuel Palomar, "COMPENDIUM: A text summarization system for generating abstracts of research papers" ELSEVIER journals of Data & Knowledge Engineering 88 (2013) 164–175J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] D. Radev, T. Allison, S. Blair Goldensohn, J. Blitzer, A. Celebi, E. Drabek, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, Z. Zhang, "MEAD a platform for multi-document multilingual text summarization"' Proceedings of the 4th International Conference on Language Resources and Evaluation, 2004, pp. 699-702.

[3] C. Sauper, R. Barzilay, "Automatically generating Wikipedia articles: a structure aware approach", Proceedings of the 47th Annual Conference of the Association of Computational Linguistics, 2009, pp. 208-216.

[4] D.M. Dunlavy, D.P. O'Leary, J.M. Conroy, J.D. Schlesinger," QCS: a system for querying, clustering and summarizing documents", Information Processing and Management 43 (6) (2007) 1588-1605.

[5] M. Litvak, M. Last, M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm", Proceedings of the 48th Annual Meeting of the Association for, Computational Linguistics, 2010, pp. 927–936.

[6] L. Plaza, A. DA az, "Retrieval of similar electronic health records using UMLS concept graphs",Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems, Lecture Notes in Computer Science, vol. 6177, Springer, 2010, pp. 296-303.

[7] A. Zamanifar, O. Kashefi, and "AZOM: a Persian structured text summarizer", Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems, Lecture Notes in Computer Science, vol. 6716, Springer, 2011, pp. 234-237.

[8] [8] A. Bossard, M. Genereux, T. Poibeau," CBSEAS, a summarization system integration of opinion mining techniques to summarize blogs", Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 5-8.

[9] Yan-Xiang He, De-Xi Liu,Dong-Hong Ji, Hua Yang, Chong Teng, (2006)"MSBGA: A Multi-Document Summarization

System based on Genetic Algorithm", In Proceedings of 5th International conference on machine learning and Cybernetics.

[10] Mohamed Abdel Fattah, Fuji Ren, (2008) "Automatic Text Summarization.In Proceedings of WASET".

[11] A.Kogilavani, Dr.P.Balasubramani"clustering and feature specific sentenceextraction based summarization of multiple documents" In Proceedings of nternational journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[12] Katja Filippova,"Multi-Sentence Compression: Finding Shortest Paths in Word Graphs" Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 322–330,Beijing August 2010