Extracting Key Sentences from Text

Mudasser Iqbal[1], Muhammad Rafi[2]

[2]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi, Pakistan*

[1]mudassar.iqb@gmail.com
[2]rafi.muhammad@gmail.com

Abstract - Automatic key sentence extraction from a text is a challenging task. It has numerous applications in text processing systems. The actual task of key sentence extraction consists of three main functionalities: (i) Identification of sentence boundary in a text, (ii) a ranking function that assigns a score between (0-1) for each correctly extracted sentence based on important semantics of the text and (iii) Deciding the most relevant sentences based on evaluation function. This study carries out a survey about the state of the art in the field. It later proposed a heuristic based extraction using lexical chain of terms. The proposed approach is evaluated by using human based evaluation criteria on some existing text datasets. Encouraging results were obtained.

Keywords - Sentence Extraction, important sentences, Document Summarization, Keyword Extraction Key phrase extraction, Summarization.

I. INTRODUCTION

Key Sentence Extraction is an autonomous technique that extract important sentences from a given text corpus. The technique uses several heuristic based features to identify and extract most significant sentences from the text. It is useful in a wide variety of text processing application such as automatic summary, central idea extraction and text reduction method. The process condensing a text while maintaining legibility and securing its content and information. The major difference between human based and automatic text summarization is that humans are able to catch and guide subtle themes that spread throughout the document. Research on renewed interest has shown on automatic text summarization. By using this technique, we identify the most important sentences in a text. This technique is used to achieve many targets. For example,

1) When the text on the first page of the journal article is printed, summarization is the first objective. The extracted sentences in this case are like an ultimate form of abstract. This allows the reader to immediately determine if the article provided is of interest in the reader's field.

2) When a text is written in the collective index for a journal, indexing is the main objective. This allows the reader to instantly find a suitable article to match the readers certain needs.

3) When a search engine has a space labeled "key word," the intent is to allow the reader to make research more definite. While looking for a document that corresponds to a provided query term in the "key word" space, it will relinquish a smaller but higher character list of hits than while looking for a similar term in the full text of a document.

In conclusion "sentence extraction" runs like a filter allowing only sentences to pass which are important. The main disadvantage of implementing sentence extraction technique is the loss of information in the final outcome but still there are valuable clues in resulting summary which are sufficient to understand the document. With the increase in the number of machines which make documents readable becoming available, automatic document summarization has become one of the main inquiry topics in IR and NLP studies.

In the domain of an automatic summarization, there are two major techniques that have been used for identification of important sentences and for automatic summarization namely Abstraction and Extraction. Abstraction involves paraphrasing of the original document. In paraphrasing, the conceptual idea of a document is shown mainly. This is observational based which is accessed through particular subject areas. This pursuit based on complex awareness of a certain subject field is helpful for confined tasks such as for example, writing and assembling a summary of a 'weather forecasts'. In extraction approach, researcher finds the key sentences from the text and copy those information deemed most important to summary. In this approach, authors mainly depend upon corpus statistics. The sentence scoring process is the main purpose of this approach. Frequently, masses are given to the specific words in a text and the complete sentence scores are established on the occurrence characteristic of excessively weighted terms (keywords) in the corresponding sentences. The weighting technique term has been extensively inspected in processing retrieval information. Many other strategies like location heuristics and title information have been suggested. Though, these methods are seen to be less dependent on the domain; however, it has been stated that it is harder to form high accuracy of retrieving by applying the weighting approach terms.

Keyword frequency is another term which is based on the weighting technique. Keyword frequency is farther less reliant on the domain than other weighting techniques, hence; well-reviewed. Main advantages which are based around keyword frequency estimate on the fact that the keywords of the articles show more often in the article but show up a few times in other articles. Methods like these seem to appear effectively in many different articles which are all characterized by a massive number of words which show up often in one article but very less in other articles. In some articles; however, the same domain such as 'weather forecasts,' one may face entirely plenty of words which may show up frequently over some articles. Therefore, extracting keywords from the particular words is a major issue in the confined subject domain.

In this paper, authors are going to propose an approach which will use an extraction technique and extract the important sentences from a document / text based on scoring criteria. The basic idea of approach is extracting key sentences from a given text corpus T, which comprises of sentences **S= {s[1], s[2], s[3], s[4], s[5]……..…s[n]}** where n is the number of sentences. The extraction routine will assign a score to each of the sentences in S. The sentences will be sorted in non- increasing order and top n/3 will be selected as key sentences.

Let's take a text T which comprises of six sentences S = {S[1], S[2], S[3], S[4], S[5], S[6]}. The developed routine will assign a score to these sentences in the following manner S={s[1]=0.7, s[2]=0.4, s[3]=0.6, s[4]=0.2, s[5]=0.1, s[6]=0.12}. The key sentence can be extracted using this scoring scheme {s[1]=0.7, s[3]=0.6}. These sentences can be used as a summary as it is exactly one- third of the text.

II. LITERATURE REVIEW

There are different approaches that have been used to find a key sentence from a document which are extraction and abstraction. In extraction approach, researcher finds key sentences from the text and copy those information deemed most important to summary. While some researcher used abstraction techniques which involves paraphrasing of the original document. In paraphrasing, they mainly show the conceptual idea of a document. Different authors have proposed different approaches to extract sentences from a text. Brief work is discussed here.

Mallet et al [1] used an extraction approach and proposed the full coverage summarizer algorithm which extracts no repeated sentences from a text for summarization desire to search for a set of unnecessary sentences devoted to the data content of the initial text. In this paper, authors claimed their proposed approach will extract sentences that will "fully covers" the perception space of the document by repeatedly measuring the parallelism of every sentence to the entire document and crossing out repeated words that have previously been used. TIME Magazine collection was used in which 5% text was lost in terms of the renewal accuracy out of the 40% size of the initial text. What is even more impressive is the DUC (Document Understanding Conferences). Compared to other, DUC opponents complete coverage access is more against the others opponents. This produces summaries which are 22% the content of the initial text with only 3% loss in renewal work.

Li and Cheng [2] have proposed an algorithm called Triangle sum for key sentences from a single document based on graph theory. This algorithm build a dependency graph for the underlying document based on co-occurrence relation as well as syntactic dependency relation. This algorithm works on any single document they have built a triangle in which node represents high frequency words and edges represents the dependency or co-occurrence relation between the nodes. These extracted sentences then used to identify the sentences that are central to the topic of the document. They introduced coefficient which is computed from each node to represent the strength of connection between the nodes and its neighborhood nodes in a graph. They have described an efficient method which does not need to build any dictionary, training data or examples before the key sentences extracted from a document to extract triangles from connected graph. The goal of key sentence extraction is to identify key sentences that best summarize the main ideas of the underlying document.

Turney [3] recommended the algorithm of naturally taking out keywords from text as a supervised learning task. They consider a document as a group of phrases which the learning algorithm must grasp to sort as negative or positive cases of key phrases. The initial set of tests utilizes the C4.5 choice of tree induction algorithm to this educational task. The GenEx algorithm is another set of tests applied to a certain task. The GenEx algorithm was developed especially for this task. The test results back up the allegation that a particular learning algorithm (GenEx) can produce better key phrases than a typical learning algorithm (C4.5) and a non-educational algorithm that are applied in commercial software like Search 97 and Word 97. To remove key phrases from the document, two methods to the learning tasks were provided.

Marujo et al [4] explored the encounter of light filtering on motorized key phrases extraction (AKE) enforced to Broadcast News (BN). Words and phrases that better characterize the formation of a text are known as key phrases. Key Phrases are usually used to mark the document or as aspects in further proceedings. This makes development in AKE efficiency particularly more important. Authors guessed that AKE accuracy would improve if hardly related sentences were filtered. Their hypothesis was confirmed true by the test. Withdrawal of very little as 10% of the text sentences leads to 2% advancement in AKE preciseness and recall. The AKE is constructed over MAUI toolkit that pursues a managed learning pathway. Eight BN programs including 110 self-operated annotated new stories were used by them to train and test their AKE plan on a gold standard. The tests were set up within a Multimedia Monitoring Solution (MMS) system for radio and TV news/programs which ran every day and monitored twelve TV and 4 radio channels. AKE is a natural language procedure that selects the most relevant phrases (key phrases) from a text. Irrelevant text is removed through light filtering and AKE represents the main concept of the text.

Table 1. Analysis of Different Techniques

Paper Title	Technique and Dataset	Drawbacks
Information-Content Based Sentence Extraction for Text Summarization (2004)	Propose the FULL-COVERAGE summarizer: which parse the input document into sentences, Determine the rank of each sentence, and generate a summary using subset of the ranked sentences and return the top ranked sentences. DATASET: SMART's TIME Magazine Collection and REC3 documents used for 2002's edition of DUC	not suitable for large document
Automatic knowledge extraction from documents (2012)	Researchers presented PRISMATIC, a large-scale lexicalized relation resource that is automatically built over massive amounts of text. They have used two stage approach, First, shallow knowledge from large collections of documents is automatically extracted and representing them as frames and slots. Second, additional semantics are inferred from aggregate statistics of the automatically extracted shallow knowledge. DATASET: Large Document for an input	
Learning to Extract Key phrases from Text (2000)	They have introduced an algorithm GenEx, Their First experiment applies the C4.5 decision tree induction algorithm to the learning task, and second experiment applies GenEx algorithm. DATASET: Five Collections of documents, with a combined total of 652 documents.	
An Automatic Extraction of Key Paragraphs Based on Context Dependency (1997)	Proposed a method for extracting key paragraphs in articles based on the degree of context dependency, their proposed method assumes that the words related to theme in an article appear throughout paragraphs. DATASET: 1988, 1989 Wall Street Journal in ACL/DCI CDROM which consists of about 280,000 part-of-speech tagged sentences	Results are not good for large paragraphs. This method is restricted to specific domain (financial articles) and Experimented on data which have less sentences in a paragraph
Extracting Sentence Segments for Text Summarization: A Machine Learning Approach (2000)	Researchers presented an approach that generates a summary by extracting sentence segments. First they broke sentences in segments then used supervised learning algorithm to train the summarizer to extract important sentence segments, based on the feature vector. DATASET: U.S. patents data	
Exploring Simultaneous Keyword and Key Sentence Extraction: Improve Graph-based Ranking Using Wikipedia (2012)	Researcher proposed a two-level graph based ranking algorithm to generate summarization and extract keywords at the same time. they used Wikipedia to build a two-level concept-based graph, instead of traditional term-based graph, to express their homogenous relationship and heterogeneous relationship. DATASET: TAC 2011	Ignore words which are not found on Wikipedia
Key Sentence Extraction from Single Document based on Triangle Analysis in Dependency Graph (2013)	Authors proposed a novel algorithm, called Triangle Sum. The algorithm builds a dependency graph for the underlying document based on co-occurrence relation as well as syntactic dependency relations. In such a dependency graph, nodes represent words or phrases of high frequency, and edges represent dependency-co-occurrence relations between them. DATASET: work on any document	Only for single document, won't work for multi documents

Okazaki et al [5] proposed a method that ranks sentences by spreading activation with assuming that "sentence which are important and are relevant to many other significant sentenc- es are also significant." Their system that range sentences by extending activation by expecting that provided systems produced a similar network from text with expressed dictionary and utilizes extending activation to range the sentences. Sentences which are related multiple ones of understanding which are also meaningful. This assumption derives from "Pages which are linked (voted) from many ones of significance are also significant" [6]. For their experimentation, they have used Mainichi Newspaper articles and Summarization Task Data [5]. Table 1 describes different sentence extraction approaches.

III. PROPOSED APPROACH

Focus of this research is on extracting key sentences from a given text corpus T, which comprises of sentences S=[1],s[2].s[3]......s[n]} where n is the number of sentences. An extraction routine was developed which assigned a score between (0-1) to each sentence from S using the following function. The sentences from S were then sorted in non- increasing order and the top n/3 sentences selected as key sentences from T.

The developed function will give the following scores to sentences:

S={s[1]=0.7,s[2]=0.4s[3]=0.6.s[4]=0.2.s[5]=0.1,s[6]=0.12}

Using this scoring scheme, the key sentences {s1=0.7,s3=0.6} can be acquired. These sentences can be used as a summary. The complete scheme is described in figure 1.

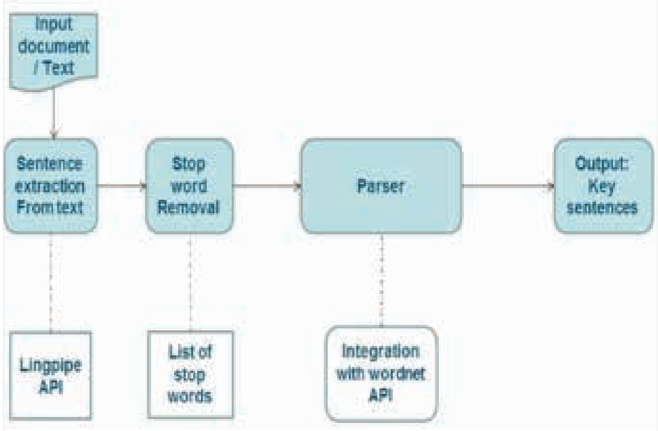

Fig. (1). Our Proposed Approach

INPUT: Any text T
OUTPUT: Important Sentences [Si]

1. Take any text as an Input.
2. Identify the boundaries of sentences Si by using LinqPipe API call. This API will return the Sentences Si.
3. Remove the stop words from each sentence. (List of stop words downloaded from internet).
4. After the stop words have been removed, all nouns and verbs from T have been identified and maintained the array of nouns "NounArray" and verbs "VerbArray".
5. The occurrence of each word have been computed from the noun and verb array, e.g. (w1:3, w2:6, w3:9, w4:8, which means w1 occurred 3 time, w2 occurred 6 times, w3 occurred 9 times, w4 occurred 8 times).
6. Wordnet API called to get the synonym set of words w[i] e.g. when Wordnet API is called for word w[1], it returned 3 synonym words (w[1],w[2].w[13]). Now the search of these words will be carried out in original text and the occurrence of these words will be computed (W[11]:3,w[12]:4,w[13]:0). So, the total occurrence of w[1] is not equal to 3, but it will be equal to W[1]=w[2]+W[1i]where i =0,1,2,3... ... n, in this case W[1] = 3 + 3+ 4 + 0 So, it can be understood that w1 word occurred 10 times in Text T. An API call has been made for each word in this way and the total occurrence of words and their synonyms have been computed.
7. The sentence counts based on the frequency have been calculated assigned to nouns and verbs of each sentence.
8. The top n/3 sentences with the highest scores have been selected and those sentences are key sentences.

IV. EXPERIMENTAL SETUP

This research work is based on experimental setup. Following steps have been taken to achieve the research objective:

A. Tools / OS

We have used following Tools in our research work.

1) Eclipse
2) LinqPipe
3) Wordnet
4) Stanford Parser

B. Procedure

Step 1:

Installation of Eclipse: Eclipse is an IDE for developing application using java programming language. All the implementation has been done in java using Eclipse IDE.

Step 2:

Integration with LinqPipe: LinqPipe .jar file have been developed from website and included in this project. This API contains many different packages and classes. The following packages have been used to identify the sentence boundaries: "com.aliasi.chunk.Chunk;
com.aliasi.sentences.MedlineSentenceModel;
com.aliasi.sentences.SentenceChunker;
com.aliasi.sentences.SentenceModel;
com.aliasi.tokenizer.IndoEuropeanTokenizerFactory;
com.aliasi.tokenizer.TokenizerFactory";

LinqPipe returned us string Array of sentences.

Step 3:

Integration with Wordnet: Wordnet API has been downloaded from website. It includes both .jar file and setup file. Setup file have been downloaded in system which basically copied a Wordnet dictionary in system and later, that dictionary have been used in project. The .jar file has been included as an external library in project and the following packages of WordNet library have been used.

“edu.smu.tspell.wordnet.SynsetType
edu.smu.tspell.wordnet.Synset
edu.smu.tspell.wordnet.WordNetDatabase
edu.smu.tspell.wordnet.WordNetException”

After integration with LinqPipe, nouns and verbs have been identified from original text T. For sentence 1, output was Contents of Nouns: [KARACHI, the, State, Bank, advance, payment, facility, importers, requirement, letters, credit, bank, guarantee]
Contents of Verbs: [has, restored, said]

Step 4:

Stanford parser has been used to check the type of input word. The type of word was required to be checked weather it is noun or verb. Wordnet API does not provide this functionality so the Stanford parser was used. The output that Stanford parser returned is following.
For sentence 1:
KARACHI/NNP :/: The/NNP State/NNP Bank/NNP has/VBZ restored/VBN advance/NN payment/NN facility/NN -LRB-/-LRB- up/IN to/TO $/$ 10000/CD -RRB-/-RRB- for/IN importers/NNS without/IN requirement/NN of/IN letters/NNS of/IN credit/NN or/CC bank/NN guarantee/NN./.
Where
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non3rd person singular present
VBZ Verb, 3rd person singular present
NN Noun, singular or mass
NNS Noun, plural
NNP Proper noun, singular
NNPS Proper noun, plural

V. RESULTS

The following result has been obtained for the text T which comprises of 15 sentences.

Input Text T: The below text T has been taken as input text T which comprises 15 (s[1],a[2]s[3]. ,,, ,,, ,,, ,,, ,s[15]) sentences and it is determined by developed routine.

“KARACHI: The State Bank has restored advance payment facility (up to $10000) for importers without requirement of letters of credit or bank guarantee. In a circular to all banks on Wednesday, it said that in order to facilitate importers to cater to their genuine small import needs, it has been decided to restore the advance payment facility up to $10000 per invoice for import of all eligible items without requirement of letters of credit or bank guarantee.”
In April 2008, “the State Bank had restricted advance payment against import invoices to specific goods or sectors only. It seems that the step has been taken in view of growing foreign exchange reserves of the country, but it has the risk of further increasing the import bill causing widening of trade imbalances”.
The Central Bank has put some condition to avail this facility provided after a gap of seven years. “The bank will take all possible measures to verify the bonfires of the importer and genuineness of the transaction while processing the advance payment request,” said the circular. “The bank will obtain an undertaking from the importer that in case goods are not received within a period of four months for any reason, the bank will recover a penalty at the rate of 1pc per month or part on the amount of advance payment from the date of remittance till the date of submission of shipping documents or repatriation of advance payment. The bank will deposit the penalty amount to the SBP on a quarterly basis along with a report on prescribed format, said the SBP. In case goods cannot be imported for any reason within the prescribed time, the bank and the importer will ensure repatriation of advance payment back. If a consistent behavior is observed on part of an importer who is unable to import goods against advance payment within four months, the bank may debar the concerned importer from making any future advance payment,” said the State Bank. “The decision is believed to have been taken in the wake of rising foreign exchange reserves (mostly because of inflows from the donor agencies, but in the presence of current account deficit and increasing imports, it could further soar the import bill. Despite higher reserves (about $16.3bn) and positive comments from Moody’s and IMF, the exchange rate is under pressure from the US dollar. Since the last week of December, the dollar has been gaining against the rupee. Currency dealers in the inter-bank market said the exchange rate has been walking on a tight rope while the State Bank has been influencing the market to keep dollar below Rs102. Since last week of December, the dollar gained Rs1.60 against the local currency. The dollar’s increasing potential to gain more is visible in the open market where the currency was traded at Rs102.60 on Wednesday. Analysts believe that the restored facility of $10,000 could be grossly misused and that could be turned into a problem for the government instead of a solution”.

Output Key Sentences:

Developed routine assigned a score to each of the sentences. On the basis of that score, top n/3 key sentences have been selected.

SENTENCE 2: In a circular to all banks on Wednesday, “it said that in order to facilitate importers to cater to their genuine small import needs, it has been decided to restore the advance

payment facility up to $10000 per invoice for import of all eligible items without requirement of letters of credit or bank guarantee.”
In April 2008,” the State Bank had restricted advance payment against import invoices to specific goods or sectors only”.

SENTENCE 9: “If a consistent behavior is observed on part of an importer who is unable to import goods against advance payment within four months, the bank may debar the concerned importer from making any future advance payment,” said the State Bank.

SENTENCE 6: “The bank will obtain an undertaking from the importer that in case goods are not received within a period of four months for any reason, the bank will recover a penalty at the rate of 1pc per month or part on the amount of advance payment from the date of remittance till the date of submission of shipping documents or repatriation of advance payment”.

SENTENCE 1: KARACHI: “The State Bank has restored advance payment facility (up to $10000) for importers without requirement of letters of credit or bank guarantee”.

SENTENCE 12: “Currency dealers in the inter-bank market said the exchange rate has been walking on a tight rope while the State Bank has been influencing the market to keep dollar below Rs102”.

VI. CONCLUSIONS

If the stated work is compared with others, it can be observed that full coverage summarizer method [1] lack the semantic consideration and proposed work does not lack semantic consideration and will produce better results. The proposed work is also not restricted to only specific domains like other work [7] (Financial domains). The proposed work will ensure that not a single word (noun / verb) will be eliminated from search while other work [8] ignores the words which do not exists on Wikipedia. The proposed work can be used to identify the important sentence extraction, for summarization purpose and for important key words extraction.

VII. ACKNOWLEDGEMENT

The authors would like to express deepest appreciation to all those who provided the possibility to complete this research work. The author appreciates the guidance provided by supervisor Muhammad Rafi. The appreciations are also for those who willingly helped authors with their abilities in conducting this study.

REFERENCES

[1] D. Mallett, J. Elding and M. A. Nascimento, "Information-Content Based Sentence Extraction for Text Summarization." In *Proceedings of International Conference on Information Technology: Coding and Computing (ITCC)*, 2004, pp: 214-218, vol. 2.

[2] K. Cheng, Y. Li and X. Wang, "Single Document Summarization Based on Triangle Analysis of Dependency Graphs" In *Proceedings of 16th International Conference on Network-Based Information Systems (NBiS)*, 2013, pp: 38-43.

[3] P. D. Turney "Learning to Extract Keyphrases from Text," *Information Retrieval*, vol. 2, no. 4, pp: 303-336, 2000.

[4] L. Marujo, R. Ribeiro, D. Martins de Matos, J. P. Neto, A. Gershman and J. Carbonell. "Key phrase extraction of lightly filtered broadcast news" in *Text, Speech and Dialogue* (Lecture Notes in Computer Science), P. Sojka, A. Horak, I. Kopecek and K. Pala, Eds. Berlin, Heidelberg: Springer, 2012, pp: 290-297.

[5] N. Okazaki, Y. Matsuo, N. Matsumura, and M. Ishizuka, “Sentence Extraction by Spreading Activation through Sentence Similarity,” *IEICE TRANSACTIONS on Information and Systems,* vol. E86-D, no. 9, pp. 1686-1694, 2003.

[6] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” *Computer Networks and ISDN Systems,* vol. 30, no. 1–7, pp. 107-117, 1998.

[7] F. Fukumoto, Y. Suzukit and J. Fukumoto: “An Automatic Extraction of Key Paragraphs Based on Context Dependency” In *Proceeding of Fifth Conference on Applied Natural Language Processing (ANLC '97),* 1997, pp: 291-298.

[8] X. Wang, L. Wang, J. Li and S. Li “Exploring Simultaneous Keyword and Key Sentence Extraction: Improve Graph-based Ranking Using Wikipedia” In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM),* Pages 2619-2622.