RESEARCH ARTICLE OPEN ACCESS

A Comparative Study on Text Summarization Methods

Fr.Augustine George[1], Dr.Hanumanthappa[2]
1Computer Science,KristuJayantiCollege,Bangalore
[2]Computer Science, Bangalore University

Abstract:

With the advent of Internet, the data being added online is increasing at enormous rate. Businesses are waiting for models that can render some useful information out of this large chunk; and hence our research holds significance. There are various statistical and NLP models that are used and each of them are efficient in their own way. Here, we will be making a comparative study so as to bring out the parameters and efficiency, thus bringing about a deduction as to when and how best we can use a particular model. Text summarization is the technique, which automatically creates an abstract, or summary of a text. The technique has been developed for many years.

Summarization is one of the research works in NLP, which concentrates on providing meaningful summary using various NLP tools and techniques. Since huge amount of information is used across the digital world, it is highly essential to have automatic summarization techniques. Extractive and Abstractive summarization are the two summarization techniques available. Lot of research work are being carried out in this area especially in extractive summarization. The techniques involved here are text summarization with statistical scoring, Linguistic Method, Graph based method, and artificial Intelligence.

Keywords — **Text Summarization, Natural Language Processing, Lexicon, Graph based.**

I. INTRODUCTION

Text summarization is a way to reduce the large amount of information into a brief form by the process of selecting important information and discarding unwanted and redundant information. It is necessary to do Automatic Text Summarization (ATS) [6] in the field of information retrieval due to the amount of textual information present in the World Wide Web (WWW). The process of summarizing a source text in to a shorter version preserving its information content called summarization. Automated summarization tools can help people to grasp main concepts of information sources in a short time. Statistical models provide a principled and mathematically sound framework in which to accomplish both of these tasks.

Automatic summary [7] can be an inductive way by collecting some parts of the original document, or in an informative way to cover all relevant information of the text.Text Summarization methods can be done in two ways. They are Extractive and abstractive summarization. An extractive summarization method works by selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form.

The importance of sentences gathered based on statistical and linguistic features of sentences.An Abstractive summarization [8][9] can be done by understanding of the main concepts in a document and then can be expressed in natural language.Extractive text summarization process further divided into two steps: Pre Processing-structured representation of the original text.Stepand processing step–extract features influencing the relevance of sentences decided and calculated and then weights assigned to these features using weight-learning method. Final score of each sentence is determined using Feature-weight equation. Preprocessing can be done using following ways Sentences boundary identification, Stop-Word Elimination and Stemming.Summary assessment [10][11][12] is a very important aspect

for text summarization. The biggestchallenge of abstractive summary is the representation problem. Systems' capabilities are constrained by the richness of their representations and their ability to generate such structures—systems cannot summarize what their representations cannot capture

II. LITERATURESURVEY

Automatic text summarization arose in the fifties and become important that suggested to weight the sentences of a document as a function of high frequency words[13], disregarding the very high frequency common words. One well-known and widely used statistical model of text is latent Dirichlet allocation [5], which is a latent variable mixture model where a document is modeled as a mixture over T clusters known as topics. Informally, a topic is a semantically focused set of words.

Informally, a topic is a semantically focused set of words. Formally, LDA represents a topic as a probability vector, or distribution, over the words in a vocabulary. Thus, the topic about "football" would give high-probability to words such as "football", "quarterback", "touchdown", etc. and give low (or zero) probability to all other non-football related words. Similarly, the topic about "traveling" would give high-probability to traveling related words, and low (or zero) probability to non-traveling related words. The following methods are also determining the sentence weights.

A. Supervised classification

The vast majority of existing work on sentence classification employs a supervised learning approach. Common classifiers include conditional random fields, naive Bayes classifiers, support vector machines, hidden Markov models and maximum entropy models.

The scope of the task refers to whether classification is performed on the abstract sentences only which is thought to be an easier task since fewer sentence types occur in the abstract– or on the entire text of the article. Alternatively, other past work has focused on a specific section within the article [2]. The second aspect in which past work differs is the annotation scheme, i.e. the set of labels used for classification. The most basic annotation scheme is modelled after the scientific method: aim, method, results, conclusion [1].

B. Semi-supervised and unsupervised classification

Guoet. al [3] use four semi-supervised classifiers for sentence classification: three variants of the support vector machine and a conditional random field model. The semi-supervised classifiers either (1) start with a small set of labeled data and choose, at each iteration, additional unlabeled data to be labeled and added to the training set (known as active learning) or (2) include the unlabeled data in the classifier formulation with an estimate of, or distribution over, the unknown labels. They perform sentence classification on biomedical abstracts using a version of the Argumentative Zones annotation scheme developed specifically for biology articles. They present experiments using only 100 labeled abstracts (approximately 700 sentences) to train the different classifiers.

Wu et al. [4] use a hidden Markov model to label sentences in scientific abstracts. They first label a set of 106 abstracts (709 sentences). They use the labeled data to extract pairs of words from sentences that are strong indicators of a particular label. They then use these word pairs and the labeled sentences to train a hidden Markov model. Again, we use less labeled data than Wu et al. Also, the annotation scheme used by Wu et al. (based on the scientific method) differs from the annotation scheme used in this paper.

C. Annotation scheme

We use an annotation scheme that is derived from Argumentative Zones [5] (AZ). There are five labels in our annotation scheme: own, contrast, basis, aim and miscellaneous. The AZ annotation scheme includes one additional label textual which describes sentences that discuss the structure of the article, e.g. "In Section 3, we show that...". We removed the label textual because it was not of obvious use for other applications. We also collapsed two of the labels in AZ – neutral and other – into one label miscellaneous. The label neutral describes sentences that refer to past work in a neutral way. The label other describes

sentences that state generally accepted background information.

III. TEXT SUMMARIZATION

A. *Steps for text summarization:*

1. ***Topic Identification:***The most prominent information in the text is identified .There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency.Methods which are based on the position of phrases are the most useful methods for topic identification.

2. ***Interpretation:***Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content.

3. ***Summary Generation:***In this step, the system uses text generation method.

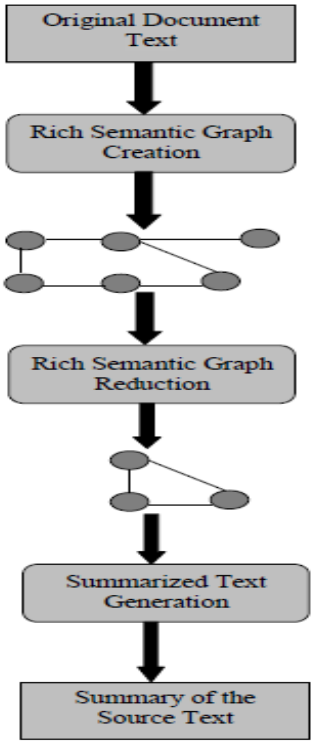

Fig.1Text Summarization

IV EXTRACTIVE TEXT SUMMARIZATION

This paper concentrates on Extractive text summarization process and is divided into two steps: Pre Processing step andprocessing step. Pre Processing is structured representation of the original text. It usually includesSentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence. b) Stop-Word Elimination—Common words with no semantics c) Stemming—The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics. It includes the following methods.

A. *Pseudo Statistical scoring methods*

1) *Title method*

This method states that sentences that appear in the title are considered to be more important and are more likely to be included in the summary. The score of the sentences is calculated as how many words are commonly used between a sentence and a title. Title method cannot be effective if the document does not include any title information.

2) *Location method*

This method states that sentences that appear in the title are considered to be more important and are more likely to be included in the summary. The score of the sentences is calculated as how many words are commonly used between a sentence and a title. Title method cannot be effective if the document does not include any title information.

3) *tf-idf method*

The term frequency-inverse document frequency is a numerical statistic which reflects how important a word is to a document. It is often used as a weighting factor in information retrieval and text mining. tf-idf is used majorly for stop words filtering in text summarization and categorization application. The tf-idf value increases proportionally to the number of times a word appears in the document. tf–idf weighting scheme are oftenused by search engines as a central tool in scoring and ranking a document's relevance given a user query.

The term frequency f(t,d) means the raw frequency of a term in a document, that i the number of times that term t occurs in document d. The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term.

4) Cue word method

Weight is assigned to text based on its significance like positive weights "verified, significant, best, this paper" and negative weights like "hardly, impossible". Cue phrases are usually genre dependent. The sentence consisting such cue phrases can be included in summary. The cue phrase method is based on the assumption that such phrases provide a "rhetorical" context for identifying important sentences. The source abstraction in this case is a set of cue phrases and the sentences that contain them. Above all statistical features are used by extractive text summarization.

Bayesian Classifier:

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S) P(s \in S)}{P(F_1, F_2 \ldots F_k)}$$

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

- Each Probability is calculated empirically from a corpus
- Higher probability sentences are chose to be in the summary

B. Linguistic Approaches

Linguistic is a scientific study of language which includes study of semantics and pragmatics. Study of semantics means how meaning is inferred from words and concepts and study of pragmatics includes how meaning is inferred from context.Linguistic approaches are based on considering the connection between the words and trying to find the main concept by analyzing the words. Abstractive text summarization is based on linguistic method, which involves the semantic processing for summarization.

Linguistic approaches have some difficulties in using high quality linguistic analysis tools (a discourse parser, etc.) and linguistic resources (Word Net, Lexical Chain, Context Vector Space, etc.).

C) Lexical chain

The concept of lexical chains was first introduced by Morris and Hirst. Basically, lexical chains exploit the cohesion among an arbitrary number of related words. Lexical chains can be computed in a source document by grouping (chaining) sets of words that are semantically related. Identities, synonyms, and hypernyms/hyponyms are the relations among words that might cause them to be grouped into the same lexical chain.

Lexical chains are used for IR and grammatical error corrections. In computing lexical chains, the noun instances must be grouped according to the above relations, but each noun instance must belong to exactly one lexical chain. There are several difficulties in determining which lexical chain a particular word instance should join. Words must be grouped such that it creates a strongest and longest lexical chain.

Generally, a procedure for constructing lexical chains follows three steps:

1. Select a set of candidate words;
2. For each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains;
3. If it is found, insert the word in the chain and update it accordingly

1) Word Net

Word Net is a on-line lexical database available for English language. It groups the English words into sets of synonyms called sys-nets. Word Net also provides a short meaning of each sys-net and semantic relation between each sys-net. Word-net also serves as a thesaurus and a on-line dictionary which is used by many systems for determining relationship between words. Thesaurus is reference

work that contains a list of words grouped together according to the similarity of meaning. Semantic relations between the words are represented by synonyms sets, hyponym trees. Word-net are used for building lexical chains according to these relations. Word Net contains more than 118,000 different word forms. LexSum is a summarization system which uses Word Net for generating the lexical chain.

D. Graph Theory

Graph theory can be applied for representing the structure of the text as well as the relationship between sentences of the document. Sentences in the document are represented as nodes. The edges between nodes are considered as connections between sentences. These connections are related by similarity relation. By developing different similarity criteria, the similarity between two sentences is calculated and each sentence is scored. Whenever a summary is to be processed all the sentences with the highest scored are chosen for the summary. In graph ranking algorithms, the importance of a vertex within the graph is iteratively computed from the entire graph.

- Higher semantic/syntactic structures
- Network (graph) based methods

1) Graph Approach

In this technique, there is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common words, in other words, their similarity is above some threshold. This representation gives two results:The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub graphs), form distinct topics covered in the documents. The second result by the graph-theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary.

Let G(V, E) be a weighted undirected graph – V - set of nodes in the graph – E - set of weighted edges • Edge weights w(u, v) define a measure of pairwise similarity between nodes u,v

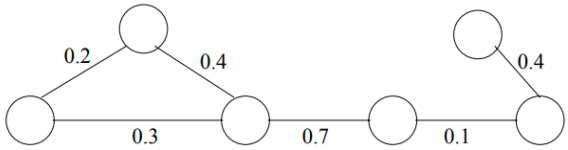

Fig.2Example of Graph-based Representations

TABLE I : GRAPH METHOD SAMPLE

Data	Directed?	Node	Edge
Web	yes	page	link
Citation Net	yes	citation	reference relation
Text	no	sent	semantic connectivity

E. Neural Network Approach

A neural network is trained on a corpus of documents. The neural network is then modified, through feature fusion, to produce a summary of highly ranked sentences in the document. Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence. The input to the neural network can be either real or binary vectors. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. We use a three-layered feed forward neural network, which has been proven to be a universal function approximate. It can discover the patterns and approximate the inherent function of any data to an accuracy of 100% as long as there are no contradictions in the data set. Our neural network consists of seven input-layer neurons, six hidden-layer neurons, and one output-layer neuron. Therefore, the unnecessary connections and neurons can be pruned without affecting the performance of the network.

Parameters	Cue method	Title method	Location method	Frequency Measures	Lexical chains	Word similarity computation
Relevance factors	Keywords, Alert-Words	Word Density Ratio	Document Structure	Individual Word count (excluding stop-words)	Homogeneity index	• Based on similarity • Based on Corpus
Accuracy	Good	Relatively good	Better than frequency measures	Comparatively lower than word-density method	Good	Average accuracy
Application	Signature-based analysis	Dynamic reporting	Dynamic reporting	Dynamic reporting	NA	Signature-based reinforcement

TABLE II :COMPARISON BASED ON PARAMETERS USED

V .CONCLUSION

This comparative survey paper is concentrating on extractive summarization methods. Extractive method is selection of important sentences from the original text based on statistical and linguistic features of sentences. Many variations of the extractive approach discussed in this paper. We found that the use of Natural Language Processing methods would provide cohesion and semantics. If texts containing multiple topics or meaning, the generated summary might not be balanced in the extractive method. Deciding proper weights of individual features is very important as quality of final summary is depending on it. The biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way. The text summarization software should produce the effective summary in less time and with least redundancy. We have shown how summarization strategies must be adapted using different methods. Our future work is to summarize web pages and journal articles, taking into account contextual information that guides sentence selection.

REFERENCES:

1. *S. Agarwal and H. Yu.Automatically classifying sentences in full-text biomedical articles into introduction, method, results and discussion.* Bioinformatics, 3(23):3174–3180, December 2009.

2. *M. A. Angrosh, S. Cranefield, and N. Stanger.Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. In Proc. of the 10th Annual Joint Conf. on Digital Libraries, pages 43–302, 2010.*

3. *Y. Guo, A. Korhonen, and T. Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In Proc. of the 2011 Conf. on Empirical Methods in Natural Lang. Proc., 2011.*

4. *W. Jien-Chen, C. Yu-Chia, H.-C.Liou, and J. Chang. Computational analysis of move structures in academic abstracts. In Proc. of the COLING/ACL on Interactive presentation sessions, COLING-ACL '06, pages 41–44, 2006.*

5. *S. Teufel and M. Moens.Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, 28(4):409–445, Dec 2002.*
6. *KarelJezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.112, ISBN 978-80-227-2827-0, FIIT STU Brarislava, UstavInformatiky a softverovehoinzinierstva, 2008*
7. *Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.*

8. *G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.*
9. *UdoHahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics , ACM, Morristown, NJ, USA , 2001.*
10. *AniNenkovaand ,RebeccaPassonneau, "Evaluating content selction in summarization: The Pyramid method", in HLT-NAACL, 145-152, 2004.*
11. *Chin-yew Lin, "A package for automatic evaluation of summaries",in Proc. ACL workshop on text summarization branches out,2004.*
12. *Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto, "Automated Summarization Evaluation with Basic Elements", In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), 2006.*
13. *H. P. Luhn, "The Automatic Creation of Literature Abstracts", Presented at IRE National Convention, New York, 159-165, 1958.*