Asian Journal of Applied Science and Technology (AJAST)
Volume 1, Issue 5, Pages 164-169, June 2017

A Novel Architecture for Agent Based Text Summarization

Dr.Amit Asthana[1], Er.Vagish Tiwari[2], Er.M.C.Pandey[3] and Er.Ankit Misra[4]

[1,2,3,4]*Swami Vivekanand Subharti University, Uttar Pradesh, India.*

Article Received: 04 June 2017 *Article Accepted: 19 June 2017* *Article Published: 23 June 2017*

ABSTRACT

Search engines deal huge volume of documents, even they output a large number of documents for a given user's query. Under these Circumstances it became very difficult for the user to find the document he actually needs, because most of the users are reluctant to make the cumbersome effort of going through each of these documents. Therefore systems that can summarize one or more documents are becoming increasingly desirable. A summary of a document is a (much) shorter text that conveys the most important information from the source document. There are a number of scripts where automatic construction of such summaries is useful.

Keywords: Agent, Text, Algorithms, Neural Network, Fuzzy and Classification.

1. INTRODUCTION

Radev et al., define a summary as –text that is produced from one or more texts that conveys important information in the original text, and that is no longer than half of the original text(s) and usually significantly less than that." This definition identifies three key factors for meaningful summaries:

The summary can be produced from one or more documents.
The summary must be less than the half of the original text.
The summary should contain important information.

Moreover, Mani defines the task of summarization as follows: –to take an information source, extract content from it & available the most important content to the user in a condensed form & in manner sensitive to the user‘s or application‘s need. Genetic Algorithm Based Approach: In Genetic Algorithm, the solutions are called individuals or chromosomes. Each run of the loop is called a generation. The quality of an individual is measured by a fitness function. To train genetic algorithm and mathematical regression models to obtain a suitable combination of feature weights. Neural Network Based Approach: A neural network is trained on a corpus of documents. The input to the neural network can be either real or binary vectors.

2. LITERATURE REVIEW

This section reviews the previous work in the area of text summarization. Interest in text summarization, arose as early as the fifties. An important paper of these days is the one in 1958,Text summarization was first studied in the late 1950s.Early works were based on the use of heuristics, such as term frequency (Luhn, 1958), lexical cues (Edmundson, 1969) and sentence location (Edmundson, 1969). Research in the late 1970s and the 1980s turned to complex text processing by exploiting techniques from artificial intelligence, including logic and production rules (Fum, Guida, & Tasso, 1985), scripts (Lehnert, 1982) and semantic networks (Reimer & Hahn, 1988).Dominant approaches since the 1990s have concentrated on finding characteristic text units with information retrieval and hybrid approaches (Hovy & Lin, 1997; Salton, Singhal, Mitra, & Buckley, 1997). Numerous large-scale competitions and workshops have been run to measure the performance of summarization systems as well. The following subsections review some approaches to extracting task.

2.1 Introduction

The advancement of information and communication technologies (ICT) has simplified the production, collection, organization, storage, and dissemination of information. On the other hand, especially with advent of internet and World Wide Web (WWW), information users are facing challenge in evaluating, filtering and selecting information that meet their information needs.

The rapid growth of the web and online electronic information services, that have supported the availability of large amount of information in a variety of format, highly initiated researches in natural language processing (NLP) field. So far, different technologies have been devised to help users to manage the problem of information overload and able to access information in multi-source, multi-format and multi-language. Text summarization is one of these technologies that help in condensing primarily textual information from one or more sources to present the most relevant information to the user.

There are many uses of summarization. It is essential for instance in order to be able to keep up with what is happening in the world. The following are some examples of uses of summarization in everyday life (Pachantouris and Dalianis, 2005).

2.2 Process of Text Summarization

According to (Alguliev and Aliguliyev, 2009) and (Moens, 1997) the process of text summarization can be decomposed in to three phases: analysis of source text, transformation, synthesis of output text. Analysis of the source text is to identify the essential content to build an internal representation. The techniques used for this task

ranges from statistical methods that search for specific key content for extraction to complex techniques that employ natural language understanding. The statistical approaches in general concerned for identification of important topic terms and the extraction of contextual sentences that contain them. On the other hand, other approaches for source analysis needs the complete understanding of the source text i.e. each sentence is processed into its propositions representing the meaning of the sentence.

3. APPROACHES OF TEXT SUMMARIZATION

Machine Learning-Based Summarization Techniques

Figure 1 illustrates the methods and techniques used in machine learning. The presented diagram is illustrative; the suggested distribution is not strict, due to the overlap between machine learning approaches. Jaqua et al. [2004] presented the ExtraNews system. They applied generation and classification to produce a very short summary from single and multiple news articles. Categorization was also proposed in the work by Diemert and Vandelle [2009]. They presented an unsupervised learning technique and cross-reference concept graphs to summarize a massive amount of knowledge derived from unstructured data (e.g., query logs and Web documents). Neural networks and Support Vector Machine (SVM) are among the methods that have been used to enhance the summarization quality by suggesting better ranking techniques.

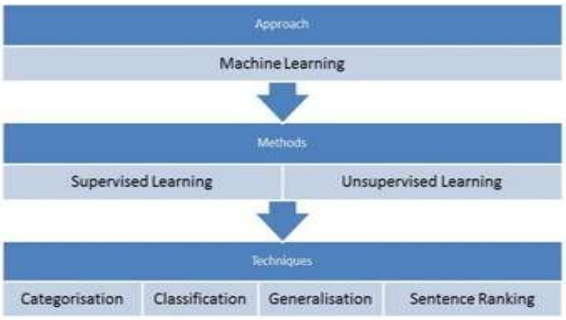

Figure 1: Machine Learning-Based Summarization Techniques

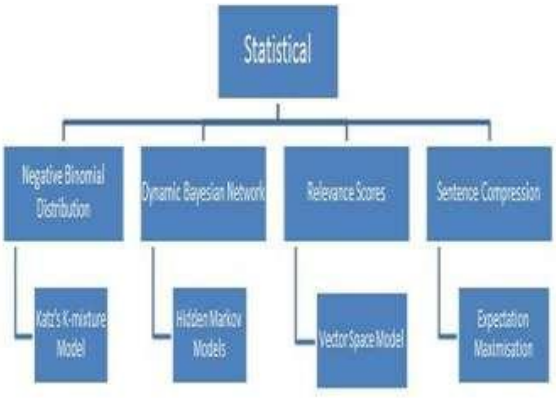

Figure 2: Statistical–Based Summarization Techniques

Statistical–Based Summarization Techniques

Figure 2 illustrates these statistical techniques, which include Hidden Markov Models, Katz's K- mixture Model [Katz, 1996], Expectation Maximization Knight and Marcu [2000] and Vector Space Model [Salton et al., 1975]. The distribution showed in the diagram is illustrative and not strict, due to the overlap between the statistical-based approaches.

Relevance scores, based on a calculation of how important the concept behind the term is to the document, are among the main factors when working with statistical-based summarization. Alguliev and Aliguliyev [2005] presented a text summarization method, which created a text summary by defining the relevance score of each sentence and extracted sentences from the original documents. The relevance score of a sentence was determined through its comparison with all the other sentences in the document and with the document title using the cosine similarity measure. The relevance scores are ranked starting with the highest score. Sentences for which the relevance score is higher than a certain threshold value are included in the summary.

Schlesinger et al. [2008] developed a system called CLASSY (Clustering, Linguistics, and Statistics for Summarization Yield).

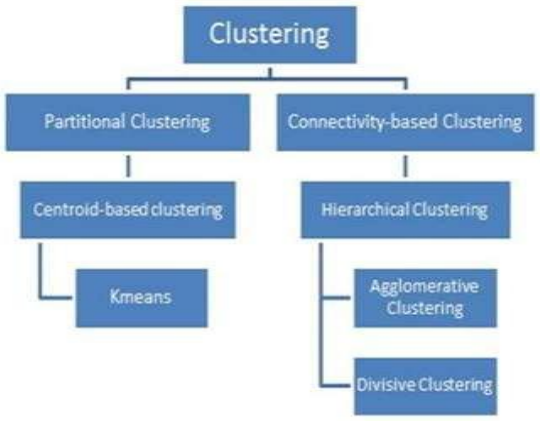

Figure 3: Cluster–Based Summarization Techniques

Cluster–based Summarization Techniques

Clustering in text summarization can be important for both selecting and extracting relevant sentences and eliminating redundancies.

Data clustering is the assignment of a set of observations into subsets, so called clusters. Clustering has received a lot of attention in the past years for improving Information Retrieval (IR) and to enhance the quality of multi-document summaries. Clustering has been applied to documents, sentences and words. As shown in Figure 3, clustering can broadly be grouped into partitional clustering and connectivity-based clustering.

As shown in Figure 4 we can see the overlap between the proposed clustering techniques. A combination of these techniques has been used in automatic text summarization to enhance the quality of the generated summaries. This method was based on clustering of sentences using language dependent techniques. A multi-document summarization technique using cluster-based link analysis. They used

three clustering detection algorithms including k-means, agglomerative and divisive clustering. They clustered sentences, using the three clustering algorithms, into different subtopics; the number of clusters was defined by taking the absolute square root of the number of all sentences in the document set.

4. EVALUATION MEASURES

The taxonomy of summary evaluation measures in [7]. Text quality is often assessed by human annotators. They assign a value from a predefined scale to each summary. For sentence extracts, it is often measured by co-selection. It finds out how many ideal sentences the automatic summary contains. Content-based measures compare the actual words in a sentence, rather than the entire sentence. Another significant group is task-based methods. They measure the performance of using the summaries for a certain task.

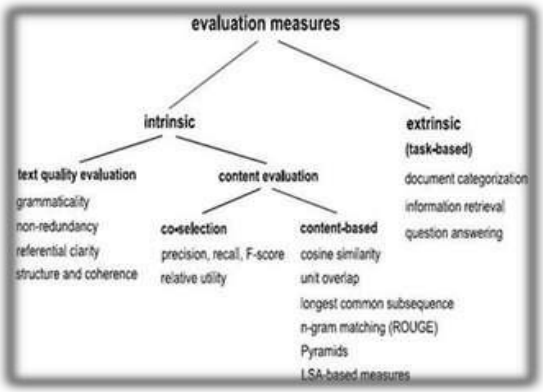

Figure 4: The Taxonomy of Summary Evaluation Measures.
Text Quality Evaluation

There are several aspects of text (linguistic) quality:

Grammaticality

This method of summary has to be no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences that make the text difficult to read. Non-Redundancy Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.

Reference Clarity

It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. Then a reference would be unclear if an entity is referenced but its identity.

Coherence and Structure

The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Co-Selection Evaluation

The main evaluation metrics of co-selection are precision, recall and f-score. Precision (P) is the number of sentences occur in system and ideal summaries divided by the number of sentences in the system. Recall denoted by R is the number of sentences coming in both system and ideal summaries divided by the number of sentences in the ideal summary. F-Score is a composite measure that combines precision and recall. The basic way how to compute the F-Score is to count a harmonic average of precision and recall:

$$F=(2*P*R)/(P+R)$$

Below is a more complex formula for measuring the F-score:

$$F=((\beta^2+1)*P*R)/(\beta^2*P+R)$$

Where β is a weighting factor that favours precision when $\beta > 1$ and favours recall when $\beta < 1$.

Content-based Evaluation

Co-selection measures can count as a match only exactly the same sentences. This ignores the fact that two sentences can contain the same information even if they are written differently. Furthermore, summaries written by two different annotators do not in general share identical sentences.

Task-based Evaluation

Task-based evaluation methods do not analyze sentences in the summary. They try to measure the prospect of using summaries for a certain task. Various approaches to task-based summarization evaluation can be found in literature. I mention the three most important tasks document categorization, information retrieval and question answering.

Document Categorization

The evaluation seeks to determine whether the generic summary is effective in capturing whatever information in the document is needed to correctly categorize the document. A corpus of documents together with the topics they belong to is needed for this task. Categorization can be performed either manually [3] or by a machine classifier [9].

If we use an automatic categorization we must keep in mind that the classifier demonstrates some inherent errors. It is often done only by comparing the system performance with the upper and lower bounds. In SUMMAC evaluation [3], apart from other tasks, 16 participating summarization systems were compared by a manual categorization task.

Given a document, which could be a generic summary or a full text source (the subject was not told which), the human subject chose a single category (from five categories, each of which had an associated topic description) to which the document is relevant, or else chose ―none of the above‖. Precision in this context is the number of correct topics assigned to a document divided by the total number of topics assigned to the document.

Information Retrieval

IR is another task for the task-based evaluation of a summary quality. Relevance correlation [50] is an IR-based measure for assessing the relative decrease in retrieval performance when

moving from full documents to summaries. Moreover, the difference between how well the summaries do and how well the full documents do should serve as a possible measure for the quality of summaries.

Suppose that given query Q and a corpus of documents D, a search engine ranks all documents in D according to their relevance to query Q. One such method is Kendall's tau and another is Spearman's rank correlation [55]. As search engines produce relevance scores in addition to rankings, we can use a stronger similarity test, linear correlation.

Question Answering
Carried out in [39], authors picked four Graduate Management Admission Test reading comprehension exercises. The exercises were multiple-choice, with a single answer to be selected from answers shown alongside each question. The authors measured how many of the questions the subjects answered correctly under different conditions.

Brevity Text Summarizer Tools
Brevity works by comparing a document to a set of similar documents. It stores this document information in a Summary Dictionary. Several dictionaries are included with Brevity. These are dictionaries designed for general categories of documents. For example to summarize a newsfeed of political news, it compares text to other political news stories of the same type.

Extractor Text Summarizer Tools
Extractor3 is a software text summarization engine. phrases found in the document together with their relative ranking (how many times was the word/phrase found in the document) along with contextual links back to the position of the key word/phrase in the document itself.

The engine returns a list of key words and Intelligent Text Summarizer Tools
It generates two summaries. Initially a summary is generated by fuzzy swarm module.

MSWord Auto Summarizer Tools
The AutoSummary Tool in Microsoft Office Word 2007 analyzes a document to identify keywords and then assign score to each word. Sentences containing the most frequent words in the document having highest scores are then selected to be included in the summary.

SweSum Text Summarizer Tools
SewSum4 is the automatic text summarizer based on statistical linguistical and heuristic Methods. The key words belong to the so called open class words. All this information is compiled and used to summarize the original text.

Pertinence Summarizer Tools
Pertinence Summarizer5 performs linguistic processing over a document and evaluates the pertinence (the relevance) of its sentences. The process takes into account not only general and/or specialized linguistic markers depending on the nature of the document analyzed, but also the user_s keywords, and optionally terminological bases, to enhance the relevance of the selected sentences.

5. PROPOSED SYSTEM

Vector Space Model
After the work of preprocessing of the whole document, we get a dictionary consisting of unique set of tokens. This dictionary can be then used to describe the characteristic features of document. In multi-document summarizer each document is converted into a numerical vector such that each cell of the vector is labeled with a word type in the dictionary and it contains its weight in the document. This weight is represented by binary value which denotes the presence or absence of the token in the document with the value 1 and 0 respectively. If the cell contains numerical value then it represents frequency (number of occurrences) of the term in the document. Thus the document is represented as an n-dimensional vector, one dimension for each possible term and hence the name [9]. We obtain a table in which the number of column is the total no of distinct word (term) and each rows correspond to the document.

It should be noted that the information about dependencies and relative position of the tokens in the document do not play any role in this representation, e.g. so "absence of light is darkness "is equivalent to "darkness is absence of light" in the vector-space model. Originally proposed by [9], vector space model is the frequently used numerical representation of text popularly used in information retrieval applications.

In single document summarization, the no of column is also representing the distinct word (term) and each rows representing the sentences. Each cell value represent whether the sentence containing that word (term) or not.

If each cell in a vector-space model is represented by term frequency (count of a type in the document) it is considered as local weighting of documents and is generally called as term frequency (tf) weighing. There are some words which occur very frequently than others. This is popularly known as the Zipf's law. This is because of the fact that there are not infinite numbers of words in a language. In 1949 in his landmark work Harvard linguist George K.

Zipf argued that the word frequency follows power law distribution $f \propto r^{\wedge}2$ with $a \approx 1$ [20], where f is the frequency of each word and r is its rank (higher frequency implies higher rank).

This law, now known as Zipf's law, states that, frequency of a word is roughly inversely proportional to its rank. To achieve this term frequency count can be weighed by the importance of a type in the whole collection.

Such weighing is called as global weighing. One of such weighing schemes is called as inverse document frequency (idf).

The motivation behind idf weighing is to reduce the importance of the words appearing in many documents and

increasing importance of the words appearing in fewer documents. Then tf model when modified with idf results in the well-known tf-idf formulation [6]. The idf of a term t is calculated as following.

$$idf(t)=log(N/N\_t)$$

In above formula N is the number of documents in the collection and indicates number of documents containing the term t. The tf-idf measure combines the weight of each term in the sentence of the document. The term frequency, number of documents and the number of documents in which the term is present and is calculated as;

$$W(t) = tf\text{-}idf(t) = tf * idf(t)$$

This vector space model provides a workspace through which we can compute various feature of each sentences.

Similarity Measures

Number of common words could be used as a measure of similarity between two texts. More sophisticated measures have been proposed which consider the number of words in common and number of words not in common and also lengths of the texts [10, 13]. Let us consider that, we want to measure similarity between two texts T1 and T2. The vocabulary consists of n terms, t1 ……...tn We use the notations tT1i and tT2i to represent the term occurrence in the text T1 and T2 respectively and can take either binary or real values.

Cosine coefficient

This is perhaps the most popular similarity measure. This measure calculates the cosine angle between two vectors in the high dimensional vector-space [1]. This is an explicit measure of similarity. It considers each document as a vector starting at the origin and the similarity between the documents is measured as the cosine of the angle between the corresponding vectors.

In the overall process, compression rate, which is defined as the ratio between the length of the summary and that of the original, is an important factor that influences the quality of the summary. While the compression rate increases, the summary will be larger; relatively, more insignificant information is contained. In fact, when the compression rate is 5–30%, the quality of the summary is acceptable [5, 6].

In our proposed method of summarization each sentence is represented as a vector of feature score, and the document is represented as matrix. This matrix is multiplied with the weight matrix computed through manually summarized text corpus to get the score of each sentences. Then according to summary factor we select the sentences in descending order of their score in their order. In statistical method [6] was described by using a Bayesian classifier to compute the probability that the sentence in the source document should be included in the summary. In [7, 8] there are various feature corresponding to the sentences measure the important of sentence in the text.

Features for Extractive Text Summarization

In this section we present various feature both for sentence level and word level which are used in calculating the importance or relevance of the sentences.

6. RESULT

Most of the summarization systems developed so far is for news articles. There are two major reasons for this: news articles are readily available in electronic format and also huge amount of news articles are produced every day. One interesting aspect of news articles is that they are written in such a way that usually most important parts are at the beginning of the text. So a very simple system that takes the required amount of leading text produces acceptable summaries. But this makes it very hard to develop methods that can produce better summaries.

Summary Evaluation

The quality of summary is varies from human to human. The summary produced by human is to select the most relevant sentence from a given document. This is different from different people. This makes the evaluation of task of automatic generated summaries is difficult and there is no standard available.

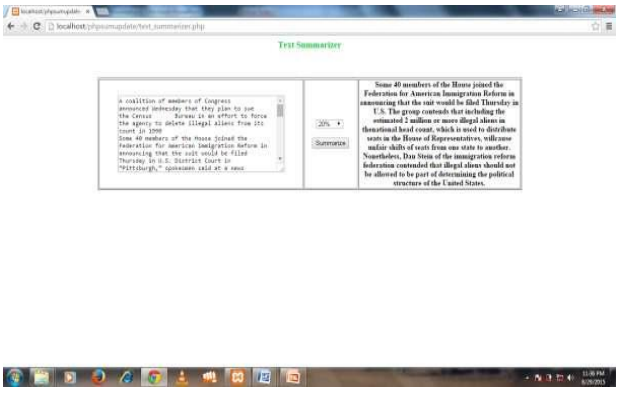

Figure 5: Snapshot of Generated Summary

7. CONCLUSION & FUTURE WORK

In our work we addressed a new extractive text summarization technique, for single documents based on Feature Extraction. We used text processing approaches as opposed to semantic approaches related to natural language. To calculate the similarity we use the well-known tf*idf model of document representation. Such graphical representation gives us a way to calculate sentence importance. Our work does not need natural language processing resources apart from a word and sentence boundary parsers and a stemmer (optional). Thus the method can be extended to other languages with little modifications.

In our system we have come up with arbitrary weights by trial and error method. We plan to implement machine learning techniques to learn these weights automatically from training data. We would like to use NLP tools such as word sense disambiguation and co-reference resolution module to obtain precise weights for the sentences in the document we also plan to extend this system to perform deeper semantic analysis of the text and add more feature to our ranking function. We

would like to extend this system for multi document summarization. Semantic information such as word sense can be utilized. Same word can mean different things in different contexts. Use of word sense information can lead to better similarity calculations. Same word can be used in different senses in different context. So using the correct word sense can lead to better similarity measurements. A more sophisticated representation that single words can be explored. A first step towards this aim could be use of multi-word units. Multi-word units can be recognized using statistical techniques. Also syntactic information such as Part-of- Speech (POS) tags might help to improve performance of the extraction algorithm.

REFERENCES

[1] Abera N. (1988), "Long vowels in Afan Oromo: A generic approach", *Master's thesis School of graduate studies, Addis Ababa University, Ethiopia.*

[2] Atif Khan, Naomie Salim,"A Review on Abstractive Summarization Methods", *Journal of Theoretical and Applied Information Technology*, 10th January 2014. Vol. 59 No.1.

[3] Aristoteles, Yeni Herdiyeni, Ahmad Ridha and Julio Adisantoso4," Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algorithm", *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 3, No 1, May 2012ISSN (Online): 1694-0814.

[4] Dipti.D.Pawar, M.S.Bewoor, S.H.Patil, ―Text Rank: "A Novel Concept for Extraction Based Text Summarization", *International Journal of Computer Science and Information Technologies*, Vol. 5 (3), 2014, 3301 – 3304.

[5] Farmin T and Chrzanowski M.J (1999), "An evaluation of text summarization Systems", 1999.

[6] M.Fachrurrozi, Novi Yusliani, and Rizky Utami Yoanita,―"Frequent Term based Text Summarization for Bahasa Indonesia", *International Conference on Innovations in Engineering and Technology (ICIET'2013) Dec. 25-26, 2013 Bangkok (Thailand).*

[7] Martin Hassel, "Evaluation of Automatic Text Summarization", KTH Numerisk analysis och datalogiSE-100 44 Stockholm Sweden.

[8] Nikita Munot, Sharvari S. Govilkar,"Comparative Study of Text Summarization Methods*", International Journal of Computer Applications (0975 – 8887)* Volume 102– No.12, September 2014.

[9] Nina Tahmasebi, Devdatt Dubhashi, "Extractive Summarization using Continuous Vector Space Models", *SE-412 96, Goteborg.*

[10] Palakorn Achananuparp, Xiaohua Hu, and Shen Xiajiong, "The Evaluation of Sentence Similarity Measures", *Drexel University, Philadelphia, PA 19104.*

[11] Regina Barzilay and Kathleen R. McKeown, "Information Fusion in the Context of Multi-Document", *New York, NY 10027, USA.*

[12] Ramesh Vaishya, Surya Prakash Tripathi, "Strategic Approach for Automatic Text Summarization", *International Journal of Computer Science and Information Security*,Vol. 9, No.5, May 2011.

[13] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, August 2010.