International Journal of Advances in Applied Sciences (IJAAS)
Vol. 4, No. 1, March 2015, pp. 6~12
ISSN: 2252-8814

An Integrated Approach for Compendium Generator using Customized Algorithms

M. Suman, Tharun Maddu, M. Mohan
Department of Electronics and Computer Engineering, K.L. University, India

Article Info

Article history:

Received Dec 12, 2014
Revised Feb 15, 2015
Accepted Feb 28, 2015

Keyword:

Coherence
Lexical similarity
Redundancy
Sentence position
Sentence resemblance
Summarizer

ABSTRACT

Text Summarization is a process that is to give the shorter version of a text document. For many research scholars who want to do their research on a specific domain has to search a lot of documents on that topic related to a specific domain. It is also difficult to go through the lot of the research papers present in that particular domain which takes a lot of time at this moment of time there are lots of chances in missing some key words present in those research papers. So that Summarizer is used to give the summary of a paper. The aim of our project is to reduce the body of the text and maintaining coherence and avoiding redundancy. Winnowing is an algorithm that gives the coherence between the multiple papers when multiple papers are given as the input. Redundancy that is the repeated words or sentences can be avoided using the MMR algorithm.

Corresponding Author:

M. Suman,
Department of Electronics and Computer Engineering,
K.L. University,
Vaddeswaram, Guntur.
Email: suman.maloji@kluniversity.in

1. INTRODUCTION

The rapid growth of the Internet yielded a massive increase of the amount of information available, especially regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.). Due to the huge volume of information in the Internet, it has become unfeasible to efficiently sieve useful information from the huge mass of documents. Thus, it is necessary to use automatic methods to understand, index, classify and present all information in a clear and concise way, allowing users to save time and resources. The need for a tool that takes a text and shortens it into a brief and succinct summary has never been greater than now. With the huge amount of information on the internet and the necessity to get the essential of this information in a short time, the need for summarizers becomes everyday pressing, especially, for people with special needs like blind or elderly people. For those people it is vital to go directly to the essential information rather than having to read through many passages. One solution is use text summarization techniques. Text summarization (TS) is the process of automatically creating a compressed version of one or more documents. It attempts to get the meaning of documents. Essentially, TS techniques are classified as Extractive and Abstractive. Extractive summaries produce a set of the most significant sentences from a document, exactly as they appear. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. It may even produce new sentences to the summary. Currently, the extractive summaries are commonly used because they are easier to create. Extractive methods are usually performed in three steps.

1. Create an intermediate representation of the original text,
2. Sentence scoring,
3. Select high scores sentences to the summary.

The first step creates a representation of the document. Usually, it divides the text into paragraphs, sentences, and tokens. Sometimes some preprocessing, such as stop word removal is also performed. The second step tries to determine which sentences are important to the document or to which extent it combines information about different topics, by sentence scoring. The score should be a measure of how significant a sentence is to the understanding of the text as a whole. The last step combines the score provided by the previous steps and generates a summary.

In order to be able to make going through IEEE papers a lot easier and a lot more effective, the compendium generator analyses the paper and shows the user details for him/her and comprehend what the paper is about. It allows the user to save this short summary in case multiple papers are being referred to. This makes it simple to keep a track of all references. Using an algorithm that combines TF/IDF, Cue-Phrases, and Resemblance to title, results are proven to be most effective. The order of the sentences are kept intact. The tool also allows the user to compare two or more papers giving an output of a joint non redundant summary, which can form the basis for a new paper. It helps us to determine coherence or how strongly the papers pertaining to the same domain are linked.

Fingerprints are generated to check how strong the relevance between two documents is. Winnowing algorithm is used to determine this. These are methods used to determine plagiarism, with a degree of modification it has been used to determine degree of relevance.

2. LITERATURE SURVEY

There are plenty of summarizers available. The online summarizers do not prove to be very effective as only sentences with more no of words are chosen, not necessarily the sentences with keywords or important sentences that resemble the title of the document. 'A Context Based Text Summarization System', explains how combining algorithms can provide more effective results [2]. Depending on the context, however, some techniques may yield better results than some others. 'Assessing sentence scoring techniques for extractive text summarization' proposes a new summarization system that easily combines different sentence scoring methods in order to obtain the best summaries depending on the context [4]. The fifteen sentence scoring methods most widely used and referenced in the technical literature in the last 10 years are applied to single document summarization. Both quantitative and qualitative measures are used to evaluate which combination of the sentence scoring methods yield better results for each context. Combining 3 to 5 specific sentences scoring methods in a certain context provides much better quality results.

The choice of those methods depend on context of the document. 'Get Only the Essential information: Text summarizer based on implicit data' was used to experiment and determine the best possible combination to summarize papers [1]. Thereby creating a customized algorithm including, Cue-Phrases, Resemblance to title and TF/IDF drastically improves accuracy. This helps us to summarize a single document without missing any important sentences and the context of the paper is also preserved. Recent research in multi-document summarization has focused on removing redundancy and statistic approaches in machine learning and language modeling to find important sentences and words in multiple documents. 'A Contextual Query Expansion Based Multi-document Summarizer for Smart Learning', provides insight on how redundancy can be removed using a technique called Maximum Marginal Relevance (MMR) [6]. This technique is proposed as a relatively better approach to tackle redundancy. [3]'A survey of text summarization techniques' explains that Precision is defined as the percentage of the relevant items in the returned set and Recall is the percentage of the relevant items in the returned set compared to those in the collection. If the whole collection is retrieved, then the Recall is maximum, but Precision is low. Most search engines suffer from this problem (high Recall and low Precision).

If search engines search only a documents primary ideas, instead of every word, then Recall will likely not be decreased but Precision will likely improve. Hence, an automated facility for summarizing documents to improve productivity is desirable. A good summarization system should include only sentences that are most important to a documents theme; it must also cover all documents topics. Using a summary instead of the whole documents as a representative of what the documents are about would mean processing a fraction (20 percent or less) of the documents text, yet yield better precision and lesser processing time. In order to determine the requirements of a good summarization system, many text summarization approaches were reviewed. An in-depth review of text summarization literature was conducted and results from this study along with a description of each algorithm. Coherence 'Winnowing: Local Algorithms for Document Fingerprinting' provides insight on plagiarism detection techniques. A technique to generate unique values for chunks of text [5].

3. PROPOSED SYSTEM

To design a compendium generator there are some specifications such as functional specifications and program specifications.

3.1. Functional Specifications

1. The compendium generator mainly aims to generate important sentences after passing through the document. Also when two or more academic papers are given as input then a combined non redundant summary is generated
2. By creating a customized algorithm that drastically improves accuracy of the summary. This helps us summarize a single document without missing any important sentences and preserving the context of the paper.
3. Maintaining correlation with the main idea, is key to providing the ideal summary. Thus multiple documents belonging to the same domain can be summarized.

3.2. Program Specifications

3.2.1. Tokenizer

1. Every word needs to be split into individual tokens, every word becomes a token.
2. PUNKT module in NLTK is used for this.

3.2.2. Stop Removal

1. NLTK stopwords package is used to remove stop words.
2. This helps improve calculation of word frequency.

3.2.3. Stemmer and Lemmatizer

1. An inbuilt lemmatizer called Wordnet is used.
2. The Stemmer used is Snowball stemmer.

3.2.4. Cue-Phrase

1. A corpus of cue phrases that are most commonly used in research papers is created.
2. In summary, in conclusion, our investigation, the paper describes, etc. are a few examples.

3.2.5. Resemblance to Title

1. A list that stores the title is created and sentences that have resemblance to these words are ranked higher.
2. This helps maintain the core essence of the paper.

3.2.6. TF-IDF

1. A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus
2. It uses the most no of occurrences as an upper end value. The other frequencies are compared to this value.
3. A custom combination of these three algorithms ranks sentences aptly for academic research papers.

3.2.7. Sentence Selection

The sentences which have a rank above the threshold rank are selected.

3.2.8. Redundancy Removal

1. Maximum Marginal Relevance algorithm is used to remove redundancy.
2. A combined non redundant summary is generated for multiple documents.

3.2.9. Fingerprinting

1. Created a hash value function using length of finger print as 20. This is an ideal number as it is low enough to provide accurate results. It is large enough to be computable.
2. A formula from the paper is used to generate unique fingerprints.

3.2.10. Winnowing

An algorithm primarily used to detect plagiarism modified to determine relevance between documents. Used to identify level of coherence between documents based on the fingerprints matched.

4. IMPLEMENTATION

4.1. Text Segmentation

Three main processes take place in this module.

4.1.1. Tokenization

Splitting a sentence into individual words. NLTK PUNKT is used.

4.1.2. Lemmatization

Converting a word to its root form. E.g. says, said, saying will all map to root form – say.

4.1.3. Stemmer

It is similar to a lemmatize, but it stems a word rather than get to the root form. eg. Laughed, laughing will stem to laugh. However, said, saying will map to sa - which is not particularly enlightening in terms of what,"sa" means. Stop word removal also takes place where constantly repeated words are removed.

4.2. Sentence Ranking

Since the words are tokenized, they are now ranked according to Cue Phrase, Sentence Position and Resemblance to title algorithms.

4.2.1. Cue Phrase

Cue-Phrases: In general, the sentences started by in summary, in conclusion, our investigation, the paper describes and emphasizes such as the best, the most important, according to the study, significantly, important, in particular, hardly, impossible as well as domain-specific bonus phrases terms can be good indicators of significant content of a text document.

4.2.2. TF-IDF

TFIDF, short for term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It uses the most no of occurrences as an upper end value. The other frequencies are compared to this value.

4.3. Sentence Selection

Sentences with rank above threshold frequency are selected.

4.4. Redundancy Removal

As multiple documents are being summarized, some documents may have points that are repeated. When a combined summary of all the documents is being displayed this redundancy continues. MMR algorithm is used to get rid of this redundancy.

4.5. Fingerprinting

Fingerprinting is a technique used to detect Plagiarism in academic documents. This method forms representative digests of documents by selecting a set of multiple substrings (n-grams) from them. So the first step is to do a text segmentation as matches should be unaffected by extra space, capitals and punctuation, etc. Then k-grams are formed where k is 20. It is found to be the ideal value.

4.6. Winnowing

This helps understand how strongly various papers pertaining to a single domain are linked. It gives us a good perspective of how the data can be organized and used. Level of similarity that needs to be matched is given a value. A lower threshold would be a noise threshold that determines if there's some amount of similarity between the documents being compared. From there on thresholds are set at custom points that determine similarity.

5. RESULTS

5.1. Module 1

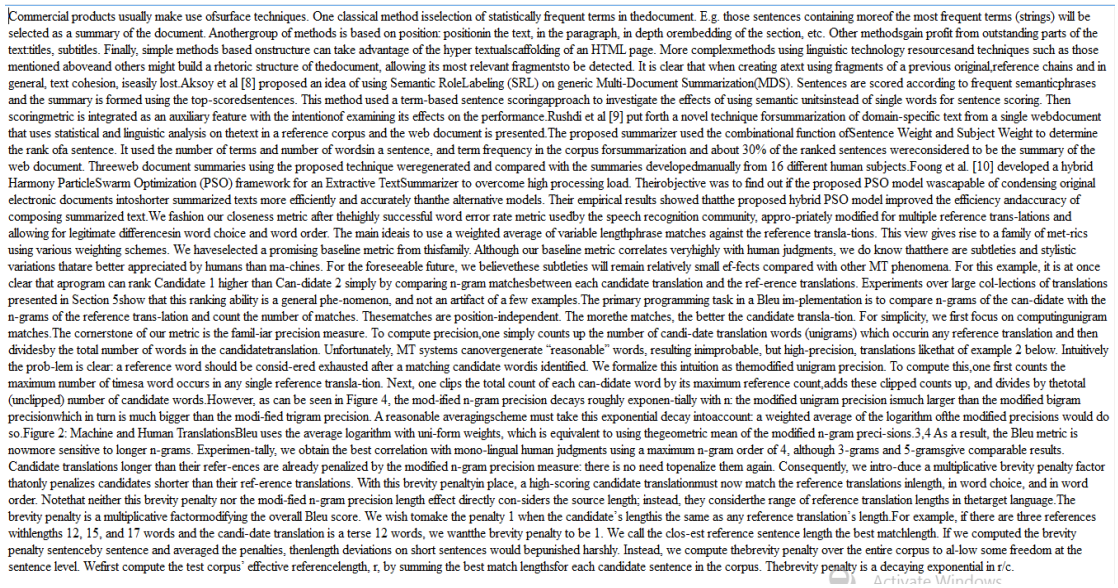
Summarization for the single or multiple IEEE papers. Enter the number of papers to summarize.

Inputs:

```
jarvis@jarvis-Inspiron-N5110:~$ cd Project
jarvis@jarvis-Inspiron-N5110:~/Project$ python base2.py
please enter no of files2
2
jarvis@jarvis-Inspiron-N5110:~/Project$
```

Figure 1. To enter the number of papers

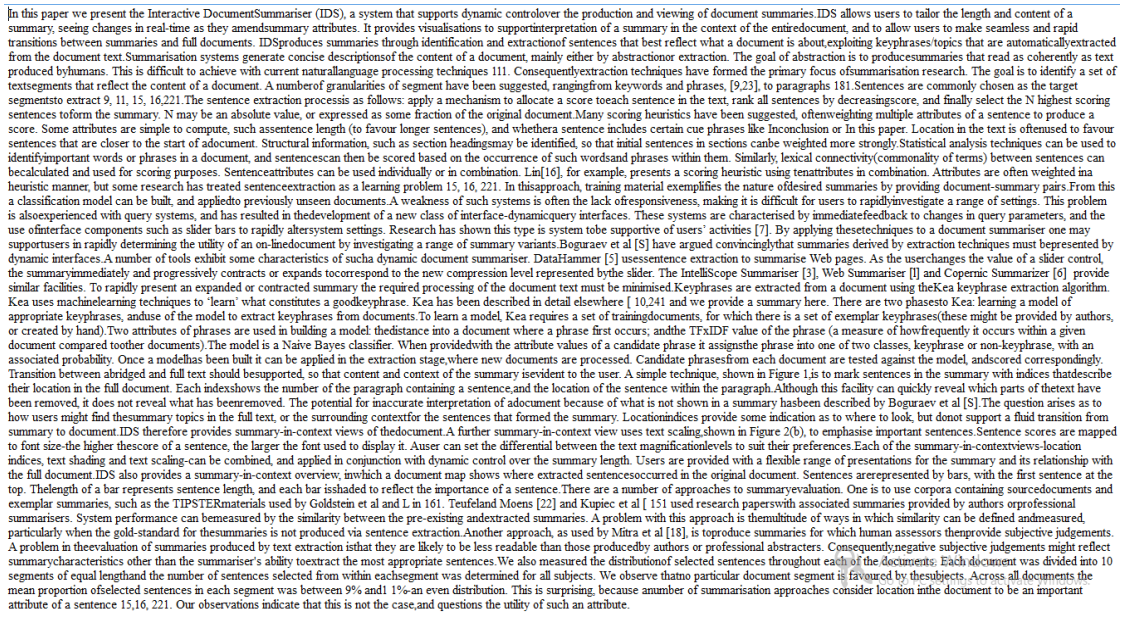
Paper 1:



Commercial products usually make use ofsurface techniques. One classical method isselection of statistically frequent terms in thedocument. E.g. those sentences containing moreof the most frequent terms (strings) will be selected as a summary of the document. Anothergroup of methods is based on position: positionin the text, in the paragraph, in depth orembedding of the section, etc. Other methodsgain profit from outstanding parts of the text:titles, subtitles. Finally, simple methods based onstructure can take advantage of the hyper textualscaffolding of an HTML page. More complexmethods using linguistic technology resourcesand techniques such as those mentioned aboveand others might build a rhetoric structure of thedocument, allowing its most relevant fragmentsto be detected. It is clear that when creating atext using fragments of a previous original,reference chains and in general, text cohesion, iseasily lost.Aksoy et al [8] proposed an idea of using Semantic RoleLabeling (SRL) on generic Multi-Document Summarization(MDS). Sentences are scored according to frequent semanticphrases and the summary is formed using the top-scoredsentences. This method used a term-based sentence scoringapproach to investigate the effects of using semantic unitsinstead of single words for sentence scoring. Then scoringmetric is integrated as an auxiliary feature with the intentionof examining its effects on the performance.Rushdi et al [9] put forth a novel technique forsummarization of domain-specific text from a single webdocument that uses statistical and linguistic analysis on thetext in a reference corpus and the web document is presented.The proposed summarizer used the combinational function ofSentence Weight and Subject Weight to determine the rank ofa sentence. It used the number of terms and number of wordsin a sentence, and term frequency in the corpus forsummarization and about 30% of the ranked sentences wereconsidered to be the summary of the web document. Threeweb document summaries using the proposed technique weregenerated and compared with the summaries developedmanually from 16 different human subjects.Foong et al. [10] developed a hybrid Harmony ParticleSwarm Optimization (PSO) framework for an Extractive TextSummarizer to overcome high processing load. Theirobjective was to find out if the proposed PSO model wascapable of condensing original electronic documents intoshorter summarized texts more efficiently and accurately thanthe alternative models. Their empirical results showed thatthe proposed hybrid PSO model improved the efficiency andaccuracy of composing summarized text.We fashion our closeness metric after thehighly successful word error rate metric usedby the speech recognition community, appro-priately modified for multiple reference trans-lations and allowing for legitimate differencesin word choice and word order. The main ideais to use a weighted average of variable lengthphrase matches against the reference transla-tions. This view gives rise to a family of met-rics using various weighting schemes. We haveselected a promising baseline metric from thisfamily. Although our baseline metric correlates veryhighly with human judgments, we do know thatthere are subtleties and stylistic variations thatare better appreciated by humans than ma-chines. For the foreseeable future, we believethese subtleties will remain relatively small ef-fects compared with other MT phenomena. For this example, it is at once clear that aprogram can rank Candidate 1 higher than Can-didate 2 simply by comparing n-gram matchesbetween each candidate translation and the ref-erence translations. Experiments over large col-lections of translations presented in Section 5show that this ranking ability is a general phe-nomenon, and not an artifact of a few examples.The primary programming task in a Bleu im-plementation is to compare n-grams of the can-didate with the n-grams of the reference trans-lation and count the number of matches. Thesematches are position-independent. The morethe matches, the better the candidate transla-tion. For simplicity, we first focus on computingunigram matches.The cornerstone of our metric is the famil-iar precision measure. To compute precision,one simply counts up the number of candi-date translation words (unigrams) which occurin any reference translation and then dividesby the total number of words in the candidatetranslation. Unfortunately, MT systems canovergenerate "reasonable" words, resulting inimprobable, but high-precision, translations likethat of example 2 below. Intuitively the prob-lem is clear: a reference word should be consid-ered exhausted after a matching candidate wordis identified. We formalize this intuition as themodified unigram precision. To compute this,one first counts the maximum number of timesa word occurs in any single reference transla-tion. Next, one clips the total count of each can-didate word by its maximum reference count,adds these clipped counts up, and divides by thetotal (unclipped) number of candidate words.However, as can be seen in Figure 4, the mod-ified n-gram precision decays roughly exponen-tially with n: the modified unigram precision ismuch larger than the modified bigram precisionwhich in turn is much bigger than the modi-fied trigram precision. A reasonable averagingscheme must take this exponential decay intoaccount: a weighted average of the logarithm ofthe modified precisions would do so.Figure 2: Machine and Human TranslationsBleu uses the average logarithm with uni-form weights, which is equivalent to using thegeometric mean of the modified n-gram preci-sions.3,4 As a result, the Bleu metric is nowmore sensitive to longer n-grams. Experimen-tally, we obtain the best correlation with mono-lingual human judgments using a maximum n-gram order of 4, although 3-grams and 5-gramsgive comparable results. Candidate translations longer than their refer-ences are already penalized by the modified n-gram precision measure: there is no need topenalize them again. Consequently, we intro-duce a multiplicative brevity penalty factor thatonly penalizes candidates shorter than their ref-erence translations. With this brevity penaltyin place, a high-scoring candidate translationmust now match the reference translations inlength, in word choice, and in word order. Notethat neither this brevity penalty nor the modi-fied n-gram precision length effect directly con-siders the source length; instead, they considerthe range of reference translation lengths in thetarget language.The brevity penalty is a multiplicative factormodifying the overall Bleu score. We wish tomake the penalty 1 when the candidate's lengthis the same as any reference translation's length.For example, if there are three references withlengths 12, 15, and 17 words and the candi-date translation is a terse 12 words, we wantthe brevity penalty to be 1. We call the clos-est reference sentence length the best matchlength. If we computed the brevity penalty sentenceby sentence and averaged the penalties, thenlength deviations on short sentences would bepunished harshly. Instead, we compute thebrevity penalty over the entire corpus to al-low some freedom at the sentence level. Wefirst compute the test corpus' effective referencelength, r, by summing the best match lengthsfor each candidate sentence in the corpus. Thebrevity penalty is a decaying exponential in r/c.

Figure 2. IEEE paper 1 as Input

Paper 2:



In this paper we present the Interactive DocumentSummariser (IDS), a system that supports dynamic controlover the production and viewing of document summaries.IDS allows users to tailor the length and content of a summary, seeing changes in real-time as they amendsummary attributes. It provides visualisations to supportinterpretation of a summary in the context of the entiredocument, and to allow users to make seamless and rapid transitions between summaries and full documents. IDSproduces summaries through identification and extractionof sentences that best reflect what a document is about,exploiting keyphrases/topics that are automaticallyextracted from the document text.Summarisation systems generate concise descriptionsof the content of a document, mainly either by abstractionor extraction. The goal of abstraction is to producesummaries that read as coherently as text produced byhumans. This is difficult to achieve with current naturallanguage processing techniques 111. Consequentlyextraction techniques have formed the primary focus ofsummarisation research. The goal is to identify a set of textsegments that reflect the content of a document. A numberof granularities of segment have been suggested, rangingfrom keywords and phrases, [9,23], to paragraphs 181.Sentences are commonly chosen as the target segmentsto extract 9, 11, 15, 16,221.The sentence extraction processis as follows: apply a mechanism to allocate a score toeach sentence in the text, rank all sentences by decreasingscore, and finally select the N highest scoring sentences toform the summary. N may be an absolute value, or expressed as some fraction of the original document.Many scoring heuristics have been suggested, oftenweighting multiple attributes of a sentence to produce a score. Some attributes are simple to compute, such assentence length (to favour longer sentences), and whethera sentence includes certain cue phrases like Inconclusion or In this paper. Location in the text is oftenused to favour sentences that are closer to the start of adocument. Structural information, such as section headingsmay be identified, so that initial sentences in sections canbe weighted more strongly.Statistical analysis techniques can be used to identifyimportant words or phrases in a document, and sentencescan then be scored based on the occurrence of such wordsand phrases within them. Similarly, lexical connectivity(commonality of terms) between sentences can becalculated and used for scoring purposes. Sentenceattributes can be used individually or in combination. Lin[16], for example, presents a scoring heuristic using tenattributes in combination. Attributes are often weighted ina heuristic manner, but some research has treated sentenceextraction as a learning problem 15, 16, 221. In thisapproach, training material exemplifies the nature ofdesired summaries by providing document-summary pairs.From this a classification model can be built, and appliedto previously unseen documents.A weakness of such systems is often the lack ofresponsiveness, making it is difficult for users to rapidlyinvestigate a range of settings. This problem is alsoexperienced with query systems, and has resulted in thedevelopment of a new class of interface-dynamicquery interfaces. These systems are characterised by immediatefeedback to changes in query parameters, and the use ofinterface components such as slider bars to rapidly altersystem settings. Research has shown this type is system tobe supportive of users' activities [7]. By applying thesetechniques to a document summariser one may supportusers in rapidly determining the utility of an on-linedocument by investigating a range of summary variants.Boguraev et al [S] have argued convincinglythat summaries derived by extraction techniques must bepresented by dynamic interfaces.A number of tools exhibit some characteristics of sucha dynamic document summariser. DataHammer [5] usessentence extraction to summarise Web pages. As the userchanges the value of a slider control, the summaryimmediately and progressively contracts or expands tocorrespond to the new compression level represented bythe slider. The IntelliScope Summariser [3], Web Summariser [l] and Copernic Summarizer [6] provide similar facilities. To rapidly present an expanded or contracted summary the required processing of the document text must be minimised.Keyphrases are extracted from a document using theKea keyphrase extraction algorithm. Kea uses machinelearning techniques to 'learn' what constitutes a goodkeyphrase. Kea has been described in detail elsewhere [10,241 and we provide a summary here. There are two phasesto Kea: learning a model of appropriate keyphrases, anduse of the model to extract keyphrases from documents.To learn a model, Kea requires a set of trainingdocuments, for which there is a set of exemplar keyphrases(these might be provided by authors, or created by hand).Two attributes of phrases are used in building a model: thedistance into a document where a phrase first occurs; andthe TFxIDF value of the phrase (a measure of howfrequently it occurs within a given document compared toother documents).The model is a Naive Bayes classifier. When providedwith the attribute values of a candidate phrase it assignsthe phrase into one of two classes, keyphrase or non-keyphrase, with an associated probability. Once a modelhas been built it can be applied in the extraction stage,where new documents are processed. Candidate phrasesfrom each document are tested against the model, andscored correspondingly. Transition between abridged and full text should besupported, so that content and context of the summary isevident to the user. A simple technique, shown in Figure 1,is to mark sentences in the summary with indices thatdescribe their location in the full document. Each indexshows the number of the paragraph containing a sentence,and the location of the sentence within the paragraph.Although this facility can quickly reveal which parts of thetext have been removed, it does not reveal what has beenremoved. The potential for inaccurate interpretation of adocument because of what is not shown in a summary hasbeen described by Boguraev et al [S].The question arises as to how users might find thesummary topics in the full text, or the surrounding contextfor the sentences that formed the summary. Locationindices provide some indication as to where to look, but donot support a fluid transition from summary to document.IDS therefore provides summary-in-context views of thedocument.A further summary-in-context view uses text scaling,shown in Figure 2(b), to emphasise important sentences.Sentence scores are mapped to font size-the higher thescore of a sentence, the larger the font used to display it. Auser can set the differential between the text magnificationlevels to suit their preferences.Each of the summary-in-contextviews-location indices, text shading and text scaling-can be combined, and applied in conjunction with dynamic control over the summary length. Users are provided with a flexible range of presentations for the summary and its relationship with the full document.IDS also provides a summary-in-context overview, inwhich a document map shows where extracted sentencesoccurred in the original document. Sentences arerepresented by bars, with the first sentence at the top. Thelength of a bar represents sentence length, and each bar isshaded to reflect the importance of a sentence.There are a number of approaches to summaryevaluation. One is to use corpora containing sourcedocuments and exemplar summaries, such as the TIPSTERmaterials used by Goldstein et al and L in 161. Teufeland Moens [22] and Kupiec et al [151 used research paperswith associated summaries provided by authors orprofessional summarisers. System performance can bemeasured by the similarity between the pre-existing andextracted summaries. A problem with this approach is themultitude of ways in which similarity can be defined andmeasured, particularly when the gold-standard for thesummaries is not produced via sentence extraction.Another approach, as used by Mitra et al [18], is toproduce summaries for which human assessors thenprovide subjective judgements. A problem in theevaluation of summaries produced by text extraction isthat they are likely to be less readable than those producedby authors or professional abstracters. Consequently,negative subjective judgements might reflect summarycharacteristics other than the summariser's ability toextract the most appropriate sentences.We also measured the distributionof selected sentences throughout each of the documents. Each document was divided into 10 segments of equal lengthand the number of sentences selected from within eachsegment was determined for all subjects. We observe thatno particular document segment is favoured by thesubjects. Across all documents the mean proportion ofselected sentences in each segment was between 9% and1 1%-an even distribution. This is surprising, because anumber of summarisation approaches consider location inthe document to be an important attribute of a sentence 15,16, 221. Our observations indicate that this is not the case,and questions the utility of such an attribute.

Figure 3. IEEE paper 2 as Input

Output:

This method used a term-based sentence scoringapproach to investigate the effects of using semantic unitsinstead of single words for sentence scoring.Rushdi et al [9] put forth a novel technique forsummarization of domain-specific text from a single webdocument that uses statistical and linguistic analysis on thetext in a reference corpus and the web document is presented.It used the number of terms and number of wordsin a sentence, and term frequency in the corpus forsummarization and about 30% of the ranked sentences wereconsidered to be the summary of the web document.We fashion our closeness metric after thehighly successful word error rate metric usedby the speech recognition community, appro-priately modified for multiple reference trans-lations and allowing for legitimate differencesin word choice and word order.For this example, it is at once clear that aprogram can rank Candidate 1 higher than Can-didate 2 simply by comparing n-gram matchesbetween each candidate translation and the ref-erence translations.To compute precision,one simply counts up the number of candi-date translation words (unigrams) which occurin any reference translation and then dividesby the total number of words in the candidatetranslation.Next, one clips the total count of each can-didate word by its maximum reference count,adds these clipped counts up, and divides by thetotal (unclipped) number of candidate words.However, as can be seen in Figure 4, the mod-ified n-gram precision decays roughly exponen-tially with n: the modified unigram precision ismuch larger than the modified bigram precisionwhich in turn is much bigger than the modi-fied trigram precision.Figure 2: Machine and Human TranslationsBleu uses the average logarithm with uni-form weights, which is equivalent to using thegeometric mean of the modified n-gram preci-sions.3,4 As a result, the Bleu metric is nowmore sensitive to longer n-grams.With this brevity penaltyin place, a high-scoring candidate translationmust now match the reference translations inlength, in word choice, and in word order.Notethat neither this brevity penalty nor the modi-fied n-gram precision length effect directly con-siders the source length; instead, they considerthe range of reference translation lengths in thetarget language.For example, if there are three references withlengths 12, 15, and 17 words and the candi-date translation is a terse 12 words, we wantthe brevity penalty to be 1.If we computed the brevity penalty sentenceby sentence and averaged the penalties, thenlength deviations on short sentences would bepunished harshly.Wefirst compute the test corpus' effective referencelength, r, by summing the best match lengthsfor each candidate sentence in the corpus.IDS allows users to tailor the length and content of asummary, seeing changes in real-time as they amend summary attributes.It provides visualisations to supportinterpretation of a summary in the context of the entiredocument, and to allow users to make seamless and rapidtransitions between summaries and full documents.IDS produces summaries through identification and extractionof sentences that best reflect what a document is about,exploiting keyphrases/topics that are automaticallyextracted from the document text.Sentences are commonly chosen as the target segmentsto extract 9, 11, 15, 16,221.The sentence extraction processis as follows: apply a mechanism to allocate a score toeach sentence in the text, rank all sentences by decreasingscore, and finally select the N highest scoring sentences toform the summary.Some attributes are simple to compute, such assentence length (to favour longer sentences), and whethera sentence includes certain cue phrases like Inconclusion or In this paper.Location in the text is oftenused to favour sentences that are closer to the start of adocument.Statistical analysis techniques can be used to identifyimportant words or phrases in a document, and sentencescan then be scored based on the occurrence of such wordsand phrases within them.Attributes are often weighted ina heuristic manner, but some research has treated sentenceextraction as a learning problem 15, 16, 221.By applying thesetechniques to a document summariser one may supportusers in rapidly determining the utility of an on-linedocument by investigating a range of summary variants.There are two phasesto Kea: learning a model of appropriate keyphrases, anduse of the model to extract keyphrases from documents.Two attributes of phrases are used in building a model: thedistance into a document where a phrase first occurs; andthe TFxIDF value of the phrase (a measure of howfrequently it occurs within a given document compared toother documents).A simple technique, shown in Figure 1,is to mark sentences in the summary with indices thatdescribe their location in the full document.Each indexshows the number of the paragraph containing a sentence,and the location of the sentence within the paragraph.The question arises as to how users might find thesummary topics in the full text, or the surrounding contextfor the sentences that formed the summary.Each of the summary-in-contextviews-location indices, text shading and text scaling-can be combined, and applied in conjunction with dynamic control over the summary length.A problem with this approach is themultitude of ways in which similarity can be defined andmeasured, particularly when the gold-standard for thesummaries is not produced via sentence extraction.A problem in theevaluation of summaries produced by text extraction isthat they are likely to be less readable than those producedby authors or professional abstracters.Consequently,negative subjective judgements might reflect summarycharacteristics other than the summariser's ability toextract the most appropriate sentences.IDS allows users to tailor the length and content of asummary, seeing changes in real-time as they amendsummary attributes.It provides visualisations to supportinterpretation of a summary in the context of the entiredocument, and to allow users to make seamless and rapidtransitions between summaries and full documents.IDSproduces summaries through identification and extractionof sentences that best reflect what a document is about,exploiting keyphrases/topics that are automaticallyextracted from the document text.Sentences are commonly chosen as the target segmentsto extract 9, 11, 15, 16,221.The sentence extraction processis as follows: apply a mechanism to allocate a score toeach sentence in the text, rank all sentences by decreasingscore, and finally select the N highest scoring sentences toform the summary.Some attributes are simple to compute, such assentence length (to favour longer sentences), and whethera sentence includes certain cue phrases like Inconclusion or In this paper.Location in the text is oftenused to favour sentences that are closer to the start of adocument.Statistical analysis techniques can be used to identifyimportant words or phrases in a document, and sentencescan then be scored based on the occurrence of such wordsand phrases within them.Attributes are often weighted ina heuristic manner, but some research has treated sentenceextraction as a learning problem 15, 16, 221.By applying thesetechniques to a document summariser one may supportusers in rapidly determining the utility of an on-linedocument by investigating a range of summary variants.There are two phasesto Kea: learning a model of appropriate keyphrases, anduse of the model to extract keyphrases from documents.Two attributes of phrases are used in building a model: thedistance into a document where a phrase first occurs; andthe TFxIDF value of the phrase (a measure of howfrequently it occurs within a given document compared toother documents).A simple technique, shown in Figure 1,is to mark sentences in the summary with indices thatdescribe their location in the full document.Each indexshows the number of the paragraph containing a sentence,and the location of the sentence within the paragraph.

Figure 4. Output of multiple papers

5.2. Module 2

To check the coherence for the multiple IEEE papers.

Input:

Paper 1

With the explosion of the World WideWeb and the abundance of text available on theInternet, the need to provide high-qualitysummaries in order to allow the user to quicklylocate the desired information also increases.Summarization is a useful tool for selectingrelevant texts and for extracting the key points ofeach text. We investigate a machine learning approach that uses Bayesian classifier to producesummaries of document. A Bayesian classifier istrained on a corpus of documents for whichextractive summary is available Document summarization is the problem ofcondensing a source document into a shorterversion preserving its information content.Document summarization can be categorized(1) Understanding content of document.(2)Identifying most important pieces ofinformation contained in it.(3) Writing of information.Given variety of available information, itwould be useful to have domain independentautomatic techniques for doing this. However,automating the first and third steps forunconstrained texts is currently beyond state ofart. Thus the process of automatic summarygeneration generally reduces to task ofextraction. Therefore current research is focusedon generating extractive summary. This paperpresents an investigation into Bayesian classierbased approach for document summarization.The paper is divided as follows: Section IIdeals with basic concepts regarding automaticDocument summarization techniques areusually classified in three families: (i) based onthe surface (no linguistic analysis is performed);(ii) based on entities named in the text (there issome kind of lexical acknowledgement andclassification); and (iii) based on discoursestructure (some kind of structural, usuallylinguistic, processing of the document isrequired). Commercial products usually make use ofsurface techniques. One classical method isselection of statistically frequent terms in thedocument. E.g. those sentences containing moreof the most frequent terms (strings) will beselected as a summary of the document. Anothergroup of methods is based on position: positionin the text, in the paragraph, in depth orembedding of the section, etc. Other methodsgain profit from outstanding parts of the text:titles, subtitles. Finally, simple methods based onstructure can take advantage of the hyper textualscaffolding of an HTML page. More complexmethods using linguistic technology resourcesand techniques such as those mentioned aboveand others might build a rhetoric structure of thedocument, allowing its most relevant fragmentsto be detected. It is clear that when creating atext using fragments of a previous original,reference chains and in general, text cohesion, iseasily lost.Based on these techniques several automaticdocument summarization methods have beendeveloped. Some of these methods include: Cutand Paste method, document summarizationusing lexical chains, pyramid method andtrainable summarizer[1,2,5,9,10,11,13,14].Most of the automatic summarizationtechniques are based on extracting significantsentences from source documents by somemeans. Therefore major idea related to documentsummarization is selection of features andlearning patterns of these features whichdetermines which sentences in source should beincluded in the summary.

Figure 5. IEEE paper 2 as input

Paper 2

This paper presents an investigation intomachine learning approach for documentsummarization. A major challenge related todocument summarization is selection of featuresand learning patterns of these features whichdetermines what information in source should beincluded in the summary. Instead of selectingand combining these features in adhoc mannerwhich would require readjustment for each newgenre, natural choice is to use machine learningtechniques. This is the basis for trainablemachine learning approach to summarization.We briefly discuss design, implementation andperformance of Bayesian classifier approach fordocument summarization.With the explosion of the World WideWeb and the abundance of text available on theInternet, the need to provide high-qualitysummaries in order to allow the user to quicklylocate the desired information also increases. Summarization is a useful tool for selectingrelevant texts and for extracting the key points ofeach text. We investigate a machine learningapproach that uses Bayesian classifier to producesummaries of document. A Bayesian classifier istrained on a corpus of documents for whichextractive summary is availableDocument summarization is the problem of condensing a source document into a shorterversion preserving its information content.Document summarization can be categorized(1) Understanding content of document. (2)Identifying most important pieces ofinformation contained in it.(3) Writing of information.Given variety of available information, itwould be useful to have domain independent automatic techniques for doing this. However,automating the first and third steps forunconstrained texts is currently beyond state ofart. Thus the process of automatic summary generation generally reduces to task ofextraction. Therefore current research is focusedon generating extractive summary. This paperpresents an investigation into Bayesian classierbased approach for document summarization.The paper is divided as follows: Section IIdeals with basic concepts regarding automaticDocument summarization techniques areusually classified in three families: (i) based onthe surface (no linguistic analysis is performed);(ii) based on entities named in the text (there issome kind of lexical acknowledgement andclassification); and (iii) based on discoursestructure (some kind of structural, usuallylinguistic, processing of the document isrequired).Commercial products usually make use ofsurface techniques. One classical method isselection of statistically frequent terms in thedocument. E.g. those sentences containing moreof the most frequent terms (strings) will beselected as a summary of the document. Anothergroup of methods is based on position: positionin the text, in the paragraph, in depth orembedding of the section, etc. Other methodsgain profit from outstanding parts of the text:titles, subtitles. Finally, simple methods based onstructure can take advantage of the hyper textual scaffolding of an HTML page. More complexmethods using linguistic technology resourcesand techniques such as those mentioned aboveand others might build a rhetoric structure of thedocument, allowing its most relevant fragmentsto be detected. It is clear that when creating atext using fragments of a previous original,reference chains and in general, text cohesion, iseasily lost.Based on these techniques several automaticdocument summarization methods have beendeveloped. Some of these methods include: Cutand Paste method, document summarizationusing lexical chains, pyramid method andtrainable summarizer[1,2,5,9,10,11,13,14].Most of the automatic summarizationtechniques are based on extracting significantsentences from source documents by somemeans. Therefore major idea related to documentsummarization is selection of features andlearning patterns of these features whichdetermines which sentences in source should beincluded in the summary. Instead of selectingand combining these features in adhoc mannerwhich would require readjustment for each newgenre, natural choice of use of machine learningtechniques. This is the basis for trainablemachine learning approach to summarization.A Machine Learning (ML) approach can beenvisaged if we have a collection of documentsand their corresponding reference extractivesummaries. A trainable summarizer can be obtained by the application of a classical(trainable) machine learning algorithm in thecollection of documents and its summaries.

Figure 6. IEEE paper 2 as input

Output:

```
[0.6777280233874308, 4.696474594492543, 1.1532500098070386, 0.45127623460302857,
 0.44446589905169276, 1.3456979748873437, 1.4597347461483423, 1.5544456074732125
, 0.6992408261806986, 0.5768353592897029, 0.8614434900451897, 2.1844669751413903
, 1.6856637955911538, 0.5066811167300784, 1.4899847365372807, 3.455070564426137,
 0.5090903909706412]


[14, 17, 30, 43, 64, 65, 70, 78, 91, 112, 114, 123, 133, 154, 155, 160]
{0.8614434900451897: 114, 1.1532500098070386: 30, 0.44446589905169276: 64, 1.685
6637955911538: 133, 1.3456979748873437: 65, 1.4597347461483423: 70, 2.1844669751
413903: 123, 4.696474594492543: 17, 1.5544456074732125: 78, 0.6992408261806986:
91, 0.5768353592897029: 112, 0.6777280233874308: 14, 0.5066811167300784: 154, 1.
4899847365372807: 155, 0.45127623460302857: 43, 3.455070564426137: 160}
27
11
there is a very strong relation
jarvis@jarvis-Inspiron-N5110:~/Project$
```

Figure 7. Output for Coherence

6. EVALUATION

Rogue method will be used to evaluate the summarizer. The official evaluation toolkit for text summarization in DUC, to evaluate the performance of our summarization system. It involves manually summarizing a document and then compare it with the automated summary. Also involves manually determining coherence between documents, and comparing it with the documents.

7.

REFERENCES

[1] H. Chorfi, "Get only the essential information: Text summarizer based on implicit data", pp. 1-4, 2013.

[2] Freitas F., *et al.*, "*A context based text summarization system*", In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop, pp. 66–70, 2014.

[3] A. Nenkova and K. McKeown, "A survey of text summarization techniques", In Mining Text Data Springer US., pp. 43-76, 2012.

[4] R. D. Lins, *et al.*, "Assessing sentence scoring techniques for extractive text summarization", Vol. 40, 2013.

[5] Wilkerson D. S., *et al.*, "*Winnowing: local algorithms for document fingerprinting*", In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 76-85, 2003.

[6] Wen D., *et al.*, "A contextual query expansion based multi-document summarizer for smart learning", In Signal-Image Technology and Internet-Based Systems (SITIS), pp. 1010-1016, 2013.

[7] I. Kupiec, *et al.*, "*A trainable document summarizer*", In Proceedings of the 18th ACMSIGIR Conference, pp. 68-73, 1995.

BIOGRAPHIES OF AUTHORS

Dr. M. Suman professor (Signals and Systems) in department of Electronics and Computer Engineering (ECM) has extended his services as HOD in ECM department, K L University. He was awarded with Ph.D. from JNTUH, Hyderabad for the thesis entitled "ENHANCEMENT OF COMPRESSED NOISY SPEECH SIGNAL". He is also the life member of Computer Society of India (CSI).

Tharun Maddu student of Electronics and Computer Engineering (ECM) pursuing 4th year of B.TECH in K L University. My previous research works are based on data mining. The present work is related to NLTK on which the present paper research is done.