A Trainable Document Summarizer Using Bayesian Classifier Approach

Aditi Sharan
SC& SS
JNU, New Delhi,India
aditisharan@mail.jnu.ac.in

Hazra Imran
SC & SS
JNU, New Delhi,India
hazrabano@gmail.com

ManjuLata Joshi
Banasthali Vidyapith,
Rajasthan,India
manjulatajoshi@gmail.com

Abstract

This paper presents an investigation into machine learning approach for document summarization. A major challenge related to document summarization is selection of features and learning patterns of these features which determines what information in source should be included in the summary. Instead of selecting and combining these features in adhoc manner which would require readjustment for each new genre, natural choice is to use machine learning techniques. This is the basis for trainable machine learning approach to summarization. We briefly discuss design, implementation and performance of Bayesian classifier approach for document summarization.

Index Terms

Automatic document summarization, machine learning, Bayesian classifier, Significant sentences Extraction.

1. *INTRODUCTION*

With the explosion of the World Wide Web and the abundance of text available on the Internet, the need to provide high-quality summaries in order to allow the user to quickly locate the desired information also increases. Summarization is a useful tool for selecting relevant texts and for extracting the key points of each text. We investigate a machine learning approach that uses Bayesian classifier to produce summaries of document. A Bayesian classifier is trained on a corpus of documents for which extractive summary is available

Document summarization is the problem of condensing a source document into a shorter version preserving its information content. Document summarization can be categorized into two categories: abstract-based and extract-based. An extractive summary consists of sentences extracted from the document while an abstractive summary may employ words and phrases that may not appear in the original document [13]. The summarization task can also be categorized as either generic or query-oriented. A query-oriented summary presents the information that is most relevant to the queries given by the user, while a generic summary gives an overall sense of the document's content [7]. In addition to single document summarization, researchers have started to work on multi-document summarization whose goal is to generate a summary from multiple documents that cover similar information.

Automated summarization dates back to 50's [12]. Different attempts in this field have shown that human quality summary generation was very complex since it encompasses understanding abstraction and language generation. Consider the process by which human accomplishes this task. Usually following steps are involved

(1) Understanding content of document.
(2)Identifying most important pieces of information contained in it.
(3) Writing of information.

Given variety of available information, it would be useful to have domain independent automatic techniques for doing this. However, automating the first and third steps for unconstrained texts is currently beyond state of art. Thus the process of automatic summary generation generally reduces to task of extraction. Therefore current research is focused on generating extractive summary. This paper presents an investigation into Bayesian classier based approach for document summarization.

The paper is divided as follows: Section II deals with basic concepts regarding automatic

978-0-7695-3267-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ICETET.2008.123

document summarization. It discusses various techniques and approaches that have been developed for automatic document summarization. Further, it discusses utility of machine learning techniques specifically Bayesian classifier in this field. Section III discusses Automatic document summarization using Bayesian classifier approach .Section IV deals with experiments and results. Finally we conclude in Section V.

2. AUTOMATIC DOCUMENT SUMMARIZATION

Document summarization techniques are usually classified in three families: (i) based on the *surface* (no linguistic analysis is performed); (ii) based on *entities named in the text* (there is some kind of lexical acknowledgement and classification); and (iii) based on *discourse structure* (some kind of structural, usually linguistic, processing of the document is required).

Commercial products usually make use of *surface* techniques. One classical method is selection of statistically frequent terms in the document. E.g. those sentences containing more of the most frequent terms (*strings*) will be selected as a summary of the document. Another group of methods is based on position: position in the text, in the paragraph, in depth or embedding of the section, etc. Other methods gain profit from outstanding parts of the text: titles, subtitles. Finally, simple methods based on *structure* can take advantage of the hyper textual scaffolding of an HTML page. More complex methods using linguistic technology resources and techniques such as those mentioned above and others might build a rhetoric structure of the document, allowing its most relevant fragments to be detected. It is clear that when creating a text using fragments of a previous original, reference chains and in general, text cohesion, is easily lost.

Based on these techniques several automatic document summarization methods have been developed. Some of these methods include: Cut and Paste method, document summarization using lexical chains, pyramid method and trainable summarizer[1,2,5,9,10,11,13,14].

Most of the automatic summarization techniques are based on extracting significant sentences from source documents by some means. Therefore major idea related to document summarization is selection of features and learning patterns of these features which determines which sentences in source should be included in the summary. Instead of selecting and combining these features in adhoc manner which would require readjustment for each new genre, natural choice of use of machine learning techniques. This is the basis for trainable machine learning approach to summarization.

A Machine Learning (ML) approach can be envisaged if we have a collection of documents and their corresponding reference extractive summaries. A trainable summarizer can be obtained by the application of a classical (trainable) machine learning algorithm in the collection of documents and its summaries. In this case the sentences of each document are modeled as vectors of features extracted from the text. The summarization task can be seen as a two-class classification problem, where a sentence is labeled as ‘‘significant’’ if it belongs to the extractive reference summary or as ‘‘insignificant’’ otherwise. The trainable summarizer is expected to ‘‘learn’’ the patterns which lead to the summaries, by identifying relevant feature values which are most correlated with the classes ‘‘significant’’ or ‘‘insignificant’’.

Bayesian learning methods are relevant to our problem for certain reasons. Firstly, Naive Bayes classifiers that calculate explicit probabilities for hypothesis are among most practical approaches to certain types of learning problems. In particular it has been widely used for solving a related problem dealing with document classification such as electronic news articles, email etc. For such learning task Naïve Bayes classifier is among most effective algorithm known. Further in bayesian classifier approach each observed training example can incrementally increase or decrease the estimated probability that hypothesis is correct. This provides more flexible approach to learning that completely eliminates the hypothesis if it is found to be inconsistent with any single example. Even in case where Bayesian method proves computationally intractable, it can provide a standard decision making against which other practical methods can be measured. However Bayesian classifier assumes that features are independent from each other. Despite this unrealistic assumption, this method presents good results in many cases and it has been successfully used in many text mining projects.

3. AUTOMATIC DOCUMENT SUMMARISATION USING BAYES CLASSIFIER

There are three phases in our process: Identification of the sentences in the document, Computation of scores of each sentences and Training of bayesian classifier. These three steps are explained in detail in the next three sections.

3.1 Identification of sentences

To identify the sentences in a document we have used following heuristics: A sentence ends with one or more points, exclamation marks and/or question marks, Sentence delimiters are optionally followed by an ending quotation mark, The final delimiter should be followed by one or multiple white spaces (space, tab, newline, etc.), The first word of the following sentence should start with a capital, The sentence delimiter should not be part of an abbreviation and occurrence of white spaces.

3.2 Score computation

When a document is given to the system, the ''learned'' patterns are used to classify each sentence of that document into either a ''significant'' or ''insignificant'' sentence, producing an extractive summary. A crucial issue is how to obtain the relevant set of features.

For each sentence different scores were calculated. These scores formed the features used for our classification task.

Edmundson Feature: The Edmundson feature assigns a score to each sentence based on the frequency of the significant words (having a frequency larger than a certain threshold and not being a common word [8]) in the sentence. Score of the sentence is calculated by adding frequencies of all significant words present in a sentence.

$$f1 = \sum_{i=1}^{n} freq_i$$

Where

$f1$ = Edmunson Feature

n = number of significant words in the sentence

$freq_i$ = frequency of i^{th} significant word.

Luhn Feature: The Luhn's method [11] does not take into account the exact frequency of the significant words instead it distinguishes significant words from non significant words. Luhn's method first generates a list of candidate terms that occur in the body of the documents in descending order of their term frequency within the document. Words with high frequency of occurrence within a document and those with very low frequency of occurrence in each document are classified as insignificant words.

In addition, a lower limit for significance needs to be defined. The lower limit for significance needs to be defined. Following the work of Trombos [3], the required minimum occurrence count for significant terms in a medium-sized document was taken to be 7; where a medium sized document is defined as containing no more than 40 sentences and not less than 25 sentences. For documents outside this range, the limit for significance is computed as

ms=7+[0.1(L - NS) for documents with NS<25

ms=7+ 0.1(NS - L) for documents with NS>40

where ms= the measure of significance i.e. the threshold for selecting significant words.

L= Limit (25 for NS<25 and 40 for NS>40)
NS= number of sentences in the document.

Luhn [11] reasoned that, the closer certain words are, the more specifically an aspect of the subject is being treated. Two significant words are considered *significantly related* if they are separated by not more than five non-significant words e.g.

'*The sentence [**scoring** process utilizes **information** both from the **structural]** organization.'*

The cluster of significant words is given by the words in the bracket ([---]), where significant words are shown in bold. The cluster significance score factor for a sentence is given by the following formula

$$f2 = \frac{SW^2}{TW}$$

Where f2 = the Luhn Feature

Thus f2 for the above sentence is 9/8= 1.125.

SW = the number of bracketed significant words
TW= the total number of bracketed words.

If two or more clusters of significant words appear in a given sentence, the one with the highest score is chosen as the sentence score.

Location Feature:

The position of a sentence within a document is often useful in determining its importance to the document. Based on this, Edmundson defined a location method for scoring each sentence based on whether it occurs at the beginning or end of a paragraph or document. Baxendale[4] demonstrated that sentences located at the beginning and end of paragraphs are likely to be good summary sentences. It is observed that short sentences are unlikely to be included in summaries [9].The first two sentences of an article are assigned a location score computed as below

$$f3 = \frac{1}{NS}$$

Where *f*3 = the location score for a sentence
NS = the number of sentences in the document.

***Cue Phrase Feature*:** The Cue Phrase feature is based on the assumption that the relevance of a sentence is based on the presence of certain pragmatic phrases like 'In this paper',' It is concluded'. Edmundson [8] introduced the Cue method in 1969. In our experiment, we have used the fixed cue phrases. Cue Phrase feature of the sentence is calculated by counting total number of cue phrases occurring in the sentences.

***Title Feature*:** Terms occurring in the title usually reveal the specific concept of a document. Therefore, sentence containing title terms are more significant .We have considered title feature as a Boolean variable. Thus, this feature has the value TRUE if the sentence contain the title word otherwise FALSE.

First sentence Feature: The first sentence of the document is always considered as a significant sentence. This is also taken as a Boolean feature.
Sentence Length Cut-off Feature: Sentence length also affects the significance of a sentence. Generally **s**hort sentences like section headings are not included in summaries. In our experiment we have taken a threshold of 5 words. The feature is TRUE for all sentences longer than the threshold and FALSE otherwise.

Occurrence of noun: Occurrence of nouns represents clues of significance of a sentence for the summary. We identified nouns in the sentence. This feature is then calculated by counting frequency of noun in the sentence.

Table 1: Scores of the document

f1	*Edmundson feature*
f2	*Luhn Feature*
f3	*Location feature*
f4	*Cue phrase feature*
f5	*Title Feature*
f6	*First sentence Feature*
f7	*Sentence Length Cut-off Feature*
f8	*Occurrence of Proper noun*

3.3 Bayesian classifier approach for Document Summarization

This section quickly reviews the basis of the Naïve Bayes classifier. We have implemented a Bayesian classifier that computes the probability that a sentence in a source document should be included in a summary.

For each sentence *s,* the probability of *s* being included in the summary is calculated based on the *k* given features Fj: j=1….k which can be expressed using the Bayers rule as follows:

$$P(s \in S \mid F_1, F_2 \ldots F_k) = \frac{P(F_1, F_2 \ldots F_k \mid s \in S)\, P(s \in S)}{P(F_1, F2 \ldots F_k)}$$

Assuming statistical independence of the features

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

$P(s \in S)$ is a constant and $P(F_j \mid s \in S)$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences. This yields a simple Bayesian classification function that assigns for each sentence *s* a score which can be used to select sentences for inclusion in the summary.

Once the classifier has been trained, it can be used as a tool to filter sentences in any document and determine whether each sentence should be included in the summary or not.

4. EXPERIMENTAL EVALUATION

A Steps in conducting the experiment

In our experiments we have used a training corpus of computational Linguistic texts from the Computation and Language E-print Archive (cmp-lg), provided in SGML form by the University of Edinburgh. The articles are between 4 to 10 pages in length and have figures, captions, references and cross references replaced by place holders. These are 198 full text articles and we have used 50 text articles in our experiment. Each document consists of 64 to 417 numbers of sentences with 216 average numbers

of sentences. The entire set consists of total 10817 sentences. Extractive summary is given for each document.

The experiment consisted of the following steps:

1. Identification of distinguished sentences: The heuristic used to identify the sentences are discussed in 3.1

2. Computation of the set of features and discretization of continuous features.: The computations are discussed in previous subsection. Refer Table 2.

3. Performing classification based on Naïve Bayes classifier.

Table 2: Features of sentence used by Bayesian classifier for classification

f1	*Edmundson feature*	*Discretize*
f2	*Luhn Feature*	*Discretize*
f3	*Location feature*	*Discretize*
f4	*Cue phrase feature*	*Discretize*
f5	*Title Feature*	*Continious*
f6	*First sentence Feature*	*Continious*
f7	*Sentence Length Cut-off Feature*	*Continious*
f8	*Occurrence of Proper noun*	*Discretize*
f9	*Class*	*Boolean*

B. Experimental Results

In bayesian classification two- third data of the entire dataset was used for training the classifier. Testing was then performed on the remaining one-third dataset. The performance of the document summarization process depends heavily given extractive summaries. The results of experiments were compared for following methods

(a) Our automatic text summarization system using Naïve Bayes classifier.
(b) Word Summarizer- It's a Microsoft text summarizer. This method produces 'almost extractive' summary from the text.
(c) First sentence - This method selects first *n* sentences.

It is important to note that our summarizer is based on ML and the two remaining methods are not trainable.Table 3 reports the results obtained by the three summarizers. We consider compression rates as 10 %. The performance is expressed in terms of precision and recall values, expressed in percentage (%). The best obtained results are shown in boldface.

Table3: ***Results for training and test sets composed by automatically-produced summaries (Compression rate is 10%)***

Summarizer	Precision	Recall
Naïve Bayes Automatic summarizer	**47.4**	**49.1**
Word-Summarizer	28.1	36.3
First sentence	23.9	27.4

Table 4: An example of Extractive and Bayesian Classifier based summary produced for a document (cmp-lg):

EXTRACTIVE SUMMARY

In previous work, the definition of the transduction relation defined by a synchronous TAG was given by appeal to an iterative rewriting process, much like the iterative rewriting of sentential forms defined by a context-free grammar except that the syntactic objects generated by the rewriting process were derived trees rather than strings .First, the weak-generative expressivity of TAGs is increased through the synchronization in the sense that the projection of the string pairs onto a single component, although the strings in that component are specified with a TAG, may not form a tree-adjoining language (TAL).The Rewriting Definition of Derivation The original definition of derivation for synchronous TAGs was based on the iterative rewriting of one derived tree pair into another. Thus, if a standard TAG parsing algorithm is used for the first step in the process (so that DL is a traditional TAG derivation tree), the second step is not well defined.

BAYESIAN CLASSIFIER BASED SUMMARY

We would have derivation trees that specify at each node an elementary tree pair, with arcs labeled by pairs of tree addresses (such that the two addresses are linked in the parent elementary tree pair). In previous work, the definition of the transduction relation defined by a synchronous TAG was given by appeal to an iterative rewriting process, much like the iterative rewriting of sentential forms defined by a context-free grammar except that the syntactic objects generated by the rewriting process were derived trees rather than strings. First, the weak-generative expressivity of TAGs is increased through the synchronization in the sense that the projection of the string pairs onto a single component, although the strings in that component are specified with a TAG, may not form a tree-adjoining language (TAL).

5. CONCLUSIONS AND FUTURE WORK

The design of an automatic document summarizer using machine learning approach is an important research area. In this paper we have explored framework for using Bayesian Classier approach for document summarization. In our proposal, we employ trainable Bayesian summarizer that uses statistical as well as linguistic features. Results of our experiments prove that Bayesian classifiers can be successfully applied for generating document summaries. In future we wish to explore other machine learning techniques for summarizing documents and compare them with Bayesian classifier approach. Further we would like to

explore feature extraction techniques and find the possibility of using machine learning algorithms to automatically learn good combination functions to combine various features.

References

1. A.Nenkova, and R. Passonneau, 2004. "Evaluating content selection in summarization: The pyramid method", In HLT/NAACL., 2004.

2. A. Nenkova, R. Passonneau and K. McKeown "The Pyramid Method: Incorporating human content selection variation in summarization evaluation", ACM New York, NY, USA ACM Transactions on Speech and Language Processing (TSLP), 2007.

3. Anastasios,Tombros," Reflecting user information needs through query Biased Summaries",thesis submitted towards the award of MSc in advance Information System in the university of Glassgow, September 1997.

4. P.B. Baxendale, "Machine-Made Index for Technical Literature: An Experiment", IBM Journal of Research and Development, vol. 2(4), pp. 354-361,1958.

5. Chan Yamin, Wang Xiaolong and Guan Yi, "Automtic Text Summarization Based on Lexical Chains", ICNC 2005, LNCS 3610, pages 947-951, 2005.

6. G. Salton and M. J. McGill. "Introduction to Modern Information Retrieval". McGraw-Hill Computer Science series. McGraw-Hill, New York, 1983.

7. Goldstein , Mark Kantrowitz,Vibhu Mittal, and Jaime Carbonell," Summarizing text documents: Sentence selection and evaluation metrics .In SIGIR,pages121-128,1999.

8. H.P. Edmundson, " New Methods in Automatic Abstracting", ACM Journal, 1969.

9. J. Kupiec, J. Pederson and F. Chen, "A Trainable Document Summarizer", Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, pp. 68-73, 1995.

10. Jen-Yuan Yeh , Hao-Ren Ke , Wei-Pang Yang , I-Heng Meng, "Text summarization using a trainable summarizer and latent semantic analysis", Information Processing and Management: an International Journal, v.41 n.1, p.75-95, January 2005

11. Luhn, "The automatic creation of literature abstracts.", IBM J. of R. and D., 2(2), 1958.

12. Mani Inderjeet, "Advances in Automatic Text Summarization.",MIT Press,Cambridge, MA, USA, 1999.

13. R.Barzilay, and M. Elhabad, "Using lexical chains for text Summarization", In Intelligent Scalable Summarization Workshop, ACL, 1997.

14. Yu Lei, Ma Jai, Ren Fuji, Shingo Kuriowa, "Automatic Text Summarization Based on Lexical Chains and Structural Features", IEEE 2007.