Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text

Amit Kumar Mondal and Dipak Kumar Maji
Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur
Kanpur, India 208016
Supervisor: Dr. Harish Karnick

***Abstract*— The problem of generating headlines for documents using purely statistical approach has been long standing. We describe here an improved extractive approach based on keywords. The insight here is that if one tries to summarize a document, one will invariably use keywords from the document itself. There are two aspects to the problem namely, finding the relevant set of keywords and finding the proper way to combine these words to reflect a coherent and grammatical headline. Keyword selection can be tackled using various statistics about keywords in the document, while the problem of generating the headline sentence from the selected keywords cannot be best solved without resorting to the knowledge about the semantics of the keywords. To enhance the accuracy of the headline in reflecting the content, the gist or topic is first identified. The results show that our extractive approach is feasible for generating informative headlines.**

I. Introduction

Text Mining is about looking for patterns in natural language text and may be defined as the ”exploration and analysis of textual (natural-language) data by automatic and semi automatic means to discover new knowledge” [1]. A special application is generating short summaries or headlines from a given document. A headline of a text, specially an article, is a succinct representation of relevant points of the input text. It differs from the task of producing abstracts, in the size of the generated text and focuses on the compressing the output. Headlines are terse while abstracts are expressed using relatively more words [2]. While headlines focus on pointing out the most relevant theme expressed in the input text, abstracts summarize the important points.

Headlines are commonly associated with news articles but application areas of headline generation range from generating table of contents for a document to providing support for interactive query refinement in search engines. Automatic headline generation tries to automate the process of providing more relevant or reflective insight into the input text rather than producing ”catchy lines”. Automating in this context has to involve some form of learning rather than an algorithmic approach given the potentially infinite stretch of natural language text. Many machine learning techniques have been explored involving varying degree of use of natural language understanding techniques.

We take a statistical approach which promises to perform well, and at the same time use very little domain knowledge.

II. Motivation

Document summarization has become a critical component in any toolkit for on-line information management. As we emerge into the 21st Century, the massive information of universe, which is mostly unstructured, that is already present makes some form of text summarization indispensable. Automatic summarization aims at providing a condensed representation of the content of an information source in a manner sensitive to the needs of the user and task.

Researchers have investigated the topic of automatic generation of brief summaries but the focus has been on different problems like sentence extraction [3] [4], processing structured templates, one sentence at-a-time compression [5] and multi-document abstracts [6]. Instead, we focus on generating a headline style summary from a text.

Text mining recognizes that complete understanding of natural language text, a long standing goal of AI, is not immediately attainable. Current techniques are based on statistical methods but do not provide efficacious results. We have tried to improve upon the existing algorithms using a mixed approach for generating representative headline level summaries of text. Here better is to be understood in terms of ’better’ quality headlines as judges by humans.

III. Application areas for Headline Extraction

- Generation of table of contents for a given document. This can be used in the categorization of large repositories of research works and storing them along with brief summaries.
- Generation of headlines for each item in the hit list obtained by web search. This will help to get a quick idea of the content of each hit item.
- Text Compression On a device with limited display or limited bandwidth, headlines can be a substitute for the full text. For example, an email message could be reduced to a set of headlines for display on pager; a web page could be reduced for display on a portable wireless web browser.

- Voice based application It can be used for giving users content information over voice based applications, providing news updates over phones.
- Interactive Query Refinement
  Narrow Hit List: Automatic headlines extraction can provide suggestions for improving a query. Often a query with a conventional search engine returns a huge lists of matching documents. The user would like to narrow the list by adding new terms to the query, but it is not clear what terms should be added. One way is to generate suggestions for refining a query is to extract headlines from the documents in the hit list for the original query. Conjunction of the new terms with the old query terms yields shorter hit lists.
  Expand Hit List: New terms can be added to a query by disjunction, instead of conjunction, which will yield a longer hit list.

IV. Organization

The rest of this report is organized as follows: First we review the various methods that have been used for creating headlines. Then we describe our keywords based approach followed by some results and discussion. In the concluding part, various problems in following a purely statistical approach and the future extensions to the project are discussed.

V. Literature Review

A lot of research in single document summarization has gone into finding out the relevant segments from the text, ranking them and finally generating the summary which expresses most of the important points. The task of title generation is strongly connected to traditional text summarization [7] and emphasizes the extractive approach which selects words, sentences or paragraphs from the document to provide a summary. Keywords and key phrases provide important clues about the category of a corpus and have been used to compose headlines.

- *Selecting title words:* The task of headline extraction can be interpreted as a twofold process. First, the system select n words from the article that best reflect its content. Second, the best grammatical ordering of these n words is determined. Witbrock and Mittal label these two tasks as *Content Selection* and *Realization*. Each of these criteria are scored probabilistically, where the probability is estimated by prior collection of corpus statistics.
  Bayesian approach [8] [9], TF*IDF (Term Frequency * Inverse Document Frequency) ranking [10], Text Model, Headline word position model [11] etc. have been used for determining the content selection probability. In Naive Bayesian approach, it tries to capture the correlation between the words in the document and the words in the title. For each training corpus, it counts the occurrence of each document-word-title-word pair where the document word and the title word is the same. C_W represents the occurrence of such a pair when both document word and title word is w. C_W can be expressed as following:

$$C_W = \sum_{j=1}^{N} doc\_tf(w,j) \times title\_tf(w,j)$$

  where $doc\_tf(w,j)$ is the term frequency of word w in document i and $title\_tf(w,j)$ is the term frequency of word w in in i^{th} title. This sum goes over all document-title pairs in the training corpus. The conditional probability $P(titlewordw|documentwordw)$ is obtained by dividing C_W by $\sum_{j=1}^{N} doc\_tf(w,j)$.
  To generate a title for a new document the generating potential G_W is computed for each word in the corpus:

$$G_W = doc\_tf(w) \times P(titlewordw|documentwordw)$$

  .
  Here doc_tf(w) is the term frequency of the word w in the new document. Those words with highest G_W are chosen to form the title.
  TF*IDF score of a term t in document D is obtained as follows:

$$W(t,D) = 0.5 \times (1 + \frac{tf(t,D)}{tf_{max}(D)} \times \log \frac{|DB|}{df(t)}$$

  where $tf(t,D)$ the term frequency of t in document D, $|DB|$ is the database size, $df(t)$ number of documents where term t appears at least once. Given a new document it calculates the tf.idf ranking of all content words in the documents and words with highest tf.idf ranking are chosen for the headline generation.
  [12] presents another novel approach for selecting headline words. This paper investigates the use of Singular Value Decomposition (SVD) as a means of determining if a word is a good candidate for inclusion in the headline. [13]describes a new approach towards title word selection, viewing title word selection as a variant of the Information Retrieval problem. The Information Retrieval (IR) problem is to find relevant documents from a text collection given a user query, while the title word selection problem is to select the representative title words from the title word vocabulary from the test document. By mapping the concepts "title word" and "title document" from the title word selection problem into "document" and "user query" in IR problem respectively, the title word selection problem becomes essentially an IR problem, i.e. finding title words, now equivalent to documents in IR, similar to test document, equivalent to the user query.
- *Sentence Extraction:*[14] [15] [16] The sentence extraction based method tries to pick the sentence which reflects the main content of the text. A number of researcher have looked into this problem of sentence extraction which carries the central idea of a corpus, as a solution of text summarization problem. The same techniques have been applied to the problem of headline extraction with a little variation [17]. Several heuristics have been designed

for ranking the sentences in the text. It uses a weighted sum of various features of the sentence like position of the sentence in the text, length of the sentence, *TF*IDF* score of the sentence, keyword density, presence of cue phrase etc. [15], [16]. The TF*IDF score of a sentence is obtained as follows:

$$Score(S_i) = \sum_{t \epsilon S_i} tf(t, s_i) \times w(t, D)$$

The density function of a sentence is calculated using the following formula:

$$Den(S_i) = \frac{\sum_{t \epsilon KW(S_i)} w(t, D)}{d(s_i)}$$

$$d(S_i) = \frac{\sum_{k=2}^{|KW(S_i)|} (dist_k)^2}{|KW(S_i)| - 1}$$

where $KW(S_i)$ is the set of keywords in the sentence S_i, $dist_k$ is the distance between the k^{th} and $(k-1)^{th}$ keywords in S_i. $W(t, D)$ is the TF*IDF score of the term t in the document D.
The score of a sentence can be calculated as a weighted sum of various features of a sentence.

$$W(S) = \alpha P(s) + \beta L(s) + \gamma T(s) + \delta K(s)$$

[18] describes a procedure to automatically acquire topic signatures and evaluates the effectiveness of applying topic signatures to extract topic related sentences. Topic signatures not only recognize related terms (topic identification), but also group together related terms under one target concept (topic interpretation). Topic identification and interpretation are two essential steps in a typical automated text summarization as well as headline extraction problem.

- *Cue Phrases method* The cue-phrase method claims that Important sentences contain phrases such as significantly, *In this paper we show*, *In conclusion*, *In summary* etc. These phrases are called bonus phrases. While non-important sentences contain phrases such as *hardly*, *impossible* etc. These phrases are called stigma phrases. These phrases can be detected and if the sentence contains bonus phrase(s), it adds to sentence score, otherwise if the sentence contains a stigma phrase, the sentence score will be decremented.

More recently, some researchers have moved towards *learning approaches* [19], [20] that take advantage of training data. [21] shows how K Nearest Neighbor algorithm can be applied for headline generation. It treats the titles in the training corpus as a set of fixed labels. For each new document, instead of creating new title it tries to find an appropriate "label", which is equivalent to searching the training document set for the closest related document. This training title is then used for the new document. [20] describes another machine learning (SVM) based summarization technique.

Another approach to construct headlines by selecting words in order from the story, removes grammatical constituents from a parse of the lead sentence until a length threshold is met. This approach is called parse-and-Trim method for headline generation [5] [22]. The first sentence of the story is passed through a parser. The parse-tree result is passed through a linguistically motivated module that selects story words to form headlines based on key insights gained from observations of human-constructed headlines.

By treating title generation as a variant of the Machine Translation problem, Kennedy and Hauptmann [23] came up with the generative approach using iterative Expectation Maximization algorithm. A number of researchers have attacked the problem of headline generation using NLP techniques. [24] [25] outline algorithms for computing lexical chains as an intermediate representation for automatic machine text summarization. A HMM [26] based summarization has been presented in [27] [28]. This idea has been extended for Headline Generation for News Stories in [29].

VI. Our Approach

Our approach is based on statistical information about keywords in a document. Based on the experiments done in the previous semester, we had collected and analyzed various statistical data about keywords in articles which were tagged with human generated headlines or titles. Analysis showed most of the documents have as many as 60% of their headline keywords taken from the respective documents.
The example in figure 1 shows the keywords obtained from the document and from the corresponding title. All the keywords except the word *diverse* is found in the document. Moreover, the lack of the word *diverse* in the headline would still make a good headline. Thus we concluded that given all the keywords present in the document, we can come up with a reasonable headline expressing the information contained in the document.

Title: **diverse confidence levels** in a **probabilistic semantics** for **conditional logics straddle**

Abstract: It has long been known that several **popular default** and **conditional logics exactly describe infinitesimal probability constructs** which **betoken virtual certainty regarding** the **truth** of one **sentence** if some other **sentence** is **true** that the **rules** for **default inference** also describe a **scheme** of **approximate inference** for all **positive-valued conditional probabilities** has also long been known more recently a **class** of **standard probability measures** has been found whose **expression** of more **likely** than not is **exactly described** by **default rules** this class can be **extended** by **straightforward algebra** so that the **range** of **confidence levels expressed** by **standard probability distributions** which are **exactly described** by the **rules** is as **complete** as for the **approximate methods**.

Fig. 1. Keywords in an example document and in the corresponding title

The flowchart in figure 2 decribes the various modules of our implementation. The various preprocessing phases are illustrated using the following example exerpt from a news article.

> For the second time in a week, the countrys technology hub had reasons to celebrate a s Wipro Ltd today joined the exclusive club of infotech firms with a billion-dollar r evenue. The IT behemoth posted record revenues of $1.2 billion from its combined IT s ervices and products business for 2003-04.
> The board has also recommended a bonus issue of two shares for each one held (2:1) fo r its shareholders. It has declared a total dividend of Rs 29 per share. An AGM will formally ratify the bonus issue in June.
> On Tuesday, Infosys Technologies Ltd celebrated its billion-dollar status with a doll ar treat for its employees across the world. TCS Ltd, an unlisted company, is the thi rd in the club.
> The global IT services business of Wipro alone generated $1 billion, accounting for 7 4 per cent of the groups total revenue of $1.35 billion (Rs 5881.2 crore), an increas e of 36 per cent from Rs 4338.3 crore in 2002-03. In 2003-04, we made a significant p rogress towards our goal of being the preferred provider of comprehensive solutions f or our customers, Wipro chairman Azim Premji said. Premji said the prospects for 2004-05 looked exciting. Looking ahead, for the quarter ending June, we expect our revenue from the global IT services business to be approximately $292 million, he added.

A. Extracting keywords

The flowchart in figure 3 describes the different preprocessing phases in the keyword extraction process

Tagging

In this phase, each of the tokens in the document is annotated with the part of speech information. This is done using TreeTagger [30] - a probabilistic part of speech tagger which uses decision trees. A binary decision tree is recursively built from a training set of trigrams using a modified version of the ID-3 algorithm. In each recursion step, the training set of trigrams gets divided into two subsets such that they differ maximally regarding the probability distribution of the third tag. The probability of a given trigram is then determined by traversing the corresponding path through the tree until a leaf is reached. The tagger has been reported to have achieved higher accuracy than the trigram tagger on the Penn-Treebank data. The sample sentence, when tagged, we have the following kind of information about the words in the sentence.

> For/**IN** the/**DT** second/**JJ** time/**NN** in/**IN** a/**DT** week/**NN** ,/**,** the/**DT** country/**NN** 's/**POS** technology/**NN** hub/**NN** had/**VHD** reasons/**NNS** to/**TO** celebrate/**VV** as/**IN** Wipro/**NP** Ltd/**NP** today/**NN** joined/**VVD** the/**DT** exclusive/**JJ** club/**NN** of/**IN** infotech/**NN** firms/**NNS** with/**IN** a/**DT** billion-dollar/**JJ** revenue/**NN** ./**SENT**

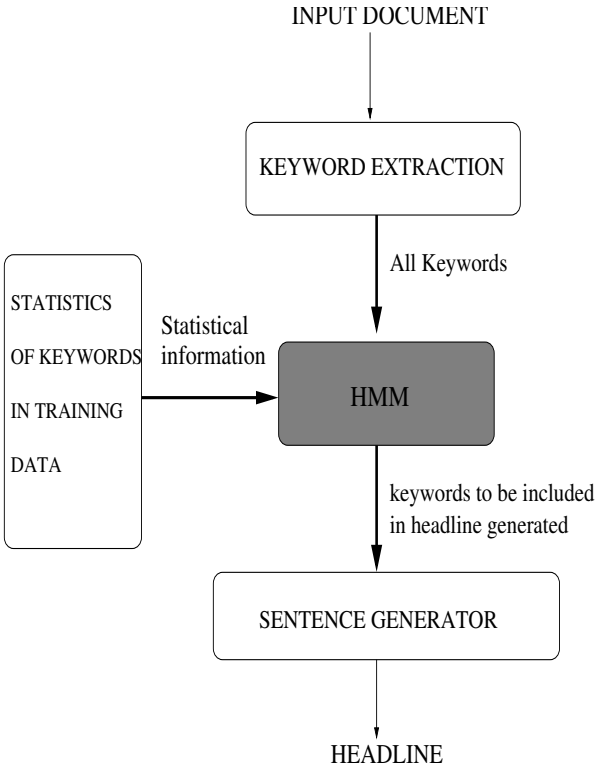

Fig. 2. The different modules of the Headline Generator

The Tree-Tagger uses the Penn-Tree bank tagset [31] for tagging words. Thus **DT** denotes Determiner, **JJ** denotes adjective while **NN** denotes noun, singular or mass.

Normalization

Normalization is the process of removing unnecessary text from the document. The unnecessary text may include text found between hyphens, after bullets, and between parentheses.

Segmentation

This phase outputs the document as a set of words by removing the unnecessary semicolons, colons, exclamation marks etc. The full-stops are retained to indicate end of sentence which is needed during further stages of headline generation.
After this stage the sample sentence would produce the following set of words :

> For the second time in a week the country's technology hub had reasons to celebrate as Wipro Ltd today joined the exclusive club of infotech firms with a billion-dollar revenue.

Stemming

The segmented words are then queried in the WordNet [32]

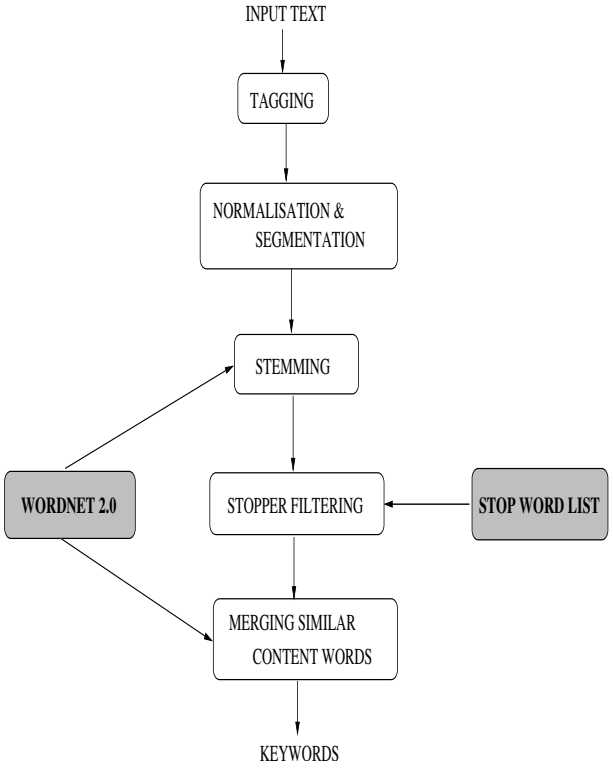

Fig. 3. Keywords Extraction Process

to get the root form of the words. For example words like *issues , relinquishing* are transformed to *issue, relinquish* resp. The part of speech information, already assigned to each word by the tagger, is used here to query the WordNet to get the root form of the word in the sense in which it is used in the document. The output is thus the set of root forms of the words of the original document as shown below for the sample sentence :

> For the second time in a week, the country technology hub have reason to celebrate as Wipro Ltd today join the exclusive club of infotech firm with a billion-dollar revenue.

Stopper Filtering

There is a precompiled set of fluff words. This stoplist consists of a list of common function words such as determiners (a, the, this), prepositions (in, from, to), conjunctions (after, since as), coordination (and, or). Also those words which occurs more frequently but contribute little meaning like *about, them, only* etc.

During this stage the segmented set of words of the document are filtered through this list of fluff words. The remaining words form the content words of the document.

The content words for the sample sentence are the following:

> time week country technology hub celebrate Wipro Ltd today join exclusive club infotech firm billion-dollar revenue

Finding keywords

The content words are then queried in the WordNet thesaurus to find out the synonyms of each word for the particular part of speech annotation with which it is used in the document. The number of important words thus get reduced by merging information about words conveying similar content. The final or condensed set of root form of these content words after this phase represents the set of keywords. In many of the cases the set of keywords are the same as the set of content words obtained from the previous phase as in the case for the sample sentence.

VII. Topic Identification

Topic identification corresponds to selecting a set of sentences that best expresses the gist of the document. Gist extraction has been tried both with complicated deep-based approaches like rhetorical structuring of source texts to select relevant information as well as simple statistical approaches based on either keywords or text mining . In the text mining approach, topic sentences are extracted using the result of measurement of the representativeness of the intra- and inter-paragraph sentences.

We have tried an approach based on simple statistics of keywords to get a set of topic sentences. To determine the ranking of the sentences of a document, a relative ”significance” score is determined and the top three are chosen for identifying the topic of the document. It is proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. Also the relative positioning of words, given their word significance, within a sentence provides useful measurement of significance of the sentence. The idea is that wherever the greatest number of frequently occurring keywords are present in close proximity the probability is very high that the information being conveyed by such clusters is most representative of the document. To be more relevant, we have identified certain clusters in a sentence and calculated the score for each of these. The highest score is assigned to the sentence. The clusters are closely knit keywords. We separate between two adjacent clusters if the number of fluff words between two adjacent keywords exceeds some value. The first of the two keywords thus mark the end of the previous cluster while the second one the start of a new cluster. The *significance factor* formula we have considered here is the following :

$$score = \frac{\sum(\text{freq of keywords in cluster})^2}{\text{length of the cluster}}$$

The frequency of occurrence of the keywords considered here takes in to account all the morphological and synonym sets of the keywords. The length of the cluster is the number of words consisting of the cluster, including the fluff words in between the keywords.

These topic sentences represent the content of the original document. Further processing through HMM to extract headline is done on these instead of the original document.

For the sample document in consideration, the topic sentences chosen are the following:

> For the second time in a week, the countrys technology hub had reasons to celebrate a s Wipro Ltd today joined the exclusive club of infotech firms with a billion-dollar r evenue.
> The IT behemoth posted record revenues of $1.2 billion from its combined IT s ervices and products business for 2003-04.
> The global IT services business of Wipro alone generated $1 billion, accounting for 7 4 per cent of the groups total revenue of $1.35 billion (Rs 5881.2 crore), an increas e of 36 per cent from Rs 4338.3 crore in 2002-03.

VIII. Automatic Headline Extraction

The problem of automatic headline extraction can be viewed as a translation problem – translating between a verbose language(of source documents) and a succinct language (of headlines). Our technique of extracting headline is based on NoisyChannelModel – with a subsequent decoder for producing headline words from story words. This NoisyChannelModel has been used for a wide range of applications including speech recognition, part-of-speech tagging, sentence generation etc. This Model treats the observed data (corpus) as a result of unobserved data (headlines) that have been distorted by transmission through noisy channel. This Noisy channel adds story words between the headline words and changes the morphology of some of the headline words. Our task is to find the headline most likely to have generated a given story. That is, each story word is taken to be generated either from the headline word or from a general story language model. Thus stories consist of headline words or morphological variants of headline words or their synonyms with many other words intermingled among them.

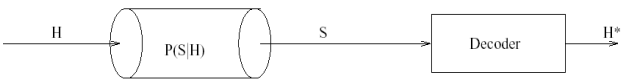

Fig. 4. Noisy Channel Model

Formally, if H is an ordered subset of the first N words of a story S, we want to find the H which maximizes the likelihood that H is the set of headline words in the story S. i. e.

$$argmax_H P(H|S)$$

It is hard to estimate $P(H|S)$, but this probability can be expressed in terms of other probabilities that are easier to compute. Using Bayes' rule :

$$P(H|S) = \frac{P(H)P(S|H)}{P(S)}$$

Since we intend to maximize this expression over H, $P(S)$ can be omitted. So this reduces to:

$$argmax_H P(H)P(S|H)$$

Let H be a headline consisting of words h_1, h_2, ..., h_n in order. The special symbols *start* and *end* represent the beginning and end of an headline. We can estimate P(H) using the bigram probabilities of the headline words used in the story.

$$P(H) = P(h_1|start)P(h_2|h_1).....P(h_n|h_{n-1})$$

To estimate $P(S|H)$, the process by which headline generates the story is to be considered. This process is represented by a Hidden Markov Model (HMM). A HMM is a weighted finite-state automaton, in which each state probabilistically emits an string. Figure 5 represents the simplest HMM to generate stories with headline words. The H state will emit words in the headline only and G state will emit all other words. The HMM switches around between H and G state as needed to generate words in the story.

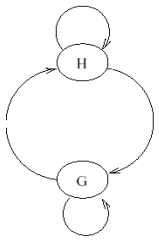

Fig. 5. Hidden Markov Model

Since we are using the bigram probabilities of the headline words, such a simple HMM will not serve our purpose. We need to have a H state for each of the headline words in the vocabulary. Since we assume that only those words in the story can occur in the headline, our vocabulary size reduces to the size of the story. Each H state will have a corresponding G state, which will emit the story words until the next headline word and remember the previously emitted headline word. The HMM of a three-word story is shown in Figure 6.

The G state emits the non-headline words in the story. A G state can emit any word in the story language model. The language story model is represented by a unigram model. For any story, the HMM consists of a start state S, end state E, an H state for each of the words in the story, and a corresponding G state for each H state. Thus, it has $2N + 2$ states. *Each H can emit only its particular word*. The G state remembers which word was emitted by its corresponding H state and can emit any word in the story language model. The HMM starts in state S. From S, it can jump to H_0 state or to G_0 state. When HMM is in an H state, it emits a headline word. From H state, the HMM may transition to the next H state or the corresponding G state. From any G state, the HMM can stay in that G state, or transition to a later H state. Any state can

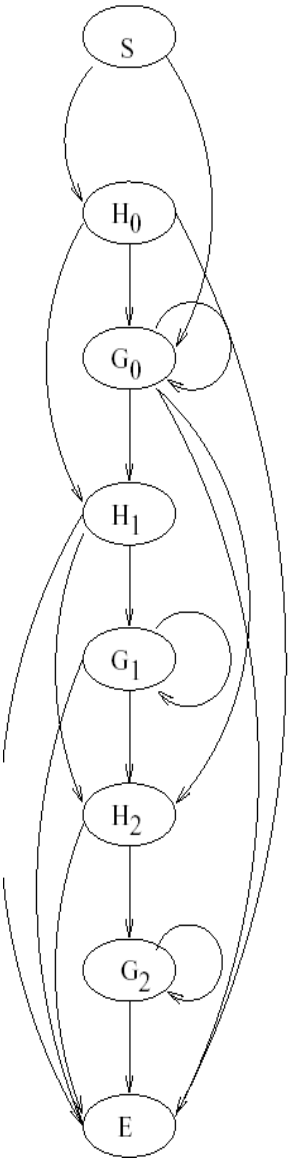

Fig. 6. Modified Hidden Markov Model for our case

transition to the end state. A headline corresponds to a path through the HMM from S to E state, that emits all the words in the story in the correct order. Instead of using all the words in the story one may use first N words, or may rank the sentence and use sentence with highest score, or the first sentence itself. N can be decided experimentally.

$P(S|H)$ is the probability of the story words that are inserted among the headline words. For a given story and headline, if $W = w_1, w_2, ...w_m$ be the words from the story which are not in the headline, and $P(w_i)$ be the unigram probability in the story language of w_i then

$$P(S|H) = P(w_1)P(w_2).....P(w_m)$$

.

Transition from H state to another H State corresponds to a clump of sequential headline words in the story. A transition from H state to G state corresponds to the end of the clump and the start of a gap i.e. headline word followed by non-headline word. Similarly, transition from G state to H state corresponds to end of a gap and start of a clump. This process generates a story and a headline simultaneously. On the other hand, we can think of headline as input to the HMM controlling the sequence of H states, and but the model is free to transition to a G state anytime. This view is similar to the Noisy Channel Model interpretation.

Consider the follwoing excerpt from a news story:

> *Story Words:* For the second time in a week, the country's technology hub had reasons to celebrate as **Wipro** Ltd today **joined** the exclusive club of infotech firms with a **billion-dollar revenue**.
>
> *Generated Headline:* Wipro joined billion-dollar revenue.

In this case, word in bold form from a fluent and accurate headline for the story. However it is often necessary to use a morphological variant of the story word to form a fluent headline. Generally, stories are written in past tense while headlines are written in present tense.

Suppose, the above story excerpt is used as an input to HMM. After passing it through the Stopper Filter, there will be 16 words in in the story. All the low content words like *for*, *the*, *in*, *a* etc. will get removed by the Stopper Filter. So there will be 16 H states, 16 G state, start state S and end state E, a total of 34 states. The above headline can generate the story as follows: The HMM will start from state S, will emit a start symbol then jump to G_0 state, where it will emit the words *time, week, country's, technology, hub, reasons and celebrate*. Then it will jump to state H_{Wipro} and emit the word Wipro. The next word in the story is not in the headline, so the HMM will go to its corresponding G state, G_{Wipro}. This corresponding G_{Wipro} state allows the HMM to remember the last emitted headline word. It will also emit the word *today* from the same G state. The word following it *joined* is in headline. So, transition to H_{joined} will occur and emit *joined*. After emitting *joined* from H state, it again goes to G state, G_{joined} and emits words until it encounter any headline word. Finally, it goes to $H_{billion-dollar}$ state and emits *billion-dollar*. Since, the next word is also in the headline, in this case there is no transition to G state. From $H_{billion-dollar}$ it jumps to $H_{revenue}$ and emits *revenue*. From this $H_{revenue}$, it will go to end state E.

Every possible headline, corresponds to a path through the HMM which emits the story successfully. The path described above is not the only path the can generate the story. Other possibilities are:

> *Headline:* Country's technoloy hub celebrate billion-dollar revenue.
>
> *Headline:* Country's technology hub celebrate.

A. Viterbi Decoding

The Viterbi algorithm selects the most likely headline for a given story. This implementation imposes a constraint that headline words are taken from the story in the order they appear. An Headline state can only emit a specific word, and all other words have zero probability. Each Headline has transitions only to the following headline state or to the corresponding G state.

A two dimensional array is constructed with a row for each state in the HMM and a column for each word in the observed story. Each cell contains a log probability and backtrace to a cell in the previous column. The cells in the first column are initialized in the following way: the log probability of the start state is set to 0 and all the others have negative infinity. The subsequent columns are filled in accordance with the previous column.

The H state cells are assigned as follows: For each H state in the previous column, add the log probability in that cell the log probability that the current story word follows the headline word emitted by that H state in the headline language model. For each G state in the previous column, add to the log probability in that cell, the log probability that the current story word follows the headline word emitted by the H state corresponding to that G state. Then select the highest log probability to store in the current cell and a backtrace to the cell in the previous column from which the log probability was calculated.

The G state cells are assigned as follows: There are only two states in the previous column from which a transition is possible to the current G state, one is the corresponding H state and second the G state itself. Do the same as in the case of H state before, i.e. select the one with highest log probability and add to it the probabability of current story word is generated by the story model and set the backtrace.

After filling up the final column, select the one with maximum log probability in the last column and follow the backtrace. All the words emitted by an H state are included in the headline.

B. Decoding Parameters

The following decoding parameters are used to mimic the actual headlinese: (1) Position penalty (2) String penalty (3) Gap penalty. These decoding parameters change the values in the cell from log probabilities to relative desirability scores. These values are set by trial and error. For improvement, one needs to estimate these values by using learning technique, like *Expectation Maximization* etc.

1) Position Penalty: In human-constructed headlines, the headline words tend to appear near the front of the story because generally topic sentence that conveys the main point of the story appears at the beginning of the story. This position penalty favors headlines which include headline words near the front of the story. The initial position penalty p is positive number less than one. The story word in the n^{th} position gets a position penalty $\log(p^n)$. The emission probabilities on H states are added with the position penalty for position of the word being considered. Thus words near the front of the story carry less of a penalty that farther along. This technique often fails in cases where stories start with a hook to get the reader's attention before getting to the main topic of the story.

2) String Penalty: In human constructed headlines, often contiguous strings of story words appear in the headline like "billion-dollar revenue". This string penalty works as a bias for clumpiness, i.e. the tendency to generate headlines composed of strings of continuous story words. The log of the string penalty is added to the desirability score with each transition from H state to G state. A string penalty lower than one is used to generate clumpy headline.

3) Gap Penalty: In human-constructed headlines, very long gap between headline words tends to be a sign of great effort to piece together a headline from unrelated words. The gap penalty is used to bias against headline gappiness, i.e. the tendency to generate headlines in which contiguous headline words correspond to widely separated story words. At each transition from G state to H state, a gap penalty, depending upon the size of the gap since last headline word was emitted, is added. This also works against spending too much time in one G state. Low gap penalties will favor headlines with few large gaps.

Below is the summary of how the cost and traceback at each cell is calculated:

- **From H state to H state:** cost of getting to previous cell + bigram probability in the headline language of current H-word given previous H-word + positionPenalty.
- **From H state to G state:** cost of getting to previous cell + bigram probability in the headline language of current H-word given previous H-word + positionPenalty + gapPenalty.
- **From G state to H state:** cost of getting to previous cell + unigram probability in the story language of current story word + stringPenalty
- **From G state to G state:** cost of getting to previous cell + unigram probability in the story language of current story word.

C. Generating the headline sentence

The output of the Viterbi Decoding stage is a set of keywords that should be present in the headline. Moreover, the ordering of the keywords is also fixed by the order of the output. To make a natural language expression out of the keywords, the fluffwords that we removed or ignored during document processing are put back. Some heuristics were applied on the keyword set to include words that would enhance grammaticality.

In a grammatically correct phrase, one expects an adjective/adverb to be present with nouns/verbs and not occur in isolation. Using this observation, if an adjective /adverb was output by the Viterbi Decoding algorithm, the following noun/verb was also included in the set of words. Also to increase fluency of the produced headline, if there are adjectives/adverbs occurring before any noun/verb keywords,

those adjectives/adverbs are included in the keywords set.
A Clustering algorithm is then applied on this set of words to produce a coherent sentence i.e the headline.

D. Clustering Algorithm

To achieve grammaticality, bigrams surrounding the chosen set of words, as present in the document, are formed. As in the sample document in Fig 7, one can clearly see clusters of words forming.

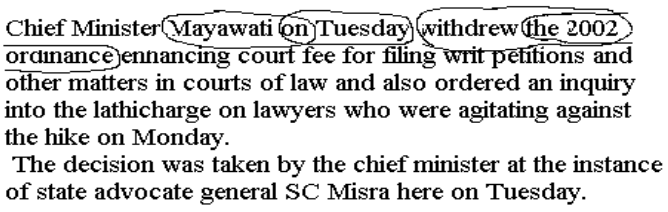

Fig. 7. Clustering example : The fluffwords which lie in the intersection zone of the brackets around keywords are included in the headline sentence. Here the words *on, the 2002* are the selected fluffwords.

Words covered by these clusters form grammatically correct expressions. The words trapped in between 2 clusters are picked along with the neighboring keywords to form phrases . Finally depending upon the length limit of the headline one or more phrases are appended to produce the final headline. If all the keywords are sufficiently close in the document, the headline sentence produced is of very good quality.

IX. Training corpus

We collected an extensive data set consisting of articles with human generated headlines. The following are the categories:

- Business Articles
  Reuters-21578 collection of business news articles.
- Scientific Articles
  General articles on science from INSPEC.
- News Articles
  We downloaded news articles from various news sites. The news texts were downloaded based on a snapshot of the links contained in main page of news sites like BBC [33], TimesOfIndia [34]. A perl program crawled the pages and collected links to news article pages. A page fetcher then downloaded all the pages listed in the collection of the links. There are around 20000 news articles with their respective headlines. Since headlines generated by our system are all words/phrases extracted from the body of the articles, we reduced the set only to only around 15000 articles, each of which contains all of its headline words.

X. Some Results

DOCUMENT:

Chief Minister Mayawati on Tuesday withdrew the 2002 ordinance enhancing court fee for filing writ petitions and other matters in courts of law and also ordered an inquiry into the lathicharge on lawyers who were agitating against the hike on Monday. The decision was taken by the chief minister at the instance of state advocate general SC Misra here on Tuesday.
The advocate general had written a letter to the chief minister on the basis of a request made by the UP Bar Council, after deliberations at a meeting held at Allahabad on January 25. Misra was also present at the meeting as an authorised representative of the state government. He submitted his recommendation for the withdrawal of the aforesaid ordinance to the government in the light of UP Bar Council resolution and representations of various Bar associations, including the High Court Bar Association, Allahabad, Oudh Bar Association, High Court Lucknow, and the Central and Lucknow Bar associations.
Considering the recommendations of advocate general, the state government on Tuesday withdrew the ordinance in question and decided that a fresh decision shall be taken on the basis of the report of UP Bar Council. The Bar association has praised Misra for his efforts and decided to honour him on Wednesday. It may be recalled that on January 25 the UP Bar Council had resolved to appeal the government for withdrawal of the ordinance and had also requested all the state Bar associations to suspended their agitation in the light of high court's interim order and state government's decision in this regard. The Bar Council had also constituted a committee to compile suggestions received from various Bar bodies and submit its recommendations.

TOPIC SENTENCES:

Chief Minister Mayawati on Tuesday withdrew the 2002 ordinance enhancing court fee for filing writ petitions and other matters in courts of law and also ordered an inquiry into the lathicharge on lawyers who were agitating against the hike on Monday.

It may be recalled that on January 25 the UP Bar Council had resolved to appeal the government for withdrawal of the ordinance and had also requested all the state Bar associations to suspended their agitation in the light of high court's interim order and state government's decision in this regard.

GENERATED HEADLINE :

Mayawati on Tuesday withdrew the 2002 ordinance enhancing court fee.

DOCUMENT :

A human reliability model for continuous tasks is explained. Human performance is increased by learning with task involvement. It is assumed that the human error rate for the model decreases monotonically according to time and approximates a certain level. The expressions for human reliability measures are given and parameter estimations derived. Method for optimal adaptation time is described with a view to system economy.

TOPIC SENTENCES:

A human reliability model for continuous tasks is explained. Human performance is increased by learning with task involvement. It is assumed that the human error rate for the model decreases monotonically according to time and approximates a certain level.

GENERATED HEADLINE :

human reliability model for continuous tasks.

XI. Evaluation

We did an informal evaluation of the headlines generated for 40 stories of from INSPEC and 20 from Times of India news article. In INSPEC articles is becomes successful in generating headlines 20% of the case. Some cases in generate totally absurd headline or no headline at all. Since we are using bigram probababilies for the headline language model and unigram of the story language model, if in some article there are words which are less common as headlines, then it produces a empty headline. These can be improved using large number of training data with extractive headline. There is a little improvement of the result in case of Times of India article. In our experiments we tried to run our algorithm for different set of values of position penalty, string penalty and gap penalty. These reveal that the decoding parameters in viterbi algorithm have an impact on the generated headlines. There parameter can be learn from the training data using Expectation Mximization algorithms and bears further study.

XII. Problems in Extractive Headline Generation

As the results show, headline generation using an extractive approach is a feasible methodolgy. In spite of this, many problems exist, specially in the final sentence generation phase. The following are a list of some of the problem areas, we encountered during the course of the project:

- Keywords extraction technique produces a concatenated list of phrases, as the headline, which is not so natural and fluent. Fusion of these keywords to form a sentence is not an easy task. Moreover, there is a need to understand the discourse to come up with correct headlines as presence or omission of words like 'not' can change the meaning of the sentence altogether.
- Domain independent headline generation is difficult as different domains of documents have different statistical traits. The scientific articles usually have headlines made out of some limited number of key phrases widely spread in the document, while news headlines consist of words, usually not many important keywords are present, quite closely found toward the beginning of the article.
- Evaluation of Headlines extraction is difficult. But, using confusion matrix and other methods helps to solve this problem to some extent.
- Statistical based extraction of headlines requires a huge amount of training data that already has headlines created from the document words itself. This is very hard to find. Lacking which the headlines created are not consistently of good quality.

XIII. Future Work

- The data collected was not enough for good results. More data of the order of few lakhs of headline tagged articles need to be assembled and used for training the HMM.
- Very few NLP constraints and heuristics were applied to produce fluent headline sentences from the set of keywords output from the HMM. More elaborate and better constraints and heuristics can be applied to produce high quality headlines.
- Better algorithms may be developed in areas of relevant topic extraction like improvements over discourse tree structure and better understanding of input text by using knowledge base for domain independent platforms.
- Improvements can be implemented for query-oriented, context-sensitive headline extraction, which is also sensitive to user feedback.

XIV. Conclusion

We have presented a system that automatically generates a headline for a single document. This uses a mixed approach of *sentence extraction* and *machine learning*. This hybrid approach enabled us to produce headline shorter than a single sentence. Also, this double process produces a good balance between informativeness, compression and readability of the generated headline. The results show the feasibility of producing reasonable headlines using carefully crafted extractive based approaches. Results also leave room for further improvement, mostly in the quality aspect. During the course of the project, we have realized only an extractive based approach cannot generate very high quality headlines. There is scope for NLP based heuristics that can enhance the sentence generation phase and thus the quality of the final output. Therefore, a careful analysis of results have to be done for the further development of the system.

Acknowledgment

We express our gratitude to Dr. Harish Karnick for his continuous guidance, suggestion and constant encouragement throughout the course of the project work. We would also like to thank Mr. David Zajic, PhD student at University of Maryland, for helping us on an issue of modifying the HMM for our requirements.

REFERENCES

[1] http://algdocs.ncsa.uiuc.edu/PR-20031116-3.ppt.

[2] Inderjeet Mani. *Automatic Summarization*. John Benjamins, 2001.

[3] T. Copeck and S. Szpzkowicz. Picking phrases, picking sentences. In *DUC*, 2002.

[4] N. Okazaki, Y. Matsuo, N. Matsumuru, and M. Ishizuka. Sentence extraction by spreading activation with refined similarity measure. In *16th Intl. FLAIRS Conference*, pages 407–411, 2003.

[5] D. Zajic, B. Dorr, and R. Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. In *Workshop on Automatic Summarization*, May 2003.

[6] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multi-document summarization. In *HLT*, 2002.

[7] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *SIGIR 99, Berkeley, CA*, August 1999.

[8] M. Witbrock and V. Mittal. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In *SIGIR 99, Berkeley, CA*, August 1999.

[9] M. Fuentes, M. Massot, H. Rodriguez, and L. Alonso. Mixed approach to headline extraction. In *DUC*, 2003.

[10] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *4th ACM conference on Digital Libraries*, pages 254–255, August 11-14 1999.

[11] L. Zhou and E. Hovy. Headline summarization at isi. In *DUC*, 2003.

[12] S. Wan, M. Dras, C. Paris, and R. Dale. Using thematic information in statistical headline generation. In *Workshop on Multilingual Summarization and Question Answering: Machine Learning and Beyond, ACL 03*, July 2003.

[13] Rong Jin and Alexander G. Hauptmann. Learning to select good titlewords: An new approach based on reverse information retrieval. In *18th International Conference on Machine Learning (ICML)*, June 28-July 1 2001.

[14] I. Demiros, H. Papageorgiou, and S. Piperidis. Sentence-based text summarization : Modelling and evaluation. In *2nd Hellenic Conf. on AI, SETN-2002*, pages 103–114, 11-12 April 2002.

[15] D. McDonald and H. Chen. Using sentence-selection heuristics to rank text segments in txtractor. In *JCDL*, pages 28–35, 2002.

[16] T.A.S. Pardo, L.H.M. Rino, and M.G.V Nunes. Gistsumm: A summarization tool based on a new extractive method. In *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, pages 210–218, 2003.

[17] W. Kraaij, M. Spitters, and A. Hulth. Headline extraction based on a combination of uni- and multi-document summarization techniques. In *Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, July 2002.

[18] C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *18th International Conference on Computational Linguistics, COLING*, 2000.

[19] W. T. Chuang and J. Yang. Extracting sentence segments for text summarization: A machine learning approach. In *SIGIR*, pages 152–159, 2000.

[20] T. Hirao, K. Takeuchi, H. Isozaki, Y. Sasaki, and E. Maeda. Ntt/naists text summarization systems for tsc-2. In *3rd NTCIR Workshop*, 2003.

[21] R. Jin and A. G. Hauptmann. Headline generation using a training corpus. In *2nd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING*, 2000.

[22] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

[23] P. Kennedy and A. G. Hauptmann. Automatic title generation for the informedia multimedia digital library. In *ACM Digital Libraries, DL-2000*, May 2000.

[24] H. Gregory Silber and Kathleen McCoy. An efficient text summarizer using lexical chains. In *1st International Conference on Natural Language Generation, INLG*, pages 268–271, June 2000.

[25] Barzilay, Regina, and M. Elhadad. Using kxical chains for text summarization. In *Intelligent Scalable Text Summarization Workshop (ISTS97)*, 1997.

[26] L. R. Rabiner. A tutorial on hidden markov models and selected apllications in speech recognition. In A. Waibel and K. F. Lee, editors, *Readings in Speech Recognition*, pages 267–296, San Mateo, CA, 1990. Kaufmann.

[27] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary. Using hmm and logistic regression to generate extract summaries. In *DUC*, 2001.

[28] D. P. O'Leary J. M. Conroy. Text summarization via hidden markov models. In *SIGIR*, pages 406–407, 2001.

[29] D. Zajic, B. Dorr, and R. Schwartz. Automatic headline generation for newspaper stories. In *Workshop on Automatic Summarization*, July 2002.

[30] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTag%ger.html.

[31] http://www.computing.dcu.ie/~acahill/tagset.html.

[32] http://www.cogsci.princeton.edu/~wn.

[33] http://www.bbcnews.com.

[34] http://www.timesofindia.com.

APPENDIX : Some more results

DOCUMENT:

An elated Ajay Jadeja, whose cricket career got a fresh lease of life following the quashing of the ban on him, on Monday said he was already looking forward to returning to international cricket. ”I am very happy with the decision. I am hoping to make a comeback soon,” Jadeja said. A High Court-appointed arbitrator on Monday quashed the five-year ban imposed on the cricketer by the cricket board for his alleged role in match-fixing, making him eligible to play domestic and international cricket again. Jadeja, who was trying his hands in films during his exile, said he had not lost his touch with the game. ”I have been playing cricket. Wherever the ban did not apply, I have been playing. So it is not that I have lost touch,” Jadeja said.

TOPIC SENTENCES :

An elated Ajay Jadeja, whose cricket career got a fresh lease of life following the quashing of the ban on him, on Monday said he was already looking forward to returning to international cricket.
”I am very happy with the decision. I am hoping to make a comeback soon,” Jadeja said.

GENERATED HEADLINE :

elated ajay jadeja, returning to international cricket

DOCUMENT:

In its verdict, the Patna High Court has held that till the liabilities of Bihar State Electricity Board (BSEB) and Jharkhand State Electricity Board (JSEB) were finally determined, they should pay pensionary dues and arrears to employees retired or retiring in the area under their respective jurisdiction. A single bench, presided by Justice Radhamohan Prasad, maintained this while disposing of the writ petitions of a bunch of five writ petitions of BSEB employees who had superannuated before and after the creation of JSEB carved out of BSEB in the wake of the division of Bihar The petitioners were aggrieved as their retiral dues were not paid. The court added that the retiral dues paid to the employees would be subject to final accounting/adjustment of all their liabilities. The court held that the respective boards should discharge their liabilities accordingly by determining the dues payable to all such pensioners, and forward the relevant papers within a week for quick payment of admitted retiral dues. The court observed that the petitioners should not be made to suffer on account of delay in fixing the liabilities of the respective boards. The court held that the cut-off point for fixing liabilities of both the boards was March 5, 2001, as per the agreement reached between Bihar and Jharkhand at a meeting held by the Central government in the wake of Bihar

TOPIC SENTENCES :

In its verdict, the Patna High Court has held that till the liabilities of Bihar State Electricity Board (BSEB) and Jharkhand State Electricity Board (JSEB) were finally determined, they should pay pensionary dues and arrears to employees retired or retiring in the area under their respective jurisdiction.
The court held that the respective boards should discharge their liabilities accordingly by determining the dues payable to all such pensioners, and forward the relevant papers within a week for quick payment of admitted retiral dues.
The court held that the cut-off point for fixing liabilities of both the boards was March 5, 2001, as per the agreement reached between Bihar and Jharkhand at a meeting held by the Central government in the wake of Bihar.

GENERATED HEADLINE :

the patna high court has held liabilities of bihar state electricity board

---

DOCUMENT:

The stand-off between the Pune Associated Cable (PAC) network, one of the largest in the city, and the Zee-Turner and Sony networks continued on Friday, as talks between the two sides remained inconclusive. Subscribers of the PAC network have not been able to view channels from the Star and Sony bouquets for more than three weeks now, due a dispute over pay channel subscriptions. Even Zee-Turner channels went off air on Tuesday. Senior executives of both the networks arrived in Pune on Friday and held a series of meetings with PAC representatives but in vain. PAC president Vasant Patwardhan told TNN, Though we have not reached to any solution on Friday, meetings will continue on Saturday. We are confident the issue can be sorted out mutually. Sudhakar Velankar, president of the cable subscribers co-operative, Grahak Hitavardhini, however, revealed a different side to the story. The blocking of transmission to the PAC network is part of deeper conspiracy on part of Star Television, which wants to have only two networks in the city the Hathway and IC. Star has stakes in the Hathway network and ICC is its official distributor. Pay channel companies have formed a cartel and it is not the interest of a the cartel to have an independent

network like PAC, which has good negotiating power because of its sheer size. ICC Ayaz Inamdar refuted the charge, saying, It is foolish to make such an allegation. Why would a company like Star, whose operations are spread across the world, try to finish-off a small network like PAC?

TOPIC SENTENCES :

The stand-off between the Pune Associated Cable (PAC) network, one of the largest in the city, and the Zee-Turner and Sony networks continued on Friday, as talks between the two sides remained inconclusive.

Subscribers of the PAC network have not been able to view channels from the Star and Sony bouquets for more than three weeks now, due a dispute over pay channel subscriptions. Even Zee-Turner channels went off air on Tuesday.

Even Zee-Turner channels went off air on Tuesday.

GENERATED HEADLINE :

pune cable network zee-turner and sony networks continued

DOCUMENT:

An analysis is made of the behavior of the Hopfield model as a content-addressable memory (CAM) and as a method of solving the traveling salesman problem (TSP). The analysis is based on the geometry of the subspace set up by the degenerate eigenvalues of the connection matrix. The dynamic equation is shown to be equivalent to a projection of the input vector onto this subspace. In the case of content-addressable memory, it is shown that spurious fixed points can occur at any corner of the hypercube that is on or near the subspace spanned by the memory vectors. Analysed is why the network can frequently converge to an invalid solution when applied to the traveling salesman problem energy function. With these expressions, the network can be made robust and can reliably solve the traveling salesman problem with tour sizes of 50 cities or more.

TOPIC SENTENCES :

An analysis is made of the behavior of the Hopfield model as a content-addressable memory (CAM) and as a method of solving the traveling salesman problem (TSP).

The analysis is based on the geometry of the subspace set up by the degenerate eigenvalues of the connection matrix.

The dynamic equation is shown to be equivalent to a projection of the input vector onto this subspace.

GENERATED HEADLINE :

behavior hopfield content-addressable memory solving the traveling salesman problem

---

DOCUMENT:

An introduction to artificial neural network models is presented, along with an overview of their practical application and potential applications in signal processing. Successful neural network implementations are described and their performances are compared to those of more traditional signal processing implementations. The Hopfield net, self-organizing feature maps, and the multilayer perceptron are reviewed. Implementation of neural nets in speech synthesis, speech recognition, target identification, image processing, pattern matching, error-correction coding, and neurocomputing are reported. Several ICs in production are briefly mentioned.

TOPIC SENTENCES :

An introduction to artificial neural network models is presented, along with an overview of their practical application and potential applications in signal processing.

Successful neural network implementations are described and their performances are compared to those of more traditional signal processing implementations.

GENERATED HEADLINE :

introduction to artificial neural network models

DOCUMENT:

A novel approach to identifying the kinematic models of redundant or dual manipulators without end-point sensing is presented. Starting from the observation that such manipulators can be made to form mobile closed kinematic chains, it is shown that these closed loops can be identified by an iterative-least-squares algorithm similar to that used in calibrating open-chain manipulators. Simulations have demonstrated that this technique is viable. The issue of the identifiability of the kinematic parameters of the closed loop is addressed.

TOPIC SENTENCES :

A novel approach to identifying the kinematic models of redundant or dual manipulators without end-point sensing is presented.

Starting from the observation that such manipulators

can be made to form mobile closed kinematic chains, it is shown that these closed loops can be identified by an iterative-least-squares algorithm similar to that used in calibrating open-chain manipulators.

GENERATED HEADLINE :

kinematic models dual manipulators closed kinematic chains

DOCUMENT:

In real-world domains, large amounts of knowledge are needed to adequately describe world behavior. With a complex domain theory, complete reasoning becomes a computationally intractable task. As a result, systems operating in these types of situations may not have complete knowledge of the world. One problem with using such a reasoning framework is that sometimes it will result in the learning of inefficient or suboptimal plans. If a system acts on incomplete information, it may make poor decisions. This paper presents methods for detecting and repairing plans which are suboptimal due to inference limitations. By noticing and analyzing fortuitous occurrences, the system can improve its plans and hence its performance. These methods cover both learning from observation and from the system's own problem-solving and represent a general framework of refinement for inference-limited systems.

TOPIC SENTENCES :

In real-world domains, large amounts of knowledge are needed to adequately describe world behavior.
With a complex domain theory, complete reasoning becomes a computationally intractable task.
As a result, systems operating in these types of situations may not have complete knowledge of the world.

GENERATED HEADLINE :

large amounts of knowledge adequately describe

DOCUMENT:

Telecom sector got some boost because of proposed decline in duties on import of capital goods to make components as well as on optical fibre, but mobile handset buyers from the legal market did not get any relief. Handsets will continue to be as expensive as they are now on account of import duties but setting up telecom networks and expanding project could get a little cheaper. The budget has also proposed addition in the service tax from 5 per cent to 8 per cent, which the consumers will have to pay for. Industry experts have said that there are already a number of taxes like revenue share, USO (universal service obligation) and service tax, which keep phones services expensive. "We believe that the government should now reduce revenue share for telecom service operators," said Bharti CMD Sunil Mittal. He welcomed other initiatives to reduce duties but said that they will have marginal impact on the sector. Pankaj Mohindroo, president of ICA (Indian Cellular Association), he was "extremely disappointed" with the finance ministry for neither reducing the basic customs duty on handsets from 10 per cent to 5 per cent, nor removing 4 per cent special addition duty (SAD). "We have projected that by 2006, the government could loose about Rs 5,000 crore revenue because of grey market, which accounts for 70 per cent of all the handsets in the country," he added. Motorola country president, Pramod Saxena welcomed the decision to reduce overall duty impact on import of telecom equipment to 15 per cent. "Extension of tax holiday to R&D centres till March 2004 supports company like ours, who are investing in developing and leveraging the local software expertise for next generation technologies," he added. Saxena said, however, that Motorola expected import duties on handsets to be lowered to curb the grey market. Ravi Sharma, MD, Alcatel India, subsidiary of the French telecom giant, which makes telecom switches and transmission equipment here, welcomed the duty cuts and said that it would boost manufacturing in the country.

TOPIC SENTENCES :

Telecom sector got some boost because of proposed decline in duties on import of capital goods to make components as well as on optical fibre, but mobile handset buyers from the legal market did not get any relief.
Handsets will continue to be as expensive as they are now on account of import duties but setting up telecom networks and expanding project could get a little cheaper.
Industry experts have said that there are already a number of taxes like revenue share, USO (universal service obligation) and service tax, which keep phones services expensive.

GENERATED HEADLINE :

telecom sector got of proposed decline in duties import mobile handset

---