Text Summarization via Hidden Markov Models *

John M. Conroy
Center for Computing Sciences
Institute for Defense Analyses
17100 Science Drive
Bowie, MD 20715
conroy@super.org

Dianne P. O'Leary †
Computer Science Dept. and
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742 USA
oleary@cs.umd.edu

ABSTRACT

A sentence extract summary of a document is a subset of the document's sentences that contains the main ideas in the document. We present an approach to generating such summaries, a hidden Markov model that judges the likelihood that each sentence should be contained in the summary. We compare the results of this method with summaries generated by humans, showing that we obtain significantly higher agreement than do earlier methods.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; G.3 [**Probability and Statistics**]: Markov Processes

Keywords

text summarization, extract summaries, hidden Markov models, automatic summarization, document summarization

1. INTRODUCTION

Automatic summarization has been studied for over 40 years [7], but with pervasive use of information retrieval systems in the last 6 years this area has been given wider attention [4]. Here, we focus on generic summaries, i.e., ones that attempt to capture the essential points of a document, although most of the ideas presented can be adapted to generate query-based summaries of text. We generate *sentence extract summaries*; i.e., the summary consists of a subset of the document's sentences. We follow [1] in the use of Natural Language Processing techniques to "go beyond the words" and instead focus on *terms*.

*A full version of this paper is available as *Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition* at Univ. of MD Comp. Sci. Tech Report, March 2001

†The work of the second author was supported by NSF Grant CCR 97-32022

SIGIR'01, September 9-12, New Orleans, Louisiana, USA..

The method we propose for text summarization is a *Hidden Markov Model (HMM)*. Jing and McKeown [5] previously proposed a HMM for decomposing a human summary, i.e., mapping the component parts of a summary generated by a human back into the document. Here, we present the first HMM for use in summarizing text.

2. THE HIDDEN MARKOV MODEL

In this section we describe an approach that given a set of features computes an a-posterior probability that each sentence is a summary sentence. In contrast to a naive Bayesian approach [6] [1], the HMM has fewer assumptions of independence. In particular, it does not assume that the probability that sentence i is in the summary is independent of whether sentence $i-1$ is in the summary. Furthermore, we use a joint distribution for the features set, unlike the independence-of-features assumption used by naive Bayesian methods.

We consider three features in the development of a Hidden Markov model for text summarization.

- position of the sentence in the document. This feature is built into the state-structure of the HMM.
- number of terms in the sentence. The value of this feature is $o_1(i) = \log(\text{number of terms} + 1)$.
- how likely sentence terms are, given the document terms $o_2(i) = \log(Pr(\text{terms in sentence} i|D))$.

We expect that the probability that the next sentence is included in the summary will differ, depending on whether the current sentence is a summary sentence or not. A first order Markov model allows such differences with marginal additional cost over a simple Bayesian classifier.

A HMM handles the positional dependence, dependence of features, and Markovity. (For more details about HMMs the reader should see [2] [8].) The model we propose has $2s+1$ states, with s summary states and $s+1$ non-summary states. A picture of the Markov chain is given in Figure 1. Note that we allow hesitation only in non-summary states and skipping of states only from summary states. This chain is designed to model the extraction of up to $s-1$ lead summary sentences and an arbitrary number of supporting sentences. Using training data we obtain a maximum likelihood estimate for each transition probability, and this forms an estimate M for the transition matrix for our Markov chain, where element (i, j) of M is the estimated probability of transitioning from state i to state j.

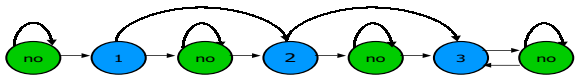

Figure 1: Summary Extraction Markov Model to Extract 2 Lead Sentences and Additional Supporting Sentences

Associated with each state i is an output function, $b_i(O) = Pr(O|\text{state } i)$, where O is an observed vector of features. We make the simplifying assumption that the features are multi-variant normal. The output function for each state can be estimated by using the training data to compute the maximum likelihood estimate of its mean and covariance matrix. We estimate $2s + 1$ means, but assume that all of the output functions share a common covariance matrix.

With this model we compute $\gamma_t(i)$, the probability that sentence t corresponds to state i. If i is even then this probability represents the probability that sentence t is the $i/2$-th summary sentence. If i is odd then it is the probability that it is a non-summary sentence. We compute the probability that a sentence is a summary sentence by summing $\gamma_t(i)$ over all even values of i. This posterior probability, which we define as g_t is used to select the most likely summary sentences.

3. RESULTS AND CONCLUSIONS

The 1304 documents we used in our test were taken from the TREC data set [9]. This included articles from the Associated Press, *Financial Times*, *Los Angeles Times*, *Washington Post*, *Wall Street Journal*, *Philadelphia Inquirer*, *Federal Registry*, and *Congressional Record*. A single person generated summaries for all of the documents.

Each of these data sets was divided into two pieces, one used to train the parameters of the model and one used for evaluation of the methods. We compared our algorithm's summaries with the human summaries, computing the following score. For each document we let k_h be the length of the human summary, k_m the length of the machine generated summary, and r the number of sentences they share in common. Then we define a metric to compare the two summaries by:

$$F_1 = 100\frac{2r}{k_h + k_m} \quad (1)$$

In the following table we report average F_1 for various data sets for the HMM and naive Bayesian method as given in [1]. These are simply defined as the mean value of the respective score over the document set.

Data	DimSum	HMM
ap-test	52	56
cr-test	35	47
fr-test	39	46
ft-test	46	53
latwp-test	45	53
pi-test	41	55
wsj-test	54	65

Table 1: F1 HMM and DimSum

We have presented a novel HMM for generating sentence abstract summaries of documents. The algorithm is quite successful in generating summaries that agree well with human-generated summaries, despite using minimal natural language processing (NLP) information, just the extraction of terms. Further results including comparisons with multiple human summarizers and sample summaries can be found in [3].

4. REFERENCES

[1] C. Aone, M. Okurowski, J. Gorlinsky, and B. Larsen. A scalable summarization system using robust nlp. *Proceeding of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73, 1997.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Stat.*, 41:164–171, 1970.

[3] J. M. Conroy and D. P. O'Leary. Text summarization via hidden markov models and pivoted qr matrix decomposition. Technical report, University of Maryland, College Park, Maryland, March 2001.

[4] U. Hahn and I. Mani. The challenges of automatic summerization. *IEEE Computer*, 33(11):29–36, 2000.

[5] H. Jing and K. R. M. Keown. The decomposition of human-written summary sentences. *Proceedings of SIGIR-99 (Melbourne, Australia)*, pages 129–136, 1999.

[6] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. *Proceedings of the 18th Annual International SIGIR Conference on Research and Development in Infomation Retrieval*, pages 68–73, 1995.

[7] H. P. Luhn. The automatic creation of literture abstracts. *IBM Journal of Research Development*, 2:159–165, 1958.

[8] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77:257–285, 1989.

[9] TREC Conference Series. Text REtrieval Conference (TREC) text research collection. Technical Report http://trec.nist.gov/, National Institute of Standards and Technology, Gaithersburg, Maryland, 1994, 1996, 1997.