Automatic Title Generation for Spoken Broadcast News

Rong Jin
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA 15213
412-268-7003
rong+@cs.cmu.edu

Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
412-268-1448
alex+@cs.cmu.edu

ABSTRACT
In this paper, we implemented a set of title generation methods using training set of 21190 news stories and evaluated them on an independent test corpus of 1006 broadcast news documents, comparing the results over manual transcription to the results over automatically recognized speech. We use both F1 and the average number of correct title words in the correct order as metric. Overall, the results show that title generation for speech recognized news documents is possible at a level approaching the accuracy of titles generated for perfect text transcriptions.

Keywords
Machine learning, title generation

1. INTRODUCTION
To create a title for a document is a complex task. To generate a title for a spoken document becomes even more challenging because we have to deal with word errors generated by speech recognition.

Historically, the title generation task is strongly connected to traditional summarization because it can be thought of extremely short summarization. Traditional summarization has emphasized the extractive approach, using selected sentences or paragraphs from the document to provide a summary. The weaknesses of this approach are inability of taking advantage of the training corpus and producing summarization with small ratio. Thus, it will not be suitable for title generation tasks.

More recently, some researchers have moved toward "learning approaches" that take advantage of training data. Witbrock and Mittal [1] have used Naïve Bayesian approach for learning the document word and title word correlation. However they limited their statistics to the case that the document word and the title word are same surface string. Hauptmann and Jin [2] extended this approach by relaxing the restriction. Treating title generation problem as a variant of Machine translation problem, Kennedy and Hauptmann [3] tried the iterative Expectation-Maximization algorithm. To avoid struggling with organizing selected title words into human readable sentence, Hauptmann [2] used K nearest neighbour method for generating titles. In this paper, we put all those methods together and compare their performance over 1000 speech recognition documents.

We decompose the title generation problem into two parts: learning and analysis from the training corpus and generating a sequence of title words to form the title.

For learning and analysis of training corpus, we present five different learning methods for comparison: Naïve Bayesian approach with limited vocabulary, Naïve Bayesian approach with full vocabulary, K nearest neighbors, Iterative Expectation-Maximization approach, Term frequency and inverse document frequency method. More details of each approach will be presented in Section 2.

For the generating part, we decompose the issues involved as follows: choosing appropriate title words, deciding how many title words are appropriate for this document title, and finding the correct sequence of title words that forms a readable title 'sentence'.

The outline of this paper is as follows: Section 1 gave an introduction to the title generation problem. The details of the experiment and analysis of results are presented in Section 2. Section 3 discusses our conclusions drawn from the experiment and suggests possible improvements.

2. THE CONTRASTIVE TITLE GENERATION EXPERIMENT
In this section we describe the experiment and present the results. Section 2.1 describes the data. Section 2.2 discusses the evaluation method. Section 2.3 gives a detailed description of all the methods, which were compared. Results and analysis are presented in section 2.4.

2.1 Data Description
In our experiment, the training set, consisting of 21190 perfectly transcribed documents, are obtain from CNN web site during 1999. Included with each training document text was a human assigned title. The test set, consisting of 1006 CNN TV news story documents for the same year (1999), are randomly selected from the Informedia Digital Video Library. Each document has a closed captioned transcript, an alternative transcript generated with CMU Sphinx speech recognition system with a 64000-word broadcast news language model and a human assigned title.

2.2 Evaluation
First, we evaluate title generation by different approaches using the F1 metric. For an automatically generated title Tauto, F1 is

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2001		**00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automatic Title Generation for Spoken Broadcast News	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**3**	

measured against corresponding human assigned title Thuman as follows:

$$F1 = 2 \times precision \times recall / (precision + recall)$$

Here, precision and recall is measured respectively as the number of identical words in Tauto and Thuman over the number of words in Tauto and the number of words in Thuman. Obviously the sequential word order of the generated title words is ignored by this metric.

To measure how well a generated title compared to the original human generated title in terms of word order, we also measured the number of correct title words in the hypothesis titles that were in the same order as in the reference titles.

We restrict all approaches to generate only 6 title words, which is the average number of title words in the training corpus. Stop words were removed throughout the training and testing documents and also removed from the titles.

2.3 Description of the Compared Title Generation Approaches

The five different title generation methods are:

1. **Naïve Bayesian approach with limited vocabulary (NBL)**. It tries to capture the correlation between the words in the document and the words in the title. For each document word DW, it counts the occurrence of title word same as DW and apply the statistics to the test documents for generating titles.
2. **Naïve Bayesian approach with full vocabulary (NBF)**. It relaxes the constraint in the previous approach and counts all the document-word-title-word pairs. Then this full statistics will be applied on generating titles for the test documents.
3. **Term frequency and inverse document frequency approach (TF.IDF)**. TF is the frequency of words occurring in the document and IDF is logarithm of the total number of documents divided by the number of documents containing this word. The document words with highest TF.IDF were chosen for the title word candidates.
4. **K nearest neighbor approach (KNN)**. This algorithm is similar to the KNN algorithm applied to topic classification. It searches the training document set for the closest related document and assign the training document title to the new document as title.
5. **Iterative Expectation-Maximization approach (EM)**. It views documents as written in a 'verbal' language and their titles as written a 'concise' language. It builds the translation model between the 'verbal' language and the 'concise' language from the documents and titles in the training corpus and 'translate' each testing document into title.

2.4 The sequentializing process for title word candidates

To generate an ordered set of candidates, equivalent to what we would expect to read from left to right, we built a statistical trigram language model using the SLM tool-kit (Clarkson, 1997) and the 40,000 titles in the training set. This language model was used to determine the most likely order of the title word candidates generated by the NBL, NBF, EM and TF.IDF methods.

3. RESULTS AND OBSERVATIONS

The experiment was conducted both on the closed caption transcripts and automatic speech recognized transcripts. The F1 results and the average number of correct title word in correct order are shown in Figure 1 and 2 respectively.

KNN works surprisingly well. KNN generates titles for a new document by choosing from the titles in the training corpus. This works fairly well because both the training set and test set come from CNN news of the same year. Compared to other methods, KNN degrades much less with speech-recognized transcripts. Meanwhile, even though KNN performance not as well as TF.IDF and NBL in terms of F1 metric, it performances best in terms of the average number of correct title words in the correct order. If consideration of human readability matters, we would expect KNN to outperform considerately all the other approaches since it is guaranteed to generate human readable title.

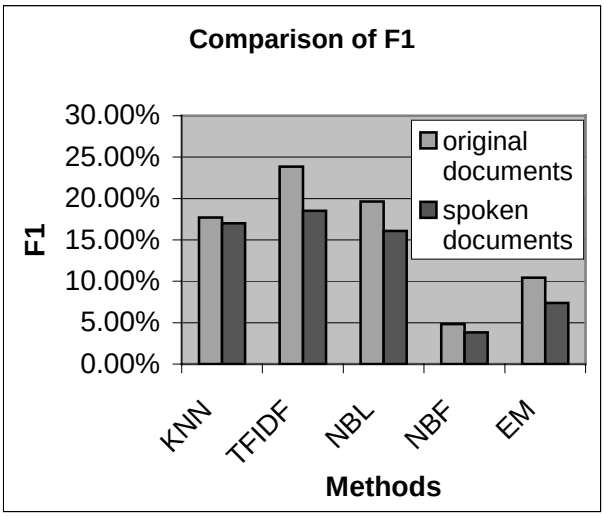

Figure 1: Comparison of Title Generation Approaches on a test corpus of 1006 documents with either perfect transcript or speech recognized transcripts using the F1 score.

NBF performs much worse than NBL. NBF performances much worse than NBL in both metrics. The difference between NBF and NBL is that NBL assumes a document word can only generate a title word with the same surface string. Though it appears that NBL loses information with this very strong assumption, the results tell us that some information can safely be ignored. In NBF, nothing distinguishes between important words and trivial words. This lets frequent, but unimportant words dominate the document-word-title-word correlation.

Light learning approach TF.IDF performances considerably well compared with heavy learning approaches. Surprisingly, heavy learning approaches, NBL, NBF and EM algorithm didn't out performance the light learning approach TF.IDF. We think learning the association between document words and title words by inspecting directly the document and its title is very problematic since many words in the document don't reflect its content. The better strategy should be distilling the document first before learning the correlation between document words and title words.

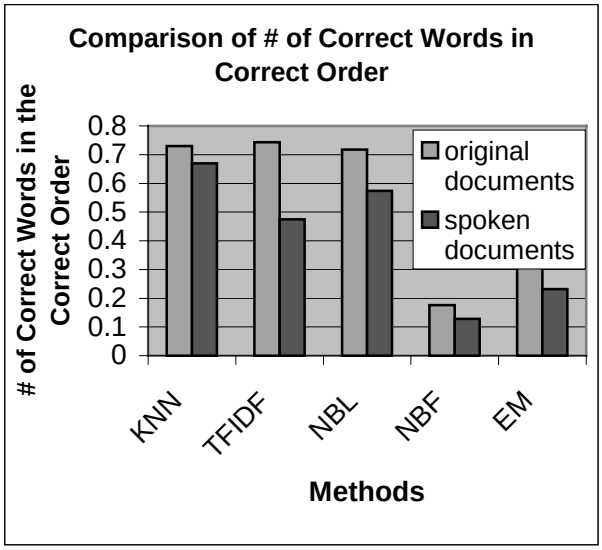

Figure 1: Comparison of Title Generation Approaches on a test corpus of 1006 documents with either perfect transcript or speech recognized transcripts using the average number of correct words in the correct order.

4. CONCLUSION

From the analysis discussed in previous section, we draw the following conclusions:

1. The KNN approach works well for title generation especially when overlap in content between training dataset and test collection is large.
2. The fact that NBL out performances NBF and TF.IDF out performance NBL and suggests that we need to distinguish important document words from those trivial words.

5. ACKNOWLEDGMENTS

This material is based in part on work supported by National Science Foundation under Cooperative Agreement No. IRI-9817496. Partial support for this work was provided by the National Science Foundation's National Science, Mathematics, Engineering, and Technology Education Digital Library Program under grant DUE-0085834. This work was also supported in part by the Advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or ARDA.

6. REFERENCES

[1] Michael Witbrock and Vibhu Mittal. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. Proceedings of SIGIR 99, Berkeley, CA, August 1999.

[2] R. Jin and A.G. Hauptmann. Title Generation for Spoken Broadcast News using a Training Corpus. Proceedings of 6th Internal Conference on Language Processing (ICSLP 2000), Beijing China. 2000.

[3] P. Kennedy and A.G. Hauptmann. Automatic Title Generation for the Informedia Multimedia Digital Library. ACM Digital Libraries, DL-2000, San Antonio Texas, May 2000.