Automatic Title Generation for Spoken Broadcast News

Rong Jin and Alexander G. Hauptmann
{rong+,alex+}@cs.cmu.edu
Language Technology Institute, School of Computer Science,
Carnegie Mellon University
Pittsburgh, PA, 15213

ABSTRACT

We implemented several statistical title generation methods using a training set of 21190 news stories and evaluated them on an independent test corpus of 1006 broadcast news documents, comparing the resulting titles based on manual transcription to the titles from automatically recognized speech. We use both F1 and the average number of correct title words in the correct order as evaluation metrics. 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.

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, title generation is strongly connected to traditional summarization because title generation can be thought of as extremely short summarization. Traditional summarization has emphasized the extractive approach, using selected sentences or paragraphs from the document to provide a summary (McKeown et al, 1995, Salton et al., 1997). The weakness of the extractive approach is its inability to take advantage of the training corpus and its difficulty in producing good summarization as a very small fraction of the complete text. Thus, extractive summarization is not ideal for title generation.

More recently, researchers have moved toward "learning approaches" that take advantage of training data. Witbrock and Mittal (1999) have used a Naïve Bayesian approach for learning the document word and title word correlation. However they limited their statistics to the case where the document word and the title word have the same surface string. Jin and Hauptmann (2001) extended this approach by relaxing the restriction. Treating title generation problem as a variant of Machine translation problem, Kennedy and Hauptmann (1999) tried an iterative Expectation-Maximization algorithm. To avoid struggling with organizing selected title words into human readable sentence, Jin and Hauptmann (2000) also tried a K nearest neighbor method for generating titles. In this paper, we contrast all these methods and compare their performance on title generation for 1006 speech recognized broadcast news 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 compare five different learning methods: a Naïve Bayesian approach with limited vocabulary, a Naïve Bayesian approach with full vocabulary, K nearest neighbors, an Expectation-Maximization approach, and a term frequency and inverse document frequency method. Details of each approach are presented in Section 2.

We decompose the generation phase as follows:

1. Choosing appropriate title words,

2. Finding the correct sequence of title words that form a readable title 'sentence'.

2 THE CONTRASTIVE TITLE GENERATION EXPERIMENT

2.1 Data Description

Our training set, consisting of 21190 perfectly transcribed documents, was obtained from the CNN.com 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), were randomly selected from the Informedia Digital Video Library. Each document has a closed captioned transcript, an alternative transcript generated with the CMU Sphinx speech recognition system using a 64000-word broadcast news language model and an original title provided by CNN. The Word Error Rate was 35%.

2.2 Evaluation

First, we evaluated title generation using an F1 metric. For an automatically generated title T_{auto}, F1 is measured against a corresponding human-assigned title T_{human} as follows:

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

Precision and recall is measured respectively as the number of identical words in T_{auto} and T_{human} over the number of words in T_{auto} and the number of words in T_{human}. Obviously, sequential word order of the generated title words is ignored by this metric.

To compare a generated title to the original human-generated title in terms of word order, we counted the number of correct title words in the hypothesis titles that were in the same order as in the reference titles.

2.3 Description of the Title Generation Approaches

The five different title generation methods are:

1. **Naïve Bayesian approach with limited vocabulary (NBL).** NBL tries to capture the correlation between the words in the document and the words in the title. For each title word TW, it counts the occurrences of document word DW, if DW = TW. For generation we apply the conditional statistics to generate new title words (Witbrock and Mittal, 1999).

2. **Naïve Bayesian approach with full vocabulary (NBF).** It relaxes the constraint (TW = DW) in the previous approach and counts all the document-word-title-word pairs. Then the full conditional probabilities will be applied to generating titles words for 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 (Rjiesbergen, 1979). The document words with highest TF.IDF are chosen as title word candidates.

4. **K nearest neighbor approach (KNN).** This algorithm is similar to the KNN algorithm applied to topic classification (Yang and Chute, 1994). It searches the training document set for the closest related document and assigns the training document title to the new document.

5. **Iterative Expectation-Maximization approach (EM)**. EM views documents as written in 'verbose' language and their titles as written in 'concise' language (Brown et al, 1990). EM builds the translation model between the 'verbose' language and the 'concise' language from the documents and titles in the training corpus and 'translates' each test document into a title.

We restricted all approaches to generate only 6 title words, which was the average training title length. Stop words were removed throughout the training and testing documents and also from the titles.

2.4 Sequencing the title word candidates

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

3 RESULTS AND OBSERVATIONS

The title-generation experiment compared the closed caption transcripts and automatic speech recognized transcripts as input. 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 worked well because both the training and test set came from CNN news of the same year. Compared to other methods, KNN degraded much less with speech-recognized transcripts. Meanwhile, even though KNN performs worse than TF.IDF and NBL in terms of the F1 metric, it performs 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 considerably outperform all the other approaches since it is guaranteed to generate human readable title.

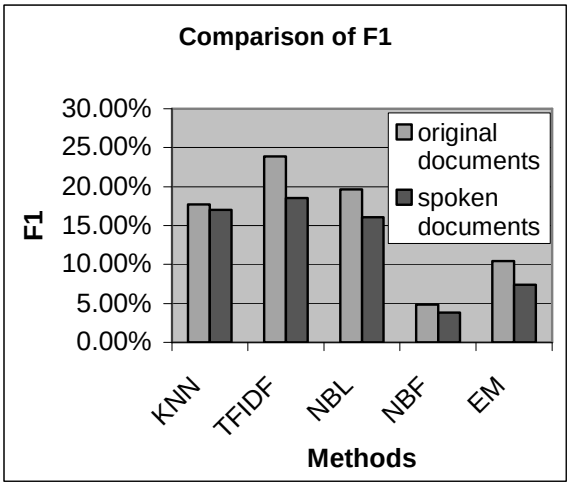

Figure 1: Comparison of Title Generation Approaches on a test corpus of 1006 documents with either perfect transcript or speech-recognized transcripts using the F1 score.

NBF performs much worse than NBL. NBF performs 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. Despite this very restrictive assumption, the results show 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.

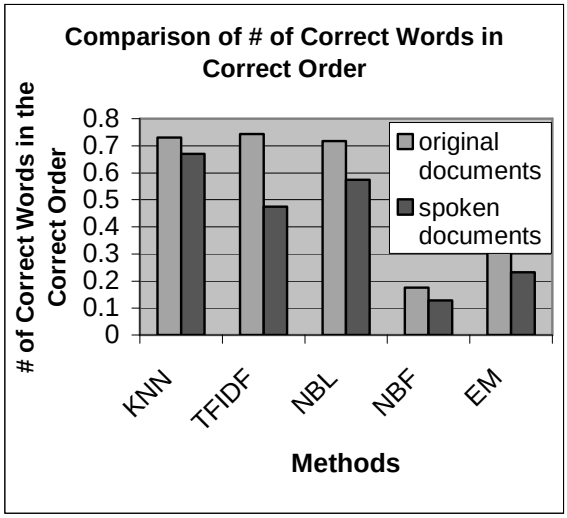

Figure 1: Comparison of Title Generation Approaches on a test corpus of 1006 documents with either perfect transcript or speech recognized transcripts using the average number of correct words in the correct order.

TF.IDF performs surprisingly well compared to the other true learning approaches. Surprisingly, the 'heavy' learning approaches, which really take advantage of the training corpus, such as NBL, NBF, EM didn't outperform the 'shallow' learning approach of TF.IDF, which selects the title words from the document without learning anything about the titles. One suspicion is that learning the association between document words and title words by directly inspecting the document and its title is very problematic. Many words in the document don't reflect its content. A better strategy may be to first distill the document so that only the essential content words are left and then compute the association between the distilled documents and their titles.

4 CONCLUSION

We conclude that KNN is desirable for title generation especially when overlap in content between training dataset and test collection is large.

The fact that NBL outperforms NBF and TF.IDF out performs NBL suggests that we need to distinguish important document words from those trivial, unimportant words.

Finally, we are encouraged that title generation from speech recognized documents is possible at a level approaching title generation from perfect text transcription.

5 ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Cooperative Agreement No. IRI-9817496 and 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 the Advanced Research and Development Activity.

6 References

M. Witbrock and V. Mittal, "Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries", *Proceedings of SIGIR 99*, Berkeley, CA, August 1999

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.

K. McKeown, J. Robin and K. Kukich, Generating Concise Natural Language Summaries, *Information Processing and Management*, 31 (5), pp.703-733, 1995.

R. Jin and A.G. Hauptmann, Title Generation Using Training Corpus, Proceedings of CICLING-2001, Mexico City, Mexico, 2001, in press.

R. Jin and A.G. Hauptmann, Title Generation for Spoken Broadcast News using a Training Corpus, Proceedings of ICSLP 2000, Beijing China. 2000.

P.R. Clarkson and R. Rosenfeld. *Statistical Language Modeling Using the CMU-Cambridge Toolkit* Proceedings ESCA Eurospeech 1997

G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summary", *Info. Proc. And Management*, 33(2): 193-207, March 1997.

Y. Yang, and C.G. Chute, "An example-based mapping method for text classification and retrieval", *ACM Transactions on Information Systems (TOIS)*, 12(3): 252-77. 1994.

V. Rjiesbergen. Butterworths, *Information Retrieval*, Chapter 7. London, 1979.

P. Brown, S. Cocke, S. Della Pietra, Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and Roossin, A Statistical Approach to Machine Translation, Computational Linguistics V. 16, No. 2, June 1990.