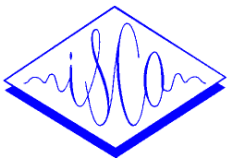

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Title Generation for Spoken Broadcast News using a Training Corpus

Rong Jin, Alex G. Hauptmann

Language Technology Institute, School of Computer Science,
Carnegie Mellon University

ABSTRACT

The problem of title generation involves finding the essence of a document and expressing it in only a few words. The results of a query to the Informedia Digital Video Library are summarized through an automatically generated title for each retrieved news story. When the document is errorful, as with speech-recognized broadcast news stories, the title creation challenge becomes even greater. We implemented a set of title word selection strategies and evaluated them on an independent test corpus of 579 broadcast news documents, comparing manual transcription results to automatically recognized speech using the CMU Sphinx speech recognition system with a 64000-word broadcast news language model. Using a training collection of 21190 transcribed broadcast news stories, we trained several systems to produce appropriate title words, i.e. Naïve Bayesian approach with full vocabulary, Naïve Bayesian approach with limited vocabulary, nearest neighbor approach and extractive approach. The F1 results shows that the nearest neighbor approach is a quick and easy way of generating good titles for speech recognized documents (F1 = 15.2%), while a Nave Bayesian approach with limited vocabulary also does well on our F1 measure (F1 = 21.6%), which ignores word order in the titles. 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. One surprising phenomenon is that extractive approach performances slightly better for speech recognized documents than for manual transcripts.

1 INTRODUCTION

To create a title for a document is to engage in a complex task: One has to understand what the document is about, one has to know what is characteristic of this document with respect to other documents, one has to know how a good title sounds to catch attention and how to distill the essence of the document into a title of just a few words. To generate a title for a spoken document becomes even more challenging because we have to deal with word errors generated by speech recognition. The speech recognition challenges include: lack of punctuation and capitalization in the transcripts, acoustical recognition errors resulting in words inserted, substituted and deleted from the transcript, as well as transcription errors introduced by the limited vocabulary (usually 64k words) of the speech system. Generating text titles for spoken documents is very attractive because it produces a very compact representation of the original spoken document, which will help people to understand the important information contained in the document quickly, without listening to the whole audio stream. From the viewpoint of machine learning, studies on how well general title generation methods can be adapted to errorful document transcriptions and which methods perform better than others will be very helpful in general understanding how to discover knowledge from noisy observations and how to apply learned knowledge in noisy environments.

In our approach to the title generation problem we will assume the following:

First, the system will be given a set of training data. Each datum consists of a document and corresponding title. After exposure to the training corpus, the system should be able to generate a title for any unseen document.

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 the learning part, we have to decide which parts of knowledge the system needs to learn and how to represent the knowledge learned from the training data. There are four pieces of knowledge that can be induced from training data:

1. Knowledge from the analysis of the document to be titled (referred to as Kd),
2. Knowledge from the analysis of the documents in the training corpus (referred to as the language model for all the documents in the training corpus, or LD),
3. Knowledge from the analysis of the titles in the training corpus (referred to as the language model for all the titles in the training corpus, or LT),
4. Knowledge from analysis of the correlation between documents and their corresponding titles (referred to as the joint document/title language model for the training corpus, or JL).

From the viewpoint of generating part, we decompose the issues involved as follows:

1. Choosing appropriate title words,
2. Deciding how many title words are appropriate for this document title,
3. Finding the correct sequence of title words that forms a readable title ‘ sentence’ .

Historically, the title generation task is strongly connected to traditional summarization [2] 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 [3,4,5]. The weakness of this approach is that most of knowledge

(referred to as LD, LT and JL above) embedded in the training corpus is ignored.

More recently, some researchers have moved toward "learning approaches" that take advantage of training data. For example, Witbrock and Mittal [1] ignore all document words that are not in the title language model LT. Only document words that effectively reappear in the title of a document are counted when they estimate the probability of a title word w_t given a document word w_d as: $P(w_t|w_d)$ where $w_t = w_d$. While the Witbrock/Mittal Naïve Bayesian approach is not in principle limited to this constraint, our experiments show that it is a very useful restriction. Kennedy tried a generative approach with an iterative Expectation-Maximization algorithm using most of the document vocabulary [10].

For the purpose of comparison over a test pool and to present contrastive results, in this paper we explore the following learning approaches:

1. Extractive summarization which selects the "best" sentence from the document as a title.

2. Naïve Bayesian approach with limited vocabulary. This closely mirrors the experiments reported in [1].

3. KNN (k nearest neighbors) which treats title generation as a special classification problem. We consider the titles in the training corpus as a fixed set of labels, then the task of generating a title for a new document is essentially the same as selecting an appropriate label (title) from the fixed set of training labels. The task reduces to finding the document in the training corpus, which is most similar to the current document to be titled. Standard document similarity vectors can be used. The new document title will be set to the title for the training document that is most similar to the current new document.

4. Naïve Bayesian approach with full vocabulary. In this approach, we compute the probability of a title word given a document word for all words in the training data, not just those that where $w_t = w_d$.

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 test set consists of 579 news story documents, where each document has a closed captioned transcript, an alternative transcript generated with automatic speech recognition and a human assigned title. The training set consists of 21190 perfectly transcribed documents. Included with each training document text was a human assigned title.

Both the test set and the training set came from 1999 CNN news. By crawling over the CNN web sites during 1999, we obtained 21190 news articles and their corresponding titles and used them as our training dataset. We randomly selected 579 CNN TV news stories for the same year (1999) from the Informedia Digital Video Library [11], which provided the audio content and closed caption transcript for each CNN TV news story, and used these to form our test suite. Using the CMU Sphinx speech recognition system [12] with a 64000-word broadcast news language model, we obtained automatic speech recognition transcripts for all the audio content in the test documents. To obtain the human assigned titles for the spoken documents in the test suite we looked through the CNN news web site (cnn.com) during the year 1999 and extracted the titles of news articles that were almost identical to the closed caption transcripts of the spoken news story documents in the test set.

2.2 Evaluation

In this paper, we evaluate title generation by different approaches using the F1 metric [7]. For an automatically generated title T_{auto}, F1 is measured against corresponding human assigned title T_{human} as follows:

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

Here, precision and recall is measured 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} respectively. Obviously the sequential word order of the generated title words is ignored by this metric. There are several other metrics that take this ordering information into consideration: string edit distance (DTW) [9] and maximal sub-string [1]. Since we are focusing here on the choice of appropriate title words, F1 is the most appropriate measure for this purpose. Furthermore, we ignore those parts of each title generation approach, which order the generated title words into a word sequence or title 'sentence'. To make a fair comparison between different approaches, all approaches only generate 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

As we mention in section 1, we will compare 4 different title generation methods. They are:

1. **Naïve Bayesian approach with limited vocabulary (NBL).** Essentially, this algorithm duplicates the work in [1], which 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).** In the previous approach, we count only the cases where the title word and the document word are same. This restriction is based on the assumption that a document word is only able to generate a title word with same surface string. The constraint can be easily relaxed by counting all the document-word-title-word pairs and apply this full statistics on generating titles for the test documents.

3. **Extractive summarization approach (TF.IDF).** We have mentioned the similarity between title generation and story summarization. In this paper, we use the "Automatic Summarization" functionality inside Microsoft WORD as the tool of extracting title sentence.

4. **K nearest neighbor approach (KNN).** This algorithm is similar to the KNN algorithm applied to topic classification in [6]. 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 document title is then used for the new document. In our experiment, we use SMART [8] to index our training documents and test documents with the weight schema "ATC" [8]. The similarity between documents is defined as the dot product between document vectors. The training document closest related to the test document is found out by computing the similarity between the test document and each training document.

3 RESULTS AND OBSERVATIONS

The experiment was conducted both on the closed caption transcripts and automatic speech recognized transcripts. The results are shown in Figure 1. The K-nearest neighbor (KNN) and Naïve Bayesian with limited vocabulary (NBL) approaches perform best for both manual transcripts and automatic speech recognized documents, 16.6% and 21.6% for manual transcripts and 15.2% and 16.6% for speech recognized documents. The extractive approach (TF.IDF) performs worst in the case of manual transcripts (5.6%) but improves slightly with the speech-recognized documents (7.4%). The Naïve Bayesian method with full vocabulary (NBF) degrades most for the speech recognized transcripts, whose F1 score drops from 8.2% for manual transcript to 3.6% for speech recognized transcripts.

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, which provides a good overlap in content coverage between training set and test set. In addition, compared to the Naïve Bayesian approach with limited vocabulary (NBF), KNN degrades much less with speech recognized transcripts than NBF (8% degradation for KNN vs. 23% degradation for NBF). We conclude that KNN is good candidate for generating titles for spoken documents if there is strong overlap in coverage between the training and test data. From the knowledge viewpoint, KNN grasps part of

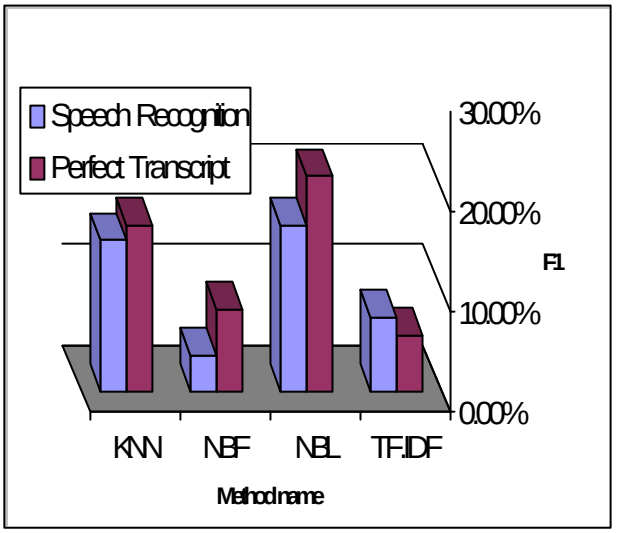

Figure 1: Comparison of Title Generation Approaches on a test corpus of 579 documents with either perfect transcript or speech recognized transcripts using the F1 score.

knowledge in LD and JL, which is much better than the sentence-based extractive approach. In addition, unlike the Naïve Bayesian approaches that decompose the documents into words and build the statistics based just on the words, KNN treats the document and its title as a single entity, which contributes to its stable performance in case of speech recognized transcripts. However, it has the adverse effect from the generating viewpoint because there is no flexibility in generating a new title. If consideration of human readability matters, which our F1 scores did not reflect, we would expect KNN to outperform all the other approaches since it is guaranteed to generate human readable title.

NBF performs much worse than NBL. NBF performances much worse than NBL in terms of both F1 measurement and

degradation with speech recognized transcripts (56% degradation for NBF vs. 23% degradation for NBL). The difference between NBF and NBL is that NBL assumes a document word can only generate a title word with the same surface string. From the knowledge viewpoint, it should be losing information with this very strong assumption. However, the results tell us that some information can safely be ignored. In NBF, nothing distinguishes between important words and trivial words, and the co-occurrence between all document words and title words is measured equally. This lets frequent, but unimportant words dominate the document-word-title-word correlation. As an extreme example, stop words show up frequently in every document. However, they are likely to have little effect on the choice of title words.

Thus, even though NBF seems to exploit more knowledge than NBL, it introduces more noise by not limiting the effects of frequent, but unimportant words, which not only pulls down the performance but also makes the method more sensitive to word errors in the transcript. This conflict suggests a strategy, which neither ignores the knowledge nor overemphasizes frequent, unimportant words.

Extractive approach improves with speech recognized transcripts. The very surprising phenomenon of the extractive approach is that the F1 score of the extractive method is improved for the case of speech recognized transcripts. This is totally against common sense because we expect to see the degradation in performance caused by the word errors from speech recognition. Actually a similar phenomenon has been found while generating titles for translated documents [13]. We believe this is not due to random effects or programming bugs/mistakes. One possible explanation is that even though the speech recognition word error rate was estimated to be about 25-35%, the essential content of the spoken document is not significantly corrupted by the speech recognition. On the contrary, the incorrect words can make the extractive approach discard unimportant sentences more easily than usual. Thus, the chance of discovering an important sentence for title generation may be increased.

4 CONCLUSION AND FUTURE WORK

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. It is much more stable in noise circumstance compared with NBL. Other advantages of KNN are that it is very simple in terms of implementation and always produces human readable titles.

2. The comparison between the performance of NBF and NBL shows that we need to treat important words and trivial words differently to limit the noise introduced by frequent, but trivial words. This kind of treatment not only brings the improvement in F1 score but also makes the methods more insensitive to word errors.

3. Word errors can bring improvement in F1 score for some methods when the essential parts of documents can still be kept intact.

Possible improvements can be done from both the learning viewpoint and the generating viewpoint:

Learn from all types of knowledge. Most methods focus on how to learn the knowledge JL, which is about the correlation between document words and title words. However, other knowledge sources, i.e. LD and LT are hardly considered. One possible improvement is to take advantage of the knowledge LT and LD when considering the document-word-document-word correlation and title-word-title-word correlation in computing document-word-title-word correlation.

Learn JL knowledge correctly For all the methods discussed in section 2, words have been treated as the basic units in collecting statistics. No attempt is made to understand the correlation between document word sequence and title word sequence. One way to improve it is to compute the statistics on multigrams, not just unigrams. In this case, we can have statistics not only on the correlation between document words and title words, but also on the correlation between the sequential order of document words and title words.

Have more flexibility on generating titles. In this paper, we actually implement the KNN approach as the single nearest neighbor (K=1). This has no flexibility to form new titles and restricts any new title to be identical to one in the training pool. We can relax this restriction by choosing K (K>1) most similar documents in the training pool and form a new title based on the K chosen titles. This may create a more appropriate title for a new document. The trade-off would be that the title may not be human readable anymore.

5 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. Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell.Summarizing, “Text Documents: Sentence Selection and Evaluation Metrics”, *Proceedings of SIGIR 99*, Berkeley, CA, August 1999.

3. T. Strzalkowski, J. Wang, and B. Wise, “A robust practical text summarization system”, *AAAI Intelligent Text*

Summarization Workshop, pages 26-30, Stanford, CA, March 1998.

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

5. M. Mitra, Amit Sighal, and Chris Buckley, "Automatic text summarization by paragraph extraction", *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

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

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

8. Gerard Salton, *The SMART Retrival System: Experiments in Automatic Document Proceeding*, Prentice Hall, Englewood Cliffs, New Jersey. 1971.

9. Nye, H., "The Use of a One Stage Dynamic Programming Algorithm for Connected Word Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. **AASP-32**, No 2, pp. 262-271, April 1984.

10. Kennedy, P and Hauptmann, A.G., "Automatic Title Generation for the Informedia Multimedia Digital Library", *ACM Digital Libraries, DL-2000*, San Antonio Texas, May 2000, in press.

11. The Informedia Digital Video Library Project http://www.informedia.cs.cmu.edu/

12. Seymore, K., Chen, S., Doh, S., Eskenazi, etc, "The 1997 CMU Sphinx-3 English Broadcast News Transcription System", *Proceedings of the DARPA Workshop on Broadcast News Understanding Systems (BNTUW-98)*, Lansdowne, VA, February 1998.

13. Jin, R. and Hauptmann, A.G, "Cross Lingual Title Generation: Initial Steps", *Workshop on Interactive Searching in Foreign-Language Collections*, Human-Computer Interaction Laboratory, University of Maryland, College Park, MD. June 1, 2000. http://www.clis.umd.edu/conferences/hcil00