Accepted to The Eighteenth International Conference on Machine Learning (ICML-2001), *Williams College, Williamsport, Maryland,* June 28-July 1, 2001

Learning to Select Good Title Words: An New Approach based on Reverse Information Retrieval

Rong Jin rong+@CS.CMU.EDU
Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA15213, USA

Alexander G. Hauptmann alex+@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA15213, USA

Abstract

In this paper, we show how we can learn to select good words for a document title. We view the problem of selecting good title words for a document as a variant of an Information Retrieval problem. Each title word is treated as a "document" and selection of appropriate title words as finding relevant "documents". Based on our training collection consisting of 40,000 document and title pairs, we learn the "document" representations for all the title words and apply these learned representations to select appropriate title words over 10,000 test documents. Compared to other learning approaches, namely K nearest neighbor approach, a Naïve Bayesian approach and a variant of a machine translation model, we find that our approach is significantly better as indicated by the F1 metric.

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. Title generation is very desirable and useful because it produces a compact representation of the original document, which helps people to quickly understand the important information contained in a document. For research, title generation is a very difficult problem from the viewpoint of machine learning and natural language processing.

The first stage for solving the title generation problem is to find title words that reflect the main content of the document, which we call the title word selection problem. Historically, the title generation task is strongly connected to traditional text summarization (Goldstein et al., 1999) and emphasizes the extractive approach which selects words, sentences or paragraphs from the document to provide a summary (Strzalkowski & Wang & Wise, 1998; Salton et al., 1997; Mitra & Singhal & Buckley, 1997). More recently, some researchers have moved toward "learning approaches" that take advantage of training data. Different from the traditional text summarization approaches, such as the extraction of key phrases (Eibe et al., 1999), which can only extract the text units from the original documents to compose the concise representation of the document, these learning approaches can actually generate titles with words not coming from the original documents. By learning the correlation between title words and document words from the document and title pairs in the training corpus, we can apply the learned association to the new unseen document and select the title words with the highest scores. Thus, we are not limited to select words only from the document. Witbrock and Mittal (1999) have used a Naïve Bayesian approach in learning the correction between document words and title words. In their approach, they ignore all document words that do not appear in the title. Only document words that effectively reappear in the title of a document are counted when they estimate the probability of generating a title word wt given a document word wd as: P(wt|wd) where wt = wd. While the Witbrock/Mittal Naïve Bayesian approach is not in principle limited to this constraint, the work of Jin and Hauptmann (2000a; 2000b) showed that it is a very useful restriction. By treating title generation as a variant of the Machine Translation problem, Kennedy and Hauptmann (2000) came up with the generative approach using iterative Expectation-Maximization algorithm. K nearest neighbor approach for selecting title words has been tried and showed its outstanding performance in (Jin & Hauptmann 2000a; Jin & Hauptmann 2000b).

In this paper, we present a novel approach for selecting good title words, which views the title word selection problem as a variant of an Information Retrieval (IR) problem. The Information Retrieval problem is to find out the relevant documents from the text collection given a user query, while the title word selection problem is to select the representative title words from the title word vocabulary for a test document. By mapping the concepts "title word" and "test document" from the title word selection problem into the concepts "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 the documents in IR, similar to the test document, equivalent to the user query in IR.

The essential difficulty with handling title word selection problem as an IR is the representation of title words. In the Information Retrieval problem, each document is represented as a feature vector, which consists of all the words from the document and usually is weighted by some term (i.e. word) weighting scheme. However, for all the title words in our title word vocabulary, there is no more information for them except their surface string. Thus, before we are going to apply the Information Retrieval techniques to the title word selection problem, we need to represent every title word as a feature vector, which consists of document words and their weights. In this paper, we present an optimization algorithm for finding these representation vectors for title words.

To assess the effectiveness of our approach, we compare it with three other learning approaches that have been used for title word selection task, namely Naïve Bayesian approach with limited vocabulary, K nearest neighbor approach, iterative Expectation-Maximization method. All the approaches are described in more detail in section 3.

The outline of this paper is as follows: In this section we gave an introduction to the title word selection problem. The explanation of our approach is presented in Section 2. The details of the contrastive experiment are presented in Section 3. The results and their analysis are presented in Section 4. Section 5 discusses our conclusions drawn from the experiment and suggests possible improvements.

2. Information Retrieval Approach toward Title Word Selection

In this section, we will first present our "Information Retrieval" model for the title word selection problem in Section 2.1. Section 2.3 describes how to obtain the vector representation of title words. Section 2.4 summarizes the procedures in our model to select title words.

2.1 "Information Retrieval" Model

In the previous section we have mentioned the similarity between Information Retrieval and title word selection, i.e. both look into their collections for objects similar to the input 'questions'. In the case of title word selection, the input question is a "Test Document" and the objects it searches for are title words. For the Information Retrieval task, the input questions are user queries and the desired objects are relevant documents. Figure 1 gives the graphic representation of frameworks for these two tasks. Shown in Figure 1, by simply mapping "Test Document" and "Title Words" in the task of selecting title words to "User Query" and "Relevant Document" in Information Retrieval task, we can treat the title word selection problem as an Information Retrieval problem and apply all the techniques developed in Information Retrieval, such as tf.idf term weighting scheme (Salton & Buckley, 1988) and pseudo-relevance feedback, to select good title words.

As pointed out in the previous section, the difficulty in applying Information Retrieval techniques to title words selection problem lies in the representation of title words, i.e. there is no convenient way to represent each title word as a vector of document words. In the next section, we will show an optimization method that is able to learn the representation vector for title words.

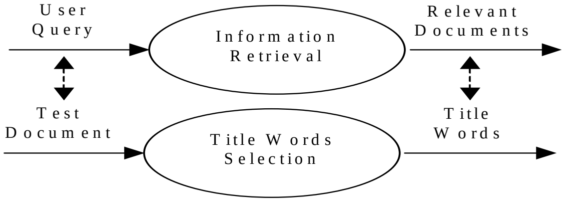

Figure 1: The graphic representation of information retrieval and title word selection, and the mapping between them.

2.2 Vector Representation of Title Words

To find out the optimum representation vectors for title words, let's assume that we have already obtained the representation vectors for all the title words. If those vectors are good for representing title words, by applying our Information Retrieval model, we should be able to generate title words similar to what human subjects have created for all the documents in training corpus, or more specifically, the difference between the human assigned title words and machine-generated title words over all the training documents will be minimized. In the following we will give the mathematical description of our method.

Let N be the number of documents in the training collection, N_{tw} be the number of distinct no-stop words in the title of the documents in the training collection and N_{dw} be the number of distinct no-stop words in the documents in the training collection.

Let $\mathbf{D}$ be a matrix with N rows and N_{dw} columns. An element $\mathbf{d}_{ij}$ in $\mathbf{D}$ is the number of occurrences in the i-th document of the j-th document word. Let $\mathbf{d}_i$ be the i-th row vector in $\mathbf{D}$ (of length N_{dw}). The row vector $\mathbf{d}_i$ characterizes the i-th document.

Let $\mathbf{T}$ be a matrix with N rows and N_{tw} columns. An element $\mathbf{t}_{ij}$ in $\mathbf{T}$ is the number of occurrences in the title of the i-th document of the j-th title word. Let $\mathbf{t}_i$ be the i-th row vector in $\mathbf{T}$ (of length N_{tw}). The row vector $\mathbf{t}_i$ characterizes the title of the i-th document.

In the standard Information Retrieval paradigm, a query $\mathbf{q}$ is represented by a row vector of length N_{dw} and a document $\mathbf{d}$ is represented by a row vector of length N_{dw}. The numbers in these vectors represent the weights of the corresponding words in the given query and document. The strength of the match between the query $\mathbf{q}$ and the document $\mathbf{d}$ is given by the inner product $\mathbf{qd}^{\mathrm{T}}$. When we issue a query to a search engine, we get back a list of documents with their similarity scores, where the scores are the inner products $\mathbf{qd}_1^{\mathrm{T}}$, $\mathbf{qd}_2^{\mathrm{T}}$, ..., $\mathbf{qd}_N^{\mathrm{T}}$ between the given query $\mathbf{q}$ and the documents in the collection $\mathbf{d}_1$, $\mathbf{d}_2$, ..., $\mathbf{d}_N$. More concisely, we can introduce a score vector $\mathbf{s}$ of length N and define $\mathbf{s}$ as $\mathbf{s}=(\mathbf{qd}_1^{\mathrm{T}}, \mathbf{qd}_2^{\mathrm{T}}, ..., \mathbf{qd}_N^{\mathrm{T}})$, or $\mathbf{s}= \mathbf{qD}^{\mathrm{T}}$.

We wish to adapt this paradigm to assign title words to documents. To do this, we need to represent every title word by a row vector of length N_{dw}, where the numbers in this vector represent the strength of the connection between each document word and the given title word. We will introduce a matrix $\mathbf{M}$ for this purpose.

Let $\mathbf{M}$ be a matrix with N_{tw} rows and N_{dw} columns. An element $\mathbf{m}_{ij}$ in $\mathbf{M}$ is an estimate of the strength of the connection between the i-th title word and the j-th document word. Let $\mathbf{m}_i$ be the i-th row vector in M. The row vector $\mathbf{m}_i$ is a vector of length N_{dw} and represents the i-th title word as a vector in document word space. Later, we will show how to calculate $\mathbf{M}$ from $\mathbf{D}$ and $\mathbf{T}$.

Let $\mathbf{d}_{\mathrm{test}}$ be the document vector (of length N_{dw}) for a document taken from the testing set. Our goal is to select title words for $\mathbf{d}_{\mathrm{test}}$. We treat the document $\mathbf{d}_{\mathrm{test}}$ as if it were a query vector in Information Retrieval, and we produce a list of title words with their scores, where the scores are the inner products $\mathbf{d}_{\mathrm{test}}\mathbf{m}_1^{\mathrm{T}}$, $\mathbf{d}_{\mathrm{test}}\mathbf{m}_2^{\mathrm{T}}$, ..., $\mathbf{d}_{\mathrm{test}}\mathbf{m}_{Ntw}^{\mathrm{T}}$ between the given document $\mathbf{d}_{\mathrm{test}}$ (analogous to $\mathbf{q}$) and the title words $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_{Ntw}$ (analogous to $\mathbf{d}_1$, $\mathbf{d}_2$, ..., $\mathbf{d}_N$). We take the K (in the following experiments, K = 6) title words with the highest scores as our chosen title words for the given document $\mathbf{d}_{\mathrm{test}}$. Similar to the treatment in Information Retrieval, we simplify the score expression using matrix multiplication. Let $\mathbf{s}$ be the score vector of length N_{tw} and $\mathbf{s}$ is defined as $\mathbf{s}=(\mathbf{d}_{\mathrm{test}}\mathbf{m}_1^{\mathrm{T}}, \mathbf{d}_{\mathrm{test}}\mathbf{m}_2^{\mathrm{T}}, ..., \mathbf{d}_{\mathrm{test}}\mathbf{m}_{Ntw}^{\mathrm{T}})$, or $\mathbf{s}=\mathbf{d}_{\mathrm{test}}\mathbf{M}^{\mathrm{T}}$. The i-th element in the score vector $\mathbf{s}$ is the score of the i-th title word for the test document $\mathbf{d}_{\mathrm{test}}$.

Therefore, for any document i in the training collection, the corresponding title word score vector $\mathbf{s}_{\mathrm{i}}$ (of length N_{tw}) can be written as $\mathbf{s}_{\mathrm{i}}=\mathbf{d}_{\mathrm{i}}\mathbf{M}^{\mathrm{T}}$. We measure the error in the score vector $\mathbf{s}_{\mathrm{i}}$ for the document i in the training collection as the sum of the squares of the differences between the author's title words $\mathbf{t}_{\mathrm{i}}$ and the machine generated title words $\mathbf{s}_{\mathrm{i}}$, i.e.

$$err = \sum_{i=1}^{N} \| \mathbf{t}_i - \mathbf{s}_i \|_2 = \sum_{i=1}^{N} \| \mathbf{t}_i - \mathbf{d}_i\mathbf{M}^T \|_2 \qquad (1)$$

where the $\| \, \|_2$ represents the Euclidean vector length, i.e. the sum of the squares of all the numbers in the vector.

We are trying to minimize the difference between the matrix $\mathbf{T}$ (N rows and N_{tw} columns) of authors' titles and the matrix $\mathbf{DM}^{\mathrm{T}}$ (N by N_{dw} times N_{dw} by N_{tw}) of mechanically assigned titles. In principle, it is possible to optimize M to minimize the error function *err*, using the Singular Value Decomposition (SVD). However, for large N_{tw}, N_{dw} and N, it will be very expensive to find the matrix $\mathbf{M}$ that minimizes the error function in Equation (1), even with Singular Value Decomposition (*SVD*) package for sparse matrices (Press, 1993). To avoid the computational complexity, we will compute matrix $\mathbf{M}$ as follows.

First, we can expand the error function in Equation (1) as following:

$$\begin{aligned} err &= \sum_{i=1}^{N} \| \mathbf{t}_i - \mathbf{s}_i \|_2 \\ &= \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{s}_i)(\mathbf{t}_i - \mathbf{s}_i)^T \qquad (2) \\ &= \sum_{i=1}^{N} \mathbf{t}_i\mathbf{t}_i^T + \sum_{i=1}^{N} \mathbf{s}_i\mathbf{s}_i^T - \sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T \end{aligned}$$

In the above equation, we can see that the first term, i.e. $\sum_{i=1}^{N} \mathbf{t}_i\mathbf{t}_i^T$, is a constant. Thus, it can be ignored. The second term, $\sum_{i=1}^{N} \mathbf{s}_i\mathbf{s}_i^T$ is a quadratic term and the last term $-\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T$ is a linear term. Therefore, most of the computation complexity is introduced by the second term. To make the computation simple and tractable, we can approximate the error function by throwing out the quadratic term and only keeping the linear term. Then, the approximated error function, named *err'* is

$$\begin{aligned} err' &= -\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T = -\sum_{i=1}^{N} 2\mathbf{t}_i(\mathbf{d}_i\mathbf{M}^T)^T \qquad (3) \\ &= -\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{M}\mathbf{d}_i^T \end{aligned}$$

The intuition behind this approximation is straightforward: The term $\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T$ measures the similarity between the human assigned titles $\mathbf{t}_i$ and machine-generated titles $\mathbf{s}_i$. Minimizing the difference between the human assigned title words $\mathbf{t}_i$ and the machine created title words $\mathbf{s}_i$ will approximately equal to maximizing the similarity between them. Therefore, to minimize the true error function in Equation (1), we can actually maximize the similarity function $\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T$, or minimize its negative, i.e. $-\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{s}_i^T$, which is exactly the approximated error function in Equation (3).

However, there is one problem with minimizing the approximated error function *err'*. Since the error function *err'* is linearly dependant on the matrix **M** and there is no constraint on the matrix **M,** the error function *err'* will have no lower bound. To avoid the case that the error function *err'* goes to negative infinity, we can enforce the Euclidean length of title word representation vector $\mathbf{m}_i$ to be 1, i.e.

$$\|\mathbf{m}_i\|_2 = \sum_{j=1}^{N_{dw}} \mathbf{m}_{ij}^2 = 1, \ \forall i \tag{4}$$

The optimum solution **M,** which minimizes the error function *err'* in Equation (3) and also satisfies the set of constraints in Equation (4), can be found using the method of *undetermined Lagrangian multiplier* (Hildebrand, 1952). For all values of $\mathbf{m}_{ij}$, we set to 0 the partial derivatives with respect to

$$-\sum_{i=1}^{N} 2\mathbf{t}_i\mathbf{M}\mathbf{d}_i^T + \sum_{i=1}^{N_{tw}} \lambda_i \sum_{j=1}^{N_{dw}} \mathbf{m}_{ij}^2$$

The result is

$$\lambda_i \mathbf{m}_{ij} = \sum_{k=1}^{N} \mathbf{t}_{ki}\mathbf{d}_{kj} \tag{5}$$

The constants λ_j is determined by the Equation (4), i.e.

$$\sum_{j=1}^{N_{dw}} \mathbf{m}_{ij}^2 = \lambda_i \sum_{j=1}^{N_{dw}} \left\{ \sum_{k=1}^{N} \mathbf{t}_{ki}\mathbf{d}_{kj} \right\}^2 = 1 \ \forall i$$

2.3 Procedures to Select Title Words

In this section, first we will summarize the steps on how to learn the representation vectors from the training corpus for all the title words, and then describe the steps on how to select good title words for a new document using the learned title word representation vectors.

To obtain the representation vectors for title words, we need to do the following:

- Build the document vector $\mathbf{d}_i$ and title vector $\mathbf{t}_i$ for every document and title pair in the training corpus. The j-th element of the vector $\mathbf{d}_i$, i.e. $\mathbf{d}_{ij}$, will be the number of occurrence of the j-th document word in the i-th document. The j-th element of the vector $\mathbf{t}_i$, i.e. $\mathbf{t}_{ij}$, is the number of occurrence of the j-th title word in the i-th title divided by the title length.
- Compute the elements in matrix **M** by

$$\mathbf{m}_{ij}' = \sum_{k=1}^{N} \mathbf{t}_{ki}\mathbf{d}_{kj}$$

- Finally, according to Equation (4) normalize $\mathbf{m}_{ij}$' as

$$\mathbf{m}_{ij} = \frac{\mathbf{m}_{ij}'}{\sqrt{\sum_j \mathbf{m}_{ij}'^2}}$$

To apply our "Information Retrieval" model to the title word selection task, we will

- Weight the title word representation vectors $\mathbf{m}_i$. Since we treat each "title word" as the "document" in the Information Retrieval problem, we can view the whole set of representation vectors for the title words as "document collection" in Information Retrieval. Thus, the standard term weighting scheme, used in the field of Information Retrieval, can be applied directly to weight the title word representation vectors. In our experiment, we use "ATC" term weighting scheme (Salton & Buckley, 1988) within the "SMART" system (Salton, 1971).
- Use standard Information Retrieval system, "SMART" in our experiment, to compute the similarity between the test document and the representation vector for each title word and retrieve the top K (K is 6 in our experiment) close title words as selected title words.

3. The Contrastive Experiment

In this section we describe the experimental setup. Our contrastive experiment is designed to evaluate the effectiveness of our learning approach compared with other learning approaches. Section 3.1 describes the data used for training and testing. Section 3.2 discusses the evaluation method. Section 3.3 gives a detailed description of all the methods, which were compared.

3.1 Data Description

The experimental dataset comes from a CD of 1997 broadcast news transcriptions published by Primary Source Media (1997). There were a total of 50,000 documents and corresponding titles in the dataset. The training dataset was formed by randomly picking four documents-title pairs from every five pairs in the original dataset. Thus, the size of training corpus was 40,000 documents with corresponding titles. The test collection consisted of 1000 documents randomly selected from the remaining 10,000 documents not used for training. By separating training data and test data in this way, we ensure strong overlap in topic content between training dataset and test dataset, which gives the learning algorithms a chance to play a significant role in the headline generation.

3.2 Evaluation Method

In this paper, we measure the quality of selected title words by comparing the human assigned title words with what machine has generated. More specifically, we use the F1 metric (Rjiesbergen, 1979). For a set of automatically generated title words T_{auto}, F1 is measured against the correspondent set of title words assigned by human subjects T_{human} as follows,

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

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. Precision shows, in the title generated by computer, the percentage of words being "correct", i.e. words are also chosen by human subjects. Meanwhile recall gives the percentage of "correct" words that computer has selected, among the title assigned by human subjects.

To make all approaches comparable (except KNN), only 6 title words were generated by each method, which was the average number of title words in the training corpus. The KNN method always uses the title of the document in the training corpus most similar to the test document as the title for the test document and the restriction of six words does not apply to titles generated by KNN. Since we wanted to emphasize content word accuracy, stop words were removed throughout the training and testing documents and titles.

3.3 Description of Learning Approaches to Compare

As we mentioned in the introduction, we compared our approach with other three learning approaches for title word selection. They were:

- **Naïve Bayesian approach with limited vocabulary** (NBL). Essentially, this algorithm duplicates the work by Witbrock and Mittal (1999), which tries to capture the correlation between the words in the document and the words in the title. It defines the conditional probability between the title word *tw* and the document word *dw*, i.e. *P(tw|dw)* as:

$$P(tw \mid dw) = \begin{cases} 0 & \text{if } tw \neq dw \\ \dfrac{C(tw \in title \wedge dw \in document)}{C(dw \in document)} & \text{if } tw = dw \end{cases}$$

  where $C(dw \in \text{document})$ is the number of documents with the word *dw* in the document text and $C(tw \in \text{title} \wedge dw \in \text{document})$ is the number of documents with the document word *dw* in the document text and the title word *tw* in the title.

  To generate title words, we merely apply the statistics *P(tw|dw)* to the test document *D* and compute the score *P(tw|D)* for every title word *tw* as

$$P(tw \mid D) = \sum_{dw \in D} C(dw, D) P(tw \mid dw)$$

  Here, *C(dw,D)* is the number of occurrence of document word *dw* in the document *D*. The title words with highest score will be selected as title words for the document *D*.

- **K nearest neighbor approach** (KNN). This algorithm is similar to the KNN algorithm applied to topic classification in (Yang & Chute, 1994). It treats the titles in the training corpus as a set of fixed labels. For each new document, instead of creating a 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" system (Salton, 1971) to index our training documents and test documents with the term weighting scheme "ATC" (Salton & Buckley, 1988). The similarity between documents is defined as the dot product between document vectors. The training document closest to the test document is found by computing the similarity between the test document and each training document.

- **Iterative Expectation-Maximization approach** (EM). This algorithm reproduces the work by Kennedy and Hauptmann (2000), which treats title generation as a translation problem. This method views a document as written in a 'verbose' language and its corresponding title as written in a 'concise' language. The approach builds a statistical translation model (Brown et al., 1990) between the 'verbose' and the 'concise' languages based on the documents and their titles in the training corpus. According to the statistical machine translation model (Brown et al., 1990), the conditional probability *P(tw|dw)* can be computed as follows:

$$P(tw \mid dw) = \lambda_{dw} \sum_{D} \frac{C(dw, D) C(tw, T(D)) P(tw \mid dw)}{\sum_{dw \in D} C(dw', D) C(tw, T(D)) P(tw \mid dw')}$$

  where *C(dw,D)* is the occurrence of the document word *dw* in the document *D* and *C(tw, T(D))* is the occurrence of the title word *tw* in the title of the document *D*. The normalization constant λ_{dw} can be computed using the constraint that the sum of the conditional probabilities *P(tw|dw)* over all possible title words *tw* equals to 1, i.e.

$$\sum_{tw} P(tw \mid dw) = 1$$

To generate a set of title words for a given document *D*, *P(tw|D)*, i.e. the score for every title word *tw*, is computed as:

$$P(tw \mid D) = \sum_{dw \in D} P(tw \mid dw) C(dw, D)$$

The title words with highest score will be selected as the chosen title word for the document *D*.

4. Experimental Results and Discussion

In this section, we will present our experimental results and their analysis. The experimental results are presented in Section 4.1. Section 4.2 gives discussion over the results.

4.1 Experimental Results

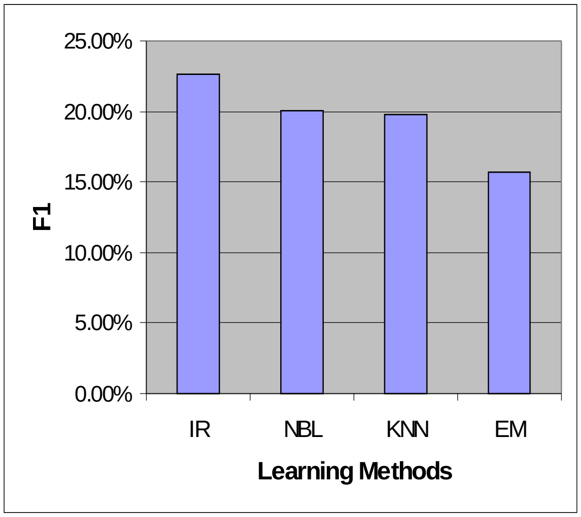

Fig. 2: Comparison of four Title Word Selection Approaches, namely "information retrieval" model (IR), Naïve Bayesian approach with limited vocabulary (NBL), K nearest neighbor approach (KNN) and iterative Expectation-Maximization approach, on a held-out test corpus of 10000 documents using the F1 score.

The F1 results for our "Information Retrieval" model (IR) and three learning methods for comparison, namely Naïve Bayesian approach with limited vocabulary (NBL), K nearest neighbor approach (KNN) and iterative Expectation-Maximization approach (EM), are shown in Figure 2. In terms of the F1 metric, our "Information Retrieval" model performs best, with F1 = 0.226. Naïve Bayesian approach with limited vocabulary and K nearest neighbor approach rank as second group, with F1 as 0.201 and 0.198 respectively. Iterative Expectation-Maximization approach performs worst, with F1 as 0.157.

We also conducted significance tests comparing our method with the other three methods over the F1 scores. We use an F-test (Myers, 1972), the F-test values for our method compared with Naïve Bayesian approach (NBL), K nearest neighbor approach (KNN) and iterative Expectation-Maximization approach (EM) are 66.7, 72.96 and 78.28 respectively. All of them exceeds the required F-test values for p = 0.001, i.e. 10.83. In other words, the statistical test shows the probability that our method is better is greater than 99.9%. Thus, we can conclude that our method is significantly better than all three other learning approaches.

4.2 Analysis and Discussion

We think the outstanding performance of our "Information Retrieval" model for title word selection can be attributed to two factors:

- **Optimized representation vectors for title words**. As discussed in the section 2.2, the representation vectors for title words, learned from the training corpus, minimized the difference between human assigned titles and machine-generated titles over all the training examples. Even though the set of title word representation vectors didn't actually optimize the true error function in Equation (1), they do minimize the error function in Equation (3). Thus, they correctly reflect the connection between title words and document words, which is crucial to title word selection. Furthermore, as we have mentioned before in Section 3.1, the training corpus has strong overlap in content with the test collection. This fact suggests that the title word representation vectors optimized for the training corpus will also be optimum for the testing collection. Without this condition, title words can have very different meaning between the testing collection and the training collection. In that case, the title word representation vectors learned from the training corpus will not be appropriate title word representation vectors for the test collection and this method will fail to find good title words for documents in the testing set.

- **Taking advantage of Information Retrieval**. Our approach toward selecting good title words for documents is to map the title word selection problem to an Information Retrieval problem. After building the representation vectors for title words, the Information Retrieval system will take care of the rest. Good term weighting schemes, such as TF.IDF (Salton & Buckley, 1988; Sparck-Jones &

Willett, 1997) and their variants, are carefully crafted to take into count the factors of the word frequency within the document (i.e. TF), the word frequency within the collection (i.e DF) and the document length (i.e. normalization factor). They have shown significant advantages in TREC conferences (Roberson & Walker, 1999).

Our algorithm benefits from the term weighting schemes of Information Retrieval in two ways: first, the TF.IDF term weights usually reflect the importance of a term related to a document. In our algorithm, the title words are represented as vectors of document words. By using the TF.IDF term weights to weight the document words in the representation vector, we promote the connection between the important document content words and the title words, and de-emphasize the connection of the trivial document words with the title words. Secondly, in Information Retrieval, the normalization factor in term weighting scheme avoid the takeover of the long documents. In our algorithm, this normalization factor helps us overcome some noise introduced by the common title words. According to our algorithm, most of the numbers in the representation vectors for the common title words will be nonzero because the common title words co-occur with most of the document words. Thus, the common title words are similar to the "long documents" in Information Retrieval. Without the normalization factor, these common title words will always be chosen because their representation vectors usually have very large overlap with the test document vector. With the help of the normalization factor, the numbers in the representation vectors for the common title words will be scaled down very much and the chance of the common title words get selected will decrease dramatically.

Therefore, even though the representation vectors for those common title words have large overlap with the test document in vocabulary, they still may get low score.

5. Conclusion and Future Work

In this paper, we present a novel approach toward title word selection, i.e. viewing title word selection as a variant of Information Retrieval. To find out a good representation vector for title words, we proposed an optimization approach, i.e. minimizing the difference between human assigned titles and machine-generated titles over the training examples. To avoid the computational complexity, we simplified the error function and found analytic solutions. We tested our approach over 10,000 documents and compared with three learning approaches. Experimental results show that our approach out-performs the three other learning approaches. We believe that the success of our approach is due to two reasons, i.e. finding optimized representation vectors for title words and taking advantage of Information Retrieval system, particularly the term weighting scheme of Information Retrieval system.

From the viewpoint of text categorization, our algorithm for the title word selection can be actually viewed as a special method of text categorization: each title word is treated as an individual class category. Then, computing the representation vectors for each title word in our algorithm corresponds to finding the representation for each class category in the text categorization. The procedure of selecting title words for documents corresponds to the task of assigning the correct class labels to the documents. Therefore, all the text categorization techniques can be applied to the problem of title word selection. In the future, we need to investigate the effectiveness of the other text categorization techniques, such as Support Vector Machine (SVM) and Decision Tree (DT), in selecting good title words for documents.

Acknowledgements

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.

References

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

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin and Craig G. Nevill-Manning (1999). *Domain-Specific Keyphrase Extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden. Morgan Kaufmann Publishers, San Francisco, CA, pp.668-673.

Goldstein, J., Kantrowitz, M., Mittal, V. & Carbonell, J. (1999). Summarizing Text Documents: Sentence Selection and Evaluation Metrics, *Proceedings of SIGIR 99*, Berkeley, CA.

Hildebrand, F. B. (1952). *Methods of Applied Mathematics*, ch. 2, Prentice-Hall, Englewood Cliffs, NJ.

Jin, R. & Hauptmann, A. G. (2000a). Title Generation for Spoken Broadcast News using a Training Corpus, *Proceedings of the 6th International Conference on Speech and Language Processing (ICSLP 2000)*, Beijing China.

Jin, R. & Hauptmann, A. G. (2000b). Title Generation Using Training Corpus, *Proceedings of CICLING-2001*, Mexico City, Mexico, 2001, in press.

Kennedy, P. & Hauptmann, A. G. (2000). Automatic Title Generation for the Informedia Multimedia Digital Library, *ACM Digital Libraries, DL-2000*, San Antonio Texas, May 2000.

Mayers, J. L. (1972). *Fundmentals of Experimental Design*. Ch. 4. Allyn and Bacon.

Mitra, M., Singhal, A. & Buckley, C. (1997). Automatic text summarization by paragraph extraction, *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1993). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Primary Source Media (1997). Broadcast News CDROM, Woodbridge, CT.

Rjiesbergen, V. (1979). *Information Retrieval*. Ch. 7. Butterworths, London.

Roberson, S. E. & Walker, S. (1999). Okapi/Keenbow at TREC-8. In E.M. Voorhees and D.K. Harmann, editor, *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, Gaithersburg.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Proceeding*, Prentice Hall, Englewood Cliffs, New Jersey.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, 24, 513—523.

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

Sparck Jones, K. & Willett P. (1997). Reading in Information Retrieval. Ch. 6. Morgan Kaufmann Publishers.

Strzalkowski, T., Wang, J. & Wise, B. (1998), A robust practical text summarization system, *AAAI Intelligent Text Summarization Workshop*, pages 26-30, Stanford, CA, March 1998.

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

Yang, Y. & Chute, C. G. (1992). A Linear Least Squares Fit method for terminology mapping. Proceedings of Fifteen International Conference on Computational Linguistics (COLING'92).

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