A New Probabilistic Model for Title Generation

Rong Jin
Language Technology Institute
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA15213, U. S. A.
rong+@cs.cmu.edu

Alexander G. Hauptmann
Department of Computer Science
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA15213, U. S. A.
alex+@cs.cmu.edu

Abstract

Title generation is a complex task involving both natural language understanding and natural language synthesis. In this paper, we propose a new probabilistic model for title generation. Different from the previous statistical models for title generation, which treat title generation as a generation process that converts the 'document representation' of information directly into a 'title representation' of the same information, this model introduces a hidden state called 'information source' and divides title generation into two steps, namely the step of distilling the 'information source' from the observation of a document and the step of generating a title from the estimated 'information source'. In our experiment, the new probabilistic model outperforms the previous model for title generation in terms of both automatic evaluations and human judgments.

Introduction

Compared with a document, a title provides a compact representation of the information and therefore helps people quickly capture the main idea of a document without spending time on the details. Automatic title generation is a complex task, which not only requires finding the title words that reflects the document content but also demands ordering the selected title words into human readable sequence. Therefore, it involves in both nature language understanding and nature language synthesis, which distinguishes title generation from other seemingly similar tasks such as key phrase extraction or automatic text summarization where the main concern of tasks is identify important information units from documents (Mani & Maybury., 1999).

The statistical approach toward title generation has been proposed and studied in the recent publications (Witbrock & Mittal, 1999; Kennedy & Hauptmann, 2000; Jin & Hauptmann, 2001). The basic idea is to first learn the correlation between the words in titles (title words) and the words in the corresponding documents (document words) from a given training corpus consisting of document-title pairs, and then apply the learned title-word-document-word correlations to generate titles for unseen documents.

Witbrock and Mittal (1999) proposed a statistical framework for title generation where the task of title generation is decomposed into two phases, namely the title word selection phase and the title word ordering phase. In the phase of title word selection, each title word is scored based on its indication of the document content. During the title word ordering phase, the 'appropriateness' of the word order in a title is scored using ngram statistical language model. The sequence of title words with highest score in both title word selection phase and title word ordering phase is chosen as the title for the document. The follow-ups within this framework mainly focus on applying different approaches to the title word selection phase (Jin & Hauptmann, 2001; Kennedy & Hauptmann, 2000).

However, there are two problems with this framework for title generation. They are:

- **A problem with the title word ordering phase**. The goal of title word selection phase is to find the appropriate title words for document and the goal of title word ordering phase is to find the appropriate word order for the selected title words. In the framework proposed by Witbrock and Mittal (1999), the title word ordering phase is accomplished by using ngram language model (Clarkson & Rosenfeld, 1997) to predict the probability P(T), i.e. how frequently the word sequence T is used as a title for a document. Of course, the probability for the word sequence T to be used as a title for any document is definitely influenced by the correctness of the word order in T. However, the factor whether the words in the sequence T are common words or not will also have great influence on the chance of seeing the sequence T as a title. Word sequence T with many rare words, even with a perfect word order, will be difficult to match with the content of most documents and has small chance to be used as a title. As the result, using probability P(T) for the purpose of ordering title words can cause the generated titles to include unrelated common title words. The obvious solution to this problem is to somehow eliminate the bias of favouring common title words from probability P(T) and leave it only with the task of the word ordering.
- **A problem with the title word selection phase.** The title word selection phase is responsible for coming up with a set of title words that reflect the meaning of the document. In the framework proposed by Witbrock and Mittal (1999), every document word has an equal vote for title words. However, title only needs to reflect the main content of a document not every single detail of that document. Therefore, letting all the words in the document participate equally in the selection of title words can cause a large variance in choosing title words. For example, common words usually have little to do with the content of documents. Therefore, allowing common words of a document equally compete with the content words in the same document in choosing title words can seriously degrade the quality of generated titles.

The solution we proposed to this problem is to introduce a hidden state called 'information source'. This 'information source' will sample the important content word out of a document and a title will be computed based on the sampled 'information source' instead of the original document. By striping off the common words through the 'information source' state, we are able to reduce the noise introduced by common words to the documents in selecting title words. The schematic diagram for the idea is shown in Figure 1, together with the schematic diagram for the framework by Witbrock and Mittal. As indicated by Figure 1, the old framework for title generation has only a single 'channel' connecting the document words to the title words while the new model contains two 'channels' with one connecting the document words to the 'information source' state and the other connecting the 'information source' state to the title words.

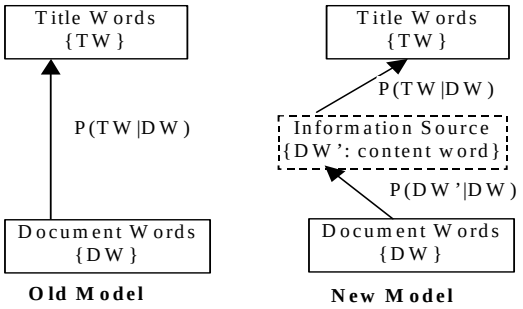

Fig. 1: Graphic representation for previous title generation model and new model for title generation.

1 Probabilistic Title Generation Model

In the language of probabilistic theory, the goal of creating a title T for a document D can be formalized as the search of the word sequence T that can be best generated by the document D, or

$$T = \arg\max_{T'} P(T'|D) \quad (1)$$

Therefore, the key of a probabilistic model for title generation is how to estimate the probability P(T|D). i.e. the probability of having a word sequence T as the title for the document D.

In this section, we will first describe the old framework using probability theory and associate the two problems of the old framework with the flaw in estimation of the probability P(T|D). Then a solution to each of the two problems will be presented and the new model

based on the old framework for title generation with the adaptation of the solutions will be described at the end of this section.

1.1 Formal Description of Old Framework for Title Generation

In terms of probability theory, the old framework can be interpreted as approximating the probability P(T|D) as a product of two terms with term P({tw∈T}|D) responsible for the title word selection and term P(T) responsible for the title word ordering and the probability P(T|D) can be written as:

$$P(T \mid D) \propto P(\{tw \in T\} \mid D)P(T) \qquad (2)$$

where {tw∈T} stands for the set of words in the title T. Since P({tw∈T}|D) stands for the probability of using the set of words tw in word sequence T as title words given the observation of the document D, it corresponds to the title word selection phase. P(T) stands for the probability of using word sequence T as a title for any document. Since word sequences with wrong word orders are rarely seen as titles for any document, the word order in word sequence T is an important factor in determining the frequency of seeing word sequence T as a title for documents and therefore it can be associated with the title word ordering phase.

1.2 Problem with the title word ordering phase

In the old framework for title generation, term P(T) is used for ordering title words into a correct sequence. However, term P(T) is not only influenced by the word order in T, but also whether words in T are common words. A word sequence T with a set of rare words will have small chance to be used as a title for any document even if the word order in T is perfectly correct. On the other side, a title T with a set of common words can have a good chance to be a title for some documents even its word order is problematic. Therefore, the probability for a word sequence T to be used as a title, i.e. P(T), is determined by both the 'appropriateness' of the word order of T and the 'rareness' of the words in T and doesn't appropriately represent the process of title word ordering whose only goal is to identify a correct word order with the given words.

In terms of formal analysis, the problem with the title word selection phase can be attributed to the oversimplified approximation for probability P(T|D). According to the chain rule in probability theory, the approximation for P(T|D) in Equation (2) is quite problematic and a more reasonable expansion for probability P(T|D) should be following:

$$\begin{aligned} P(T \mid D) &\approx P(\{tw \in T\} \mid D)P(T \mid \{tw \in T'\}) \\ &= P(\{tw \in T\} \mid D)P(T) / P(\{tw \in T\}) \end{aligned} \qquad (3)$$

where P({tw∈T}) stands for the probability of using the set of word {tw∈T} in titles without considering the word order. The difference between Equations (3) and (2) is that, Equation (2) uses term P(T) directly for title word ordering phase while Equation (3) divides term P(T) by term P({tw∈T}) and uses the result of division for title word ordering process. Because term P({tw∈T}) concerns only with the popularity of the words tw in sequence T, dividing P(T) by P({tw∈T}) has the effect of removing the bias of favouring popular title words from term P(T). Therefore, term P(T)/P({tw∈T}) is determined mainly by the word order in T and not influenced by the popularity of title words in T.

1.3 Problem with title word selection phase

As already discussed in the introduction section, the old framework for title generation allows all the words in the document equally participate in selecting title words and therefore, the final choice of title words may be influenced significantly by the common words in the document which have nothing to do with the content of the document. Thus, we suggest a solution to this problem by introducing a hidden state called 'information source' which is able to sampled the important content words from the original document. To find an optimal title for a document, we will create the title from the 'distilled information source' instead of the original document.

To allow titles being generated from the 'distilled information source' instead of the original document, we can expand the probability P(T|D) as the sum of the probabilities P(T| 'information source' S) over

all the possible 'information sources' S, where probability P(T|S) stands for the probability of using the word sequence T as the title for the 'information source' S. Formally, this idea can be expressed as:

$$P(T \mid D) = \sum_{S} P(T \mid S) P(S \mid D) \quad (4)$$

where symbol S stands for a possible 'information source' S for the document D. In Equation (4), term P(T|S)P(S|D) represents the idea of two noisy channels, with term P(S|D) corresponding to the first channel that samples 'information source' S out of the original document D and term P(T|S) corresponding to the second noisy channel that creates title T from the 'distilled information source' S. Since the first noisy channel, i.e. P(S|D), is new to the old framework for title generation, we will focus on the discussion of the noisy channel P(S|D).

Since the motivation of introducing the hidden state 'information source' S is to strip off the common words and have important content words kept, we want the noisy channel P(S|D) to be a sampling process where important content words have higher chances to be selected than common words. Let function g(dw,D) stands for the importance of the word dw related to the document D. Then, the word sampling distribution should be proportional to the word importance function g(dw,D). Therefore, we can write the probability P(S|D)

$$P(S \mid D) \propto \prod_{dw \in S} g(dw, D) \quad (5)$$

As indicated by Equation (5), the probability for 'information source' S to represent the content of the document D, i.e. P(S|D), is proportional to the product of the importance function values for all the words selected by 'information source' S.

1.4 A New Model for Title Generation

The new model is based on the old framework with the proposed solutions to the problems of the old framework. As the summary of discussions in the previous two subsections, the essential idea of this new model is in two aspects:

- **Creating titles from the distilled 'information source'**. To prevent the common words in the document from voting for title words, in the new model, titles will be created from the estimated 'information source' which has common document words stripped off.
- **Subtract the influence of the 'commonness' of title words from P(T).** In the old framework for title generation, term P(T) is associated with the title word ordering phase. Since both the word order and the word 'commonness' can influence the occurring probability of the sequence T, i.e. P(T), we need to subtract the factor of word 'commonness' from term P(T), which results in term P(T)/P({tw∈T}) for the title word ordering phase.

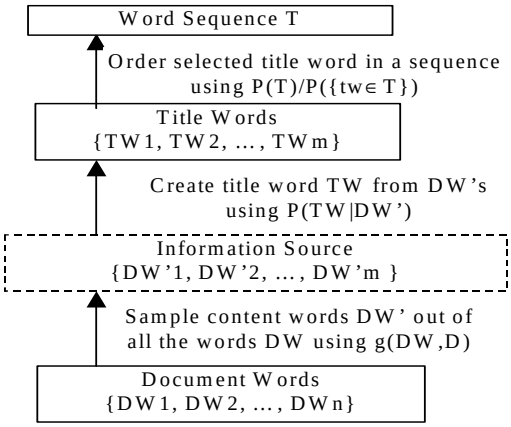

Fig. 2: Representation of the title generation scheme used by the new model. n is the number of words in the document and m is the number of words in the title.

Therefore, by putting Equations (2), (4) and (5) together, our new model for title generation can be expressed as

$$P(T \mid D) \propto \frac{P(T)}{P(\{tw \in T\})} \sum_{S} P(\{tw \in T\} \mid S) \prod_{dw} g(dw, D) \quad (6)$$

By further assuming that the number of words in any 'information source' S is equal to the number of words in the title T and, words in title T are created from the 'information source' S by first aligning every title word with a different word in the 'information source' S and then generating every title word tw from its aligned document word dw according to the probability distribution P(tw|dw), Equation (5) can be simplified as

$$P(T \mid D) \propto \frac{P(T)}{P(\{tw \in T\})} \prod_{tw \in T} \sum_{dw \in D} P(tw \mid dw) g(dw, D) \quad (7)$$

Equation (7) is the center of the new probabilistic model for title generation. There are three components in Equation (7). They are

word importance function g(dw,D), title-word-document-word translation probability P(tw|dw) and the word ordering component P(T)/P({tw∈T}). A schematic diagram in Figure 2 shows how a title is generated from a document in the new model through the three components. As shown in Figure 2, a sampling process based on the word importance function g(dw,D) will be applied to the original document to generate the 'information source' set containing most content words. Then, a set of title words will be scored according to probability P(tw|dw') based on the words dw' selected by the 'information source'. Finally, the word ordering process is applied to the chosen title words tw using P(T)/P({tw∈T}).

1.5 Estimation of Components

To implement the new model for title generation, we need to know how to estimate each of the three components.

- *The word importance function g(dw,D).* In information retrieval, normalized tf.idf value has been used as the measurement of the importance of a term to a document (Salton & Buckley, 1988). Therefore, we can adapt normalized tf.idf value as the word importance function g(dw,D). Therefore, function g(dw,D) can be written as

$$g(dw,D) = tf(dw,D)idf(dw) / \sum_{dw} tf(dw,D)idf(dw) \qquad (8)$$

- *The title-word-document-word 'translation' probability P(tw|dw).* The title-word-document-word 'translation' probability can be estimated using statistical translation model. Similar to the work of Kennedy and Hauptmann (2000), we can treat a document and its title as a 'translation' pair with the document as in 'verbose' language and the title as in 'concise' language. Therefore, title-word-document-word 'translation' probability P(tw|dw) can be learned from the training corpus using statistical translation model (Brown et al., 1990).
- *Word ordering component P(T)/P({tw∈T}).* There are two terms in this component, namely P(T) and P({tw∈T}). As already used by the old framework for title generation, P(T) can be estimated using a ngram statistical language model (Clarkson & Rosenfeld, 1997). The term P({tw∈T}), by assuming the independence between words tw, can be written as the product of the occurring probability of each tw in T, i.e. $P(\{tw \in T\}) \approx \prod_{tw \in T} P(tw)$.

With the expressions for g(dw,D) and P({tw∈T}) substituted into Equation (6), we have the final expression for our model, i.e

$$P(T \mid D) \propto \left(\frac{P(T)}{\left(\sum_{dw \in D} tf(dw,D)idf(dw) \right)^{|T|} \prod_{tw \in T} P(tw)} \right) \times \left(\prod_{tw \in T} \sum_{dw \in D} P(tw \mid dw) tf(dw,D) idf(dw) \right) \qquad (9)$$

2 Evaluation

In this experiment, we introduce two different of evaluations, i.e. a F1 metric for automatic evaluation and human judgments to evaluate the quality of machine-generated titles.

F1 metric is a common evaluation metric that has been widely used in information retrieval and automatic text summarization. Witbrock and Mittal (1999) used the F1 measurement (Rjiesbergen, 1979) as their performance metric. For an automatically generated title T_{auto}, F1 is measured against the correspondent human assigned title T_{human} as follows:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (10)$$

Here, precision and recall is measured as the number of identical words shared by title T_{auto} and T_{human} over the number of words in title T_{auto} and the number of words in title T_{human} respectively.

Unfortunately, this metric ignores syntax and human readability. In this paper, we also asked people to judge the quality of machine-generated titles. There are five different quality categories, namely 'very good', 'good', 'ok', 'bad', 'extremely bad'. A simple score scheme is developed with score 5 for the category 'very good', score 4 for 'good', score 3 for 'ok', score 2 for 'bad' and score 1 for 'extremely bad'. The average score of human judgment is used as another evaluation metric.

3 Experiment

3.1 Experiment Design

The experimental dataset comes from a CD of 1997 broadcast news transcriptions published by Primary Source Media [PrimarySourceMedia, 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. Only 1000 documents randomly selected from the remaining 10,000 documents are used as test collection because of computation expensiveness of applying language model to sequentialize the title words.

To see the effectiveness of our new model for title generation, we implemented the framework proposed by Witbrock and Mittal (1999) and conducted a contrastive experiment. The length of generated titles was fixed to be 6 for both methods and all the stop words in the title are removed.

3.2 Examples of Machine-Generated Titles

Table 1 and 2 give 5 examples of the titles generated by the old framework and the new probabilistic model, respectively. The true titles are also listed in Table 1 and 2 for the purpose of comparison.

As shown in Table 1, one common problem with this set of machine-generated titles is that common title words are highly favoured. For example, the phrase "president clinton" is a common title phrase and appears in 3 out of 5 titles and frequently is not necessary. As already discussed in previous sections, the problem of over-favouring common title words in the old framework can be attributed to the use of term P(T) for the title word ordering phase. The other problem with the set of generated titles in Table 1 is that, sometimes machine-generated titles contain words that have nothing to do with the content of the document. For example, the third machine-generated title in Table 1 is "president clinton budget tax tobacco settlement" while the original corresponding title is "senate funds fight against underage smoking". By the comparison of the two titles, we can see that the word "budget" has little to do with the content of the story and shouldn't be selected as title words. We think this problem is due to the fact that in the old framework for title generation, all the words in the document have an equal chance to vote for their favourite title words and the votes of common words in the document can cause unrelated title words to be selected.

Table 1: Examples of titles generated by the old framework. Stopwords are removed

Original Titles	Machine-generated Titles
bill lann lee	president clinton affirmative action supreme court
researchers say stress can cause heart disease	stress heart disease medical news day
senate funds fight against underage smoking	president clinton budget tax tobacco settlement
reaction to john f. kennedy jr. speaking out about his family	joe kennedy family reaction entertainment news
clinton's fast track quest and other stories	vice president clinton gore campaign fundraising

As shown in Table 2, the titles generated by the new model appear to be more relevant to the content of the document by comparison to the original titles. Furthermore, the titles in Table 2 appear to 'smoother' than the titles listed in Table 1 and don't have unnecessary common words in titles. We believe it is due to the effects of both modified process for the title word ordering and dual noisy channel model. By replacing term $P(T)/P(\{tw \in T\})$ with term $P(T)$, we make the title word selection phase concentrate on finding the correct word order and therefore avoid the problem of overly favouring common title words. With the introduction of the hidden state 'information source', the title words will be selected based on the sampled important content words and therefore the noise introduced by common words in the document is reduced dramatically.

Table 2: Examples of titles generated by new probabilistic model. Stopwords are removed

Original Titles	Machine-generated Titles
bill lann lee	civil rights nominee bill lann lee
researchers say stress can cause heart disease	study links everyday stress heart disease
senate funds fight against underage smoking	companies settlement tobacco deal tax laws
reaction to john f. kennedy jr. speaking out about his family	george magazine discusses joe kennedy family
clinton's fast track quest and other stories	senate vote fast track trade authority

3.3 Results and Discussions

The F1 score of each method is computed based on the comparison of the 1000 generated titles to their original titles using Equation (10). To collect human judgments for machine-generated titles, we randomly chose 100 documents out of the 1000 test documents and sent the machine-generated titles by both methods to the assessor for the quality judgment. The F1 scores and the average scores of human judgments for the old framework and the new probabilistic model are listed in Table 3.

Table 3: Evaluation results of the old framework and the new probabilistic model

	F1	Human Judg.
Old model	0.21	2.09
New model	0.26	3.07

As seen from Table 1, the F1 score for the new probabilistic model is better than the score for the old model with 0.26 for the new model and 0.21 for the old model. Since the F1 metric basically measures the word overlapping between machine-generated titles and the original titles, the fact that the new model is better than the old model in terms of F1 metric indicates that the new model does a better job than the old model in terms of finding title words appropriate for documents. More important, in terms of human judgments, the new model also outperforms the old model significantly, which implies that titles generated by the new model is more readable than the titles generated by the old model. Based on these two observations, we can conclude that the new probabilistic model for title generation is effective in generating human readable titles.

Conclusion

In this paper, we propose a new probabilistic model for title generation. The advantages of the new model over the old framework are on the modification of the title word ordering phase and the introduction of the hidden state 'information source'. In the contrastive experiment, the new model outperforms the old model significantly in terms of both the automatic evaluation metric and the human judgments of the qualities of the generated titles. Therefore, we conclude that our new probabilistic model is effective in creating human readable titles.

Acknowledgements

The authors are grateful to the anonymous reviewers for their comments, which have helped improve the quality of the paper. 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

I. Mani and M. T. Maybury (1999) *Advances in Automatic Text*. MIT press, pp 51—53.

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

R. Jin and A. G. Hauptmann (2001) *Learn to Select Good Title Word: A New Approach based on Reverse Information Retrieval*, ICML 2001.

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

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

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

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

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