Automatic Title Generation for Chinese Spoken Documents Considering the Special Structure of the Language

Lin-shan Lee and Shun-Chuan Chen

Speech Lab, College of EECS, National Taiwan University, Taipei, Taiwan, ROC
lslee@gate.sinica.edu.tw, hypno@speech.ee.ntu.edu.tw

Abstract

The purpose of automatic title generation is to understand a document and to summarize it with only several but readable words or phrases. It is important for browsing and retrieving spoken documents, which may be automatically transcribed, but it will be much more helpful if given the titles indicating the content subjects of the documents. On the other hand, the Chinese language is not only spoken by the largest population of the world, but with very special structure different from western languages. It is not alphabetic, with large number of distinct characters each pronounced as a monosyllable, while the total number of syllables is limited. In this paper, considering the special structure of the Chinese language, a set of "feature units" for Chinese spoken language processing is defined and the effects of the choice of these "feature units" on automatic title generation are analyzed with a new adaptive K nearest-neighbor approach, proposed in a companion paper also submitted to this conference as the baseline.

Keyword: title generation, Chinese spoken documents

1. Introduction

Automatic generation of titles for spoken documents is believed to be very important for future network era. When many multi-media information become available over the Internet and most of them include voice information, automatically generated titles (in texts) for segments of such materials will be very helpful in browsing and retrieval. This is just a simple example application scenario. A title can be regarded as an abstract or a brief summary. Unlike the traditional task of summarization, title generation needs to "learn" from the documents more and try to express the concepts carried by the documents in only several words or phrases. Sometimes a title may even include words that didn't appear in the document. Title generation is therefore usually considered as a difficult and challenging problem.

The Informedia project [1] at CMU started some research on title generation several years ago for English language. They tried to find out the relationship between the documents and the corresponding human-generated titles for the training corpus, and extend such relationship to evaluation documents. Such work has been done on text documents as well as broadcast news. But there is not much work reported on the title generation for Chinese spoken documents yet. In a companion paper also submitted to this conference [2], a new approach for title generation for Chinese spoken documents is developed based on a new adaptive K nearest-neighbor concept.

2. The Baseline Approach — The New Adaptive K Nearest-Neighbor Approach for Title Generation

The general framework for title generation includes two parts: the learning part and the generating part. The basic idea is to try to find out in the learning part the relationship between the training documents and the corresponding human-generated titles, and in the generating part extend such relationship to evaluation documents. Assume $D=\{d_j, j=1,2,...,N\}$ is the set of N training documents where d_j is the j-th document, and $T=\{t_j, j=1,2,...,N\}$ is the set of the corresponding human-generated titles, all in text form. On the other hand, $\bar{D}=\{\bar{d}_i, i=1,2,..., \bar{N}\}$ and $\bar{T}=\{\bar{t}_i, i=1,2,..., \bar{N}\}$are respectively the sets of evaluation or new documents transcribed from speech form into text form and the desired corresponding automatically generated titles.

The new adaptive K nearest-neighbor approach to title generation, proposed in the companion paper and taken as the baseline for this paper, is very briefly summarized here. It tries to integrate the nice properties of quite several approaches previously proposed [2].

In the K nearest-neighbor approach (KNN) [3], for each new document $\bar{d}_i \in \bar{D}$, instead of creating a new title $\bar{t}_i$ for it, we try to find an appropriate title in the training corpus $t_j \in T$ for a training document $d_j \in D$ which is nearest to $\bar{d}_i$. In the learning part, the documents in the training corpus, $d_j \in D$, are indexed by some document weighting parameter evaluated from $tf(w_k, d_j)$, the term frequency of the word w_k in the document d_j. In the generating part, the similarity measure $S(\bar{d}_i, d_j)$ between the input evaluation document $\bar{d}_i$ and each training document d_j in the training corpus is calculated based on these document weighting parameters as well as $tf(w_k, \bar{d}_i)$. The title $\bar{t}_i$ automatically generated for $\bar{d}_i$ is the title t_j corresponding to the document d_j with maximum $S(\bar{d}_i, d_j)$.

In the Naive Bayesian approach with limited vocabulary (NBL) [4], we try to find out the correlation between the words in the document d_j and its title t_j. Limited vocabulary implies that we only consider those words co-occurring in the document d_j and its title t_j. In the learning part, for all document-word/title-word pairs w_k co-occurring in all document/title pairs (d_j, t_j) in the training corpus, where $w_k \in d_j$, $d_j \in D$, and $w_k \in t_j$, $t_j \in T$, we compute the conditional probability $P(w_k, T \mid w_k, D)$. In the generating part, the words used in the title $\bar{t}_i$ of a new document $\bar{d}_i$ are selected based on a generating potential $G(w_k, \bar{d}_i)$ obtained from $tf(w_k, \bar{d}_i)$ and $P(w_k, T \mid w_k, D)$, for all words $w_k \in d_j$. We then choose the top L words $w_k \in \bar{d}_i$ ranked with $G(w_k, \bar{d}_i)$ to be used in the title $\bar{t}_i$, where L is the average length of all titles $t_j \in T$.

In the extractive summarization approach using TF/IDF (TF/IDF) [5], we simply select the sentence in the document with the highest total TF/IDF score as the title. Therefore in the learning part, we simply compute the IDF values *idf(*w_k*)* for all words w_k in all documents $d_j \in D$. In the generating part, we pick up the sentence that has the highest total TF/IDF score, which is the sum of the TF/IDF values of all words w_k in the sentence.

In the above approaches, KNN has the nice property that it uses the original structure of human-generated titles in the training corpus, but with the fatal problem that it needs to have some training documents highly correlated to the new evaluation documents. It cannot perform well at all for a document telling a completely new story. The new Adaptive KNN approach (AKNN) developed in the companion paper has a block diagram of the generating part as shown in Figure 1 [2]. We first obtain the top K training documents nearest to the input evaluation document. The titles of these K nearest training documents are then rescored using the NBL approach, and the best training document is chosen. But the title t_j^* for this selected document d_j^* is still for a training document, not necessarily good enough for the new evaluation document. We therefore assume that the words w_k with the highest TF/IDF values in the TF/IDF approach in the new evaluation document are the key named entities. So the words that appear in the selected training title t_j^* yet do not appear in the new evaluation document are replaced by those words in the new evaluation documents with the highest TF/IDF values. The title for the new evaluation document is thus produced.

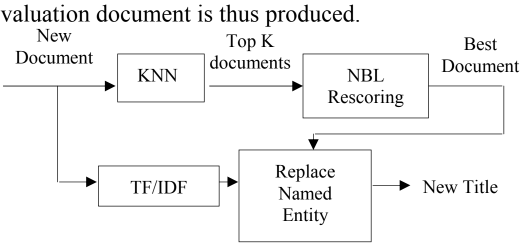

Figure 1. The generating part of the AKNN approach.

3. Considerations for the Special Structure of Chinese Language

The Chinese language is not alphabetic. Because each of the large number of characters is pronounced as a monosyllable, and is a morpheme with its own meaning, new words are very easily generated everyday by combining a few characters or syllables. For example, the combination of the characters "電(electricity)" and "腦(brain)" gives a new word "電腦(computer)", and the combination of the characters "股(stock)", "市(market)", "長(long)", and "紅(red)" gives a new word "股市長紅(stock price remains high for long)" in business news. In many cases the meaning of these words more or less has to do with the meaning of the component characters. Examples of such new words also include many proper nouns such as personal names and organization names which are simply arbitrary combinations of a few characters, as well as many domain specific terms just as the examples mentioned above. Many of such words are in fact very often the right key in the title generation for spoken documents, because they usually carry the core information, or characterize the subject topic. But in many cases these important words for our purposes are simply not included in any lexicon. It is therefore believed that the out-of-vocabulary (OOV) problem is especially important for title generation for Chinese spoken documents, and this is a very important reason why syllables make great sense in the problem here. In other words, the syllables represent characters with meaning, and in analyzing the spoken documents they do not have to be decoded into words which may not exist in the lexicon.

Actually, there are more considerations for the monosyllabic structure and character/syllable mapping relationships of the Chinese language. Although there exist more than 10,000 commonly used Chinese characters, a nice feature of the language is that all Chinese characters are monosyllabic and the total number of phonologically allowed Mandarin syllables is very limited. So a syllable is usually shared by many homonym characters with completely different meanings. Each Chinese word is then composed of from one to several characters (or syllables), thus the combination of the syllables actually gives an almost unlimited number of Chinese words. In other words, each syllable may stand for many different characters with different meanings, while the combination of several specific syllables very often gives only very few, if not unique, homonym polysyllabic words. As a result, when analyzing the Chinese spoken documents, segments of several syllables may provide a very good indication about the subject topic.

In fact, there exist also important reasons to use characters. Because almost every Chinese character is a morpheme with its own meaning, thus very often plays quite independent linguistic roles. As a result, the construction of Chinese words from characters is very often quite flexible. One example phenomenon is that in many cases different words describing the same or similar concepts can be constructed by slightly different characters, e.g., both "中華文化(Chinese culture)" and "中國文化(Chinese culture)" means the same, but the second characters used in these two words are different. Another example phenomenon is that a longer word can be arbitrarily abbreviated into shorter words, e.g., "國家科學委員會 (National Science Council)" can be abbreviated into "國科會", which includes only the first, the third and the last characters. Furthermore, an exotic word in foreign languages can very often be translated into different Chinese words based on its pronunciation, e.g., "Kosovo" may be translated into "科索沃", "柯索佛", "克索夫", "科索伏", "科索佛" and so on, but these words usually have at least one or two characters in common. Therefore, an intelligent title generation system needs to be able to handle such wording flexibilities, such that when the new spoken documents include some important words, the correct title can be generated even if in the training corpora the similar concepts are described by words in some other different forms. The analysis of spoken documents by characters or segments of characters does provide such flexibilities to some extent, because different forms of words describing the same or similar concepts very often do have some characters in common.

Based on the above considerations, the roles of the words w_k mentioned previously in the adaptive K nearest-neighbor approach may be replaced by quite several different choices of "feature units". The first will be a set of character-level units, including overlapping character segments with length N (C(N), N=1,2,3) and character pairs separated by n characters (P_c(N), N=1,2). Considering a character sequence of ten characters $c_1c_2c_3 \ldots c_{10}$, examples of the former are listed on the upper half of Table I, while examples of the latter on the lower half of Table I. For example, overlapping character segments of length 3 (C(N), N=3) include such segments as ($c_1c_2c_3$) , ($c_2c_3c_4$), ($c_3c_4c_5$), etc., while character pairs separated by 1 character (P_c(n), n=1) include such pairs (c_1c_3), (c_2c_4), (c_3c_5), etc. Similarly, a set of syllable-level units can be defined in exactly the same way, i.e., overlapping syllable segments with length N (S(N), N=1,2,3) and syllable pairs separated by n syllables (P_s(n), n=1,2). They can be obtained in exactly the same way as in Table 1, except the character sequence is replaced by a syllable sequence $s_1s_2s_3…s_{10}$. These are referred to as "feature units" here in this paper. Initial experimental results when the role of the words w_k is replaced by those "feature units" are presented below.

Table I

VARIOUS CHARACTER-LEVEL FEATURE UNITS FOR AN EXAMPLE CHARACTER SEQUENCE $c_1c_2c_3…c_{10}$

Character Segments	Examples
C(N), N=1	$(c_1)(c_2)...(c_{10})$
C(N), N=2	$(c_1,c_2)(c_2,c_3)...(c_9,c_{10})$
C(N), N=3	$(c_1,c_2,c_3)(c_2,c_3,c_4)...(c_8,c_9,c_{10})$
Character Pairs Separated by n Syllables	Examples
P_c(n), n=1	$(c_1,c_3)(c_2,c_4)...(c_8,c_{10})$
P_c(n), n=2	$(c_1,c_4)(c_2,c_5)...(c_7,c_{10})$

4. Initial Experimental Results

The training corpus used in the tests includes 151,537 pieces of Chinese news in text form with human-generated titles offered by the Central News Agency (CNA) at Taipei. 210 spoken news stories broadcast by FM News98 radio at Taipei were used as the evaluation documents. The reference titles for these evaluation spoken documents were produced by the students of the Graduate Institute of Journalism of National Taiwan University. These reference titles were used in the performance measures presented below. The objective performance measure used is the F1 scores,

$$F1 = \frac{2\times precision\times recall}{(precision + recall)} \qquad (1)$$

where precision and recall are calculated from the number of identical Chinese characters in computer-generated and human-generated titles. The speech recognition accuracy for words, characters and syllables for the evaluation spoken documents are 63.58%, 76.83% and 79.22% respectively.

The initial experimental results for the various "feature units" replacing the role of words w_k in the AKNN approach are shown in Figure 2. The first bar on the far left is the baseline using words in the AKNN approach, the others are those using the feature units discussed here, where S'(1), S'(2) and S'(3) are for the syllable segments S(1), S(2), S(3) but disregarding the tone information. Many interesting observations can be made here. First, the words carry much clearer meanings than the characters, but with significantly lower recognition accuracy due to the OOV problem. The characters have much higher recognition accuracy and can somehow avoid the OOV problem to a good extent. As a result, C(1) performs almost as well as the words. In addition, about 91% of the top 5000 most frequently used Chinese words are bi-character. It is thus natural that C(2) is able to catch most of the words in the documents and gives significantly better results. The number of tri-character words is much less, so C(3) actually are quite noisy. This is why C(3) gives lower F1 score. For the syllables, on the other hand, S(1) is much worse than C(1) even with a higher syllable accuracy, because a monosyllable is usually shared by many homonym characters with different meanings which causes serious ambiguity. But S(2) gives almost exactly the same performance as C(2), because with the many bi-character words the concatenation of two syllables substantially reduces this ambiguity. For similar reasons as C(3), S(3) is not good. On the other hand, the tones certainly differentiate the characters and words better, but the tone behavior is very complicated in fluent continuous speech, and the relatively higher error rates for tone recognition may also degrade the performance. The next three cases, S'(1), S'(2) and S'(3), are for those in which the tone information in S(1), S(2), S(3) was disregarded. S'(1) is lower than S(1), S'(2) almost the same as S(2), while S'(3) becomes significantly higher than S(3). These results are reasonable. The last two cases, P_c(1) and P_s(1), also make sense. Considering the fact that longer words may be arbitrarily abbreviated into shorter words, e.g., "國家科學委員會 (National Science Council)" may be abbreviated into "國科會" by including the first, the third and the last characters, character pairs separated by 1 character or syllable pairs separated by 1 syllable, P_c(1) and P_s(1), may be useful here. Also, substitution, deletion and insertion errors are inevitable in speech recognition process, and P_c(1) and P_s(1) may be able to handle such errors in some cases.

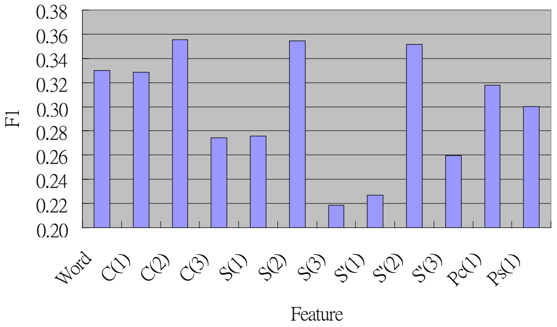

Figure 2. Initial results for the various "feature units"

5. Integration of the Functions for Different Feature Units

The results in Figure 2 indicated that there exist quite several useful "feature units" whose performance are very good, some even significantly better than the baseline unit of words. For example, C(2), S(2) and S'(2) are all much better than the words, and so on. Note that many of these "feature units" in fact carry different levels of information as mentioned above. For example, a syllable stands for many different homonym characters with different meanings, but the recognition accuracy is higher because it doesn't suffer from the OOV problem. Also, the character has an accuracy between the word and the syllable, but it can differentiate the meanings for the same syllable appearing in different context. It is therefore reasonable to consider the possibility of integration of the functions for different "feature units", such that extra information can be extracted when extra feature units can be used. One such example is shown in Figure 3 for S’(2)+S’(3), because it is possible that the noisy information brought by S’(3) can be clarified by S’(2), as shown in Figure 3. In this figure, the F1 scores for such an integration are plotted as a function of the relative weight, α, between S’(2) and S’(3). So α= 0 or 1 represents respectively the cases of using S’(3) alone or S’(2) alone. From Figure 3 it is easy to see α= 0.8 offered the best result, better than either S’(2) alone or S’(3) alone. This is reasonable because S’(2) performs much better than S’(3), therefore should play the major role in the integration.

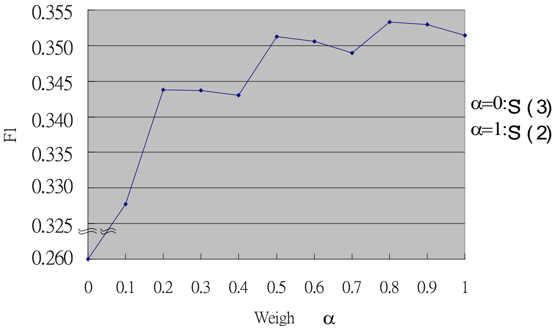

Figure 3. The choice of the relative weight α when S’(2) and S’(3) are integrated.

Several further examples are plotted in Figure 4 for such integration. In this figure, the first bar at the far left is for the baseline using words only, which is exactly the same baseline in Figure 2. In each integration case in the middle of the figure the relative weight, α, has been optimized empirically as in the above. It can be found that all the integration cases can provide significant improvements as compared to the baseline, and the difference among these different combinations of “feature units” seems to be relatively less significant. It is interesting to note that all these cases offering good F1 scores include the integration of the overlapping segments of two characters (or syllables) and three characters (or syllables). As mentioned previously, the majority of commonly used Chinese words are bi-character (or bi-syllabic), therefore C(2), S(2) or S’(2) do bring most important information. C(3), S(3) or S’(3) can bring some extra information, but also mixed with noisy information. That’s why C(3), S(3) or S’(3) alone gives only relatively low performance. But when C(3), S(3) or S’(3) is integrated with one of C(2), S(2) or S’(2), the noisy information may be clarified and therefore the overall performance improved. The last bar on the far right, on the other hand, is the one based on words but with all key phrases properly extracted as new words. As explained previously, the primary problem with words is the high percentage of out-of-vocabulary (OOV) words, and thus the low recognition accuracy, and many of the OOV words are very often the most important words for title generation. In this case, special efforts were made and special algorithms were used [6] to try to extract automatically most of the key phrases in the training documents as well as the corresponding human-generated titles, and then add all these newly extracted key phrases as the new words into the lexicon. As can be seen from Figure 4, including the key phrases do give very significant improvements as compared to the baseline at the far left of the figure, since it is along the right direction of solving the problem. However, when compared with the six integration cases in the middle of Figure 4, it is clear that by proper choosing the right “feature units”, much better performance can be obtained without extracting the key phrases, or the difficult problem of OOV words can be easily bypassed with the right choice of “feature units”, and better results can be obtained.

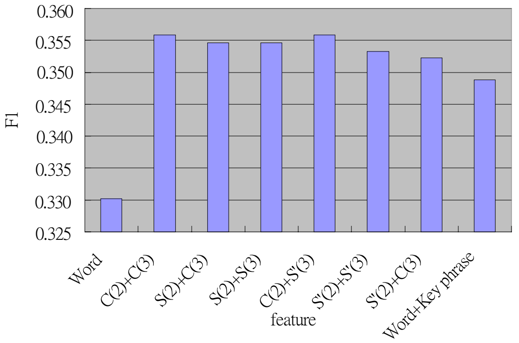

Figure 4. Integration of the functions for different sets of “feature units”.

6. Conclusion

In this paper, a set of “feature units” is proposed and the performance of automatic title generation for Chinese spoken documents when different “feature units” are chosen are analyzed and discussed considering the special structure of the Chinese language. The discussions are based on a new adaptive K nearest-neighbor approach proposed in a companion paper also submitted to this conference.

7. Acknowledgement

The help offered by the professors and students of the Graduate Institute of Journalism of National Taiwan University in producing the human-generated titles for the 210 spoken news stories, to be used in the performance evaluation presented in this paper, is highly appreciated.

8. Reference

[1] Rong Jin and Alexander G. Hauptmann, "Title Generation for Spoken Broadcast News using a Training Corpus", ICSLP 2000, Beijing, China, October 16-20, 2000.

[2] Shun-Chuan Chen and Lin-shan Lee, “Automatic Title Generation for Chinese Spoken Documents using an Adaptive K-Nearest Neighbor Approach”, submitted to Eurospeech2003.

[3] Yang, Y., Chute, “An example-based mapping method for text classification and retrieval,” C.G. ACM Transaction on Information Systems (TOIS), 12(3):252-77.1994.

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

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

[6] George Saon and Mukund Padmanabhan, "Data-Driven Approach to Designing Compound Words for Continuous Speech Recognition", IEEE TRANSACTIONS ON SPEECH AN D AUDIO PROCESSING, VOL. 9, No. 4, MAY 2001.