NSC92-2213-E-002-035-

92 08 01 93 07 31

93 5 31

E-mail: lslee@gate.sinica.edu.tw

Abstract

The Chinese language is not only spoken by the largest population in the world, but quite different from many western languages with a very special structure. It is not alphabetic: large number of Chinese characters are ideographic symbols and pronounced as monosyllables. The open vocabulary nature, the flexible wording structure and the tone behavior are also good examples within the special structure. It is believed that better results and performance will be obtainable in developing Chinese spoken language processing technologies, if this special structure can be taken into account. In this paper, a set of "feature units" for Chinese spoken language processing is identified, and the retrieval, segmentation and summarization of Chinese spoken documents are taken as examples in analyzing the use of such "feature units". Experimental results indicate that by careful considerations of the special structure and proper choice of the "feature units", significantly better performance can be achieved.

1. Introduction

The Chinese language is spoken by the largest population in the world. With the fast development of networks and wireless technologies, the access of network information by Chinese people may rely on Chinese spoken language processing technologies in the near future. When the majority of information activities have evolved from personal-computer -based to network-based, people may simply access information services over networks via hand-held devices such as handsets and PDA's with spoken interfaces. As the size of such hand-held devices shrinks, the conventional interfaces for personal computers such as keyboards and mice won't be useful any longer. Human voice is apparently one of the few most convenient interactive interfaces across all different hand-held devices [1]. When this is realized some day in the future, the population of 1.2 billion Chinese people (may even more at that time) may spend a vast amount of money purchasing computing and networking facilities with spoken language processing technologies. The demand is there, the market will someday be huge, and the potential impact on related areas is almost unlimited.

When huge quantities of multi-media information become available over the networks and many of them include voice information, retrieval of spoken documents may become the key for retrieving such multi-media information, because the voice usually carries the core information. Also, spoken documents include primarily audio signals which are usually not well divided into separate files or paragraphs as are text documents. This may cause various difficulties in further processing. Automatic segmentation of spoken documents into short passages, paragraphs or stories with focused subject topic is therefore important. On the other hand, contrary to text documents, it takes quite a long time for a person to "go through" a spoken document by listening to it from the beginning to the end. Automatic summarization of spoken documents to produce brief summaries will therefore be necessary when a user wishes to "browse" many spoken documents. As a result, the retrieval, segmentation and summarization of spoken documents will all be very important spoken language processing technologies.

The Chinese language is quite different from many western languages with a very special structure, as will be further discussed below. In this paper, a whole set of "feature units" for Chinese spoken language processing is identified, and the retrieval, segmentation and summarization of Chinese spoken documents are taken as examples in analyzing the use of such "feature units". Experimental results indicate that by careful considerations of the special structure and proper choice of the "feature units", significantly better performance can be achieved.

2. Special Structure of Chinese Language

The Chinese language is quite different from many western languages with a very special structure [2]. It is not alphabetic: large number of Chinese characters are ideographic symbols. Almost each Chinese character is a morpheme with its own meaning. A "word" is composed of one or several characters, with a meaning which is very often somehow related to the meanings of the component characters. A nice feature is that all the characters are pronounced as monosyllables, and the total number of phonologically allowed syllables is limited. Chinese is also a tone language, with a tone assigned to each syllable. There are 4 lexical tones plus a neutral tone for Mandarin. When the difference in tones is disregarded, the total number of syllables is further reduced. The small number of syllables also implies large number of homonym characters sharing the same syllable. As a result, each syllable represents many characters with different meanings, and the combination of these syllables (or characters) gives unlimited number of words and sentences. This is referred to in this paper as the "monosyllabic structure" of Chinese language.

The wording structure in Chinese is extremely flexible. For example, a long word can be arbitrarily abbreviated, such as " (Taiwan University)" being abbreviated as " ", i.e., including only the first and the third characters, and new words can be easily generated every day, such as the characters " (electricity)" and " (brain)" forming a new word " (computer)". These have to do with the fact that every character has its own meaning, and thus can play some linguistic role very independently. Furthermore, there are no "blanks" in written or printed Chinese sentences serving as word boundaries. As a result, the "word" in Chinese language is not very well defined, the segmentation of a sentence into a string of words is definitely not unique, and there never exists a commonly accepted lexicon. This is referred to in this paper as the "open vocabulary nature" and "feasible wording structure" of the Chinese language. For western alphabetic languages, since the words are well defined, speech processing is primarily

word-based, such as based on a lexicon of words and word based language models. For the Chinese language , since the words are not easy to identify, the out-of-vocabulary(OOV) problem is especially serious, and special measures are usually needed. In particular, many of the OOV words, or unknown words, are the names of people, organizations or events, or the special terms for the subject domain or task. They are very often the key to identify the content or meaning of the spoken documents. Their correct identity is usually necessary for spoken document retrieval, segmentation and summarization.

With this special structure, some extra efforts will also be needed for multi-tier annotation of spoken language corpora in Chinese, as discussed in another paper also presented in this session.

3. Spoken Document Retrieval

The problem of spoken document retrieval via speech queries has been investigated extensively and very important results have been obtained [3-5]. It was found that syllable level statistical characteristics are especially useful in such problems. A whole class of syllable-level indexing terms was defined, including 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). Considering a syllable sequence of 10 syllables s_1 s_2 s_3…. s_{10}, examples of the former are listed on the upper half of Table 1, while examples of the latter on the lower half of Table 1. For example, overlapping syllable segments of length 2 (S(N), N=2) include such segments as (s_1 s_2), (s_2 s_3), (s_3 s_4), etc., while syllable pairs separated by 1 syllable ($P_s(n)$, n=1) include such pairs as (s_1 s_3), (s_2 s_4), (s_3 s_5), etc. These indexing terms make good sense for various processing purposes including retrieval. For example, as mentioned previously, each syllable represents some characters with meaning, and thus very often different words with similar or relevant concepts have some syllables in common, even if some of such words are out-of-vocabulary. Therefore syllable segments with length 1 (S(N), N=1,non-overlapping monosyllables in this case) make sense in identifying the content or meaning of spoken documents.

However, because each syllable is also shared by many homonym characters each with a different meaning, syllable segments with length 1 (S(N), N=1) alone also cause serious ambiguity. In fact, about 91% of the top 5,000 most frequently used Chinese polysyllabic words are bi-syllabic, i.e., they are pronounced as a segment of two syllables. Therefore, the syllable segments with length 2 (S(N), N=2) definitely carry a plurality of linguistic information, and it makes great sense to use them as important indexing terms. Similarity, if longer

Syllable Segments	Examples
S(N), N=1	(s_1) (s_2) …(s_{10})
S(N), N=2	(s_1 s_2) (s_2 s_3)…(s_9 s_{10})
S(N), N=3	(s_1 s_2 s_3) (s_2 s_3 s_4)…(s_8 s_9 s_{10})
Syllable Pairs Separated by n Syllables	Examples
$P_s(n)$, n=1	(s_1 s_3) (s_2 s_4) …(s_8 s_{10})
$P_s(n)$, n=2	(s_1 s_4) (s_2 s_5) …(s_7 s_{10})

TABLE 1: Various syllable-level indexing terms for an example syllable sequence $s_1,s_2,s_3…s_{10}$

syllable segments such as S(N), N=3, are matched between a document and the query, very often very important information for retrieval may be captured in this way. On the other hand, because of the flexible wording structure in the Chinese language as described previously, syllable pairs separated by n syllables are also helpful. Considering the example that the word " (Taiwan University)"may be arbitrarily abbreviated as" " including only the first and the third characters, syllable pairs separated by n syllables become

apparently useful in such cases. Furthermore, because substitution, insertion and deletion errors always happen during the syllable recognition process, such indexing terms as syllable pairs separated by n syllables are also helpful in handling such syllable recognition errors.

4. A Whole Set of "Feature Units" for Chinese Spoken Language Processing

The syllable-level indexing terms including 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) as shown in Table 1 were shown to be useful for Chinese spoken document retrieval. The above concept can be extended to 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), as well as overlapping word segments with length N (W(N) , N=1, 2, 3) and word pairs separated by n words ($P_w(n)$, n=1, 2). Table I can be easily modified to include these cases, as long as the syllable sequence S_1 S_2 S_3…S_{10} is replaced by a character sequence C_1 C_2 C_3…C_{10} or a word sequence W_1 W_2 W_3…W_{10}, and everything else remains the same. Because every character has its own meaning, the character-level information carried by C(N) and $P_c(N)$ apparently makes sense. Because many homonym characters may share the same syllable, the characters can certainly identify the content or meaning of the spoken documents better than the syllables, if the characters are recognized correctly. If not, the syllables are more helpful. This is similar for words. The words bring clearer meaning than either characters or syllables, if they are recognized correctly. If not, the characters and syllables are more helpful. Due to the serious OOV problem, the word error rate for speech recognition is very often the highest, the character error rate lower, and the syllable error rate the lowest. Therefore, all the different unit levels are helpful. They are apparently useful not only in spoken document retrieval, but in various spoken document processing tasks having to do with the content or meaning of the spoken documents, for example, segmentation, summarization, title generation, key phrase extraction of spoken documents, etc. They are therefore referred to as the "feature units" for Chinese spoken language processing in this paper. Some preliminary investigation on the application of these "feature units" for spoken document segmentation and summarization are briefly presented below as examples.

5. Spoken Document Segmentation

Here we briefly summarize some preliminary work on automatic segmentation of Chinese spoken documents and analyze the use of the "feature units" presented above. In general the automatic speech recognition process transforms the spoken documents into word sequences without punctuations, and it is even difficult to identify where a sentence is. In this research, we simply assume all silence periods with duration exceeding a threshold is a sentence boundary, and try to identify a most probable topic cluster for each sentence. When two adjacent sentences belong to two different topic clusters, a topic segmentation boundary is identified.

The hidden Markov model (HMM) based segmentation approach [6, 7] was adopted and shown in Figure 1. A total of N topic clusters, T_1, T_2, … T_N, form an HMM, in which each topic cluster is a state. The sentences composed of recognized word sequences are taken as the observations. Each topic cluster (state) has equal transition probabilities P_1 for transition to a different topic cluster, and P_2 for remaining in the same topic cluster. N-gram probabilities, i.e., $P(w_k)$, $P(w_k| w_{k-1})$, $P(w_k| w_{k-2}, w_{k-1})$, are used to evaluate the score for each sentence in each topic cluster, where w_k is the k-th word in a sentence, and these N-gram probabilities are trained for each topic cluster with a training corpus. The topic clusters are obtained from the training

corpus by a K-means clustering algorithm. In this way the total number of topic clusters, N, can be empirically or arbitrarily set, while there is no need to label the stories and topics for the training corpus. Everything can be performed automatically. Viterbi algorithm can then be performed to segment the spoken documents into stories suitable for different topic clusters. This is the baseline segmentation scheme (BSL). Further improvements are in fact achievable by including more information; for example, a story length duration model can be developed using the story length histogram for the training corpus [7], the pause duration cue can be used by modeling the duration distributions for boundary and non-boundary pauses [8], and confidence measures can be used to weight more those words recognized more reliably. All these can be integrated into the baseline scheme (BSL) to form an improved scheme (IMP). Preliminary tests were performed with TDT 2001 evaluation data [9]. TDT-2 was used as the training corpus, including 3,320 min of audio signals, or 2,936 news stories, while TDT-3 as the testing corpus, including 7,620 min of audio signals, or 4,578 news stories. All of these were in Mandarin Chinese only. The segmentation cost defined by TDT evaluation was also used here which includes the cost for false alarm and missing [9]. The speech recognition was performed with the Dragon recognizer, with word, character and syllable error rates being 36,97%, 19.78% and 15.06% respectively for the testing corpus, and similar rates for the training corpus. Initial results are shown in Figure 1, where in each case the left bar is for the baseline scheme (BSL) and the right bar for the improved scheme (IMP). The first set of data on the left labeled by W(1) represents the conventional approach, in which the words w_k were used in evaluating the scores for each sentence in different topic clusters as mentioned above, or exactly the same as those used for alphabetic languages such as English. The next several sets of data are then the results when the various "feature units" as defined above were used to replace the role of the words, including overlapping segments of characters, C(1), C(2), C(3), overlapping segments of syllables, S(1), S(2), S(3), as well as character pairs separated by a character, $P_c(1)$. Various observations can be made. First, the character-level feature units including C(1), C(2), C(3) and $P_c(1)$ all performed better than the conventional unit of words popularly used for alphabetic languages such as English, for both the BSL and IMP cases. C(2) is especially better, actually gives the best performance among all units tested here, because most frequently used words are bi-character and thus C(2) carry most linguistic information. C(3) and $P_c(1)$ are also good, except that they also bring some noisy information. On the other hand, the syllable-level feature units are also good. S(2), S(3) both offer better performance than the conventional unit of words for both BSL and IMP cases for various reasons mentioned previously, but S(1) did worse for a very clear reason. A single syllable is shared by many homonym characters with many different meanings, thus brings ambiguity. It is believed that proper integration of more than one carefully chosen "feature units" as discussed here will give even better results. Such investigation is currently under progress and will be reported in the near future.
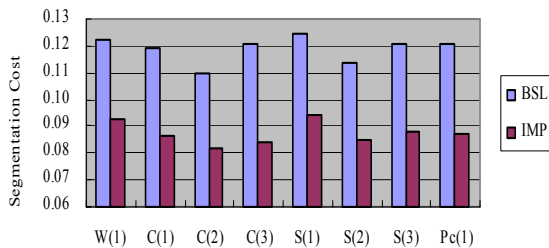
Figure 1 Preliminary results for segmentation cost when different "feature units" were used.

6. Spoken Document Summarization

Here we briefly present some preliminary work on automatic summarization of Chinese spoken documents [10]. In recent reports, the summarization of spoken documents may be achieved by two stages: important sentence extraction and sentence compaction [11]. In the preliminary work to be presented here, however, only the importance sentence extraction was performed, i.e., the most important sentences in the documents were automatically selected and concatenated to form a summary.

Two approaches were used to choose the most important sentences. The first approach uses the term frequency (TF) and inverse document frequency (IDF) as well as the vector space model (TF/IDF) popularly used in information retrieval [12]. In this approach, a feature vector $\vec{s}$ is defined for each sentence and also a feature vector $\vec{D}$ is defined for each document. Each component of these vectors is the TF/IDF score for a word w_k in the sentence and the document. The similarity score between a sentence and the whole document is then

$$S(\vec{s}, \vec{D}) = (\vec{s} \cdot \vec{D}) / |\vec{s}||\vec{D}|, \qquad (1)$$

and the sentences with the highest similarity scores are chosen to be concatenated to form the summary. The other approach used the significance score (SIG) of each word W_k in the sentence and in the document,

$$I(w_k) = f_k \log \frac{F_A}{F_K} \qquad (2)$$

where f_k and F_K are respectively the occurrence frequencies of the word w_k in the recognized sentence and in the training corpus [11,13].The significance score of a sentence is then the sum of the significance scores for all words in the sentence. The sentences with the highest scores are then chosen to be concatenated to form the summary.

In the preliminary experiments, the training corpus included roughly 150,000 news stories in text form from Jan to Dec 2000 provided by the Central News Agency of Taipei. The average length of each story was about 510 characters. They were used to calculate the IDF and F_A, F_K parameters mentioned above. The testing corpus included 200 news stories broadcast in Aug 2001 by a few radio stations at Taipei. The average length of each story was about 29 sec. The speech recognition accuracy for the testing corpus for words, characters and syllables are 66.46%, 74.95% and 81.70% respectively. Three human subjects (students of National Taiwan University) were requested to do the human summarization to be taken as the references in two forms: the first simply to rank the importance of the sentences in each transcribed news story from the top to the middle (since here we simply try to select the most important sentences as the summary), and the second to write a summary for the news story by himself with a length being roughly 25% of the original news story. Two cases of summarization ratios, 20% and 30%, were performed in the tests, which is the ratio of summary length to the total length. In each case the two human-produced summaries were used in the evaluation. The first, u_1, is the concatenation of the top several important sentences selected by the students, while the second, u_2, is simply the one he wrote by himself. The summarization accuracy for the j-th news story, A_j, is then the average similarity score [14] for the machine-produced summary, $\bar{u}$, with u_1 and u_2,

$$A_j = \frac{1}{2}[S(\bar{u}, u_1) + S(\bar{u}, u_2)] \qquad (3)$$

where the similarity score $S(\bar{u}, u_1)$, $S(\bar{u}, u_2)$ are calculated in exactly the same way as in equation (1) based on the feature vectors of the TF/ IDF scores. In this way, higher accuracy

will be obtained if more words that are important in the news story are included in the machine-produced summary, and both types of human-produced summaries, u_1 and u_2, are considered. The final summarization accuracy is then the average of A_j in equation (3) over all the 200 news stories and all the three human subjects. The experimental results for summarization accuracy of 20% and 30% are shown in Figure 2 (a) and (b) respectively. In each case the first set of data on the left for W(1) are for the two approaches TF/IDF and SIG mentioned above with words used as the basic units. The others are the results when other different "feature units" were used to replace the words, including some combinations. Quite interesting observations can be made here. Considering the case of 20% summarization ratio in Figure 3(a). First, in some cases TF/IDF is better and in some other cases SIG is better. TF/IDF gave the highest accuracy, probably because in calculating the summarization accuracy TF/IDF scores were used, which may not be very fair when comparing the two approaches. Second, as expected, the words conventionally used in alphabetic languages such as English gave relatively low accuracy, while some other "feature units" proposed here in this paper can offer significantly better results. In the case of TF/IDF, even if S(1) performed very poorly (since a single syllable is shared by many homonym characters with different meanings, thus causing ambiguity), S(2) offered very good accuracy and the combination of S(1)+S(2) did even better. As mentioned previously, S(2) carries plenty of linguistic information which helps to clarify the ambiguity caused by S(1). C(1)+S(2) is also good, though not as impressive as S(1)+S(2). The words may be helpful in some cases, if integrated with some better units. For example, W(1)+C(1)+S(2) is better than without words, i.e., C(1)+S(2), but W(1)+S(1)+S(2) becomes significantly worse than that without words, S(1)+S(2). The relatively low accuracy for words (primarily due to the OOV problem) may be a good reason for such phenomena. A similar situation can be found for the case of SIG, for which C(1)+S(2) gives the best results. For SIG C(1) alone is the best for all single "feature unit" cases, i.e. among W(1),C(1) and S(1), which is reasonable considering the relatively high character accuracy and the fact that it clarifies the ambiguity caused by single syllables. As a result, C(1)+S(2) becomes the best is natural. Similar trends can be observed for 30% summarization ratio in Figure 2(b).
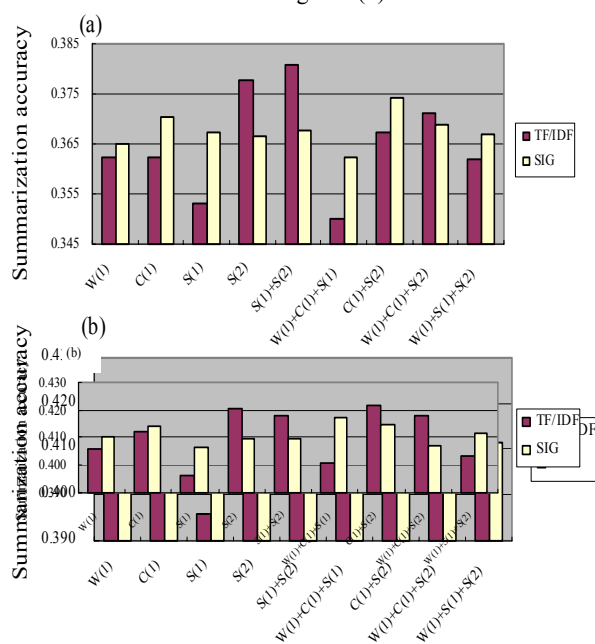
Figure 2 Preliminary results for the summarization ratio of (a) 20% and (b) 30% respectively when different "feature units" were used.

7. Conclusion

In this paper, a whole set of "feature units" for Chinese spoken language processing is identified, and the retrieval, segmentation and summarization of Chinese spoken documents are taken as examples in analyzing the use of such "feature units". The experimental results indicated that by careful considerations of the special structure and proper choice of the "feature units", significantly better performance can be achieved.

8. References

[1] Lin-shan Lee, Yumin Lee, "Voice Access of Global Information for Broadband Wireless: Technologies of Today and Challenges of Tomorrow", Proceedings of the IEEE, Jan. 2001,pp. 41-57.

[2] Lin-shan Lee, "Voice Dictation of Mandarin Chinese", IEEE Signal Processing Magazine, Vol.14, No.4, July 1997, pp.63-101.

[3] Berlin Chen, Hsin-Min Wang and Lin-shan Lee, "Discriminating Capabilities of Syllable-based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese", IEEE Transactions on Speech and Audio Processing, Vol.10, No.5, July 2002, pp.303-314.

[4] Berlin Chen, Hsin-min Wang, Lin-shan Lee, "An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval", 2001 European Conference on Speech Communication and Technology, Aalborg, Denmark, Sept 2001, CD-ROM.

[5] Chun-Jen Wang, Berlin Chen, Lin-shan Lee, "Improved Chinese Spoken Document Retrieval with Hybird Modeling and Data-driven Indexing Features", International Conference on Spoken Language Processing, Denver, Co, USA, Sept 2002, CD-ROM.

[6] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, P. van Mulbreget, "A Hidden Markov Model Approach to Text Segmentation and Event Tracking, " ICASSP, 1998.

[7] W. Grei, A. Morgan, R. Fish, M. Richards, A. Kundu, "Fine-grained hidden markov modeling for broadcast-news story segmentation," Human Language Technology Conference, 2001.

[8] G. Tür, D. Hakkani-Tür, A. Stolcke, E. Shriberg, "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation, "Computational Linguistics, vol. 27, March 2001.

[9] "The Topic Detection and Tracking 2001(TDT-2001) Evaluation Plan", http://www.nist.gov/speech/tests/tdt/tdt2001/evalplan.htm .

[10] Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu and Alex Waibel ,"Automatic Speech Summarization Applied To English Broadcast News Speech", ICASSP 2002.

[11] T. Kikuchi, S.Furui, C. Hori, "Two-stage Automatic Speech Summarization by Sentence Extraction and Compaction", IEEE and ISCA Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, April 2003, pp.207-210.

[12] Klaus Zechner, "Automatic Generation of Concise Summaries of Spoken Dialogues in Unrestricted Domains", SIGIR 2001.

[13] Chiori Hori and Sadaoki Furui, "Automatic Speech Summarization Based On Word Significance And Linguistic Likelihood", ICASSP 2000.

[14] Eduard Hovy and Daniel Marcu, "Automated Text Summarization Tutorial", COLING/ACL 1998.