ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

AN INITIAL PROTOTYPE SYSTEM FOR CHINESE SPOKEN DOCUMENT UNDERSTANDING AND ORGANIZATION FOR INDEXING/BROWSING AND RETRIEVAL APPLICATIONS

Lin-Shan Lee, Shun-Chuan Chen, Yuan Ho, Jia-Fu Chen, Ming-Han Li, Te-Hsuan Li
National Taiwan University, Taipei
E-Mail: lslee@gate.sinica.edu.tw

Abstract

In the future, the network content will include all knowledge, information, services relevant to our daily life. The most attractive form of future network content will be multi-media, which usually includes voice information. As long as the voice information is included, it usually carries the core concepts for the content. As a result, the spoken documents associated with the multi-media content very possibly can serve as the key for indexing/browsing and retrieval. However, unlike the written documents, the multi-media or voice information are very often just audio/video signals. They are very difficult to index, browse or retrieve, since the users can't go through each of them from the beginning to the end during browsing. A possible approach then may be to segment the audio/video signals automatically into short paragraphs, each with a central concept or topic, and then automatically generate a title and/or a summary for each of these short paragraphs, in either speech or text form. The topics and central concepts described in the segmented short paragraphs are then further analyzed and organized into some graphic structures describing the relationships among these topics and central concepts. In this way, the multi-media content can be much more efficiently indexed automatically and browsed and retrieved by the user based on the title, summary and the graphic structure. This is referred to as the understanding and organization of spoken documents here. In this paper, an initial prototype system for such functions with broadcast news taken as the example multi-media content was presented. The graphic structure used to describe the relationships among the topics and central concepts are 2-dimensional tree structures developed based on the probabilistic latent semantic analysis.

1. Introduction

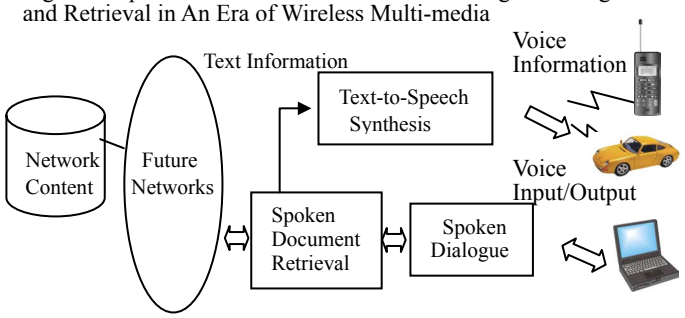
In the future network era, the digital content over the network will include all the information activities for human life, from real-time information to knowledge archive, from working environment to private services, etc. Apparently, most attractive form of the network content will be in multi-media, which usually includes speech information. As long as the speech information is included in the network content, it usually tells the subjects, topics and concepts of the multi-media content. As a result, speech information will become the key for indexing, browsing and retrieval. On the other hand, the fast development of wireless technologies will make it possible for people to access the network content at any time, from anywhere via simple hand held devices such as cell phone units or PDA's. Personal Computers (PC's), which used to be the core of information activities in the past, will get less and less important, and keyboards and mice, which used to be the most convenience user interface for PC's, won't be convenient any longer. When all these become true, it is believed that speech interface will become one of the few most important and convenient user interface across all user terminals for users to access the network content at any time, from anywhere.

Today, the network access is primarily text-based. The users enter the instructions by words or texts, and the network or search engine offers text materials for the user to select. The users interact with the network or search engine and obtain the desired information. In the future, it can be imagined that almost all such roles of text can be directly replaced by speech without any problem as shown in Figure 1. The users' instructions can be entered by speech. The network content may be index/browsed and retrieved by its speech information. Here spoken document retrieval with speech queries will become a key. The users interact with the network or the search engine via spoken dialogues. There is always some information expressed in text form. Text-to-speech synthesis can be used to transform the text information into speech. The user terminals can always include a small display window if needed, such that some of the information can be shown in text form to help the inadequacy of pure speech scenario. In such a speech-based network content indexing/browsing /retrieval environment, using speech instructions to access the network content whose key concepts are specified by speech information will be natural.

2. Spoken Document Understanding and Organization

When considering the above speech-based network content indexing/browsing/retrieval environment, we need to keep in mind that unlike the written documents which are better structured and easier to browse, multi-media/spoken documents are just video/audio signals, or a sequence of words if transcribed. For example, a 3-hour video of interview, a 2-hour movie, a 1-hour news stories, or a very long sequence of

Figure 1. Speech-based Network Content Indexing/Browsing and Retrieval in An Era of Wireless Multi-media



transcribed words. The user simply can't go through each one from the beginning to the end during browsing. As a result, better approaches for understanding and organization of multi-media/spoken documents for easier indexing/browsing/retrieval thus become necessary. Such spoken document understanding and organization should include at least the following:

(1) Spoken Document Segmentation
Automatically segmenting the spoken documents into short paragraphs, each with some central topic.

(2) Named Entity Extraction for Spoken Documents
Named entities are usually the keywords in the spoken documents, and therefore the key in understanding the subject topics of the documents. However, in many cases such named entities are in fact out-of-vocabulary (OOV) words creating difficulties in the recognition of the spoken documents.

(3) Information Extraction for Spoken Documents
Information extraction usually referrers to the extraction of the key information such as who, when, where, what and how for the events described in the documents. They are usually the relationships among the named entities extracted. Such information is definitely important for indexing/browsing and retrieval.

(4) Spoken Document Summarization
Automatically generating a summary (in text or speech form) for each segmented short paragraph of the spoken documents.

(5) Title Generation for Spoken Documents
Automatically generating a title (in text or speech form) for each segmented short paragraph of the spoken documents.

(6) Topic Analysis
Automatically analyzing the central concepts and topics of the segmented short paragraphs and organizing them into some graphic structures describing the relationship among these topics and central concepts.

When all the above can be properly performed, the spoken documents (or network content) are in fact better understood and re-organized in a way that indexing/ browsing/retrieval can be performed easily. For example, the spoken documents (or network content) are now in form of short paragraphs, properly organized in graphic structures with titles/summaries as indices for browsing. They can be retrieved either based on the full content, or based on the summaries/titles/concepts, or both.

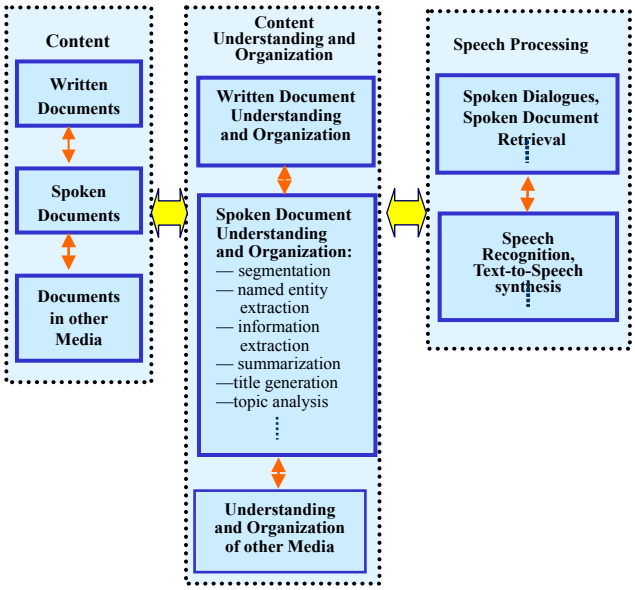
Fig.2 is a block diagram for the user/content scenario for the speech-based network content indexing/browsing and retrieval as mentioned here. The network content is on the left of the figure in which the multi-media content includes at least written parts (written documents), spoken parts (spoken documents) plus other parts (in other media such as video). They all need certain degree of understanding and organization as shown in the middle column of the figure, referred to as content understanding and organization here, in which the spoken document understanding and organization as discussed above are shown in the middle. The users are on the right of the figure, trying to use speech processing technologies including speech synthesis/recognition, spoken dialogues and spoken document retrieval to access the network content. Note that the concept of speech understanding is not new. But in the past it usually referred to understanding the speaker's intention in spoken dialogues within specific task domains, such as asking for weather, or air travel information or reservation, etc. But here the domains for network content can be arbitrary and almost unlimited. It is definitely not limited to very specific tasks.

3. Brief Summary for the Technologies Used in the Initial Prototype System

The initial prototype system is very briefly summarized here. For spoken document segmentation, the hidden Markov model (HMM) based segmentation approach [1, 2] was adopted. A total of N topic clusters 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 some transition probabilities for transition to a different topic cluster, or remaining in the same topic cluster. N-gram probabilities are used to evaluate the score for each sentence in each topic cluster [3]. The transition from one topic cluster to another is a segmentation point.

For named entity extraction from spoken documents, the spoken documents are first transcribed into word graphs, on which words or monosyllables with higher confidence measures are identified. Temporal/topical homogeneous reference text corpora are also automatically retrieved and selected to be used for named entity matching to find some named entities which can't be correctly transcribed in the word graphs. Both forward and backward Pat-Trees [4] are constructed to develop complete data structures for the context information for both the spoken documents and the selected temporal/topical homogeneous reference text corpora. The context information beyond sentences is

Figure 2. User/Content Scenario for Speech-based Network Content Indexing/Browsing and Retrieval in An Era of Wireless Multi-media



very often very helpful to identify named entities in both text and spoken documents. Multi-layered Viterbi search was performed based on a class-based language model [5], class generation models and a class context model, in order to handle the situation that a named entity may be composed of several named entities. In this way, the named entity extraction, word segmentation and spoken document transcription can be accomplished simultaneously.

For spoken document summarization, 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 [7]. The other approach used the significance score (SIG) of each word in the sentence and in the document, which is based on the occurrence frequencies of the word in the recognized sentence and in the training corpus [8, 9]. The sentences with the highest scores are then chosen to be concatenated to form the summary and played in audio form [3].

For automatic title generation, a corpus of 150,000 news stories (in text form) with human- generated titles was used in the training process. For a new spoken document, it was first transcribed into recognized word sequences. We then try to construct a title for the new spoken document based on the relationship between the news stories and their human-generated titles learned in the training process [10, 11].

Topic analysis for spoken documents, on the other hand, was performed based on probabilistic latent semantic analysis (PLSA). Latent semantic analysis (LSA) has been a very powerful approach to develop a latent semantic space for the semantic concepts hidden in a large corpus by reducing the word-document relationships into much lower dimensionalities with singular value decomposition [12]. The probabilistic latent semantic analysis (PLSA), on the other hand, tries to construct a statistical framework onto LSA by incorporating the probabilities for the words, documents and latent classes to build an "aspect model" [13]. In this way, the semantic concepts or the topic information regarding each segmented paragraph can be properly analyzed. Two-dimensional "topic maps" are then developed based on the idea very similar to the "self-organization maps" [14] previously developed, such that the relationships among the semantic concepts (or topics) can be displayed on an N*N map [15]. The distance between two blocks on the map has to do with the relationship between the semantic concepts (or topics) represented by the blocks. The shorter the distance, the closer the relationships. Each block of a semantic concept (or topic) can then be further analyzed and represented as another L*L map in the next layer, in which the blocks again represent the fine structures of the more detailed semantic concepts (or topics), etc.. In this way, the concept/topic relationships among the segmented spoken documents can be organized into a two-dimentional tree-structure, much easier for indexing, browsing and retrieval [16].

4. Brief Summary for the Functions of the Initial Prototype System

In the initial prototype system, the broadcast news (some including video parts) are taken as the example multi-media content. A typical example is shown in Figure 3. In this figure the waveform of a typical broadcast news story, taken as an example spoken document, is shown in Figure 3(a). The waveform of the automatically generated summary, which is the concatenation of a few selected sentences, is shown in Figure 3 (b). The automatically generated title, in text form, is printed in Figure 3 (c). Such process was actually performed on a large corpus of roughly 130 hrs of about 7,000 broadcast news stories. They were all recorded from radio/TV stations in Taipei from Nov 2002 to Oct 2003.

The initial broadcast news browsing prototype system is described below. The homepage of the system lists 20 categories of news (e.g. international political news, local business news, etc.). All the 7000 broadcast news stories mentioned above were actually automatically classified into these 20 categories. When the user click the item of "local political news", for example, a two-dimensional graphical structure of 4*4 topic maps appears as shown in Figure 4, in which the 16 blocks

Figure 3 A typical example: (a) the waveform of the spoken document (a broadcast news story) (b) the waveform of the automatically generated summary (a few selected sentences) (c) the automatically generated title (in text form)

(a)

(b)

(c)

聯合國祕書長安南強烈譴責聖多美軍事政變

Figure 4. The 16 Blocks for major semantic concepts or topics in the category of "local political news"

亞太地區 台灣 安全 外交 美國	兩岸關係 陸委會 大陸 統獨 美中	內閣 行政院 游錫堃 內政部 交通 建設	陳水扁 總統府 中央 祕書長 蘇貞昌
外交部長 西非 中南美 友邦 經貿	李登輝 國民黨 索羅門群島 瓜地馬拉	三通 包機 台商 春節 海基會	經濟部 景氣 股市 金融 央行 林信義
副總統 印尼 丹麥 日本 交流 全球	謝長廷 縣長 市政 馬英九 委員會	立法院 黨員 國民黨 執政黨 台聯 衝突	市場 價格 民生物資 油品 台電 董事長
連戰 宋楚瑜 合作 協商 反彈 基層	賄選 議長 民情 下鄉 立法委員	黑金 財團 承包商 工程款 金主 執政黨	台積電 高科技 竹科 產業 資訊 精密

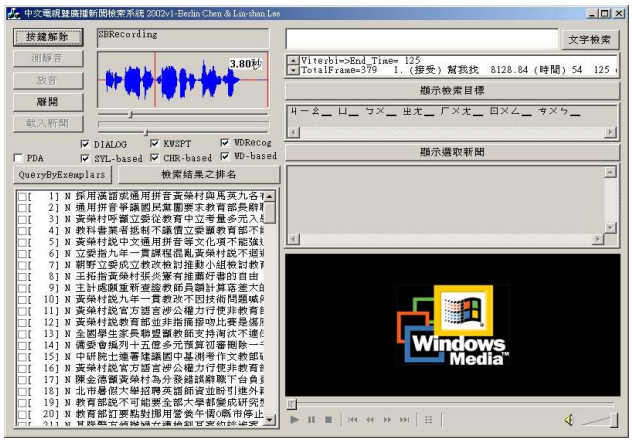
represent the major semantic concepts or topics in the area of "local political news" for the broadcast news corpora. Each block (or major semantic concept) is here characterized by the top several words with the highest probabilities. As can be found that the distance among the blocks has to do with the relationships among the major semantic concepts or topics. The user can click one of the blocks to see the next layer 2-dimensional graphical representation for the fine structure of the more detailed semantic concepts (or topics). So the broadcast news are organized in a two-dimensional tree-structure for better indexing and easier browsing. When the user decides to see all the broadcast news items within a node in this two-dimensional tree, whether a leaf-node or not, he can click a button for that node, and the automatically generated titles for all news stories categorized into that semantic concept or topic with high enough probabilities are shown in a list. The user can further click the "summary" button after each title to listen to the automatically generated summaries. The two-dimensional tree structure and titles/summaries are very helpful to browse the news stories.

The broadcast news retrieval function is shown in Figure 5. The retrieval was primarily based on the combination of syllable/character/word-level indexing features [17]. But now all retrieved news stories as listed in the left lower part in the figure have automatically generated titles and summaries. The user can select the news stories by the title, or by listening to the summaries, rather than listening to the whole story and then found it is not the one he is looking for. The user can also select to click another functional button to see how the semantic concepts or topics of these retrieved news stories are located within the two-dimensional tree structures, which will also be very helpful for the user to identify the desired new items.

5. Conclusion

In this paper an initial prototype system for Chinese spoken document understanding and organization for future speech-base network content indexing/browsing/ retrieval environment is presented.

Figure 5. The broadcast news retrieval system with automatic title/summary generation functions (with query: 請幫我找教育部長黃榮村 (please retrieve the news regarding the minister of education, prof. R.T. Huang))



6. References

[1] 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.

[2] 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.

[3] Lin-shan Lee, Yuan Ho, Jia-fu Chen, Shun-Chuan Chen, "Why Is the Special Structure of the Language Important for Chinese Spoken Language Processing- Examples on Spoken Document Retrieval, Segmentation and Summarization," ISCA Eurospeech 2003, Geneva, Switzerland

[4] Lee-Feng Chien, "PAT-Tree Based Keyword Extraction for Chinese Information Retrieval," ACM SIGIR '97.

[5] J. Sun, et al, "Chinese Named Entity Identification Using Class-based Language Model," COLING 2002, Taipei, Taiwan.

[6] Yu-In Liu, "An Initial Study on Named Entity Extraction from Chinese Text/Spoken Documents and Its Potential Applications," Master Thesis, National Taiwan University, July 2004.

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

[8] 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.

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

[10] Shun-Chuan Chen and Lin-shan Lee, "Automatic Title Generation for Chinese Spoken Documents Using an Adaptive K Nearest-Neighbor Approach," ISCA Eurospeech 2003, Geneva, Switzerland

[11] Lin-shan Lee and Shun-Chuan Chen, "Automatic Title Generation for Chinese Spoken Documents Considering the Special Structure of the Language," ISCA Eurospeech 2003, Geneva, Switzerland

[12] S. Deerwester et al, "Indexing by Latent Semantic Analysis," Proceeding of the American Society for Information Science, 1990.

[13] T.Hofmann, "Probabilistic Latent Semantic Indexing," ACM SIGIR 1999.

[14] T. Kohonen, "Self-Organizing Maps," Springer 1995.

[15] T. Hofmann, "ProbMap-A Probabilistic Approach for Mapping Large Document Collections," Intelligent Information System Journal, 2000.

[16] Shun-Chuan Chen, "Initial Studies on Chinese Spoken Document Analysiss-Topic Segmentation, Title Genevation and Topic Orgnization," Master Thesis, National Taiwan University, July 2004.

[17] 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.