Lin-shan Lee and Berlin Chen

The key to future efficient retrieval/browsing applications

Spoken Document Understanding and Organization

Speech is the primary and most convenient means of communication between individuals [1]. 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 archives, from working environments to private services. Apparently, the most attractive form of the network content will be in multimedia, including speech information. Such speech information usually provides insight concerning the subjects, topics, and concepts of the multimedia content. As a result, the spoken documents associated with the network content will become key for retrieval and browsing.

On the other hand, the rapid development of network and wireless technologies is making it possible for people to access the network content not only from the office/home, but from anywhere, at any time, via small handheld devices such as personal digital assistants (PDAs) or cell phones. Today, network access is primarily text based. The users enter the instructions by words or texts, and the network or search engine offers text materials from which the user can select. The users interact with the network or search engine and obtain the desired information via text-based media. In the future, it can be imagined that almost all such functions of text can also be performed with speech. The user's instructions can be entered not only by text but possibly through speech as well since speech is a convenient user interface for a variety of user terminals, especially for small handheld devices. The network content may be indexed/retrieved and browsed not only by text but possibly also by the associated spoken documents as mentioned above. The users may

1053-5888/05/$20.00©2005IEEE

also interact with the network or the search engine via either text-based media or spoken/multimodal dialogs. Text-to-speech synthesis can be used to transform the text information in the content into speech when required. This is the general environment of retrieval/browsing applications for multimedia content with associated spoken documents.

SPOKEN DOCUMENT UNDERSTANDING AND ORGANIZATION

When considering the above network content access environment, we must keep in mind that, unlike the written documents that are better structured with titles and paragraphs and thus easier to retrieve and browse, multimedia/spoken documents are merely video/audio signals, or a very long sequence of words including errors even if automatically transcribed. Examples include a three-hour video of a course lecture, a two-hour movie, or a one-hour news episode. In general, they are not segmented into paragraphs and no titles are mentioned for the paragraphs. Thus, they are much more difficult to retrieve and browse because the user simply cannot browse through each from the beginning to the end. As a result, better approaches are required for the understanding and organization of spoken documents (or the associated multimedia content) for easier retrieval/browsing. This should include at least the following:

1) *Named-entity extraction from spoken documents*. Named entities (NEs) are usually the keywords in the spoken documents (or associated multimedia content). They are the key to understanding the subject matters of the documents, although they're usually difficult to extract from the spoken documents.

2) *Spoken document segmentation*. Spoken documents (or the associated multimedia content) are automatically segmented into short paragraphs, each with some central concept or topic.

3) *Information extraction for spoken documents*. This step involves automatically extracting the key information (such as who, when, where, what, and how) for the events described in the segmented short paragraphs. Very often, the relationships among the NEs in the paragraphs are extracted.

4) *Spoken document summarization*. This step involves automatically generating a summary (in text or speech form) for each segmented short paragraph of the spoken documents (or associated multimedia content).

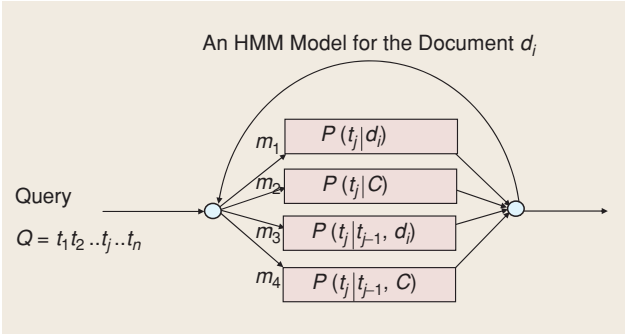
5) *Title generation for spoken documents*. This step involves automatically generating a title (in text or speech form) for each short paragraph of the spoken documents (or the associated multimedia content).

6) *Topic analysis and organization*. This step involves automatically analyzing the subject topics of the segmented short paragraphs of the spoken documents (or the associated multimedia content), clustering them into groups with topic labels, constructing the relationships among the groups, and organizing them into some hierarchical visual presentation that is easier for users to browse.

When all the above tasks can be properly performed, the spoken documents (or associated multimedia content) are in fact better understood and reorganized in a way that retrieval/browsing can be performed easily. For example, they are now in the form of short paragraphs, properly organized in some hierarchical visual presentation with titles/ summaries/topic labels as references for retrieval and browsing. The retrieval can be performed based on either the full content, the summaries/titles/topic labels, or both. In this article, this is referred to as spoken document understanding and organization for efficient retrieval/browsing applications.

Note that while some of the areas mentioned above have been studied or explored to a good extent, some of them have not at this point in time. Yet, most of the work has been performed independently within respective individual scopes and reported as separated analyses and results. Great efforts have been made to try to integrate several of these independent projects into a specific application domain, and in recent years several well-known research projects and systems have been successfully developed towards this goal. Examples include the Informedia System at Carnegie Mellon University [2], the Multimedia Document Retrieval Project at Cambridge University [3], the Rough'n'Ready System at BBN Technologies [4], the Speech Content-based Audio Navigator (SCAN) System at AT&T Labs-Research [5], the Broadcast News Navigator at MITRE Corporation [6], the SpeechBot Audio/Video Search System at Hewlett-Packard (HP) Labs [7], and the National Project of Spontaneous Speech Corpus and Processing Technology of Japan [8]. All of these successful projects and systems have evidenced the importance and potential of improved technologies towards the common goal of efficient retrieval/browsing for massive quantities of spoken documents and multimedia content.

The purpose of this article, however, is to present a concise, comprehensive, and integrated overview of these related areas, (including relevant problems and issues, general principles, and basic approaches) in a unified context of spoken document understanding and organization for efficient retrieval/browsing applications. In addition, we present an initial prototype system we developed at National Taiwan University as a new example of integrating the various technologies and functionalities. Note that similar efforts can be made based on information of other media in the multimedia content, for example based on the text, graphic, or video information. Here we focus only on those based on the associated spoken documents. As mentioned previously, speech information is usually the key compared to other information. Also, note that speech understanding has existed as a research topic for a long time; statistical approaches to speech understanding have been reviewed in the article by Wang et al. [9]. Previously, speech understanding usually referred to understanding the speaker's intention in such applications as spoken dialogues within specific task domains (for example, for air travel reservation or conference registration). Here, the domains for network content can be arbitrary and almost unlimited. Therefore, the technologies



[FIG1] An illustration of the HMM-based retrieval model.

needed will have to be domain independent and quite different from those used in spoken dialogues. Of course, for retrieval/browsing purposes, the required accuracy for understanding is also different from that required in applications such as spoken dialogs, where any error in understanding may directly lead to a wrong system response.

The majority of the approaches and technologies to be reviewed and discussed in this article were analyzed and verified using various broadcast news archives. This is apparently due to the availability of huge quantities of broadcast news archives and the fact that many of the statistical approaches mentioned here have to be based on huge quantities of corpora. Therefore, in this article, if not otherwise mentioned, broadcast news archives are taken as the default example spoken documents. Other types of spoken documents, for example technical presentations, will be mentioned when specific approaches for such tasks are discussed.

BRIEF REVIEW OF INFORMATION RETRIEVAL AND AUDIO INDEXING

Although this article is focused on spoken document understanding and organization, the purpose is to explore efficient retrieval and browsing. So we will start with a brief review of information retrieval (IR) and audio indexing. In the past two decades, most of the research efforts in IR were focused on text document retrieval, and the Text Retrieval Conference (TREC) evaluations [10] of the 1990s are good examples. In the conventional text document retrieval, a collection of documents $D = \{d_i, i = 1, 2, \ldots, N\}$ is to be retrieved by a user's query Q. The retrieval is based on a set of indexing terms specifying the semantics of the documents and the query, which are very often a set of keywords or all the words used in all the documents. The document retrieval problem can then be viewed as a clustering problem, i.e., selecting the documents out of the collection that are in the class relevant to the query Q. The documents are usually ranked by a retrieval model (or ranking algorithm) based on the relevance scores between each of the documents d_i and the query Q evaluated with the indexing terms (i.e., those documents on the top of the list are most likely to be relevant). The retrieval models are usually characterized by two different matching strategies, namely, literal term matching and concept matching. These two strategies are briefly reviewed below.

LITERAL TERM MATCHING

The vector space model (VSM) is the most popular example for the literal term matching [11]. In VSM, every document d_i is represented as a vector $\vec{d}_i$. Each component $w_{i,t}$ in this vector is a value associated with the statistics of a specific indexing term (or word) t, both within the document d_i and across all the documents in the collection D,

$$w_{i,t} = f_{i,t} \times \ln(N/N_t), \tag{1}$$

where $f_{i,t}$ is the normalized term frequency (TF) for the term (or word) t in d_i used to measure the intradocument weight for the term (or word) t, while $\ln(N/N_t)$ is the inverse document frequency (IDF), where N_t is the total number of documents in the collection that include the term (or word) t, and N is the total number of documents in the collection D. IDF is to measure the interdocument discriminativity for the term t, reflecting the fact that indexing terms appearing in more documents are less useful in identifying the relevant documents. The query Q is also represented by a vector $\vec{Q}$ constructed in exactly the same way, i.e., with components $w_{q,t}$ in exactly the same form as in (1). The cosine measure is then used to estimate the query-document relevance scores:

$$R\left(\vec{Q}, \vec{d}_t\right) = \left(\vec{Q} \cdot \vec{d}_t\right) \Big/ \left(\|\vec{Q}\| \cdot \|\vec{d}_i\|\right), \tag{2}$$

which apparently matches Q and d_i literally based on the terms. This model has been widely used because of its simplicity and satisfactory performance.

The literal term matching can also be performed with probabilities, and the N-gram-based [12] and hidden Markov model (HMM)-based [13] approaches are good examples. In these models, each document d_i is interpreted as a generative model composed of a mixture of N-gram probability distributions for observing a query Q, while the query Q is considered as observations, expressed as a sequence of indexing terms (or words) $Q = t_1 t_2 \ldots t_j \ldots t_n$, where t_j is the jth indexing term in Q and n is the length of the query (as illustrated in Figure 1). The N-gram distributions for the terms t_j, for example $P(t_j|d_t)$ and $P(t_j|\bar{t}_{j-1}, \bar{d}_i)$ for unigrams and bigrams, are estimated from the document d_i and then linearly interpolated with the background unigram and bigram models estimated from a large outside text corpus C, $P(t_j|C)$ and $P(t_j|t_{j-1}, C)$. The relevance score for a document d_i and the query $Q = t_1 t_2 \ldots t_J \ldots t_n$ can then be expressed as (with unigram and bigram models)

$$\begin{aligned} P(Q|d_i) = {} & [m_1 P(t_1|d_i) + m_2 P(t_1|C)] \\ & \times \prod_{j=2}^{n} \big[m_1 P(t_j|d_i) + m_2 P(t_j|C) \\ & + m_3 P(t_j|t_{j-1}, d_i) + m_4 P(t_j|t_{j-1}, C)\big], \end{aligned} \tag{3}$$

which again matches Q and d_i based literally on the terms. The unigram and bigram probabilities, as well as the weighting parameters, $m_1, \ldots, m_4$, can be further optimized. For example, they can be optimized by the expectation-maximization (EM) or minimum classification error (MCE) training algorithms, given a training set of query exemplars with the corresponding query-document relevance information [14].

CONCEPT MATCHING

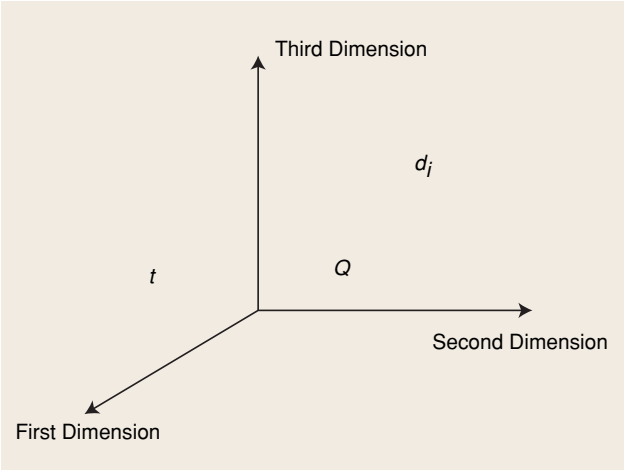
Both approaches mentioned above are based on matching the terms (or words) and thus very often suffer from the problem of word-usage diversity (or vocabulary mismatch) when the query and its relevant documents are using quite different sets of words. In contrast, the concept matching strategy tries to discover the latent topical information inherent in the query and documents, based on which the retrieval is performed; the latent semantic indexing (LSI) model is a good example [15]. LSI starts with a "term-document" matrix W, describing the intra- and interdocument statistical relationships between all the terms and all the documents in the collection D, in which each term t is characterized by a row vector and each document d_i in D is characterized by a column vector of W. Singular value decomposition (SVD) is then performed on the matrix W to project all the term and document vectors onto a single latent semantic space with significantly reduced dimensionality R:

$$W \approx \hat{W} = U \Sigma V^T, \tag{4}$$

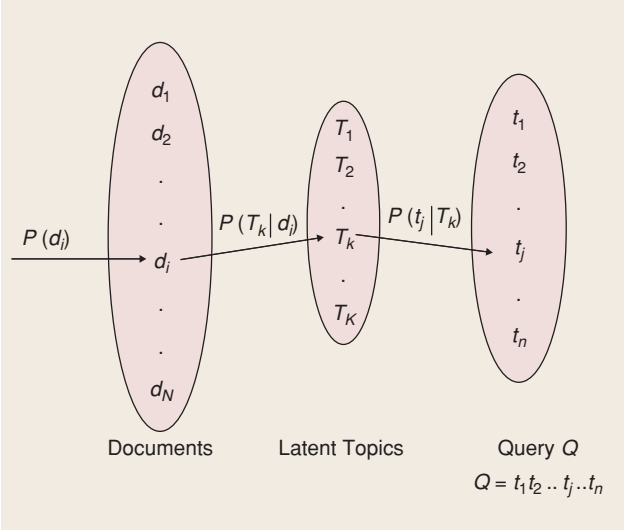
where $\hat{W}$ is the rank-R approximation to the term-document matrix W, U is the right singular matrix, Σ is the $R \times R$ diagonal matrix of the R singular values, V is the right singular matrix, and T denotes matrix transposition. In this way, the row/column vectors representing the terms/documents in the original matrix W can all be mapped to the vectors in the same latent semantic space with dimensionality R. As shown in Figure 2, in this latent semantic space, each dimension is defined by a singular vector and represents some kind of latent semantic concept. Each term t and each document d_i can now be properly represented in this space, with components in each dimension related to the weights of the term t and document d_i with respect to the dimension, or the associated latent semantic concept. The query Q or other documents that are not represented in the original analysis can be folded in, i.e., similarly represented in this space via some simple matrix operations. In this way, indexing terms describing related concepts will be near to each other in the latent semantic space even if they never cooccur in the same document, and the documents describing related concepts will be near to each other in the latent semantic space even if they never use the same set of words. Thus, this is concept matching rather than literal term matching. The relevance score between the query Q and a document d_i is estimated by computing the cosine measure between the corresponding vectors in this latent semantic space. A more detailed elucidation of LSI and its applications can be found in [16].

In recent years, new efforts have been made to establish the probabilistic framework for the above latent topical approaches, including improved model training algorithms. The probabilistic latent semantic analysis (PLSA) or aspect model [17] is often considered as a representative of this category. PLSA introduces a set of latent topic variables $\{T_k, k = 1, 2, \ldots, K\}$ to characterize the term-document cooccurrence relationships, as shown in Figure 3. A query Q is again treated as a sequence of observed terms, $Q = t_1 t_2 \ldots t_j \ldots t_n$, while the document d_i and a term t_j are both assumed to be independently conditioned on an associated latent topic T_k. The conditional probability of a document d_i generating a term t_j thus can be parameterized by

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|T_k) P(T_k|d_i). \tag{5}$$



[FIG2] Three-dimensional schematic representation of the latent semantic space and the LSI retrieval model.



[FIG3] Graphical representation of the PLSA-based retrieval model.

When the terms in the query Q are further assumed to be independent given the document, the relevance score between the query and document can be expressed as

$$P(Q|d_i) = \prod_{j=1}^{n} \left[\sum_{k=1}^{K} P(t_j|T_k) P(T_k|d_i) \right]. \quad (6)$$

Notice that this relevance score is not obtained directly from the frequency of the respective query term t_j occurring in d_i but instead through the frequency of t_j in the latent topic T_k as well as the likelihood that d_i generates the latent topic T_k. A query and a document thus may have a high relevance score even if they do not share any terms in common; this, therefore, is concept matching. The PLSA model can be optimized by the EM algorithm in either an unsupervised manner using each individual document in the collection as a query exemplar to train its own PLSA model or in a supervised manner using a training set of query exemplars with the corresponding query-document relevance information.

SPOKEN DOCUMENTS/QUERIES

All the retrieval models mentioned above can be equally applied to text or spoken documents with text or spoken queries. The primary difficulties for the spoken documents and/or queries are the inevitable speech-recognition errors, including the problems with spontaneous speech (such as pronunciation variation and disfluencies) and the out-of-vocabulary (OOV) problem for words outside the vocabulary of the speech recognizer. A principal approach to the former, in addition to the many approaches for improving the recognition accuracy, is to develop more robust indexing terms for audio signals. For example, multiple recognition hypotheses obtained from N-best lists, word graphs, or sausages can provide alternative representatives for the confusing portions of the spoken query or documents [18]. Improved scoring methods using different confidence measures (for example, posterior probabilities incorporating acoustic and language model likelihoods) or other measures considering relationships among the recognized word hypotheses [19]–[20], as well as prosodic features including pitch, energy stress, and duration measure [21], can also help to properly weight the term hypotheses. The use of subword units (e.g., phonemes for English and syllables for Chinese) or segments of them, rather than words as the indexing terms t mentioned above has also been shown to be very helpful [18], [19], [22]. Because the incorrectly recognized spoken words may include several subword units correctly recognized, matching on the subword level has the advantages of partial matching. Moreover, all words, whether within the vocabulary of the recognizer or not, are composed of a few subword-level units. Therefore, matching on the subword level is very often an effective approach for bypassing the OOV problem mentioned above because the words in the spoken query/documents may be reasonably matched even if they are not in the vocabulary and cannot be correctly recognized. Furthermore, because word-level terms possess more semantic information, whereas subword-level terms are more robust against the speech recognition errors and the OOV problem, there are good reasons to fuse the information obtained from the two different levels of terms t. In addition, another set of approaches tries to expand the representation of the query/document using the conventional IR techniques (such as pseudo-relevance feedback). Techniques based on the acoustic confusion statistics and/or semantic relationships among the word- or subword-level terms derived from some training corpus have been shown to be very helpful as well [20], [23]. More detailed issues regarding retrieval of spoken documents will also be reviewed by Koumpis and Renals [24].

TECHNOLOGY AREAS FOR SPOKEN DOCUMENT UNDERSTANDING AND ORGANIZATION

As mentioned previously, a number of technology areas are involved in spoken document understanding and organization for efficient retrieval/browsing. Each of these areas will be briefly reviewed in this section.

NAMED ENTITY EXTRACTION FROM SPOKEN DOCUMENTS

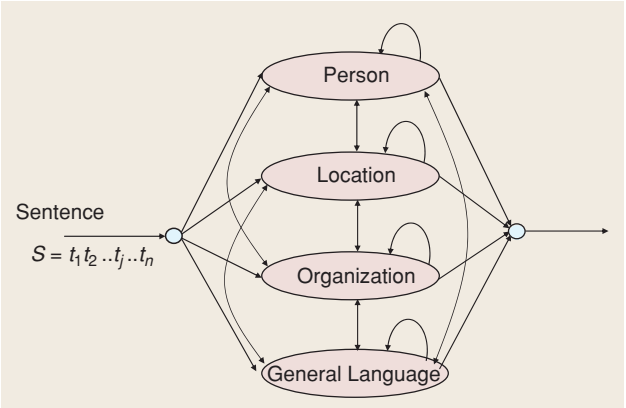
Named entities (NEs) include 1) proper nouns such as names for persons, locations, organizations, artifacts, and so on, 2) temporal expressions such as "Oct. 10 2003" or "1:40 p.m.," and 3) numerical quantities such as "fifty dollars" or "thirty%." They are usually the keywords of the documents. The temporal expressions and numerical quantities can be easily modeled and extracted by rules; therefore, they will not be further mentioned here. The person/location/organization names, however, are much more difficult to identify in text or spoken documents. Sometimes it is even difficult to identify which kind of names they are, for example, "White House" can be either an organization name or a location name based on the context [25]. The NEs, as well as their categories, are the first and most fundamental knowledge required for understanding and organizing the content of the documents. The task of automatic extraction of NEs originated from the Message Understanding Conferences (MUC) sponsored by a U.S. Defense Advanced Research Projects Agency (DARPA) program [26] in the 1990s; this program was aimed at the extraction of information from text documents. More research work on NE extraction from spoken documents has been carried out on American English broadcast news (also under DARPA-sponsored programs)since the late 1990s, and it has been extended to many other languages in the past several years [27], [28]. Substantial work has been done in developing rule-based approaches for locating the NEs [29]. For example, the cue-word "Co." possibly indicates the existence of a company name in the span of its predecessor words and a cue-word "Mr." possibly indicates the existence of a person name in the span of its successor words. Such an approach has been very useful. However, the rules may become very complicated when we wish to cover all different possibilities. It will be very time consuming and difficult to handcraft all the rules, especially when the task domain becomes more general or when new sources of docu-

ments are being handled. This may be the reason that model-based (or machine-learning-based) approaches were proposed later on, such that some model can be initially trained using a sufficient quantity of data annotated with the types and locations of the NEs, and the model can then be steadily improved through use [25]. In many cases, the rule-based and model-based approaches can be properly integrated to take advantage of the nice properties of both.

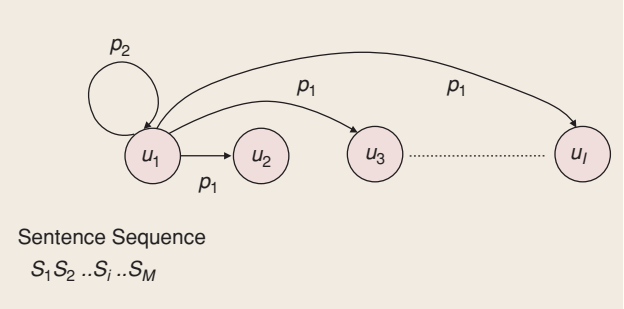
In model-based NE extraction, the goal is usually to find the sequence of NE labels (person name, location name, or other words), $E = n_1 n_2 \ldots n_j \ldots n_n$, for a sentence or term sequence, $S = t_1 t_2 \ldots t_j \ldots t_n$, that maximizes the probability $P(E|S)$. S can be a sequence of recognized words with recognition errors and incorrect sentence boundaries in the transcription of spoken documents and n_j is the NE label for the term t_j. The HMM, as depicted in Figure 4, is probably the best typical representative model used in this category [25]. This model consists of one state modeling each type of the NE (person, location, or organization), plus one state modeling other words in the general language (non-named-entity words), with a possible transition between states. Each state is characterized by a bigram or trigram language model estimated for that state, and state-transition probabilities can be similarly trained. The Viterbi algorithm can thus be used to find the most likely state sequence, or NE label sequence E, for the input sentence; the segment of consecutive words in the same NE state is taken as an NE. As another important representative, in the recent past the maximum entropy (ME) modeling approach has been used in NE extraction. In this approach, many different linguistic and statistical features, such as part-of-speech (POS) information, rule-based knowledge, and term frequencies, can all be represented and integrated in the framework of a statistical model with which the NEs can be identified. It was shown that very promising results can be obtained with this approach [28], [30].

At National Taiwan University, we recently developed a new approach for NE extraction with the help of the global information from the entire document using the PAT tree data structure [31]. Very often, some NEs are difficult to identify in a single sentence. However, if the scope of observation can be extended to the entire document, it will be found that this entity appears several times in several different sentences; it has a higher likelihood to be an NE when all those occurrences in different sentences can be considered jointly. The PAT tree is an efficient data structure successfully used in information retrieval. It is especially efficient for indexing a continuous data stream, such that the frequency counts of all segments of terms in a document can be obtained easily from a tree structure constructed with that document. It was shown that by incorporating the framework of various approaches of NE extraction with a PAT tree for the entire document, significantly improved performance can be achieved. In all cases, the NEs are very often OOV, or unknown, words. A typical approach for dealing with this problem, for example in the HMM modeling mentioned above, is to divide the training data into two parts during training. In each half, every segment of terms or words that does not appear in the other half is marked as "unknown," such that the probabilities for both known and unknown words occurring in the respective NE states can be properly estimated. During testing, any segment of terms that is not seen before can thus be labeled "unknown," and the Viterbi algorithm can be carried out to give the desired results [25].

All the approaches mentioned above are equally useful for NE extraction from both text and spoken documents. Usually, the spoken documents are first transcribed into a sequence of terms or words, and the same approaches for text can then be applied. However, NE extraction from spoken documents is much more difficult than that from text documents, not only because of the presence of recognition errors and problems of spontaneous speech but also due to the absence of the textual clues such as punctuations or sentence boundaries (in particular the capital letters indicating proper nouns). In fact, a more difficult problem is that many NEs are OOV words and thus cannot be correctly recognized in the transcriptions in the first place [27]. Extra measures are, therefore, definitely needed in the case of spoken documents. Performing the extraction on multiple recognition hypotheses, such as word graphs or sausages, and using confidence scores have been two typical approaches. Class-based language models with different POS as classes for the state of general language words to be used in the above HMM approach were also found useful [31]–[33]. In recent years, several approaches were developed to try to utilize the rich resource of the Internet as the background knowledge for extracting from spoken documents NEs that are OOV words. In such approaches, selected words with higher confidence measures in the transcription obtained with the initial recognition on the spoken document can be used to construct queries to retrieve from the Internet text documents relevant to the spoken document being considered. NE extraction can then be performed on these retrieved text documents, the obtained NEs can be added to the recognition vocabulary, and the language model can be adapted using the retrieved text documents. Second-pass transcription of the spoken document based on the updated vocabulary and language model may then be able to correctly recognize the OOV words. In some similar approaches, for the recognized words in the initial transcription with lower



[FIG4] An illustration of HMM-based named-entity extraction.



[FIG5] HMM-based spoken document segmentation.

confidence measures, the associated phone (or syllable) sequence can also be directly matched with the relevant text documents or the newly obtained NEs, such that the incorrectly recognized OOV words may possibly be recovered. In both cases, the achievable improvements depend heavily on the accurate retrieval of the highly relevant documents that can successfully include the OOV words for the spoken documents, which in turn relies on good choice of the indexing terms and the construction of good queries [31], [34].

SPOKEN DOCUMENT SEGMENTATION

Automatic spoken document segmentation is used to set the boundaries between different small topics being mentioned in long streams of audio signals and divide the spoken documents into a set of cohesive paragraphs sharing some common central topic. This is less critical in text documents, which are usually well segmented into paragraphs, sections, or chapters by the authors, although much work has also been done on text document segmentation. For multimedia or spoken documents, such as a three-hour video of a course lecture, a two-hour movie, or a one-hour news episode, such segmentation becomes important for efficient retrieval and browsing. In such cases, the automatic segmentation is to be performed on long streams of transcribed words with recognition errors, even without sentence boundaries. Substantial efforts have been made on feature-based segmentation approaches, in which cue words or phrases indicating the beginning/ending of a topic or a switching point of the topics can be identified as features for segmentation either statistically or with the help of human knowledge [35]. For example, a set of discourse markers consisting of cue words obtained with word statistics was successfully used in conjunction with pause information for the task of detecting paragraph breaks in a technical presentation archive [36]. Approaches based on similarity among sentences have also been extensively investigated. Here, a "sentence" may be a short sequence of transcribed words with fixed length, or the sentence boundaries may be assumed at longer pause durations in the acoustic signal. As an example of such approaches, the similarity between two sentences can be estimated by the number of terms commonly used or cosine measures for the vector representations of the sentences obtained by VSM or LSI in IR. This similarity measure can be evaluated for each pair of sentences within a paragraph hypothesis, as well as for each pair of sentences taken from two adjacent paragraph hypotheses across a segmentation point candidate. By comparing these two different sets of similarities while shifting the segmentation point candidate, the segmentation point can be identified. As another example, assuming a sequence of sentences $S_1S_2 \ldots S_i$ has been determined to describe the same topic, the sentences can be considered as a document and the relevance scores between the next sentence S_{i+1} and this document can be evaluated with any of the above IR approaches (VSM or LSI); based on this, it can be decided if S_{i+1} belongs to the same paragraph or is the beginning of a new paragraph.

Recent work on segmentation seems to be more focused on model-based approaches, primarily the HMM approach [37], [38]. In this approach (as shown in Figure 5), a total of I topic clusters, $\{u_k, k = 1, 2, \ldots, I\}$, form the I states of the HMM. These topic clusters are trained with a training corpus of segmented text documents with labeled topics, or by, for example, K-mean algorithm if the training segments are not labeled. The unsegmented transcription of spoken documents is taken as the observations in the form of a sequence of sentences, $S_1S_2 \ldots S_i \ldots S_M$. Similar to the above, the "sentences" here may not have correct boundaries. The probability for each "sentence" $S_i = t_1 t_2 \ldots t_j \ldots t_n$, where t_j is the jth term, to be observed for each topic cluster (or state) u_k can then be evaluated by N-gram probabilities trained from the training documents in the topic cluster u_k

$$P(S_i|u_k) = m_1 P(t_1|u_k) \times \prod_{j=2}^{n} \left[m_1 P(t_j|u_k) + m_2 P(t_j|t_{j-1}, u_k) \right], \tag{7}$$

assuming uni- and bigram probabilities are used, and m_1 and m_2 are weighting parameters. The above equation is very similar to (3) for HMM for IR, except the document d_i in (3) is replaced by the topic cluster u_k here, and we may not need an outside corpus C for smoothing the uni- and bigram models if enough training corpus is available. Each topic cluster (state) may have a fixed transition probability p_1 for transition to a different state as well as another p_2 for remaining in the same state. The transition probabilities may also be estimated using the frequency counts of transitions between clusters in the training set. Various approaches can be developed to improve this model. For example, p_1 can be further modified by a pause duration model (so a longer pause in the audio signal implies higher probability to transit to a different topic), and p_2 can be further modified by a paragraph-length model (e.g., p_2 can be smaller when the present paragraph has been long enough), and so on [38]. Viterbi algorithm can then be performed on the input observations, and the state transition obtained is taken as a segmentation point. This HMM-based approach has been recently extended to embed the PLSA model in the representation of the state observation probabilities, and a considerable performance gain was indicated [39].

INFORMATION EXTRACTION FOR SPOKEN DOCUMENTS

Information extraction (IE) has long been considered a very important area in natural language understanding. It usually refers to the processes of extracting the salient facts about some prespecified templates of information regarding the entities or events (in particular the relationships among them) from the documents (or paragraphs of documents as discussed here) and then organizing some semantic representation for such information for efficient use. The approaches used in IE vary significantly for different cases, but the core processes include lexical processing, syntactic analysis, semantic analysis, and output presentation [40], [41]. POS tagging is usually an important component in lexical analysis. In this component, a most likely lexical class tag is automatically assigned to each word in a sentence. The NE extraction discussed previously is commonly considered as a part of this component, because the person/location/organization names are very important lexical classes that can be tagged to the words in addition to other lexical classes. In fact, in many cases, the final output of IE is exactly the relationships among the NEs in the documents. Syntactic analysis, on the other hand, is very often accomplished by syntactic parsing, in which the core constituent structures of sentences (e.g., the noun, verb, and prepositional phrases) are identified based on a grammar, an example of which is a set of finite state rules with or without probability distributions. Then, the head words in each of the identified constituents usually give the key information needed. For example, the head words in the subject-noun phrase and the object-noun phrase are very often some domain-specific NEs, the verb that takes them as arguments may represent the relationship between them, and all these together may describe some event. The prepositional phrases may reveal the temporal and location information for the objects or events. Semantic analysis then tries to further explore some additional linguistic cues via semantics. As a simple example, it is often necessary to identify the various forms of the same objects or NEs throughout a given context (e.g., acronyms or aliases) and to resolve the references of pronouns or demonstratives in most cases based on the earlier mentioned terms. The final step of IE is to organize or present the extracted information in some form of appropriate templates (or abstract data structures) for efficient use, for example being automatically entered into a database to be used in indexing/retrieval or question-answering, or being used in composing a summary or a title with natural language.

All the processes mentioned above can be accomplished by either rule- or model-based approaches or the combination of both. Take the component of POS tagging as an example [42]. As mentioned above, POS tagging can be considered as an extended or generalized version of the NE extraction problem because other lexical class tags in addition to the NEs are also to be assigned to all the words in a sentence. It can therefore be imagined that approaches and issues similar to those for NE extraction equally apply here. For example, the rule-based approaches are apparently useful, in which a large set of disambiguation rules derived either manually or statistically are able to solve many of the problems. On the other hand, HMM-based tagging is a good example for model-based approaches. In this approach, the best sequence of tags $Q = p_1 p_2 \ldots p_j \ldots p_n$ for a sentence $S = t_1 t_2 \ldots t_j \ldots t_n$, where p_j is the POS tag for the word (or term) t_j, can be chosen by maximizing the probability $P(Q|S)$ based on a probabilistic generative model or an HMM. This model has an extended form similar to that in Figure 4 and can be trained by a pretagged corpus. There is also the so-called transformation-based tagging, which is in fact an integration of both rule- and model-based approaches. In this case, a set of templates (or transformations) that condition the tag specification of a given word on the context of its preceding and/or succeeding words was developed for deducing the rules that can capture the interdependencies between the words and the corresponding tags. With the aid of a pretagged training corpus, these rules can then be incrementally learned by selecting and sequencing the transformations that can transform the training sentences to a best set of POS tag sequences closest to the set of the corresponding training tag sequences.

All these approaches mentioned above apply equally to text and spoken documents, although at the moment most work on IE is focused on text documents and relatively limited work on spoken documents have been reported [32], [33]. In the case of spoken documents, apparently more difficult problems (such as those due to speech recognition errors and spontaneous speech) must be addressed.

SPOKEN DOCUMENT SUMMARIZATION

Research work in automatic summarization of text documents dates back to the late 1950s, and the efforts have continued through the decades. The World Wide Web not only led to a renaissance in this area but extended it to cover a wider range of new tasks, including multidocument, multilingual, and multimedia summarization [43]. The summarization can, in general, be either extractive or abstractive. Extractive summarization tries to select a number of indicative sentences, passages, or paragraphs from the original document according to a target summarization ratio and then sequence them to form a summary. Abstractive summarization, on the other hand, tries to produce a concise abstract of desired length that can reflect the key concepts of the document. The latter seems to be more difficult, and recent approaches have focused more on the former. As one example, the VSM model for IR can be used, respectively, to represent each sentence of the document as well as the whole document in vector form. The sentences that have the highest relevance scores to the whole document, as evaluated with (2), are selected to be included in the summary. When it is desired to cover more important but different concepts in the summary, after the first sentence with the highest relevance score is selected, indexing terms in that sentence can be removed from the rest of sentences and the vectors can be reconstructed; based on this, the next sentence can be selected, and so on [44]. As another example, the LSI model for IR can be used to represent each sentence of a document as a vector in the latent semantic space for that document, which is constructed by performing SVD on the "term-sentence" matrix W' for that document. The right singular vectors with larger singular values represent

dimensions for more important latent semantic concepts in that document. Therefore, the sentences with vector representations having the largest components in these dimensions can be included in the summary [44]. As still another example, each sentence in the document, $S = t_1 t_2 \ldots t_j \ldots t_n$, represented as a sequence of terms, can be simply given a score $I(S)$, which is evaluated based on some statistical measure $s(t_j)$ (such as TF/IDF in (1) or similar) and linguistic measure $l(t_j)$ (e.g., NEs and different parts-of-speech (POSs) are given different weights, function words are not counted) for all the terms t_j in the sentence,

$$I(S) = \frac{1}{n} \sum_{j=1}^{n} \left[\lambda_1 s(t_j) + \lambda_2 l(t_j) \right], \tag{8}$$

where λ_1 and λ_2 are some weighting parameters, and the sentence selection can be based on this score [45]. These selected sentences in all the above cases can also be further condensed by removing some less important terms if a higher compression ratio is desired.

The above approaches equally apply to both text and spoken documents. However, the spoken documents bring added difficulties such as recognition errors, problems with spontaneous speech, and lack of correct sentence or paragraph boundaries. To avoid the redundant or incorrect parts while selecting the important and correct information, multiple recognition hypotheses, confidence scores, language model scores, and other grammatical knowledge have been utilized [46]. As an example, (8) may be extended as

$$I(S) = \frac{1}{n} \sum_{j=1}^{n} \left[\lambda_1 s(t_j) + \lambda_2 l(t_j) + \lambda_3 c(t_j) + \lambda_4 g(t_j) \right] + \lambda_5 b(S), \tag{9}$$

where $c(t_j)$ and $g(t_j)$ are obtained from the confidence score and N-gram score for the term t_j, $b(S)$ is obtained from the grammatical structure of the sentence S, and λ_3, λ_4, and λ_5 are weighting parameters. In addition, prosodic features (intonation, pitch, energy, or pause duration) can be used as important clues for summarization, although reliable and efficient approaches to use these prosodic features are still under active research [46]. The summary of spoken documents can be in either text or speech form. The text form has the advantages of easier browsing and further processing but is inevitably subject to speech recognition errors as well as the loss of the speaker/emotional/prosodic information carried only by the speech signals. The speech form of summary can perverse the latter information. It is free from recognition errors but with the difficult problem of smooth concatenation of speech segments.

TITLE GENERATION FOR SPOKEN DOCUMENTS

A title is different from a summary, in addition to being especially short. The summary is supposed to tell the key points, major conclusion, or central concepts of the document. A title only tells what the document is about, using only several words or phrases concatenated in a readable form. Generally, a title is at most a single sentence, if not shorter. But the title exactly complements the summary during browsing. The user can easily select the desired document with a glance at the list of titles. Title generation seems to be more difficult than summarization, and much less work has been reported. A few years ago, the Informedia Project at CMU compared several basic approaches based on a framework assuming the availability of some training corpus of text documents with human-generated titles [47]. This assumption is reasonable at least for broadcast news because all text news stories do have human-generated titles. Let $\bar{D} = \{\bar{d}_j, j = 1, 2, \ldots, L\}$ be the set of L training documents and $\bar{Z} = \{\bar{z}_j, j = 1, 2, \ldots, L\}$ be the set of corresponding human-generated titles, all in text form. The basic idea is then to learn the relationships between $\bar{d}_j$ and its corresponding human-generated title $\bar{z}_j$ using $\bar{D}$ and $\bar{Z}$ during training and try to extend such relationships to the given spoken document d_i to obtain the automatically generated title z_i. Such a framework is also quite different from document summarization.

There are several basic approaches for title generation. In the K-nearest-neighbor (KNN) approach [48], for each given spoken document d_i, rather than creating a new title z_i, one may try to find an appropriate title in the training corpus $\bar{z}_j \in \bar{Z}$ for a training document $\bar{d}_j \in \bar{D}$ that is the nearest to d_i. The distance measure between documents can be evaluated with statistics of the indexing terms such as TF/IDF in (1). Because $\bar{z}_i$ is actually generated by a human, it is well readable. Yet, this approach has the fatal problem of requiring some training documents highly correlated to the given spoken documents. It cannot perform well at all for a document telling a complete new story. In the native Bayesian approach with limited terms (NBL) [49], one tries to identify the document-term/title-term pair t cooccurring in all pairs of documents and its corresponding human-generated title $(\bar{d}_j, \bar{z}_j)$ in the training corpus, where $t \in \bar{d}_j$, $\bar{d}_j \in \bar{D}$ and $t \in \bar{z}_j$, $\bar{z}_j \in \bar{Z}$. Then, one estimates the probability for a term t in the training document $\bar{d}_j$ to be selected and used in its human-generated title $\bar{z}_j$. This estimation was achieved by some statistical measures based on the term frequencies for all such document-term/title-term pairs t in its respective title $\bar{z}_j$ and document $\bar{d}_j$ in the training corpus, as well as some other statistics from the training corpus. For a given spoken document, the terms to be used in the automatically generated title are then selected from the terms in the transcription based on the probabilities mentioned above and other statistical parameters, such as the term frequencies for all the terms t in the transcription. In still another approach, those terms to be used in the titles can be simply selected from the extracted NEs with higher scores, and the TF/IDF scores can also be used. But in these latter approaches, the selected terms may not be in good order; therefore, they may need to be resequenced to produce a readable title. One way to achieve this purpose is to perform a Viterbi algorithm on an HMM for the selected terms, in which each selected term is a state and N-gram probabilities are assigned as state transition probabilities.

At National Taiwan University, we successfully developed a improved approach [50], [51] that tries to integrate the nice

properties of the above approaches. We first extract the NEs when transcribing the given spoken document d_i and obtain the top j training documents in the training corpus $\bar{D}$ nearest to this transcription using the KNN approach. The human-generated titles of these j training documents are then rescored based on the term scores obtained from the probabilities estimated in the NBL approach mentioned above, and the best training title $\bar{z}_i^*$ for the given spoken document d_i is chosen. But this title is still for a training document. So the terms that appear in the selected training title $\bar{z}_i^*$ but do not yet appear in the transcription of the given spoken document d_i should be replaced by the NEs extracted from d_i with the highest scores but that do not yet appear in the selected title $\bar{z}_i^*$. All the terms in the title thus obtained should be resequenced by an Viterbi algorithm for the terms, as mentioned above, to produce the final title for the given spoken document. This new approach was shown to produce better titles than the previous approaches. The basic structure of the title is human generated and, therefore, well readable. The keywords of the given spoken documents, very often NEs or OOV words, can be selected and used in the title thanks to the NE extraction and the probabilities for terms to be used in the titles estimated by the NBL approach.

TOPIC ANALYSIS AND ORGANIZATION

Topic analysis and organization refers to analyzing the inherent topics discussed in each document and offering an overall knowledge of the semantic content of the entire document collection in some form of hierarchical structure with concise visual presentation. The purpose is to enable comprehensive and efficient access to the document collection and to derive a best set of query terms for each topic cluster. This is usually some kind of data-driven organization. It can help the users to compose better queries and browse across the documents efficiently as well as to better grasp the relationships among the documents. BBN's Rough'n'Ready system [4] may represent one of the few early efforts for spoken documents in this direction. In this system, each spoken document (broadcast news story) was automatically given a short list of topic labels describing the main themes of the document (serving the purposes of the titles as mentioned above), and all documents were organized in a tree-structure hierarchy classified by dates, sources for the news, and so on.

The WebSOM approach [52], [53] is another typical example of data-driven topic organization for a document collection. In this approach, each document in the collection can be represented by a vector using the LSI model in IR. These vectors are then taken as the input to derive the self-organizing map (SOM) for the documents. SOM is a well-known neural network model that can be trained in an unsupervised way [54]. The documents can thus be clustered based on the latent semantic concepts, and the document clusters and the relationships among the clusters can be presented as a two-dimensional (2-D) map. On this map, each topic cluster is represented as a lattice point (or a neuron), and the closely located lattice points (or neurons) in nearby regions on the map represent related topic clusters. Each document can be assigned a unique coordinate corresponding to the closest cluster representative. In this way, the users are able to browse the semantically related documents on the map, either within the same lattice point (or topic cluster) or across neighboring lattice points.

The ProbMap [55] is yet another typical example with purposes similar to the above but extended from the PLSA model. The documents are organized into latent topic clusters, and the relationships among the clusters can be presented as a 2-D map on which every latent topic cluster is a lattice point. In this approach, an additional set of latent variables $\{Y_k, k = 1, 2, \ldots, K\}$ is introduced with respect to the latent topics $\{T_k, k = 1, 2, \ldots, K\}$. Each latent variable Y_k defines a probability distribution $\{P(T_l|Y_k), l = 1, 2, \ldots, K\}$ that represents the statistical correlation between the latent topic T_k and each of the other latent topics T_l. This distribution not only describes the semantic similarity among the latent topics, but it can blend in additional semantic contributions from related latent topics T_l to a given latent topic T_k. In this way, the conditional probability of observing a term t_j in a document d_i, $P(t_j|d_i)$, previously expressed by (5) in PLSA, can be modified as

$$P(t_j|d_i) = \sum_{k=1}^{K} P(T_k|d_i)P(t_j|T_k) = \sum_{k=1}^{K} P(T_k|d_i)\left[\sum_{l=1}^{K} P(T_l|Y_k)P(t_j|T_l)\right]. \quad (10)$$

The probability distribution $P(T_l|Y_k)$ can be expressed as a neighborhood function in terms of some distance measure between the location of the latent topic T_k and those for other latent topics T_l on the 2-D map. This model can be trained in an unsupervised way by maximizing the total log-likelihood L_T of the document collection $\{d_i, i = 1, 2, \ldots, N\}$ in terms of the unigram $P(t_j|d_i)$ of all terms t_j observed in the document collection, using the EM algorithm:

$$L_T = \sum_{i=1}^{N}\sum_{j=1}^{N'} c(t_j, d_i)\log\ P(t_j|d_1), \quad (11)$$

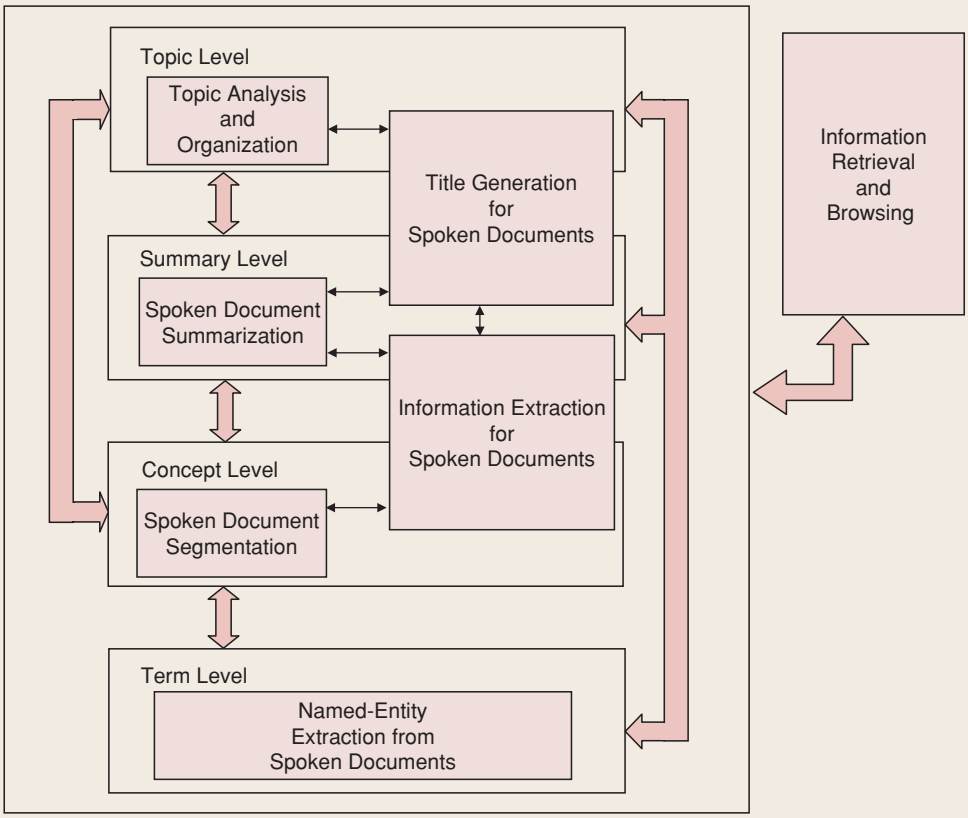
where N is total number of documents in the collection, N' is the total number of different terms observed in the document collection, $c(t_j, d_i)$ is the frequency count for the term t_j in the document d_i, and $P(t_j|d_i)$ is the probability obtained in (10). In this way, the probability of observing a term t_j in a latent topic T_k is further contributed to by the probabilities of observing this term t_j in all different latent topics T_l but weighted by the distance between T_k and T_l on the map, as indicated in (10). By maximizing the total likelihood L_T in (11), those latent topics with higher statistical correlation are to be located close to each other on the map. Different from the WebSOM mentioned above in which each document d_i is assigned to a single topic cluster, here a document d_i can be assigned to many different latent topics with different probabilities; this intuitively seems to be more reasonable. Also, with the total log-likelihood adopted in training,

quite a few theoretically attractive machine learning algorithms can be applied.

INTEGRATION RELATIONSHIPS AMONG THE INVOLVED TECHNOLOGY AREAS

As can be understood, all the various technology areas discussed above are closely related to one another and should be properly integrated to achieve the desired functionalities and the best possible performance. Figure 6 is a diagram for such integration relationships among these areas in a hierarchical structure. As in the figure, the various technology areas may be considered as being on four different levels from bottom to top, including the term, concept, summary, and topic levels. NE extraction belongs to the lowest term level. It provides the most important terms to be used by all the other processes; therefore, it is the base for all higher level processes. Spoken document segmentation divides a long spoken document stream into short paragraphs based on the concepts mentioned. Information extraction further identifies the exact concepts within each paragraph. Therefore, they both belong to the concept level. Precise NEs are certainly important for both of these processes. On the other hand, it is believed that the information extraction actually offers a very good framework helpful in composing good summaries. Yet, it is still not well known today how this can be achieved because most of the spoken document summarization techniques today are actually structured in a different way, for example by selecting the important sentences rather than via information extraction. Similarly, it is also believed that information extraction can offer some key knowledge very helpful in producing good titles, although today again the titles are not generated in this way. The terms used in the titles may naturally be the key terms in the summaries, and the titles may be considered as very concise summaries as well. Therefore, not only can the summarization and title generation both be considered to be on the summary level, but the information extraction can also be considered to extend from the concept level up to the summary level; thus, the concepts identified on the concept level can be further concretized into summaries and titles on the summary level. Finally, the titles usually include key terms indicating the topic area, so the title generation can also be considered to extend from the summary level up to the topic level. On the topic level, the topic analysis and labels may help to produce good titles, and the titles may help in the topic analysis. Topic analysis/organization and title generation on the topic level are again helpful to each other.

The above description provides bottom-up relationships in Figure 6. The complete relationships in Figure 6 also include those that are top down. For example, the knowledge about the topic areas of the documents is definitely helpful to all-lower level processes, for example, to the NE extraction in the lowest



[FIG6] Integration relationships among the involved technology areas for spoken document understanding and organization for retrieval/browsing applications.

term level and the spoken document segmentation and information extraction in the second lowest concept level. In fact, it is easy to see that the understanding achieved on each level is helpful to all other levels, both upward and downward. As a result, an efficient interaction among the different processes on the different levels will be very beneficial, and a good approach incorporating both bottom-up and top-down integration will be highly desired. One possible way to achieve this may be to begin the processes bottom-up and then feed the information or knowledge obtained in some higher-level processes back to some lower levels. The lower-level processes can then be restarted and the bottom-up processes can be reexecuted. This can be carried out recursively in several iterations to achieve the best results, although it is still not quite clear at the moment how this can be exactly accomplished. Finally, all these processes are to serve the purposes of efficient retrieval/browsing applications, and can therefore be considered as extended areas for today's information retrieval.

AN INITIAL PROTOTYPE SYSTEM

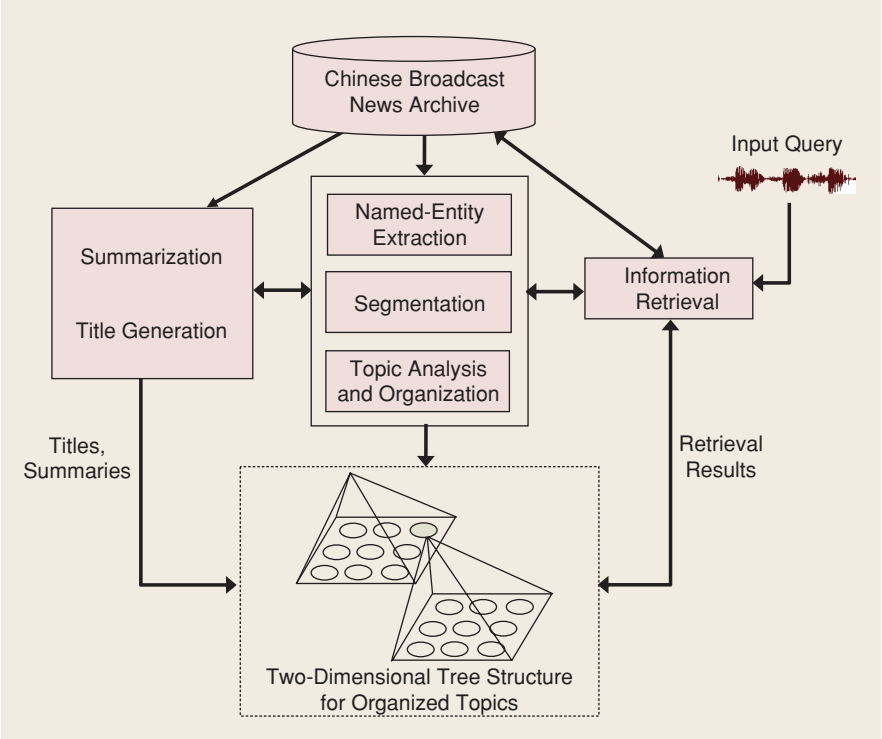
An initial prototype system of spoken document understanding and organization for efficient retrieval/browsing applications, as discussed above, has been successfully developed at National Taiwan University. This system will be briefly summarized here. The broadcast news segments (primarily in speech form only, but some including video parts) are taken as the example spoken/multimedia documents. The document collection to be retrieved and browsed, $D = \{d_i, i = 1, 2, \ldots, N\}$, includes roughly 130 hours of about 7,000 broadcast news stories, all in Mandarin Chinese. They were all recorded from radio/TV stations in Taipei from February 2002 to May 2003. Because of the special structure of the Chinese language, it has been found that special efforts in selecting the indexing terms t mentioned above (usually words for western languages, but they can be segments of one or more syllables, characters, or words in Chinese) may result in significantly better performance in both IR [14], [18] and spoken document understanding and organization [50], [51], [56].

BRIEF SUMMARY FOR THE TECHNOLOGIES USED IN THE INITIAL PROTOTYPE SYSTEM

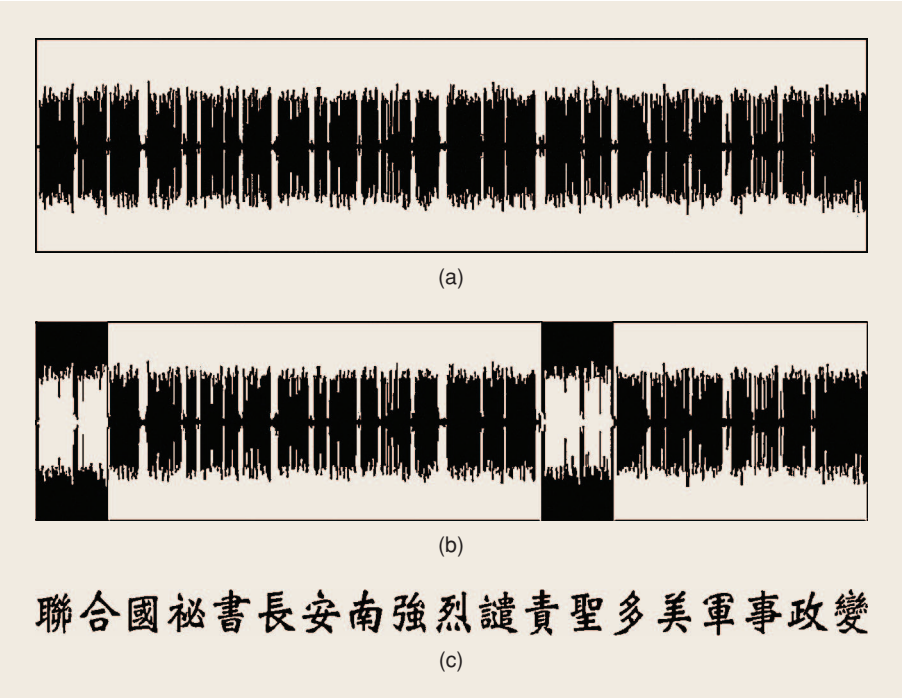
The spoken document retrieval system was implemented using a combination of several different syllable/word-level indexing terms [14], [18]. Only the VSM model for literal term matching for IR was used for simplicity, although it has been shown that the hybrid retrieval model or those with a similar structure as PLSA for concept matching gave better performance for Chinese spoken document retrieval [57], [58].

For NE extraction, the broadcast news segments were first transcribed into word graphs, on which words with higher confidence scores were identified. The words were used to construct queries for retrieval of the text news in the same time period (available over the Internet) to obtain text documents relevant to the spoken documents being considered. The retrieved relevant text documents were used for NE extraction and vocabulary/language model updating for second-pass transcription. Selected parts of these corpora were also used for directly matching NEs, as mentioned previously. This is the way to extract the many NEs in Chinese news that are OOV and cannot be correctly transcribed in the word graphs. Both forward and backward PAT trees were constructed to provide complete data structures for the global context information for each of the spoken documents as well as the retrieved relevant text documents. Both rule- and HMM-based approaches for NE extraction were incorporated. A multilayered Viterbi search was performed with a class-based language model [59] to handle the problem that an NE in Chinese may be a concatenation of several NEs of different types. In this approach, a lower-layer Viterbi search was first performed within those longer NEs, while a top-layer Viterbi search finally performed over the whole sentence [31].

For spoken document segmentation, the purpose is to segment those one-hour-long news episodes into short news stories. The HMM-based segmentation approach mentioned



[FIG7] Block diagram of the initial prototype system for Chinese broadcast news.



[FIG8] A typical example of the spoken document (a broadcast news story) in the collection: (a) the complete audio signal waveform, (b) the waveform of the automatically generated summary (a few selected sentences), and (c) the automatically generated title (in text form).

previously was adopted. The approaches based on sentence similarities were also implemented and tested but found to have slightly worse performance. The total number I for topic clusters (or states) of the HMM was carefully chosen during the K-means training using a training corpus without topic labeling. It was found that $I = 200$ offered reasonably good results. Adaptive algorithms to adjust the transition probabilities were also implemented, including that for p_1 for transition to a different topic cluster based on a pause duration model, and that for p_2 for remaining in the same topic cluster based on a story length model. Confidence scores for the recognized terms were also considered [56], [60]. The information extraction, on the other hand, has not yet been implemented.

For spoken document summarization, every segmented news story was automatically given a summary, but only some preliminary work was completed. The most important sentences in the documents were automatically selected and directly concatenated to form a summary. The selected sentences were not further condensed. Different approaches for choosing the most important sentences were tested, and the two with the best performance were finally selected. The first approach used the term frequency and inverse document frequency as well as VSM in IR. The other approach was in fact very similar, except a simplified version of (9) mentioned above was used [61], [62]. In both cases, the sentences with the highest scores were concatenated with the original order and played in audio form as the summary such that no recognition errors would be perceived [63]. The audio form of summaries also complements the titles in text form. For automatic title generation, the improved approach developed at National Taiwan University was used. It generated a title in text form for each segmented news story in addition to a summary.

Topic analysis for spoken documents, on the other hand, was performed with the ProbMap approach based on the PLSA model. All the news stories in the document collection were first automatically divided into 20 categories, such as international political news and local business news. All the news stories within a category were then clustered into $m \times m$ latent topics T_k based on the PLSA model and then organized as an $m \times m$ map. Every latent topic T_k was then characterized by the several terms (words including NEs) selected with the probabilities $P(t_j|T_k)$. These terms (words) served as the topic labels for the cluster. This map was displayed as a square of $m \times m$ blocks, with the topic labels shown on each block to indicate what the documents in this cluster are all about. As mentioned previously, the distance between two blocks on the map has to do with the relationship between the latent topics represented by the blocks. All the documents in each block (or cluster) could then be further analyzed and organized into another $l \times l$ smaller latent topic cluster and represented as another $l \times l$ map in the next layer, in which the blocks or clusters again represent the fine structures of the smaller latent topics. In this manner, the relationships among the topics of the segmented news stories can be organized into a 2-D tree structure or a multilayer map, both of which enable much easier retrieval and browsing [64].

BRIEF SUMMARY FOR THE FUNCTIONALITIES OF THE INITIAL PROTOTYPE SYSTEM

Figure 7 is the block diagram of the initial prototype system for Chinese broadcast news. There are three parts in the system: the automatic generation of titles and summaries is on the left, the IR is on the right, and the rest of the modules are in the middle. The output below is the 2-D tree structure for the organized topics of the broadcast news.

A typical example broadcast news story d_i within the document collection is shown in Figure 8. The waveform of the complete news story (with length 63.0 s) is shown in Figure 8(a). The waveform of the automatically generated summary, which is the concatenation of a few selected sentences (with length 8.5 s), is shown in Figure 8(b). The text of this summary is "西非國家聖多美發生軍事政變，總理及政府首長被扣押，聯合國秘書長安南發表

聲明加以強烈譴責(A military coup took place in Sao Tome, a country in West Africa. The prime minister and many government officers were detained. UN Secretary General Annan issued a statement of strong condemnation.)," which includes several character errors (not shown here). As the speech waveform was simply extracted from the original audio signal, the text is completely correct. The automatically generated title in text form is printed in Figure 8(c), the English translation of which is, "UN Secretary General Annan strongly condemned the military coup in Sao Tome." This title is not a sentence in the original news story but automatically put together with a set of the most important NEs. It should be pointed out that all the NEs here are OOV words; yet they are correctly recovered, and the title here is actually smoothly readable. Note that the summary and title here may not be very good for the original news story. While certainly worse than those generated by humans, they may be reasonably good for retrieval/browsing purposes. The algorithms for automatically generating the summary and the title shown in Figure 8 were performed on all the 7,000 news stories in the collection as mentioned above. The generated summaries and titles were all stored in the collection together with the original news stories. Figure 8 is a typical example out of the 7,000 stories.

[FIG9] **The top-down browsing and button-up retrieval functionalities of the initial prototype system.**

The functionalities of the initial broadcast news retrieval/browsing prototype system are shown in Figure 9 and described below. First consider the top-down browsing functionalities. The home page of the browsing system lists the 20 news categories as in Figure 9(a) (not completely shown). When the user clicks the first category of "international political news," as shown here, a 2-D map of 3×3 latent topic structure (with nine blocks) appears as shown in Figure 9(b) (only four blocks are shown here). Each block represents a major latent topic in the area of "international political news" for the news collection, characterized by roughly four topic labels (terms selected with probabilities) shown in the block. As can be found, the block on the upper-right corner has labels "以色列 (Israel)," "阿拉法特 (Arafat)," "巴勒斯坦 (Palestine)" and "迦薩市 (Gaza City)," which clearly specify the topic. The block to its left, on the other hand, has labels "伊拉克 (Iraq)," "巴格達 (Baghdad)," "美軍 (American Amy)" and "陸戰隊 (Marine Corps)," whereas the block below it in the middle-right has labels "聯合國 (United Nations)," "安理會 (Security Council)," "武檢人員 (military inspectors)," and "武器 (weapons)." Clearly, these are all different but related topics, and the distance between the blocks has to do with the relationships between the latent topics. The user can then click one of the blocks (for example, the one on the upper-right corner as shown here) to see the next layer 3×3 map for the fine structure of smaller latent topics for this cluster, as shown in Figure 9(c). As can be found in Figure 9(c), the block on the upper-right corner now has labels "以色列 (Israel)," "夏隆 (Shilom)," "約旦河 (Jordan River)" and "美國 (USA)," while the block below it has labels "中東 (Middle East)," "鮑威爾 (Powell)," "和平 (peace)," and "路線 (roadmap)," and so on. Apparently, the collection of broadcast news stories is now organized in a 2-D tree structure or a multilayer map for better indexing and easier browsing. Here the second-layer clusters are in fact the leaf nodes, and the user may wish to see all the news stories within such a node. With just a click, the automatically generated titles for all news stories clustered into that node are shown in a list, as in Figure 9(d) for the upper-middle small block in Figure 9(c) labeled with "阿拉法特 (Arafat)." This list includes the automatically generated titles for five news stories clustered into this block, together with the position of this node within the 2-D tree as shown in the lower-right corner of the screen. The user can further click the "summary" button after each title to listen to the automatically generated summaries, or click the title to listen to the complete news story. This 2-D tree structure with topic labels and the titles/summaries is therefore very helpful for browsing the news stories.

The retrieval functionalities, on the other hand, are generally bottom-up. The screen of the retrieval system output for an input speech query (can be in either speech or text form), "請幫我找以色列與阿拉法特相關的新聞 (Please find news stories relevant to Israel and Arafat)" is shown in Figure 9(e). A nice feature of this system is that all retrieved news stories, as listed in the upper half of Figure 9(e), have automatically generated titles and summaries. The user can therefore select the news stories by browsing through the titles or listening to the summaries, rather than listening to the entire news story and then finding that it was not the one he was looking for. The user can also click another functional button to see how a selected retrieved news item is located within the 2-D tree structure, as mentioned previously in a bottom-up process. For example, if he selected the second item in the title list of Figure 9(e), "阿拉法特反對以色列所提結束包圍條件 (Arafat objected to Israel's proposal for conditions of lifting the siege)," he can see the list of news titles in Figure 9(d), including the titles of all news stories clustered in that smaller latent topic (or leaf node). Alternatively, he can go one layer up to see the structure of different smaller latent topics in Figure 9(c) or go up one layer further to see the structure of different major latent topics in Figure 9(b), and so on. This bottom-up process is very helpful for the user to identify the desired news stories or find the related news stories, even if they are not retrieved in the first step as shown in Figure 9(e).

We also successfully implemented a prototype subsystem that allows the user to retrieve the broadcast news via a PDA using speech queries. A small client program was implemented on the PDA to transmit the acoustic features of the speech query to the IR server. The retrieved results are then sent back to the PDA, and the user can use the PDA to browse through the titles and the 2-D tree described above, listen to the summaries of the retrieved documents, and click to play the audio files for the complete news story sent from the audio streaming server.

PERFORMANCE EVALUATION FOR THE INITIAL PROTOTYPE SYSTEM

The performance evaluation for the initial prototype system, especially for each of the individual modules, is reported below.

TRANSCRIPTION AND RETRIEVAL OF THE BROADCAST NEWS

For transcription of the broadcast news, a corpus of 112 hours of radio and TV broadcast news collected in Taipei from 1998 to 2004 was used in acoustic model training. This is different from the 130 hours of document collection $D = \{d_i, i = 1, 2, \ldots, N\}$ used in the initial prototype system. Out of this corpus, 7.7 hours of speech were equipped with orthographic transcriptions, in which 4.0 hours were used to bootstrap the acoustic model training and the other 3.7 hours were used for testing. The other 104.3 hours of untranscribed speech were used for lightly supervised acoustic model training [65]. The language model used consisted of bi- and trigram models estimated from a news text corpus of roughly 170 million Chinese characters with Katz smoothing. The character and syllable error rates of 14.29% and 8.91%, respectively, were achieved for the transcription task [65]. Note that word error rate (WER) is not a good performance measure for the Chinese language in general due to the ambiguity in segmenting a Chinese sentence into words; therefore, it is not used here.

In the retrieval experiments, a set of 20 simple queries with length of one to several words, in both text and speech forms,

was manually created. Four speakers (two males and two females) produced the 20 queries using an Acer n20 PDA with its original microphone in an environment of slight background noise. To recognize these spoken queries, another read speech corpus consisting of 8.5 hours of speech produced by an additional 39 male and 38 female speakers over the same type of PDAs was used for training the speaker-independent HMMs for recognition of the spoken queries. Significantly higher character and syllable error rates of 27.61% and 19.47%, respectively, were obtained for the spoken queries as compared to those for broadcast news mentioned above. The retrieval experiments were performed with respect to a collection of about 21,000 broadcast news stories, all recorded in Taipei from 2001–2004 (including the 7,000 used in the initial prototype system presented here). The results in terms of the mean average precision at a document cutoff number of ten were 0.8038 and 0.6237 for text and spoken queries, respectively. At a document cutoff number of 30, the mean average precision were 0.6692 and 0.5232 for text and spoken queries, respectively.

A TITLE IS DIFFERENT FROM A SUMMARY, IN ADDITION TO BEING ESPECIALLY SHORT.

PERFORMANCE EVALUATION FOR NE EXTRACTION

The performance evaluation for NE extraction was performed with both text documents and the broadcast news (spoken documents). For the test with text documents, the Chinese test corpus of Multilingual Entity Task (MET-2) of MUC-7 [26] was used, which included 100 Chinese documents on the same topic with manually identified reference NEs. The evaluation was based on the recall rate r_1, precision rate r_2, and F1 score

$$F1 = \frac{2\, r_1\, r_2}{r_1 + r_2}, \tag{12}$$

as defined by the MUC-7 test references. The results are listed in the upper half of Table 1, in which the baseline (BSL) was based on a recently published, very successful algorithm including the multilayered Viterbi search with some special approaches [59]. The improved approach (IMP) was that described earlier using PAT trees to consider global context information but not including those specifically developed for spoken documents. Very significant improvements can be found with IMP as compared to BSL; with IMP, the recall/precision rates and F1 scores for the three types of NEs are all well above 0.90. In fact, these numbers for IMP in Table 1 represent the best published results for this MET-2 task for the Chinese language up to this moment. For the test with broadcast news, 200 news stories broadcast in September 2002 and recorded in Taipei were used as the test corpus. The manually identified NEs were taken as references, including 315 person names, 457 location names, and 500 organization names. Many of them are actually OOV words. The text news corpora searched to recover the NEs in the broadcast news that are OOV words were the Chinese text news available from the "Yahoo! Kimo News Portal" [66] for the whole month of September 2002, including about 38,800 text news stories. The results are listed in the lower half of Table 1, in which BSL are those with exactly the same approach as BSL used for text documents mentioned above, performed directly on the transcriptions of the broadcast news. IMP are those with exactly the approaches reported earlier, including using the knowledge obtained from the retrieved text news corpora, using PAT trees to analyze the global context information, and considering confidence scores on word graphs. There was a clear gap between the results for spoken documents and those for text documents, but the proposed IMP approach offered very significant improvements as compared to BSL in almost all cases. The overall F1 score for IMP reached exactly 0.80, which was very satisfactory for Chinese broadcast news that included many OOV words.

PERFORMANCE EVALUATION FOR BROADCAST NEWS SEGMENTATION, SUMMARIZATION, AND TITLE GENERATION

Preliminary tests for segmentation of Chinese broadcast news were performed with TDT 2001 evaluation data [67]. TDT-2 was used as the training corpus, including 55.3 hours of audio signals, or about 3,000 news stories. TDT-3 was employed as the testing corpus, including 127.0 hours of audio signals, or about 4,600 news stories, all in Mandarin Chinese. The segmentation cost defined by TDT evaluation was used here and included the cost for false alarm and missing [67]. Very good initial results were obtained. It was found that the confidence measures and the adaptive adjustment of the transition probabilities p_1 and p_2 (by a pause duration model and a story length model, respectively) for the HMM model as mentioned previously can offer very significant improvements. Proper choice of the terms t (e.g., segments of two

[TABLE 1] PERFORMANCE EVALUATION FOR CHINESE NAMED ENTITY EXTRACTION FROM TEXT DOCUMENTS AND BROADCAST NEWS.

EXPERIMENTS		NAMED ENTITIES	RECALL (r_1)	PRECISION (r_2)	F1 SCORE
TEXT DOCUMENTS	BSL	PERSON NAME	0.94	0.66	0.775
		LOCATION NAME	0.90	0.77	0.831
		ORGANIZATION NAME	0.75	0.89	0.814
	IMP	PERSON NAME	0.95	0.96	0.955
		LOCATION NAME	0.95	0.92	0.935
		ORGANIZATION NAME	0.91	0.96	0.934
BROADCAST NEWS	BSL	PERSON NAME	0.65	0.73	0.688
		LOCATION NAME	0.81	0.62	0.702
		ORGANIZATION NAME	0.77	0.44	0.560
		OVERALL	0.74	0.53	0.618
	IMP	PERSON NAME	0.73	0.85	0.785
		LOCATION NAME	0.87	0.91	0.890
		ORGANIZATION NAME	0.67	0.95	0.740
		OVERALL	0.74	0.88	0.800

syllables or two characters), considering the structure of the Chinese language, to replace the role of the words commonly used for western languages certainly made the difference. The lowest segmentation cost is only slightly above 0.08 [56], [60].

In the preliminary tests for broadcast news summarization, the training corpus included roughly 150,000 news stories in text form from January to December 2000 provided by the Central News Agency of Taipei. These samples were used to calculate the IDF and other statistical parameters. The testing corpus included 200 news stories broadcast in August 2001 by a few radio stations at Taipei. Three human subjects (students at National Taiwan University) were requested to do the human summarization in two forms: first, 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 second, write an abstract for the news story with a length being roughly 25% of the original news story. Two summarization ratios, 20% and 30%, were tested; these are the ratios of summary length to the total length. In each case, the two human-produced summaries were used. The first, y_1, was the concatenation of the top several important sentences selected by the human subject. The second, y_2, was simply the one he wrote by himself. The summarization accuracy for the jth news story, A_j, was then the averaged similarity score [56], [63], [68] for the machine-produced summary, $\bar{y}$, with respect to y_1 and y_2,

$$A_j = \frac{1}{2}\left[S(\bar{y}, y_1) + S(\bar{y}, y_2)\right], \tag{13}$$

where the similarity scores $S(\bar{y}, y_1)$, $S(\bar{y}, y_2)$ were calculated based on the vectors of the TF/IDF values. In this way, higher accuracy would be obtained if more words that were important in the news stories were included in the machine-produced summaries. The final summarization accuracy was then the average of A_j in (13) over all the 200 news stories and all the three human subjects. The final summarization accuracy was found to be slightly above 0.381 and 0.422 for 20% and 30% summarization ratios, respectively. Proper choice and reasonable combination of a few word- and subword-level terms t during the process of automatic summarization were certainly key to achieving better accuracy [56], [63].

In the preliminary tests for title generation, the same training corpus used in summarization experiments (including roughly 150,000 news stories in text form) was used in training, except here the human-generated titles for all the text news stories were used in training as well; 210 broadcast news stories recorded in 2001 were used in testing. The reference titles for these broadcast news stories 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 was based on F1 scores in (12), where the precision and recall rates were calculated from the number of identical Chinese characters in automatically generated and human-generated titles. In addition, five-level, subjective human evaluation was also performed, where five was the best and one was the worst. Two different metrics were used in the subjective human evaluation: "relevance," calibrating the correlation between the automatically generated titles and the broadcast news, and "readability," indicating if the automatically generated title was readable. In performing the subjective human evaluation, each subject was given in advance the example titles with reference scores for both "relevance" and "readability" of five, three, and one for some example broadcast news stories. They were then asked to follow the calibration of the examples to give the scores between five and one, so the results could be more consistent for different subjects. The best results were obtained with proper choice of the terms t (segments of two or three syllables/characters to replace the role of words in western languages) and careful integration of them considering the structure of the Chinese language. It was found that the new approach developed at National Taiwan University performed much better than the few previously proposed approaches. An F1 score slightly above 0.356 was obtained, with average scores of 3.294 and 4.615 for "relevance" and "readability," respectively, for the subjective human evaluation [50], [51], [64].

SPEECH IS THE PRIMARY AND MOST CONVENIENT MEANS OF COMMUNICATION BETWEEN INDIVIDUALS.

PERFORMANCE EVALUATION FOR TOPIC ANALYSIS AND ORGANIZATION FOR BROADCAST NEWS

Very rigorous performance evaluation for the ProbMap approach has been performed based on the TDT-3 Chinese broadcast news corpus. A total of about 4,600 news stories in this corpus were used to train the 2-D tree structure for the topics. A total of 47 different topics have been manually defined in TDT-3, and each news story was assigned to one of the topics manually, or assigned as "out of topic." These 47 classes of news stories with given topics were used as the reference for the two evaluation measures as defined below.

Intuitively, those news stories manually assigned to the same topic should be located on the map as close to each other as possible, while those manually assigned to different topics should be located on the map as far away from each other as possible. We therefore define the "between-class to within-class" distance ratio R as in

$$R = \frac{\bar{h}_B}{\bar{h}_w}, \tag{14}$$

where $\bar{h}_B$ is the average of the distance on the map for all pairs of news stories manually assigned to different topics (thus is the "between-class distance"), and $\bar{h}_w$ is the similar average, but over all pairs of news stories manually assigned to identical topics (thus

the "within-class distance"). So the ratio R in (14) describes how distant the news stories with different manually defined topics are on the map. Apparently, the higher the values of R, the better.

On the other hand, for each news story d_i, the probability $P(T_k|d_i)$ for each latent topic T_k, $k = 1, 2, \ldots, K$, was given by the model. Thus, the total entropy for topic distribution for the whole document collection with respect to the organized topic clusters can be defined as

$$H = \sum_{i=1}^{N} \sum_{k=1}^{K} P(T_k|d_i) \log \frac{1}{P(T_k|d_i)}, \qquad (15)$$

where N is the total number of news stories used in the evaluation. Apparently, lower total entropy means the news stories have probability distributions more focused on fewer topics.

Table 2 lists the results of the two performance measures proposed above. There are several choices of the terms considering the special structure of the Chinese language, i.e., W (words), $S2$ (segments of two syllables), $C2$ (segments of two characters), and combinations. As we can see, the words [W in row 1] were certainly not a good choice of terms for the purposes of topic analysis here. Segments of two syllables [$S2$ in row 2] were apparently better, with much higher distance ratio R and much lower total entropy H. Segments of two characters [$C2$ in row 3] turned out to be even better. The last row indicated that integration of $S2$ and $C2$ may be another good choice, with better distance ratio R, though slightly higher total entropy H. This is consistent with the results in retrieval, segmentation, summarization, and title generation. In all these cases, the word is not a good choice of term for Chinese spoken documents, considering the structure of the Chinese language.

CONCLUSION

Speech usually carries the core concepts for the ever-increasing multimedia content in the network era, and therefore spoken document understanding and organization will be the key for efficient retrieval/browsing applications in the future. This article presents a concise, comprehensive, and integrated overview of various technology areas reaching towards such a goal in a unified context. The involved technology areas covered here include NE extraction, segmentation, and information extraction for the spoken documents as well as automatic summarization, title generation, and topic analysis and organization. The relevant problems and issues, general principles, and basic approaches for each area were briefly reviewed. A framework for properly integrating all these different technology areas was proposed, in which four different levels of processes were defined (term, concept, summary, and topic levels) and bottom-up and top-down relationships were discussed. An initial prototype system for such purposes recently developed at National Taiwan University was also presented. This system used broadcast news in Mandarin Chinese as example spoken documents. Preliminary performance results for the various functionalities for the initial prototype system were reported as well.

[TABLE 2] PERFORMANCE EVALUATION FOR THE *PROBMAP* APPROACH FOR TOPIC ANALYSIS AND ORGANIZATION OF BROADCAST NEWS.

CHOICE OF TERMS	DISTANCE RATIO R	TOTAL ENTROPY H
1) W	2.34	5135.62
2) $S2$	3.38	4637.71
3) $C2$	3.65	3489.21
4) $S2 + C2$	3.78	4096.68

AUTHORS

Lin-shan Lee received a B.S. degree from National Taiwan University in 1974 and a Ph.D. from Stanford University in 1977. He has been a professor at National Taiwan University since 1982. He also holds a joint appointment with Academia Sinica as a research fellow. His research areas include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world, including text-to-speech systems, natural language analyzers, voice dictation systems, spoken document retrieval systems, and spoken dialogue systems. He is a member of the Permanent Council of the International Conference on Spoken Language Processing (ICSLP) and a board member of International Speech Communication Association (ISCA). He was the vice president for the International Affairs and the Awards Committee chair of IEEE Communications Society. He is a Fellow of IEEE.

Berlin Chen received his B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1994 and 1996, respectively. He received his Ph.D. in computer science and information engineering in 2001 from National Taiwan University, Taipei. He is currently an assistant professor with National Taiwan Normal University. His research interests include acoustic and language modeling, search algorithms for large-vocabulary continuous speech recognition, and speech IR and organization.

REFERENCES

[1] B.H. Juang and S. Furui, "Automatic recognition and understanding of spoken language—a first step toward natural human-machine communication," *Proc. IEEE*, vol. 88, no. 8, pp. 1142–1165, 2000.

[2] CMU Informedia Digital Video Library project [Online]. Available: http://www.informedia.cs.cmu.edu/

[3] Multimedia Document Retrieval project at Cambridge University [Online]. Available: http://mi.eng.cam.ac.uk/research/Projects/Multimedia_Document_Retrieval/

[4] D.R.H. Miller, T. Leek, and R. Schwartz, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[5] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "SCAN: Designing and evaluating user interface to support retrieval from speech archives," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 26–33.

[6] A. Merlino and M. Maybury, "An empirical study of the optimal presentation of multimedia summaries of broadcast news," in *Automated Text Summarization*, I. Mani and M. Maybury, Eds. Cambridge, MA: MIT Press, 1999, pp. 391–401.

[7] SpeechBot Audio/Video Search at Hewlett-Packard (HP) Labs [Online]. Available: http://www. speechbot.com/

[8] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1–6.

[9] Y.Y. Wang, L. Deng, and A. Acero, "Spoken language understanding," *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 16–31, Sept. 2005.

[10] Text Retrieval Conference [Online]. Available: http://trec.nist.gov/

[11] G. Salton and M.E. Lesk, "Computer evaluation of indexing and text processing," *J. ACM*, vol. 15, no. 1, pp. 8–36, 1968.

[12] J.M. Ponte and W.B. Croft, "A language modeling approach to information retrieval," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1998, pp. 275–281.

[13] D.R.H. Miller, T. Leek, and R. Schwartz, "A hidden Markov model information retrieval system," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 214–221.

[14] B. Chen, H.M. Wang, and L.S. Lee, "A discriminative HMM/N-gram-based retrieval approach for Mandarin spoken documents," *ACM Trans. Asian Lang. Inform. Processing*, vol. 3, no. 2, pp. 128–145, 2004.

[15] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R. Harshman, L.A. Streeter, and K.E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1988, pp. 465–480.

[16] J.R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Mag.,* vol. 22, no. 5, pp. 70–80, Sept. 2005.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 50–57.

[18] B. Chen, H.M. Wang, and L.S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 303–314, 2002.

[19] K. Ng and V.W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Commun.*, vol. 32, no. 3, pp. 157–186, 2000.

[20] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 2000, pp. 81–87.

[21] B. Chen, H.M. Wang, and L.S. Lee, "Improved spoken document retrieval by exploring extra acoustic and linguistic cues," in *Proc. European Conf. Speech Communication and Technology*, 2001, pp. 299–302.

[22] E. Chang, F. Seide, H. Meng, Z. Chen, Y. Shi, and Y.C. Li, "A system for spoken query information retrieval on mobile devices," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 8, pp. 531–541, 2002.

[23] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 34–41.

[24] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Mag.,* vol. 22, no. 5, pp. 61–70, Sept. 2005.

[25] D.M. Bikel, R. Schwartz and R.M. Weischedel, "An algorithm that learns what's in a name," *Mach. Learn.*, vol. 34, no. 1–3, pp. 211–231, 1999.

[26] Message Understanding Conference [Online]. Available: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

[27] M. Federico, N. Bertoldi, and V. Sandrini, "Bootstrapping named entity recognition for Italian broadcast news," in *Proc. ACL Conf. Empirical Methods in Natural Language Processing*, 2002, pp. 296–303.

[28] A. Kobayashi, F.J. Och, and H. Ney, "Named entity extraction from Japanese broadcast news," in *Proc. European Conf. Speech Communication and Technology*, 2003, pp. 1125–1128.

[29] D.E. Appelt, R.H. Jerry, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson, "SRI international FASTUS system MUC-6 test results and analysis," in *Proc. 6th Message Understanding Conf. (MUC-6)*, 1995, pp. 237–248.

[30] A. Borthwick, "A maximum entropy approach to named entity recognition," Ph.D. dissertation, New York Univ., NY, 1999.

[31] Y.I. Liu, "An initial study on named entity extraction from Chinese text/spoken documents and its potential applications," M.S. thesis, National Taiwan Univ., July 2004.

[32] D.D. Palmert, M. Ostendorf, and J.D. Burger, "Robust information extraction from spoken language data," in *Proc. European Conf. Speech Communication and Technology*, 1999, pp. 1035–1038.

[33] D.D. Palmert, "Modeling uncertainty for information extraction from speech data," Ph.D. dissertation, Univ. Washington, 2001.

[34] A. Fujii, K. Itou, and T. Ishikawa, "A method for open-vocabulary speech-driven text retrieval," in *Proc. 2002 Conf. Empirical Methods in Natural Language Processing*, 2002, pp. 188–195.

[35] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1–3, pp. 1–34, 1999.

[36] T. Kawahara, M. Hasegawa, K. Shitaoka, T. Kitade, and H. Nanjo, "Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 4, pp. 409–419, 2004.

[37] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbreget, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1998, vol. 1, pp. 333–336.

[38] W. Grei, A. Morgan, R. Fish, M. Richards, and A. Kundu, "Fine-grained hidden Markov modeling for broadcast-news story segmentation," in *Proc. Human Language Technology Conf.*, 2001.

[39] D.M. Blei and P.J. Moreno, "Topic segmentation with an aspect hidden Markov model," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 2001, pp. 343–348.

[40] D.E. Appelt and D.J. Israel, "Introduction to information extraction technology," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1999.

[41] R. Engels and B. Bremdal, "Information extraction: State-of-the-art report," On-To-Knowledge Consortium, CognIT a.s, Asker, Norway, IST Project IST-1999-10132, 2000.

[42] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[43] I. Mani and M.T. Maybury, Eds., *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.

[44] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 2001, pp. 19–25.

[45] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell, "Summarizing text documents: Sentence selection and evaluation metrics," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 121–128.

[46] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[47] R. Jin and A.G. Hauptmann, "Title generation for spoken broadcast news using a training corpus," in *Proc. Int. Conf. Spoken Language Processing*, 2000, pp. 680–683.

[48] Y. Yang, C.G. Chute, "An example-based mapping method for text classification and retrieval," *ACM Trans. Inform. Syst.*, vol. 12, no. 3, pp. 252–77, 1994.

[49] M. Witbrock and V. Mittal, "Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 315–316.

[50] S.C. Chen and L.S. Lee, "Automatic title generation for Chinese spoken documents using an adaptive K-nearest-neighbor approach," in *Proc. European Conf. Speech Communication and Technology*, 2003, pp. 2813–2816.

[51] L.S. Lee and S.C. Chen, "Automatic title generation for Chinese spoken documents considering the special structure of the language," in *Proc. European Conf. Speech Communication and Technology*, 2003, pp. 2325–2328.

[52] T. Kohonen, S. Kaski, K. Lagus, J. Salojvi, J. Honkela, V. Paatero, and Saarela A, "Self organization of a massive document collection," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 574–585, 2000.

[53] M. Kurimo, "Thematic indexing of spoken documents by using self-organizing maps," *Speech Commun.*, vol. 38, no. 1, pp. 29–45, 2002.

[54] T. Kohonen, *Self-Organizing Maps*, 2nd ed. Berlin: Springer, 1997.

[55] T. Hofmann, "ProbMap—A probabilistic approach for mapping large document collections," *J. Intell. Data Anal.*, vol. 4, no. 2, pp. 149–164, 2000.

[56] L.S. Lee, Y. Ho, J.F. Chen, and S.C. Chen, "Why is the special structure of the language important for Chinese spoken language processing?—Examples on spoken document retrieval, segmentation and summarization," in *Proc. European Conf. Speech Communication and Technology*, 2003, pp. 49–52.

[57] C.J. Wang, B. Chen, and L.S. Lee, "Improved Chinese spoken document retrieval with hybrid modeling and data-driven indexing features," in *Proc. Int. Conf. Spoken Language Processing* [CD-ROM], 2002, pp. 1985–1988.

[58] B. Chen, J.W. Kuo, Y.M. Huang, and H.M. Wang, "Statistical Chinese spoken document retrieval using latent topical information," in *Proc. Int. Conf. Spoken Language Processing*, 2004, pp. 1621–1625.

[59] J. Sun, J. Gao, L. Zhang, M. Zhou, and C. Huang, "Chinese named entity identification using class-based language model," in *Proc. Int. Conf. Computational Linguistics*, Taipei, 2002, pp. 967–973.

[60] J.F. Chen, "Chinese spoken document segmentation with consideration of features, language models and extra information—Examples using broadcast news," M.S. thesis, National Taiwan Univ., July 2003.

[61] T. Kikuchi, S. Furui, and C. Hori, "Two-stage automatic speech summarization by sentence extraction and compaction," in *Proc. IEEE and ISCA Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 207–210.

[62] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000, vol. 3, pp. 1579–1582.

[63] Y. Ho, "An initial study on automatic summarization of Chinese spoken documents," M.S. thesis, National Taiwan Univ., July 2003.

[64] S.C. Chen, "Initial studies on Chinese spoken document analysis—Topic segmentation, title generation, and topic organization," M.S. thesis, National Taiwan Univ., July 2004.

[65] B. Chen, J.W. Kuo, and W.H. Tsai, "Lightly supervised and data-driven approaches to Mandarin broadcast news transcription," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2004, vol. 1, pp. 777–780.

[66] Yahoo! Kimo News Portal [Online]. Available: http://tw.news.yahoo.com/

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

[68] E. Hovy and D. Marcu, "Automated text summarization tutorial," in *Proc. 36th Ann. Meeting Association for Computational Linguistics and 17th Int. Conf. Computational Linguistics*, Montreal, Quebec, Canada, 1998. SP