A MULTI-MODAL DIALOGUE SYSTEM FOR INFORMATION NAVIGATION AND RETRIEVAL ACROSS SPOKEN DOCUMENT ARCHIVES WITH TOPIC HIERARCHIES

*Yi-cheng Pan, Chien-chih Wang, Ya-chao Hsieh, Te-hsuan Lee, Yen-shin Lee,
Yi-sheng Fu, Yu-tsun Huang and Lin-shan Lee*

Graduate Institute of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

speech@speech.ee.ntu.edu.tw

ABSTRACT

Unlike the written documents, the spoken documents are difficult to be shown on the screen and browsed by the user during retrieval. In this paper, we propose to use multi-modal dialogues to help the user to “navigate” across the spoken document archives and retrieve the desired documents based on a topic hierarchy constructed by the key terms extracted from the retrieved spoken documents. An initial prototype system for such functions has been developed, in which the broadcast news in Mandarin Chinese was taken as the example spoken documents, and the Named Entities (NEs) are taken as the key terms to construct the topic hierarchy.

1. INTRODUCTION

The most attractive form of future network content will be multi-media including speech information, and such speech information usually carries the core concepts for the content. As a result, the spoken documents associated with the multi-media content very possibly can serve as the key for retrieval and browsing. On the other hand, the wireless networks have made it possible for users to access the network resources easily at any time from anywhere using small hand-held devices. Very substantial research efforts have been made in recent years, and very successful techniques and systems have been developed in the area of Spoken Document Retrieval (SDR). Carefully designed robust features and retrieval models were used to handle the high degree of variation in signal characteristics of the spoken queries and documents produced under different acoustic conditions, as well as the complicated concepts and knowledge carried by the spoken or multi-media documents. But many difficult problems still remain unsolved.

The first problem is that unlike the written documents which are well structured with paragraphs and titles and easy to browse with human eyes, the spoken documents are simply audio signals. The user can’t just listen to each of the retrieved document from beginning to the end during browsing. Also, the query given by the user is usually very short and thus not specific enough, and very often gives large number of retrieved documents. But the screen on the hand-held devices is very small, not able to show enough number of retrieved documents.

In this paper, we propose to use multi-modal dialogues to help the user to “navigate” across the spoken document archives and retrieve the desired documents based on a topic hierarchy constructed by the key terms extracted from the retrieved spoken documents. An initial prototype system for such functions has been developed, in which the broadcast news in Mandarin Chinese was

taken as the example spoken documents, and the Named Entities (NEs) are taken as the key terms to construct the topic hierarchy.

2. PROPOSED APPROACH

In this paper, we propose to solve the above difficult problems by multi-modal dialogues to help the user to “navigate” across the spoken documents archives and find the desired spoken documents. In this approach, for a query given by the user, the retrieval system produces a topic hierarchy constructed from the retrieved spoken documents to be shown on the screen of the hand-held devices. The user can then expand his query easily by choosing the key terms or key phrases within the topic hierarchy by a simple click or a second spoken query to specify more clearly what he is looking for. This is achievable because the system knows the archives much better than the user. This is a multi-modal dialogue process because the system response is in form of a topic hierarchy displayed on the screen, and the user input may be given by clicks or spoken queries. With a few dialogue turns, the small set of spoken documents desired by the user can be found by a more specific query precisely expanded during the dialogue process. This is the way the system guides the user to “navigate” across the spoken archives to find the desired documents.

In the dialogue process the user works with the system and utilizes the knowledge of the system about the archives. Such a retrieval procedure can in fact be modeled very similar to a conventional dialogue system with some “hidden slots” to be filled up. For example, when the initial spoken query is very short such as “President George Bush”, the number of relevant documents will be many, but can be significantly reduced if an extra key term, “Israel”, was used to expand the query. This key term can be considered as a “filler” for the “hidden slots” of the system. Here the key term “Israel” can be a node on the topic hierarchy mentioned above constructed based on the large number of spoken documents retrieved by the short query “President George Bush”. In this way the Spoken Document Retrieval (SDR) task can be modeled and analyzed as a conventional dialogue task as given below in this paper. The only difference is that in the retrieval task here the number of “hidden slots” needed is not fixed. It is different from case to case. However, here the system interactively help the user to formulate more specific queries, just as the conventional dialogue system interactively helps the user to complete a transaction.

In the prototype system presented in this paper, the user can ask the system to show the small set of finally retrieved spoken documents on the screen with automatically generated titles and

summaries[1][2]. The user can then browse across the titles, click to listen to the summaries in speech form, and find the desired spoken documents without listening to all of them and find most of them are not what he is looking for. In this way the user can “navigate” across the spoken document archives and retrieve the desired information efficiently. This is the basic concept proposed in this paper.

The complete dialogue system for the prototype broadcast news retrieval system is shown in the block diagram in Fig. 1. The automatic generation of titles and summaries (on the left) will not be further discussed here due to space limitation. The system discussed here (in the dotted lines) includes four parts: Named Entity (NE) recognition, broadcast news retrieval, topic hierarchy construction from the retrieved broadcast news, and the discourse and dialogue manager. The Named Entity (NE) recognition produces the key terms for the broadcast news. The broadcast news retrieval system tries to make use of the extra knowledge obtained from the Named Entities (NEs) to improve the retrieval performance. The topic hierarchy construction is the core of the system here, producing a hierarchical structure with the Named Entities (NEs) in the retrieved broadcast news to help the user to further expand his query. The discourse analysis, usually playing important roles in conventional dialogue systems, becomes relatively simple here, because the query is simply continuously expanded by adding new Named Entities (NEs) chosen by the user, and the “hidden slots” are directly filled up one by one, which is the discourse. The discourse and dialogue manager therefore simply produces two types of outputs to the user after each dialogue turn entered by the user: the topic hierarchy so that the user can choose any node on the hierarchy to expand his query, or the titles, summaries or the complete spoken documents, for the user to browse across the retrieved results with limited size.

Fig. 1. The block diagram of the multi-modal dialogue system for information navigation and retrieval presented in this paper.

The complete system in Fig. 1 is complicated including quite several modules and apparently the system performance is dependent on the performance of each individual module. In order to achieve better system analysis and design, performance measure parameters were also defined for this dialogue system, and quantitative simulation approaches [3] were used to estimate these pa-

rameters in terms of various variables to be chosen in the respective modules such as Named Entity (NE) recognition, topic hierarchy construction and broadcast news retrieval. In this way not only the performance of the dialogue system can be measured, but the various variables for the different modules can be properly chosen in optimizing the system performance. Below, the major models in Fig. 1 will be first very briefly summarized individually, and the analysis and experimental results for the complete system then follow.

3. NAMED ENTITY (NE) RECOGNITION FROM BROADCAST NEWS

Here we propose to use NEs as extra feature elements for broadcast news navigation and retrieval. This is not only because such names are usually the key for the content of the news and many of them are OOVs, but because many heuristic rules and carefully designed algorithms are available to recognize named entities from spoken documents [4]. As a result, the NEs recognized from broadcast news are usually more reliable feature elements than other terms extracted from broadcast news transcription.

In our NE recognition module in Figure 1 two special approaches were developed. The first is to recognize the NEs from a text document (or the transcription of a spoken document) using global information extracted from the entire documents in addition to the local (internal and external) evidences. The basic idea is that very often an NE is difficult to identify in a single sentence. But 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, and has higher likelihood to be an NE when all those occurrences in different sentences can be considered jointly. The PAT tree data structure was found very useful in recording such global information for the entire text documents. It was shown that by incorporating the global information obtained with the PAT tree with the various approaches of NE recognition for text documents, significant improved performance can be achieved.

The second special approach used here is for spoken documents, to recover the OOV NEs using external knowledge. Each broadcast news story was first transcribed into word graphs, on which not only the NE extraction approaches for text documents including using PAT trees as mentioned above can be applied, but words with higher confidence scores can be identified. Possibly relevant text news documents published in the same time period available over the Internet were then automatically retrieved using queries constructed from those words with higher confidence measures on the transcribed word graphs. Named entity recognition was then performed on these retrieved text news stories including using the global information extraction from the PAT tree as mentioned above. In this way a set of named entity candidates can be obtained for NE matching. The basic idea for NE matching is that those word segments in the transcribed word graphs of the spoken documents with relatively lower confidence measure are likely to be recognition errors due to OOV NEs. So we can match the recognized phone lattices for these segments with the NE candidates obtained from the retrieved relevant text documents using dynamic programming. If the similarity measure is higher than a threshold, we then include the matched NE as a confirmed NE candidate for the spoken document to go through the standard NE verification/classification procedure. In order to perform the matching between two phone sequences, we defined a phone similarity matrix. The phone sequence matching is then based on the total distance

normalized with the number of phones in the sequence. In this way some OOV NEs can be recovered.

4. BROADCAST NEWS RETRIEVAL ENHANCED BY NAMED ENTITIES

The NEs recognized from broadcast news are apparently extra indexing features for SDR, not only because with the special approaches mentioned above they are more robust compared to normal terms recognized from spoken documents, but because they are very often the key terms carrying the core semantics of the broadcast news. Below we briefly summarize the procedures to enhance the SDR process using recognized NEs.

We first recognized off-line all the NEs from the entire broadcast news archives. We then define the vector representation for each news story using the conventional vector space model for information retrieval while using all the recognized NEs as the indexing terms. Let V be the vocabulary for all the NEs recognized from the spoken document archives. Each news story d is thus represented as a $|V|$ -dimensional vector v_d with components being the *tf-idf* scores for each NE, where $|\cdot|$ is the total number of elements in a set. For a spoken query q , we performed the matching between its phone lattice with the phone sequences for all the recognized NEs using exactly the same approach as mentioned in the above section, including using a phone similarity matrix. After an utterance verification process for the matched NEs, the query q is also represented by a $|V|$ -dimensional feature vector v_q , with most components having values of zero, but those for the matched NEs having values being the normalized confidence score for the specific NEs. After the user enters other NEs selected from the topic hierarchy as the extra query terms during the dialogue process, the corresponding components in the vector v_q are assigned non-zero values as well.

Two approaches were used here to enhance the broadcast news retrieval using the NEs. The first is based on the Latent Semantic Analysis (LSA) using NEs. A NE-document matrix W was constructed for all the recognized NEs with respect to all the news stories in the archives, and Singular Value Decomposition was performed on this matrix. Each NE is then represented as a vector in the latent semantic space, and the correlation between each pair of NEs is defined as the cosine measure of the angle between the vectors for them in this space. The query vector v_q mentioned previously having non-zero components only for matched NEs in the spoken query or extra query terms entered by the user can now be expanded. Those NEs with zero components in the vector v_q but with correlation with the NEs with non-zero components higher than a threshold are assigned values based on the correlation. The expanded query vector is then folded into the latent semantic space as a pseudo-document. All news stories are represented as vectors in the latent semantic space too. So the relevance between the vector for the expanded query and those for the news stories can be calculated using the cosine measure of the angle between the vectors.

The second approach is based on the Probabilistic Latent Semantic Analysis (PLSA), again based on NEs. Here every NE is taken as a term t , and a set of "latent topics" $\{T_k, k = 1, 2, \dots, l\}$ was trained from the broadcast news archives using EM algorithm by maximizing a total likelihood function. For each spoken query q , the retrieval is based on the probability of observing the query q given all news stories d , $P(q|d)$, which is in turn obtained by the probability of observing the NEs or terms t_j either matched in the

spoken query q or entered by the user, $P(t_j|d)$. Each probability $P(t_j|d)$ is then further expanded as

$$P(t_j|d) = \sum_k P(t_j|T_k)P(T_k|d), \quad (1)$$

where $P(t_j|T_k)$ and $P(T_k|d)$ are respectively the probabilities of observing the NE or term t_j in the latent topic T_k and of observing the topic T_k in the news stories d . So the relevant news stories d for a query q can be found by the probabilities $P(q|d)$.

The above two approaches were integrated with a baseline broadcast news retrieval system based on Mandarin syllable-level indexing terms with vector space model [5]. For each news story the baseline system and the LSA and PLSA approaches respectively produce a score for the given query q . The weighted sum for these scores are then used to select the retrieved news stories.

5. TOPIC HIERARCHY CONSTRUCTION FROM THE BROADCAST NEWS

Although the hierarchical organization of retrieved documents in text form to help the user to browse through the relevant documents has been well studied [6] [7], the extension to spoken documents is not straightforward because of the many recognition errors in the transcriptions. Here we propose to use the relatively reliable key elements in broadcast news, the NEs recognized with the special approaches discussed in Section 4, to construct a topic hierarchy by properly clustering the NEs based on the statistics they appear in the retrieved broadcast news. These NEs not only appear on the topic hierarchy as the names of the nodes to guide the user to choose the directions to proceed further, but serve as the suggested extra query terms for the user to expand his query. There are more important reasons to choose NEs rather than other terms or phrases to play this role, i.e., NEs provide high coverage for the broadcast news (i.e., they cover almost all news stories) and high discriminative ability (i.e., they easily separate news stories addressing different topics) and thus are very useful augmented query terms. The approach we used here for topic hierarchy construction is the Hierarchical Agglomerative Clustering and Partitioning algorithm (HAC+P) recently proposed for text documents [8], but here performed on NEs recognized from broadcast news. This algorithm is briefly summarized below.

With the vector representation v_d for each spoken document d in terms of all the recognized NEs as mentioned previously for all the retrieved news stories, for each involved NE or key term t appearing in these news stories, we built a feature vector v_t for it by averaging the vector representations for all news stories including this NE t , normalized by the term frequencies of t in these news stories and the lengths of the documents. The Hierarchical Agglomerative Clustering and Partitioning (HAC+P) algorithm was then performed on-line in real time using these feature vectors v_t to cluster all the involved NEs into a balanced hierarchy. This algorithm consists of two phases: an HAC-based clustering to construct a binary-tree hierarchy and a partitioning (P) algorithm to transform the binary-tree hierarchy to a balanced and comprehensive m-ary hierarchy, where m can be different integers at different splitting nodes. The principles of this algorithm is briefly summarized below.

The HAC algorithm is based on the similarity between two

clusters C_i and C_j of NEs, $S(C_i, C_j)$,

$$S(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{v_t \in C_i} \sum_{v_s \in C_j} c(v_t, v_s), \quad (2)$$

where $c(v_t, v_s)$ is the cosine measure of the angle between the vectors v_t and v_s for NEs t and s . The HAC algorithm is performed bottom-up. Assume there are n NEs in the retrieved news stories, the initial clusters, $C_1, C_2, \dots, C_n$, are exactly the n NEs. Let C_{n+i} be the new cluster created at the i -th step by merging two clusters. The output binary-tree hierarchy can be expressed as a list, $C_1, \dots, C_n, C_{n+1}, \dots, C_{2n-1}$. An example is in Figure 2.(a), where $n = 5$, and $C_6, \dots, C_9$ are created by HAC.

Fig. 2. An illustrative example for the HAC+P algorithm.

The second phase of partitioning is top-down. The binary-tree is partitioned into several sub-hierarchies first, and then this procedure is applied recursively to each sub-hierarchy. The point is that in each partitioning procedure the best level at which the binary-tree hierarchy should be cut in order to create the best set of sub-hierarchies has to be determined based on the balance of two parameters: the cluster set quality and the number preference score, as will be explained below. As shown in Figure 2(a), partitioning can be performed on 4 possible levels by a cut through the binary tree, $l = 1, 2, 3$, and 4. If a cut was performed at the level $l = 2$, the result will be three sub-hierarchies, C_5, C_6 , and C_7 as shown in Figure 2.(b).

Cluster Set Quality is based on the concept that each cluster should be as cohesive and isolated from the other clusters as possible. Therefore, the cluster set quality of a cluster set H is defined as

$$Q(H) = \frac{1}{|H|} \sum_{C_i \in H} \frac{S(C_i, \overline{C_i})}{S(C_i, C_i)}, \quad (3)$$

where $\overline{C_i} = \bigcup_{k \neq i} C_k$ is the complement of C_i , and here C_i are the clusters for the sub-hierarchies produced after the partitioning cut. Apparently the smaller the value of $Q(H)$ the better.

Number Preference Score. In principle, when a node in the hierarchy is split into m sub-hierarchies, the number m should be neither too small nor too large, considering both the efficiency and the convenience of the user. In this algorithm, a preferred number m_0 is defined as a design parameter to be assigned when constructing the hierarchy. It can be either a constant value or a variable. A gamma distribution function is then used as an approximated score to measure the degree of the user's preference regarding the number of sub-hierarchies,

$$f(m) = \frac{1}{\alpha! \beta^\alpha} m^{\alpha-1} e^{-m/\beta}, \quad (4)$$

where m is the number of sub-hierarchies at the splitting node, α is a positive integer, and the constraint $(\alpha - 1)\beta = m_0$ gives $f(m_0) \geq f(m)$, or $f(m)$ is maximized when m is equal to the assigned parameter m_0 . In the experiments below we set m_0 to be the largest integers smaller than the square root of the number of leaf nodes for the partitioning being considered.

With the two parameters $Q(H)$ and $f(m)$ defined in equations 3 and 4, the best level of partitioning cut is then chosen as the one which minimizes the following parameter,

$$\eta = Q(H)/f(m), \quad (5)$$

with which $Q(H)$ is minimized and $f(m)$ is maximized simultaneously. After the entire hierarchy structure is constructed, we need to name each cluster. For example, for the cluster C_7 in Figure 2.(b), the NE with the highest $tf \cdot idf$ score in the news stories in its children, C_1 and C_2 , are chosen as the name. An example of such a topic hierarchy in terms of NEs is shown in Figure 3.

Fig. 3. The topic hierarchy constructed for the retrieved news stories obtained with the query: "US and Middle East".

6. DIALOGUE MANAGER

The discourse and dialogue manager here are relatively simple. Once the user enters his query in speech or text form, the system constructs a topic hierarchy in terms of NEs for the retrieved news documents shown on the screen. Each node in the hierarchy represents possible extra query term. The user then chooses one or more nodes (or NEs) to expand his query by clicks or speech input. Once the query is changed, the retrieved news stories are different and a different topic hierarchy is shown. This process repeats such that the user continuously refines his query and makes the query more and more specific, which is the discourse. The user can always choose at any time to take a look at the automatically generated titles or listen to the automatically generated summaries in speech form or the complete news stories for all news stories within a given cluster.

7. DIALOGUE SYSTEM PERFORMANCE ANALYSIS BY QUANTITATIVE SIMULATION

The dialogue system performance analysis is based on a system operation and user scenario model for the dialogues. Assume a user has a certain desired subject or event in mind and there are K relevant news stories for it. K is a random variable with a certain distribution. The first query given by the user, however, is assumed to be very short. The number of retrieved news stories for it is L . L is another random variable with another distribution. Very often L

is much larger than K . The small screen on the hand-held devices can show only up to the top N news titles for the user to browse even if much more news stories are retrieved and available. The user's response pattern for the dialogue is modeled as shown in Fig. 4. Assume M out of the K news stories desired by the user as mentioned above are among the top N shown on the screen (M may be zero). The recall rate is then M/K and the precision rate is M/N . Assume the user is satisfied if $M/K > r_0$, where r_0 is a pre-defined threshold, and the transaction is success and completed in that case. If not, a topic hierarchy is constructed and shown on the screen for the user to select one or more nodes to enter. The process then repeats recursively.

Fig. 4. The user's response pattern for system performance analysis.

There are a few key parameters for the topic hierarchies which determines the system performance in the above model. They are defined below. The correctness (C) is the measure if all the NEs in the topic hierarchy is correctly located at the right node position. It can be evaluated by counting the number of NEs in the topic hierarchy which have to be moved manually to the right node position to produce a completely correct topic hierarchy, and obtaining the ratio of this number to the total number of nodes in the hierarchy. C is then one minus this ratio. The coverage ratio (P) is the percentage of the retrieved news stories which can be retrieved using the NEs in the topic hierarchy. For example, if there are 100 retrieved news stories given a query and there are 15 out of them simply can't be retrieved even using all the NEs in the topic hierarchy. The coverage ratio P is then 85%. Discriminative ratio (d) of an NE in the topic hierarchy with respect to its parent node is then how efficient it is in reducing the size of the relevant news stories when the NE is selected as an additional query term. For example, for a certain query given by the user there are 100 relevant news stories. After the user selects a specific NE, the relevant news stories are reduced to 40. The discriminative ratio d for this NE is then 40%. Given the above system operation and user scenario model plus the parameters, the quantitative simulation can be performed by entering huge number of simulated queries each with a random number of relevant news stories randomly sorted and so on. Average number of dialogue turns needed for success transactions can then be evaluated as a good performance measure. However, there should be a fourth parameter for the topic hierarchies, the size (J), or average number of NEs used in the topic hierarchies. In practice the larger the size J is, the higher the coverage ratio P and the smaller the discriminative ratio d will be, and thus the smaller the average number of turns will be. But a topic hierarchy with large size is not only difficult to show on a small screen, but difficult for the user to find an appropriate node to enter. So the size J should be kept within a reasonable range.

8. INITIAL PROTOTYPE SYSTEM

An initial prototype system has been successfully developed at National Taiwan University. The broadcast news are taken as the example spoken/multi-media documents. The broadcast news archives to be navigated across and retrieved includes roughly 130 hours of about 7,000 news stories, all in Mandarin Chinese. They were all recorded from radio/TV stations in Taipei from Feb 2002 to May 2003. The character and syllable error rates of 14.29% and 8.91% respectively were achieved in the transcriptions. All the modules shown in Fig.1 and presented in Sections 3, 4, 5 and 6 were successfully implemented.

9. PRELIMINARY EXPERIMENTS AND SIMULATION RESULTS

9.1. Performance of individual modules

For NE recognition 200 Chinese broadcast news recorded from radio stations at Taipei in Sept. 2002 were used as the test corpus. NEs manually identified were taken as references. The external knowledge sources to be retrieved for relevant documents for OOV recovery are the Chinese text news available at "Yahoo! Kimo News Portal" for the whole month of Sept. 2002. The results are listed in Table 1. The results in part (A) are obtained with

Experiments	NE	Recall	Precision	F1 score	Overall F1
(A) Baseline	PER	71	86	77.8	77.6
	LOC	86	91	88.4	
	ORG	64	95	76.5	
(B) Special Approaches	PER	76	87	81.1	80.9
	LOC	87	90	88.5	
	ORG	68	95	79.3	

Table 1. Experimental results for NE recognition: person names (PER), location names (LOC), and organization names (ORG).

a baseline NER system directly performed on the transcriptions of the 200 news stories. The results using the special approaches presented in Section 3 are listed in part (B). Significant improvements in almost all cases can be observed.

For broadcast news retrieval enhanced by NEs, a total of 1708 distinct NEs recognized from a subset of the broadcast news archives were used in the LSA or PLSA training. A total of 350 latent topics were used in either LSA or PLSA. The results are listed in Table 2. Apparently, incorporating NEs as extra indexing features are helpful, and the improvements achieved by PLSA (row(c)) are more significant than those by LSA (row(b)).

Experiments	Precision	Recall	F1 score
(a) baseline	38.986	50.542	44.019
(b) baseline+LSA	47.027	59.704	52.613
(c) baseline+PLSA	48.649	60.437	54.723

Table 2. Experimental results for broadcast news retrieval enhanced by NEs.

For the topic hierarchy construction, we used 20 queries to generate 20 sets of retrieved news stories and constructed 20 topic hierarchies. The average values of correctness (C), coverage ratio

(P), and discriminative ratio (d) as defined in Section 7 were obtained with some manual efforts and listed in (1)-(4) in Table 3, where row (1) is for all NEs and rows (2)-(4) are for each individual class of NEs. As a comparison, we also extract the same number of 1708 terms or phrases with highest $tf \cdot idf$ scores (but not necessarily NEs) and evaluate the corresponding coverage ratio (P) and discriminative ratio (d), as listed in row (5). As can be found, the NEs offered much more lower discriminative ratio (d), although with slightly degraded coverage (P) as compared to the terms and phrases selected simply by $tf \cdot idf$ scores. Also, for each individual class of NEs (person/organization/location names) the discriminative ratios d are roughly equally low, and the coverage ratios P are on the order of 70%.

	correctness (C)	coverage ratio (P)	discriminative ratio (d)
(1) All NEs	0.91	0.97	0.15
(2) PER	—	0.66	0.12
(3) ORG	—	0.71	0.13
(4) LOC	—	0.67	0.17
(5) terms or phrases by $tf \cdot idf$ scores	N/A	1	0.35

Table 3. Experimental results for topics hierarchy construction: person names(PER), organization names(ORG), location names(LOC).

9.2. Simulation Analysis of the whole system

With the parameters correctness(C), coverage ratio(P) and discriminative ratio(d) obtained in Table 3 and the system operation and user scenario model presented in Section 7, we simulated the dialogue system and evaluate it in terms of number of turns needed for transaction success. Figure 5(a)(b) are two example results, in which the two horizontal scales are L (number of retrieved news stories with the initial query) and K (number of news stories desired by the user). Here L is taken as a random variable with uniform distribution between [200,500], and K is assumed to be another independent random variable with uniform distribution between [1, $L/6$]. The vertical scale is the number of turns needed for the transaction success. In both cases N (number of news titles shown on the screen) was set to be 30 and r_0 (the recall rate threshold for user being satisfied) was set to 0.5. The dots represent simulated cases. Figure 5(a) is for $d = 0.40$ (close to the case for terms or phrases in row (5) of Table 3), and Figure 5(b) for $d = 0.15$ (the case for NEs in row (1) of Table 3). It can be found that with $d = 0.4$ all transactions were successful in 5 turns, most in 4 turns, while others in 3, 2, or 1 turns. With $d = 0.15$, however, all transactions were completed within 3 turns. Apparently the lower distribution ratio(d) of NEs have successfully helped the user to navigate across the news archives and find the desired news stories very efficiently. When we averaged all the dots in these figures, we obtained the average number of turns which were plotted as a function of the discriminative ratio (d) for two different recall rate threshold r_0 . Such results will be very helpful in system design and performance analysis. As can be seen, for NEs as proposed here with $d = 0.15$, the difference between $r_0 = 0.50$ and 0.75 is actually negligible.

Fig. 5. Number of dialogue turns simulated for random L and K , $N = 30$, $r_0 = 0.5$, and (a) $d=0.4$ (b) $d=0.15$.

Fig. 6. Average number of turns needed for a successful transaction for different discriminative ratio d and recall threshold r_0

10. CONCLUSION

In this paper we presented a concept of using dialogues to guide the user to navigate across spoken document archives using a topic hierarchy. A prototype system has been successfully developed, and a simulation approach was also proposed for performance analysis.

11. REFERENCES

- [1] S.-C. Chen and L.-S. Lee, "Automatic title generation for chinese spoken documents using an adaptive k nearest-neighbor approach," in *Proc. of Eurospeech*, 2003, pp. 2813–2816.
- [2] 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 *EUROSPEECH*, 2003, pp. 49–52.
- [3] B.-S. Lin and L.-S. Lee, "Computer-aided analysis and design for spoken dialogue systems based on quantitative simulations," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 534–548, 2001.
- [4] Y.-Y. Liu, "An initial study on named entity extraction from chinese text/spoken documents and its potential applications," M.S. thesis, National Taiwan University, 2003.
- [5] B.-L. Chen, H.-M. Wang, and L.-S. Lee, "Retrieval of broadcast news speech in mandarin chinese collected in taiwan using syllable-level statistical characteristics," in *IEEE Int. Conf. Acoustics, Speech, Signal processing*, 2000, vol. 3, pp. 1771–1774.
- [6] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J.-W. Ma, "Learning to cluster web search results," in *ACM SIGIR*, 2004, pp. 210–217.
- [7] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in *WWW*, 2004, pp. 658–665.
- [8] S.-L. Chuang and L.-F. Chien, "A practical web-based approach to generating topic hierarchy for text segments," in *ACM SIGIR*, 2004, pp. 127–136.