Internet Chinese Information Retrieval Using Unconstrained Mandarin Speech Queries Based on A Client-Server Architecture and A PAT-tree-based Language Model

Lee-Feng Chien[1],Sung-Chien Lin[2], Jenn-Chau Hong[2], Ming-Chiuan Chen[1]
Hsin-Min Wang[1], Jia-Lin Shen[3], Keh-Jiann Chen[1], Lin-Shan Lee[1,2,3]

[1]Institute of information Science, Academia Sinica
[2]Dept. of Computer Science and Information Engineering, National Taiwan University
[3]Dept. of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
lfchien@iis.sinica.edu.tw

ABSTRACT

In order to pursue high performance of Chinese information access on the Internet, this paper presents an attractive approach with a successful integration of efficient speech recognition and information retrieval techniques. A working system based on the proposed approach for speech retrieval of real-time Chinese netnews services has been implemented and tested. Very exciting performance has been achieved.

INTRODUCTION

With the rapid growth of the electronic resources published and distributed over the Internet, the fast increasing demand for efficient, high-performance networked information retrieval (IR) systems providing a convenient and user-friendly interface is obvious [1,2]. Many efficient Internet search tools have been developed to allow users formulating a request subject with unconstrained quasi-natural language queries, while it is always highly desired that such systems are capable of accepting queries with unconstrained speech [3-5]. In particular, information retrieval systems with speech recognition capabilities are specially needed in the Chinese community, because of the difficulties of entering Chinese characters into computers since Chinese language is not alphabetic.

This paper presents an attractive approach for Internet Chinese information retrieval using unconstrained Mandarin speech queries based on the monosyllabic structure of Chinese language. There are more than 10,000 commonly used Chinese characters, but each character is monosyllabic and is almost always a morpheme with its own meaning, and the total number of phonologically allowed Mandarin syllables is only 1345. The combination of these monosyllabic characters or 1345 syllables provides almost unlimited number of Chinese words. In such monosyllabic structure of Chinese language, Mandarin syllables become very special linguistic units carrying plurality of linguistic information. The approach proposed in this paper includes a syllable-based client-server architecture for efficient information retrieval, a character-based language model using PAT tree data structure, and a set of reliable character/syllable-level statistical feature parameters for efficient information retrieval with speech queries. The reliable character/syllable-level parameters make it possible to move the language model to the server side to simplify the client end requirements. The PAT tree data structure provides very powerful character-based language model at the server easily adapted based on the dynamic network resources. The client-server architecture can thus properly utilize the monosyllabic structure of Chinese language and the plurality of linguistic information carried by the Mandarin syllables. A successful working system for speech retrieval of real-time Chinese news services automatically obtained from the Internet news groups has been implemented using the proposed approach. Very exciting performance has been achieved and the development of conversational information retrieval system which allows spoken dialogue [6] can be also extended.

OVERVIEW OF THE PROPOSED APPROACH

The basic client-server architecture of the proposed approach is shown in Fig. 1. The server part includes a resource discovery subsystem, an information retrieval subsystem, and a linguistic decoding subsystem, while the client end includes a user interface and an acoustic processing subsystem. The resource discovery subsystem automatically extracts pieces of relevant information (records) and uses them to construct the network resource databases. Signatures (statistical indices) for each of the records in the network resource databases are also generated to be stored in the signature file [7]. The PAT-tree-based language model is also constructed based on a PAT tree data structure [8]. The linguistic decoding subsystem produces the quasi-natural language queries based on the syllable lattice obtained from the client end and the language model parameters obtained from the PAT tree data structure. The information retrieval subsystem then receives these quasi-natural language queries and retrieves the desired records from the network resource databases to be sent to the user interface at the client end by evaluating the statistical relevance between the queries and the record signatures in the signature file. The specially designed signature for each record includes both character-based and syllable-level statistical characteristics of the record content specially considering the monosyllabic structure of Chinese language [2]. These signatures provide the full-text indexing of the network resource databases. Also as shown in Fig. 1, the acoustic processing and linguistic decoding subsystems for very-large-vocabulary Mandarin speech recognition [9] are in fact separated, one at the client end and the other at the server side. Without the space overhead necessary for linguistic decoding at the client end, it is easier for the acoustic processing subsystem to be combined with navigation tools such as Netscape to allow many users to enter their speech queries simultaneously. On the other hand, the linguistic decoding subsystem at the server side can have sufficient space to store a large number of statistical parameters for a powerful language model, and it is easier for this language model to be adapted according to the dynamic network resources. The key element here, the character-based N-gram language model using the PAT tree data structure, will be described in the following.

PAT-TREE-BASED LANGUAGE MODEL

The PAT-tree-based language model is developed for more reliable performance in speech recognition. The design of the language model is pursued to model all of the searching words or patterns which might be used in formulating query subjects in a specific database domain. In order to increase the accuracy of robust speech recognition, especially on the recognition of proper nouns which are often out-of-vocabulary words (OOV‘s) in speech recognition but keywords in terms of information retrieval, all of the Markovian parameters in the language model are obtained directly from the content of the target database instead of from general corpus. In addition, to allow the composed characters of OOV‘s that can be correctly recognized, characters instead of words are taken as basic units of language modeling. The language model is then formed as a kind of statistical character N-gram model which is used to estimate the probability of any string of Chinese characters instead of words in the searching database domain.

PAT tree is an efficient data structure successfully used in the area of information retrieval. It was developed in 1987 by Gonnet [8] from Morrison‘s PATRICIA algorithm (Practical Algorithm to Retrieve Information Coded in Alphanumeric) [10] for indexing a continuous data stream and locating every possible position of a prefix in the stream. The PAT tree is conceptually equivalent to compressed digital search tree but smaller. The superior features of the PAT tree data structure is most resulted from the use of so-called semi-infinite strings [11] in storing the substream values in the nodes of the PAT tree. Using this data structure for indexing full-text content of network resource databases, all possible character strings including their frequency counts in the databases can be retrieved and updated in a very efficient way, but not every character string with arbitrary length need to be stored. In this way, the PAT tree can be extended to serve as an efficient data structure for the representation of N-gram language models.

Statistical N-gram language models are often used in speech recognition systems to estimate the probability of any string of words or characters. For reducing the complexity of model representation, bigram or trigram models are frequently used as an approximation. Instead of keeping all neighboring relations for 2 or 3 words or characters as in the conventional bi- or tri-gram word or character based language models, the PAT tree actually provides effective indices to all possible segments of characters with an arbitrary length N, where N can be significantly larger than 2 or 3, together with the frequency counts for these segments in the network resource databases. Each character segment is represented in the PAT tree as a bit string using some coding scheme (e.g. ASCII for English characters) for its component characters, and these bit strings are then used to construct the PAT tree as the indices for all character segments. In searching for a character segment and its frequency counts, the desired character segment is also represented as a bit string to be matched with this PAT tree. The time complexity of finding a character segment and its frequency counts from the PAT tree is proportional to the length of its corresponding bit string. In this way, the PAT tree effectively provides statistical parameters for all possible character segments with an arbitrary length N to appear in the network resource databases, to be used as approximated values for the character-based N-gram language model parameters with the network resource databases being the training text corpus. Since N can be significantly larger than 2 or 3, such a language model is much more powerful than character-based bi- or tri-grams, and in fact the PAT tree keeps all character segments appearing in the network resource databases even without any information loss. Because the input queries always include character segments also appearing in the network resource databases, such a PAT-tree-based language model is most efficient for linguistic decoding for the queries. Although this PAT tree requires relatively large memory space, this is fine because it is stored at the server side. Also, because it is at the server side, it is very easy to change the content of the PAT tree for language model adaptation to take care of the dynamic nature of the network resource databases.

THE LINGUISTIC DECODING SUBSYSTEM

With the PAT-tree-based language model, the linguistic decoding system is also extended. When the linguistic decoding subsystem receives the syllable lattice produced by the acoustic processing subsystem at the client end, this syllable lattice is first matched with the PAT tree to generate a lattice of character strings, in which each node is a character string, with acoustic recognition score obtained from the acoustic processing subsystem and frequency counts in the network resource database obtained from the PAT tree. An example of such lattice is shown in Fig. 2. A Viterbi search is then performed over this lattice of character strings. The relevant character strings for the paths on the lattice with maximal scores can thus be found and forwarded to the information retrieval subsystem. In some cases, if a syllable within a word in the query is not recognized correctly, it will be difficult for the word including this syllable to be recognized by conventional word-based linguistic decoding techniques. But with this character-based approach the character segments within the word not including the incorrectly recognized syllables can still be found, so the loss in information due to acoustic recognition errors can be minimized, and more robust speech recognition and more precise retrieval results can be obtained.

While the PAT-tree-based language model mentioned above can be used with any language, it is specially useful for Chinese language. As mentioned previously, all Chinese characters are almost morphemes with its own meaning, new words can be easily generated everyday by putting together several characters, and the proper nouns (very often the keywords in information retrieval) can be arbitrarily generated but very difficult to be included in the lexicon. Furthermore, because there do not exist blanks between words as marks for word boundaries in Chinese sentences, a Chinese sentence is a sequence of words but also a sequence of characters. As a result, the words in Chinese are not well defined, and when a sentence is given the identification of the words in the sentence are very often not unique. So the word-based N-gram language models are not only difficult to train, but very often not sufficient in linguistic decoding for Chinese language. This is why character-based N-gram language models are highly desired for Chinese language, if models with large enough N are feasible to construct. Note that in this case a character segment can be a word, a concatenation of several words, a part of a word, or a concatenation of words and parts of words, etc. So that PAT-tree-based language model proposed here is very powerful in Chinese information retrieval. In our experiments, in comparison with the performance of using general language model, the average recognition accuracy of the proposed system has an improvement with a rate of 3~8%. In special, the recognition of searching terms is very reliable. This obvious improvement encourages us toward the development of a Mandarin speech interface for networked information retrieval.

EXPERIMENTAL SYSTEM AND RESULTS

The proposed approach has been successfully implemented into a working system which provides unconstrained speech retrieval for real-time Chinese news services obtained from the Internet news groups. The continuous Mandarin speech recognition techniques previously developed [9] is used in the acoustic processing subsystem. The client end including the user interface and the acoustic processing subsystem is implemented on Pentium PCs under WIN 95, and the server implemented on a SPARC 20 workstation. The network resource database contains more than 100,000 real-time news items. In the preliminary experiments, 200 speech queries produced by 10 speakers were tested. The character recognition rates for these queries is roughly 90%, and very high accuracy can be obtained even for the proper nouns never included in the lexicon in conventional speech recognition systems. The precision rate for top 10 retrieved news items is about 82% on average, which is almost the same as the precision rate for typed text queries, apparently due to the excellent functions of the PAT-tree-based language model.

CONCLUDING REMAKRS

In this paper an attractive approach with a successful integration of efficient speech recognition and information retrieval techniques has been proposed. A working system based on the proposed approach for speech retrieval of real-time Chinese netnews services has been implemented and tested. Very exciting performance has been achieved.

REFERENCE

[1] Lycos Home Page (http://www.lycos.com)

[2] L-F. Chien, et. al. “Natural Language Information Retrieval with Speech Recognition Techniques for Chinese Network Resources Discovery,” International Workshop on Information Retrieval with Oriental Language, Korea, June, 1996.

[3] S-C. Lin, L-F. Chien, K-J. Chien, and L-S. Lee, “An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model,” ICASSP96, pp. 287-290, Atlanta, USA, May, 1996.

[4] K. Sparck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, “Experiments in Spoken Document Information Retrieval”, Information Processing and Management, Vol. 32, No. 4, pp. 399-417, 1996.

[5] Julian Kupiec, Don Kimber and Vijay Balasubramanian, “Speech-Based Retrieval Using Semantic Co-Occurrence Filtering”, Proc. of the Human Knowledge Technology Workshop, pp. 373-377, 1994.

[6] Mona Singh and Jim Barnett, Designing Spoken Dialogue Systems for Text Retrieval, Proceedings on the 1996 International Symposium on Spoken Dialogue, Philadelphia, USA, 73-76.

[7] L-F Chien, “Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts”, ACM SIGIR‘95.

[8] G.H. Gonnet, R.A. Baeza-Yates, and T. Sinder, “New Indices for Text: PAT Trees and PAT Arrays,” Information Retrieval: Data Structure and Algorithms, edited by W.B. Frakes and R. A. Baeza-Yates, pp. 66-82.

[9] H-M. Wang, L-S. Lee, et. al., “Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary,” ICASSP95, pp. 61-64, Detroit, USA, May, 1995.

[10]Morrison, D., "PATRICIA :Practical Algorithm to Retrieve Information Coded in Alphanumeric", JACM, pp. 514-534, 1968.

[11]Manber, U. and R. Baeza-Yates, "An Algorithm for String Matching with a Sequence of Don't Cares", Information Processing Letters, 37, pp.133-136, 1991.

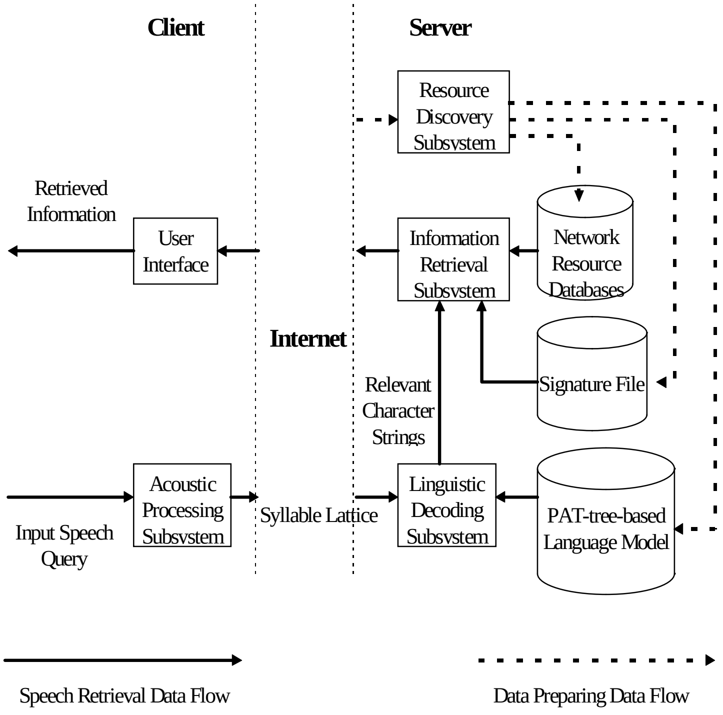
Client
Server
Internet
Resource Discovery Subsystem
Retrieved Information
User Interface
Information Retrieval Subsystem
Network Resource Databases
Signature File
Relevant Character Strings
Acoustic Processing Subsystem
Syllable Lattice
Linguistic Decoding Subsystem
PAT-tree-based Language Model
Input Speech Query
Speech Retrieval Data Flow
Data Preparing Data Flow

Fig. 1 the client-server architecture of the proposed approach for speech retrieval of Internet Chinese information.

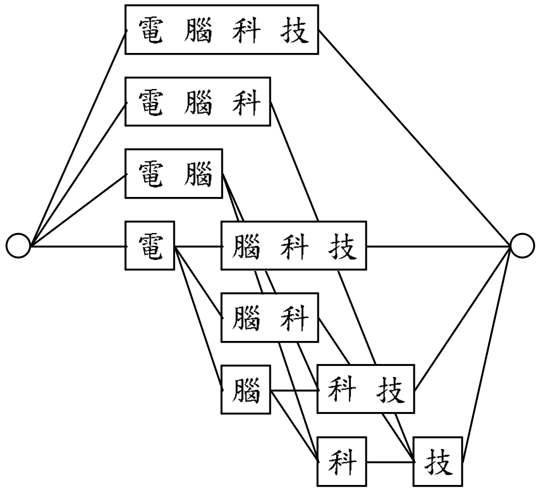

Fig. 2 an example lattice of the character segments, where each block indicates a character segment.