LIVE LEXICONS AND DYNAMIC CORPORA ADAPTED TO THE NETWORK RESOURCES FOR CHINESE SPOKEN LANGUAGE PROCESSING APPLICATIONS IN AN INTERNET ERA

Lin-Shan Lee^{1,2} and Lee-Feng Chien¹ ¹Institute of Information Science, Academia Sinica ²Dept. of Electrical Engineering, National Taiwan University Taipei, Taiwan, Republic of China { lsl, lfchien } @iis.sinica.edu.tw

ABSTRACT

In the future network era, huge volume of information on all subject domains will be readily available via the network. Also, all the network information are dynamic, ever-changing and exploding. Furthermore, many of the spoken language processing applications will have to do with the content of the network information, which is dynamic. This means dynamic lexicons, language models and so on will be required. In order to cope with such a new network environment, automatic approaches for the collection, classification, indexing, organization and utilization of the linguistic data obtainable from the networks for language processing applications will be very important. On the one hand, high performance spoken language technology can hopefully be developed based on such dynamic linguistic data on the network. On the other hand, it is also necessary that such spoken language technology can be intelligently adapted to the content of the dynamic and the ever-changing network information. Some basic concept for live lexicons and dynamic corpora adapted to the network resources has been developed for Chinese spoken language processing applications and briefly summarized here in this paper. Although the major considerations here are for Chinese language, the concept may equally apply to other languages as well.

I. INTRODUCTION

It is generally realized today that speech recognition technology has matured to a point where the achievable scope of tasks, accuracy, speed, and the cost of such systems are almost simultaneously crossing the threshold for practically usable systems. Many applications have already been developed and used, and many others are being contemplated currently [1,2]. On the other hand, the improved computer technology and new technical environment such as the Internet today has very naturally extended the speech recognition technology to a much wider scope, i.e., the spoken language processing including such areas as linguistic knowledge processing and integration, spontaneous speech processing, telephony applications, speech understanding, dialogues, and information retrieval.

For the Mandarin Chinese, however, the input of Chinese characters into computers is still a very difficult and unsolved problem. Speech recognition or spoken language processing is believed to be the perfect solution to this problem, but with many very challenging technical issues yet to be solved. The 1.2 billion Chinese people would spend a vast amount of money purchasing computers, peripherals, networks, software packages, and other relevant products to computerize their communities - if their language could be conveniently entered into computers just as western alphabetic languages are. The demand is there, the market will someday be huge, and the potential impact on related areas is almost unlimited.

On the other hand, the development of the Internet and WWW has created a completely new environment for spoken language processing technologies. Speech interfaces become highly desired in many applications, but the dynamic nature of the Internet information also requires the spoken language processing applications to be intelligent and adaptive enough to handle the wide variety of ever-changing information activities over the Internet. Live lexicons, dynamic language models developed from the dynamic corpora automatically collected from the Internet will be very important in such an environment. These will be the focus of this paper. Although the major result here are for Chinese language, the concepts may equally apply to other languages as well.

II. THE NEW ENVIRONMENT FOR SPOKEN LANGUAGE PROCESSING IN A NETWORK ERA

As the network technology will be connecting everywhere globally in the future network era, in which huge volume of information can be disseminated across the globe in microseconds. Almost all information-related activities and services will be performed on the network. Digital libraries, information retrieval, electronic commerce and distant learning are good examples of them. Such a network environment actually creates a completely new space with many difficult new challenges for spoken language processing technology developments. In many cases speech interfaces for users to access the network services will be highly attractive, and speech recognition may play very special roles. On the other hand, the unlimited volume of linguistic resources available via the network, in both text and audio form, also provide a variety of new opportunities for spoken language processing technology developments. The multi-lingual and multi-media nature of the network information actually requires much more active interaction between the speech recognition processes and many other information processing technologies such as information retrieval, natural language processing, databases, multimedia, and networking technologies. The dynamic and ever-changing nature of the networked information also requires the speech recognition approaches to be intelligent and adaptive, so as to be able to handle the "live" information over the network. For example, the Internet-based resources are not only huge, unlimited, produced globally and disseminated world-wide, on almost all possible subject domains, but are dynamic, updating, ever-growing, exploding, and ever-changing. Such network environment actually makes spoken language processing a completely new research area significantly different from how it has been before.

The vision of future spoken language processing technology under such a network era is shown in Fig.1. The future integrated networks with all informationrelated activities and services available via the networks such as the digital libraries and intelligent offices are shown on the left part of the figure, while the users shown on the right of the figure try to use a variety of terminal equipments such as personal computers, TV sets, telephone sets and handsets to access the network services, through all possible spoken language processing application tasks as shown in the middle. In some cases the relatively simple voice command and control will be enough, in some cases a dictation system other cases the voice conversational interfaces will be very helpful, because the users can receive the services only after some voice dialogues with the network. Other applications may include voice retrieval of network resources, voicedirected network information processing, etc. In order to develop the needed application tasks, a set of basic technology will be required, i.e., the integrated language and speech technology, in which the various speech and language processing technologies will be properly developed in an integrated way. Such a vision will be the base for the discussions in the rest of this paper.

III. LINGUISTIC KNOWLEDGE FOR SPOKEN LANGUAGE PROCESSING IN FUTURE NETWORK ERA

In the future network era, huge volume of information on all subject domains will be readily available via the network. Also, all the network information are dynamic, ever-changing and exploding. Furthermore, many of the spoken language processing applications will have to do with the content of the network information, which is dynamic. In order to cope with such a new network environment, automatic approaches for the collection, classification, indexing, organization and utilization of the linguistic data obtainable from the networks will be very important. On the one hand, high performance spoken language processing technology can hopefully be developed based on such huge volume of linguistic data on the network. On the other hand, it is also necessary that such spoken language processing technology can be intelligently adapted to the content of the dynamic and ever-changing network information. Considering all the above, some basic concepts have been developed and briefly summarized here.

Fig.2 depicts a simple comparison between the linguistic knowledge (lexicon, language model, plus others) used in existing spoken language processing technology and that for possible future spoken language processing technology in the network era under the proposed concept. As can be observed in Fig2(a), in the existing processes, the linguistic knowledge required, such as the lexicon, the language model, and other possible knowledge such as the morphological rules and the grammar rules, are primarily encoded from some static corpora, primarily by human efforts and human knowledge, although some of them may be obtained automatically with statistical approaches. On the other hand, in order to handle the "live" language in the network era, under the proposed framework as shown in Fig 2(b), unlimited dynamic corpora are collected from



Fig 1. The Vision of Spoken Language Processing in the Future Network Era

automatically, such as the multiple dynamic lexicons and multiple dynamic language models for different subject domains. The dynamic collections of linguistic knowledge on all possible subject domains provide a much more powerful base to handle the "live" language in the network era for future spoken language processing technology. Although there may always exist some linguistic knowledge which can't be easily generated completely automatically and human efforts are still needed, automatic generation of linguistic knowledge will always be an attractive goal to pursue in the future.



Fig. 2. Comparison between the Linguistic Knowledge Used in Existing and Future Spoken Language Processing Technology Environments

IV. FEATURES OF CHINESE LANGUAGE AND PAT-TREE BASED DYNAMIC LANGUAGE MODELING

Chinese language is quite different from many western languages in various structural features[3]. It is not alphabetic. Large number of Chinese characters are ideographic symbols. Almost each Chinese character is a morpheme with its own meaning. A "word" is composed of one to several characters, with meaning sometimes related somehow, sometimes completely different from the meaning of the component characters. Some of the words are "compositional", i.e., the meaning of the word have to do with the meaning of the component characters, such as the characters "[[(big)]" and "[](learning)" forming a word "III(university)", but some other words are "ideographic", such actually as the characters "[(harmony)" and "[(prefer)" forming а word "DD(monk)", i.e., the meaning of the word is completely different from the meaning of the component characters. A nice feature is that all the characters are pronounced as monosyllables, and the total number of phonologically allowed syllables is limited, 1345 for Mandarin.

The wording structure in Chinese language are extremely flexible. For example, a long word can be arbitrarily abbreviated, such as "DDDD(Taiwan University)" being abbreviated as "00", and new words can be easily generated every day, such as the characters "[(electricity)" and "[(brain)" forming a new word "DD(computer)". These have to do with the fact that every character has its own meaning, and thus they can play some linguistic role very independently. Furthermore, there are no "blanks" in written or printed Chinese sentences serving as word boundaries. As a result, the "word" in Chinese language is kind of not very well defined, the segmentation of a sentence into a string of words is definitely not unique, and there never exists a commonly accepted lexicon. Such situations make the Chinese linguistic processing very special and challenging. For western alphabetic languages, since the words are well defined, linguistic processing is primarily word-based, such as based on a lexicon of words and word (or word class) based language models. For Chinese language, since the words are not easy to identify, special measures are usually needed.

PAT tree is an efficient data structure successfully used in the area of information retrieval[4,5]. In this data structure, all possible character segments including subsegments of a text collection can be stored, retrieved and updated in a very efficient way. As a result, instead of keeping all neighboring relations for 2 or 3 words or characters as was done 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 dynamic text collection obtained from the Internet or the training corpus. In this way, the PAT tree effectively provides statistical parameters for all possible character segments with an arbitrary length N to appear in the training corpus, to be used as N-gram language model parameters. 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. This is one approach to handle the flexible wording structure of Chinese language in the Network environment as mentioned above. So the PAT-tree-based language model here is very helpful in Chinese language modeling.

V. PAT-TREE BASED LIVE LEXICON

The purpose here is to extract automatically the keywords/ key phrases or special terms for each of the subject domains or applications from the dynamic corpora collected from the Internet. These keywords/ key phrases or special terms can then be combined with some general domain lexicon to form "live lexicons" for all the different subject domains or applications, to be used in spoken language processing. Relevant research for western languages have been in good progress. For Chinese language, due to the feature that Chinese words are not well defined and wording structure is very flexible, the problem of extracting key words or key phrases becomes much more challenging. Here one such approach is briefly summarized [6,7]. The basic idea is that all possible segments of characters in the given text are taken as candidates for keywords or key phrases in the beginning, and those who can't be keywords or key phrases are then deleted by statistical approaches. Assume $W = w_1, w_2, \dots, w_n$ is a segment of n characters, let W_s and We be the longest starting and ending sub-segments of W, i.e., $W_s = w_1, w_2, \dots, w_{n-1}$ and $W_e = w_2, w_3, \dots, w_n$, as shown in Fig.3. The frequency counts of W, W_s and W_e in the text are then used to check carefully if W is more "complete" in meaning than its sub-segments W_s and W_e. For example, if W_s and W_e appears only in W in the text but not in any other way, then W is more "complete" in meaning. On the other hand, let W₁ be a segment appearing next to the left of W somewhere in the text and W_r be a segment appearing next to the right of W somewhere in the text, as also shown in Fig.3. The numbers of such different segments W₁ and W_r and their frequency counts in the text are then used to check carefully if the segment is used freely in the text. For example, if W is only the left part of a longer phrase W' = $w_1, w_2, \dots w_n, w_{n+1}, w_{n+2}$, then very often only $W' = w_1$, $w_2, \, \ldots \, w_n, \, w_{n+1}, \, w_{n+2}$ can be found in the text and the right context of W is highly limited. In this way many segments W can be deleted because they can't be a word or a phrase with "complete" and "significant" meaning. The segments obtained in this way very often include many frequently used words or phrases not specific for any subject domain. They can be found and removed by checking with a lexicon with commonly used words or phrases, or some data structure constructed from a large collection of texts with many different subject domains. Note that all the above processes become very straightforward if the target texts are used to construct a PAT tree as mentioned above, because all the segments and corresponding frequency counts are readily available from the PAT tree, and the PAT tree can be easily updated any time.

Some preliminary test results are listed in Table 1. The news abstracts published by the Chinese News Agency of Taiwan from June 1 to 20 were entered as the raw test corpora for extracting keywords and key terms, out of which a total of 2580 terms were selected manually as a reference for comparison. Different thresholds can be selected for the significance values of the terms to be taken as key terms. For example, in the second row of Table 1(a), when the threshold of significance value is



Fig. 5. Automatic Key Phrase Extraction Based on the Association for the Sub-segments and the Right/Left Context Dependency

taken to be 2.0, a total of 1,455 terms were extracted, out of which 532 are in a general domain dictionary but 923 are not. 1,135 terms out of the 1,435 extracted are in fact correct, i.e., within the 2580 manually selected. This gives a precision rate of 0.78(1,135/1,455) and a recall rate of 0.44(1,135/2,580). The results for the on-line extraction are in Table 1(b), in which the terms are extracted continuously when the news abstracts are entered one by one. The term length in Table 1(b)is the number of characters in a term. It can be found from Table 1(b) that roughly half of the key terms have only 2 characters, and there are much less key terms with more characters. Also, in average a new key term was extracted when 2.41 news abstracts were entered, and each key term appears 28.95 times in average in the entire news abstract database. Furthermore, when a new key term is extracted, it has appeared 9.25 times in average.

VI. PAT-TREE BASED ON-LINE DOCUMENT CLASSIFICATION

On-line Document classification is another highly desired functionality. The purpose is to classify the documents collected form the Internet on-line automatically. It was found that the PAT-tree is again very useful. The approach is very briefly summarized below. Assume a total of M subject domains is defined, each is given a set of training documents. M PAT-trees, {T₁, l = 1, 2, 3, ..., M}, can thus be constructed for these subject domains with the given training documents respectively. Assume an unknown document d has N sentences, $d = \{ s_1, s_2, s_3, ..., \}$ s_N }, where s_k is the k-th sentence. Assume the sentence s_k in d has n characters, $s_k = \{ c_1, c_2, c_3, ..., c_n \}$, where c_m is the m-th character. Now define P_{ij} is a segment of character string with (j-i+1) characters in a sentence s_k , $P_{ii} = (c_i, c_{i+1}, c_{i+2}, \dots, c_{i-1}, c_i)$, which is a subset of s_k . With the above definitions, the score of a document d with respect to a subject domain 1 can be evaluated as the following:

$$S(d; 1) \sum_{K=1}^{N} \sum_{1 \le i < i \le n} F[f_d(P_{ij}), f_1(P_{ij}), \omega(j-i+1)]$$

Where $f_d(P_{ij})$ is the frequency count of the segment P_{ij} in the document d, $f_1(P_{ij})$ is that in the PAT-tree for the subject domain 1, $\omega(j-i+1)$ is a weighting factor depending on the length of the segment, j-i+1, and

 $F[\bullet, \bullet, \bullet]$ is a carefully designed function. The domain 1 giving the highest score S(d; 1) is the domain the document d should be classified to.

(a)

Significance Value	Total Number of Extracted Terms (A)	Number of Extracted Terms in a Large Dictionary (B)	Number of Extracted Terms Outside the dictionary (C)	number of correct terms (D)	Precision (D/A)	Estimated Recall (out of 2580)
>1.5	2,291	977	1,314	1,374	0.70	0.53
>2.0	1,455	532	923	1,135	0.78	0.44
>2.5	723	208	515	593	0.82	0.23
>3.5	214	39	175	184	0.86	0.07

(b)

Domain	PAT Tree Size (K bytes)	Training	Size	Testing Size		Speed on P•266PC (#Doc/sec)	Precision	Recall
		Doc #	K bytes	Doc #	K bytes			
Politics	27,310	12,587	4,490	1,262	382.0	10.51	51.20%	92.50%
Congress	15,852	10,114	2,490	1,173	227.0	15.45	69.40%	61.00%
Business	7,071	7,337	1,800	1,082	219.0	15.03	91.40%	45.20%
Education	4,894	2,942	775	424	87.4	11.13	62.30%	53.30%
Sports	2,571	1,626	413	248	49.3	13.47	85.20%	78.60%
Local	11,117	6,580	1,690	968	201.0	14.57	73.80%	43.40%
Weather	867	2,896	3,130	435	474.0	8.97	100.00%	100.00%
Events	5,844	989	1,690	132	193.0	3.8	39.10%	81.80%
Total/Avg	75,526	45,071	16,478	5,724	1832.7	12.07	72.00%	65.50%

Table 1 Preliminary Results for Live Lexicon Construction from New Abstracts (a) precision and recall rates (b) on-line results

Term length	Total number of terms	Averaged number of documents to find a new term	Average frequency for the extracted new terms	Average accumulated frequency when a new term is extracted
2	776	3.93	34.22	9.37
3	416	6.04	24.60	9.09
4	171	12.16	19.22	8.97
5	51	37.28	20.35	9.18
6	17	109.81	27.00	8.65
7	15	23.60	27.40	11.20
8	6	274.67	13.00	9.83
9	3	205.67	18.00	11.33
Total	1455	2.41	28.95	9.25



Fig. 4. The Block Diagram for the initial Framework for Network Corpora Organization

Some preliminary results are listed in Table 2. These results are for news abstracts published by the Chinese News Agency of Taiwan in January to June 1997, in which the news for January to May are taken as training documents, and those for June as testing documents. The news abstracts have been manually classified by the Chinese News Agency into 8 domains, such as politics, business, sports, weather, etc. So 8 PAT-trees for the 8 domains were constructed, and the test results can be easily compared with the manual classification results. In Table 2 the sizes of the training/testing documents and the PAT-trees are listed for all the 8 domains, together with the precision and recall rates. For example, the precision rate for the sports news is 85.20%, i.e., out of those classified as sports news 85.2% are correct, and the recall rate is 78.60%, i.e., 78.60% of sports news are correctly classified into the sports domain. It can be found that the average precision rate is 72.00%, while the average recall rate is 65.50%. Also, the weather reports give the highest precision/recall rates, 100% for both cases, apparently due to the limited vocabulary and sentence structures.

VII. AN INITIAL FRAMEWORK FOR NETWORK CORPORA ORGANIZATION

With the technologies mentioned above, an initial framework for network corpora organization was

developed, with a block diagram \rightarrow shown in Fig 4. The major part of this figure includes a set of document processors with different modules required to process the unlimited documents collected from the Internet, and a set of linguistic information databases constructed from the collected documents. An information spider is used as the first module for the document processors to discover and collect all desired documents from the network automatically, and a document classifier is then used to divide the collected documents into different categories of classified corpora. A PAT tree generator is then used to construct the PAT trees for the classified corpora including all documents classified to each corpus. These PAT trees for the classified corpora can then be used as pattern indexing files for information retrieval, N-gram language models in speech recognition, live lexicon construction and new documents classification, etc.. All these will be very useful in future spoken language processing technology.

VIII. CONCLUSION

This paper presents results obtained for automatic network corpora organization, such that live lexicon, dynamic corpora and language models can be automatically obtained from on-line classified network documents. In this way the lexicons and the language models can be automatically adapted to the dynamic network resources, and as a results the spoken language processing technologies can also reflect the dynamic network resource and services.

REFERENCES

- [1] "Fundamentals of Speech Recognition", Lawrence Rabiner and Biing-Hwang Juang, Prentice-Hall Inc.
- [2] 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, USA, Dec.1999.
- [3] Lin-shan Lee, "Voice Dictation of Mandarin Chinese", IEEE Signal Processing Magazine, Vol.14, No.4, July 1997, pp.63-101.
- [4] Gaston H. Gonnet, Ricardo A. Beaza-yates and Tim Snider, "New Indices for Text : Pat Trees and Pat Arrays", Information Retrieval data Structure & Algorithms, Prenticce Hall, pp.66-82, 1992.
- [5] Morrison, D., "PATRICIA : Practical Algorithm to Retrieve Information Coded in Alphanumeric", JACM, pp.514-534, 1968.
- [6] Min-Jer Lee and Lee-Feng Chien, "Automatic Acquisition of Phrasal Knowledge for English-Chinese Bi-lingual Information Retrieval", ACM SIGIR-98, Melbourne, Australia, Aug, 1998
- [7] B.-R. Bai, et al., "Intelligent Retrieval of Dynamic Networked Information from Mobile Terminals Using Spoken Natural Language Queries", IEEE Trans. Consumer Electronics, Feb 1998, pp.62-72.