THE BROAD STUDY OF HOMOGRAPH DISAMBIGUITY FOR MANDARIN SPEECH SYNTHESIS

Wern-Jun Wang **, Shaw-Hwa Hwang** and Sin-Horng Chen***
email:wjwang@ms.tl.gov.tw

* Telecommunication Laboratories, DGT, MOTC, R.O.C.
** National Chiao Tung University, R.O.C.

ABSTRACT

How to increase the intelligibility and naturalness of synthetic speech have drawn much attentions in the recent Mandarin text-to-speech(TTS) researches. They have always been treated as bottleneck due to their effects are explicit for human perception. However, as qualities of synthetic speech increase for syllables, words or phrase, there is also an increasing need to improve the various components of the text processing. One of these desired improvements for Mandarin speech synthesis is the accuracy of character-to-sound(CTS) process. From the viewpoint of application, the purpose of speech synthesis should be aimed at making the synthetic speech understandable by human and minimize the misunderstanding between them. It thus is very important to increase the accuracy of CTS process. Such process is designed to predict phonetic pronunciations from a coarse surface text input and the difficulty mainly result from ambiguous homograph characters. In this paper, we proposed some effective analysis method incorporated with linguistic knowledge to resolve homograph ambiguity. The methods we used in the following experiments are discriminating lexical association and tree-based language model. From the experiment results, we can get about 10% more improvement on the average accuracy rate than traditional maximum frequency guess approach for most ambiguous homograph character.

1. INTRODUCTION

Text processing, including word identification, syntactic analysis and semantic analysis, etc. plays a very important role in a TTS system. To improve the naturalness and intelligibility of synthetic speech, text processing must be able to provide sufficient and accurate information for the succeeding acoustic process like prosody generation and synthesis unit selection. On the other hand, there are less discussions about the improvement of various component of text processing. One of them is the accuracy of CTS process and the difficulty mainly due to homograph ambiguity. Just like many languages, full of homograph exist in surface text, different meaning with different pronunciation but with the same ideogram. To resolve such ambiguity, a powerful text understanding system that can apprehend the meaning of every homograph character or word should be the best solution. The pronunciation of those homograph can be chosen after the meaning of them being decided. However, the system with such text understanding ability is still not available and is not allowed for a TTS system that with real-time processing requirement. Therefore, we are looking for a more effective and accurate homograph disambiguity algorithm as a substitution.

Many ideograms in Chinese have several kinds of pronunciations, some have only one meaning no matter what pronunciation they have, some have different meanings with different pronunciation. The latter case is what we want to solve and the pronunciation decision is a disambiguity problem. By using word identification process for all kinds of Chinese text, some of the pronunciation problem of ambiguous ideograms no longer exist when they are part of multi-ideogram words because we can easily find their pronunciation from a large lexicon. As for the ambiguous mono-ideograms, their pronunciation are still left unresolved because the CTS process for Mandarin is not entirely a simple dictionary look-up. Furthermore, Chinese language has many characteristics which make homograph disambiguity difficult to be attained[1]: a) Chinese exhibits more free word order, which allows many possible part-of-speech(POS) in a fixed context. and b) Chinese is weakly marked language with little inflection. A Chinese word may play many different grammatical functions in different contexts without morphological changes. Such characteristics result in many ambiguities in the POS tagged corpus. The purpose of this research will not discuss how to improve the accuracy of word identification and POS tagging at all, but build a robust homograph pronunciation decision system. To get a measure of the difficulty of this problem, we will discuss our proposed method for some frequent homograph characters. More specifically, we will focus on those homograph characters that are mono-ideograms word and discuss the possible solutions.

2. DISCRIMINATING LEXICAL ASSOCIATION

In this section, we will discuss how to make effective use of the information afforded from surrounding characters, to realize how

those characters in the left and right context of homograph characters can be helpful for pronunciation decision. This is so called "lexical association". By considering the interrupted sequence condition, we will also take into consideration about the header ideogram of every word. The related researches for English that have ever been discussed [2][3][4] are focusing on the collocation problem. Collocations are a wide variety of interesting linguistic phenomena and can be simply considered as a sequence of words (n-gram) among millions of other possible sequences. Collocation with lexical approach is an element of generating dictionary among a few thousand of other lexical approach. Obviously, they are difficult to be produced for second language learners. In this paper, we will introduce how to identify the significant collocation from corpus-based computational linguistics methods. We therefore expect to construct a pronunciation prediction model based on these collocations. As pointed out in [2], [3]and [4], the important things that must be bewared while handling collocation problems are the length of collocation, the window of observation, the interrupt sequence and the decision of significant collocation.

The window size parameter allows us to look at different scales, small window size will identify fixed expressions(idioms) and other relations that hold over short ranges; long window size will highlight semantic concepts and other relations that hold over larger scales. In theory this window size can be of any length. For the remainder of this paper, the window size will be set to 4 words as a compromise. The significant collocation terms will be decided by using some objective measurements such as association ratio, strength, spread and likelihood ratio that are deduced from the statistics of a very large corpus.

According to [2], if two words, x and have probabilities f(x) and f(y), then their association ratio, I(x,y), is defined as follows:

$$\mathrm{I}(\mathrm{x},\mathrm{y}) \;=\; log_2 \;\frac{\mathrm{f}(\mathrm{x},\mathrm{y})}{\mathrm{f}(\mathrm{x})\;\;\mathrm{f}(\mathrm{y})} \qquad (1)$$

where f(x) and f(y) are estimated by counting the number of observations of x and y in a corpus and normalized by the total words count of this corpus, f(x,y) can be regarded as joint probability and is estimated by counting the number of times that x is followed by y in a window. From the definition of I(x,y), we can realize that the variance of f(x,y) has not been considered in their approach. To include the variance of co-occurrence terms in decision procedure, the following strength k_i and spread U_i are proposed in [3].

$$k_i \;=\; \frac{freq_i \;-\; \overline{f}}{\sigma} \qquad (2)$$

$$U_i \;=\; \frac{\sum_{j=1}^{8}\left(p_i^j \;-\; \overline{p_i}\right)^2}{8} \qquad (3)$$

where $freq_i$ is the occurrence count of i-th evidence, $\overline{f}$ is the mean of all evidences, p_i^j is the occurrence count of i-th evidence on j-th position and $\overline{p_i}$ is the mean of all p_i^j. The strength and spread measurements have been adopted for collocation identification and work well for language generation and lexicography. As for the homograph problems, we must pay attention to another point which is the discriminating ability between all evidences. This comparison has not been included in the above measurements. To accomplish it, we have defined the likelihood ratio as follows[5]:

$$lr_i \;=\; \log_2 \;\frac{p(E_i \;\mid\; proA)}{p(E_i \;\mid\; proB)} \qquad (4)$$

where *pro*A and *pro*B are two possible pronunciation regarding to the specific evidence. The threshold for all above measurements are decided by heuristics. We have to do some statistics and analyses to find the suitable threshold for significant collocation decision and homograph disambiguity.

3. TREE-BASED ANALYSIS ALGORITHM

The use of POS for homograph disambiguity will be investigated in this section. The straight-forward procedure is to collect n-gram statistics on the training data, however, such n-gram statistics on the characters are probably not ideal for this problem since the contextual effects that we are trying to model are very huge. Although bigram or trigram context would be the possible choices with available training data, they are too rough and can not describe the effect within long sentence. For these reasons, we proposed a tree-based analysis model by using a probabilistic approach. By implementing such approach, a decision tree is used to capture lexical and contextual information according to a probability distribution. It relies on the statistics of adjacent POS context. However, as have been raised in Section 1, there are many ambiguous and errorous terms in our training and testing corpus and make the homograph disambiguity job more difficult. Consequently, we are seeking the possibility that building a theoretical framework that with error tolerance capability.

Considerable successes have been achieved in many speech processing tasks by using tree-based algorithm to effectively represent the speech production model. The advantages of tree-based analysis model lie in the ability to include the variety of questions. Furthermore, a extensive introduction of the standard tree growing procedures, known as the CART algorithm[6], enables a tree-based model to be estimated from training data in a very straightforward way. The CART algorithm assume that the source can be recurrsively splitted into some homogeneous node according to a questions set. To sufficiently describe any decision tree algorithm, we need to define the specific question set and the splitting algorithm. In this paper, we will use composite question algorithm and gini splitting criteria for tree

generation. The simple questions set is constructed from POS context with the relation of position in sentence. By using composite question, we can investigate the effectiveness of POS to some extent. The basic diagram of composite question is shown in Fig. 1[7]. The larger tree you need, the more complicated combination of questions you get.

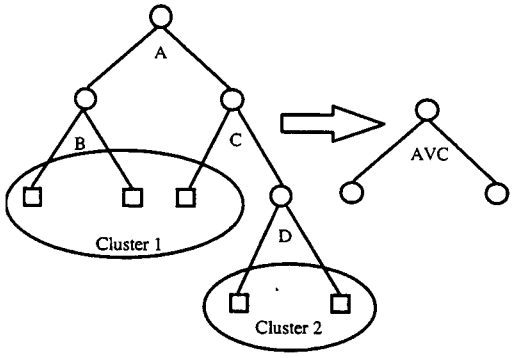

Figure 1: The composite question generated from the clustering of simple questions.

4. EXPERIMENTS

The corpus we used include 4 million ideograms and are extracted from news. We first apply the word identification and POS tagging processes to the corpus with the help of a 80000-token dictionary. The sentences for each homograph was collected to be disambiguated. Each sentence will include at least one homograph character. The number of frequent homograph ideograms under investigation are currently about 35. The actual number of homograph ideogram is much larger than this value. However, some ideograms are no longer ambiguous when they are mono-ideogram word, i.e. those homograph ideogram with different pronunciation has different POS and will be easily disambiguated. Besides, some ideograms are only with few occurrence count and are not worth of complicated analyses for them.

The corresponding text corpus are labeled using 46 POS categories not including punctuation symbols. The simple questions set will be generated from these 46 tags. The simulation procedures can be organized as follows:

1) Apply word identification to corpus and extract those sentences including homograph characters that are mono-ideogram word

2) Hand-label pronunciation of every homograph character in training and testing data.

3) Apply POS tagging.

4) Calculate the strength, spread, association ratio and likelihood ratio of all collocation and decide what are the significant ones that we need.

5) Generate decision tree, including questions set design, splitting and stopping criteria decision.

5. DISCUSSION AND EVALUATION

big5	(1)	(2)	(3)	(4)
的	2.874	15.091	33360.250	-0.499
在	4.291	12.108	97776.250	-2.386
年	3.696	1.611	836.750	-0.867
人	2.795	1.178	260.984	1.178
爲	2.553	1.155	77.609	-0.461
場	5.331	1.010	206.500	1.789
獎	4.661	-0.267	36.234	3.833
槍	4.382	-0.267	18.484	7.641
球	4.573	-0.331	4.750	7.708
考	4.806	-0.337	13.859	4.249
投	4.113	-0.354	19.750	7.348
抽	4.681	-0.365	37.188	8.772

Table 1: The different measurements for a list of surrounding characters of homograph ideogram "中" from the statistics of corpus.

To filter out the inappropriate collocation, we need the help of above four measurements defined in section 2. As shown in Table 1, the first column is a list of surrounding characters of "中" from the training corpus. The following columns are the corresponding values calculated according to (1), (2), (3) and (4). From this table, the first 6 characters are particular rich in making combinations with "中" owing to the higher strength and spread value. These collocation terms are improper word associations reflecting some spurious aspect of the training corpus that did not stand for true collocation. The actural discriminating collocation terms we need are the last six characters and is easily to see that the likelihood ratio can provide a good judgement. To investigate this phenomenon in detail, we can find one possible reason is that some Chinese words or characters always occurs in the surface texts simply because they have higher frequency. In order to choose the significant collocation terms that can resolve ambiguities, we

therefore realize the importance of the discriminating ability of all evidences. Another characteristic of Chinese that have to be pointed out and have been considered in our experiments is that collocation may exist in the form of head ideogram of word, that is we must investigate the relation between words and the head of words as well.

To check the validity of the tree-based algorithm, Table 2 and 3 will show the results of the prediction abilities. Owing to the limited samples numbers, only three homograph characters with more than 7000 sentences have completed the close and open tests as shown in Table 2. We have conducted only close test for the other homograph characters and the results of five are shown in Table 3.

	train			test		
big5	proA	proB	correct	proA	proB	correct
	3109	1891	4020	2504	1496	3083
	4910	90	4912	3930	70	3926
	4892	108	4914	3299	53	3310

Table 2: The close and open tests results of three homograph characters. The proA and proB are two possible pronunciations of the corresponding character and the numbers stand for the sample sentences numbers.

big5	proA	proB	correct
	1893	118	1943
	1607	383	1757
	383	311	502
	219	166	328
	190	75	213

Table 3: The close test results of 5 homograph characters.

6. CONCLUSION

The accuracy of CTS process can be regarded as a basic requirement for a TTS system. Comparing with 408 basic syllables and 5401 frequent characters in Mandarin Chinese, the homograph is just small amount. However, the wrong pronunciation of homograph characters will form a barrel for understanding in speech communication and is a fatal defect for a TTS system. In this paper, we have proposed the method by combining lexical association and tree-based language model. The lexical association is used for discriminating collocation identification and the tree-based model is expected to effectively use POS information.

Although accurate POS tagger is necessary for homograph disambiguity, such tagger with complete syntactic analysis is time-consuming and unpractical for a real time requirement. A more effective and accurate algorithm is still to be the better choice at present. According to the experiment results, the average prediction accuracy is about 10% higher than traditional maximum frequency assignment for the most ambiguous characters. We will continue improving the accuracy of word segmentation and POS tagging processes. They can be helpful for assigning the right POS to homograph and surrounding characters and helpful for homograph disambiguity. For the more complicated problems, we will introduce the semantic resolution in the future work to the sentences like “ ”(She wrote one line of characters), “ ”(He led a line of people), both with same POS string “PN VR Q N”[8].

7. REFERENCES

1. J. Chen, S. H. Liu, L. P. Chang and Y. H. Chin, “A Practical Tagger for Chinese Corpora”, ROCLING VII, pp.111-126, 1994.

2. Church and P. Hanks, “Word Association Norms, Mutual Information, and Lexicography”, Proceedings of 27th Meeting of the ACL, pp.76-83, 1989.

3. Frank Smadja, “Retrieving Collocations from Text: Xtract”, Computer Speech and Language, Vol.19, No.1, pp.143-177, 1993.

4. Michael D. Riley, “A Statistical Model for Generating Pronunciation Networks”, ICASSP, pp.737-740, 1991.

5. Richard Sproat, “Recent Work on Text-to-Speech Synthesis for English at AT&T Bell Laboratories”, Seminar materials at Telecommunication Laboratories, Aug. 1992.

6. Breiman, J. H. Friedman, R. A. Olsen and C. J. Stone, “Classification And Regression Trees”, Monterey. CA: Wadsworth, 1984.

7. K. F. Lee, S. Hayamizzu, H. W. Hon, C. Huang, J. Swartzz and R. Weide, “Allophone Clustering for Continuous Speech Recognition”, ICASSP, pp.749-752, 1990.

8. W. J. Wang, S. H. Hwang, C. H. Lee and C. S. Liu, “The Pronunciation Prediction Rules of Homograph Characters for Mandarin”, ROCLING, pp.225-231, 1995.