Resolving Category Ambiguity of Non-Text Symbols in Mandarin Text

Feng-Long Hwang **Ming-Shing Yu**

Text-To-Speech System Laboratory
Department of Applied Mathematics, National Chung-Hsing University
Taichung, Taiwan, R.O.C. Tel: (+886)-4-2860133 ext. 609
E-mail:flhwang@amath.nchu.edu.tw, msyu@dragon.nchu.edu.tw

Abstract

Usually there are some non-text symbols (e.g., "/" and ":") within the Mandarin texts (such as newspaper, magazine and files in Internet). Such symbols in sentence may have more than one possible oral expression. For instance, "1/2" can be pronounced as "January Second" or "one half"; and "10:15" may be pronounced as "ten versus fifteen" or "fifteen minutes past ten". In contrast to 2-gram, 3-gram and n-gram language models, this paper proposes an approach of multiple layer decision classifiers, which can resolve the category ambiguity of oral expression efficiently. Currently, the approach is composed of two layer decision classifiers: the first layer decision classifier is constructed under the linguistic knowledge and plays a binary characteristic function to predict all the promising categories. The second decision classifier is based on the corpus-based statistical method with two voting criterions. There are three principal phases in proposed approach: training phase, testing phase and translating phase. We can predict the correct category then translate non-text symbols into correct oral expression further. Based on the context, the evaluation precision rates for non-text symbol "/" on inside test and outside test are 96.3% and 91.2% respectively.

1 Introduction

The purpose of a Mandarin Text-To-Speech (TTS) system is to translate the text input into correct Mandarin speech. There are three principal phases in a TTS system: 1)text analysis , 2)prosody generation and 3)speech synthesis phase. The task of text analysis is to analysis the syntax and semantic information of text and to generate the phonetic transcription and part-of-speech (POS). The prosody generating is to generate the prosodic feature of text, such as duration, speech energy and pitch contours. The phase of speech synthesis, which should merges the prosodic feature and synthesis units in the acoustic inventory , is to generate the output of Mandarin speech with clear intelligibility and nice comprehensibility. The acoustic inventory may contain about 407 synthesis units with monotone or 1345 synthesis units with 4 tones (tone 1, 2, 3 and 4) in Mandarin speech.

Within the process for translating text to speech output, one situation is frequently encountered: because of existence of homograph words or non-text symbols, there are several possible different oral expressions based on its context for these words and non-text symbols in sentence. Usually there are some non-text symbols (e.g., "/" and ":") within the Mandarin texts (such as newspaper, magazine and files in Internet). For instance, "1/2" can be translated into "January Second" or "one half"; and "10:15" may be translated into "ten versus fifteen" or "fifteen minutes past ten". Two such sentences are displayed in sentence (A) and (B) . "3/5" in (A) is distributed into *date* category, whereas "3/5" in (B) into *fraction* category. Sentence (A') and (B') are the oral expression with respect to (A) and (B). Some major types of homographs are listed in [Yarowsky 1997].

The Academic Sinica Balanced Corpus version 3.0 (ASBC) [黃居仁等] includes 317 text files distributed in different fields, occupying 118MB memory and 5.22 millions of words(詞) totally. In ASBC, sentences have been segmented into several words (詞 or so-called lexicons) based on corpus of Academia Sinica Chinese Electronic Dictionary (ASCED). There are some non-text symbols (such as /, %, :, X ,..., and so on) in texts. Each non-text symbol may have different meanings subject to the syntax and semantics, such situation is so-called oral ambiguity. Different categories of each symbol should be translated into related oral expression. Whether the real meanings of non-text symbols can be expanded into its oral expression or not will affect seriously the correct output of Mandarin speech in TTS system. On the other hand, there is a one-to-many possible correspondence between a non-text lexical symbol and its possible semantic meaning translation. Referring to the linguistic knowledge and usage of prosody in TTS systems, the possible semantic categories of non-text symbol slash "/" are classified in Table 1.

Usually the non-text symbols in sentence will not affect human being to generate the correct speech of oral expression for each category of the symbol, because of the knowledge of linguistics and experience on text reading. In the paper, the so-called **non-text symbol** is defined as follows: the symbol in sentence that has several different semantic meanings and oral expressions. such symbols including some punctuation (such as ":"," • " , "-", etc) will be found in text frequently.

(A) 3／5, 電算中心出版使用手冊。
March 5th, Computer Center publish the users' manual.
(A') 三月五日, 電算中心出版使用手冊。
Suan1 yue4 wu3 r4, dian4 suan4 jung1 shin1 chu1 bian3 shi3 yuan4 shou3 che4.
(B) 產品價格比台灣的價格便宜3／5左右。
Products' price is less about three-fifth than that in Taiwan.
(B') 產品價格比台灣的價格便宜五分之三左右。
Chan2 pin3 jia4 ge2 bi3 tai2 wan1 de1 jia4 ge2 pian2 yi2 wu3 fen1 jr1 suan1 tzuo3 you4.

Table 1: some categories and its related oral expression of non-text symbol slash "/".

Category	Lexical item in sentence	Oral expression in Mandarin
1. date	3／4	三月四日
2. fraction	3／4	四分之三
3. tempo	3／4	四分之三拍
4. path, directory	／d e v／n u l l	根目錄d e v斜線n u l l
5. computer words	I／O	silence(or 斜線)
6. production version	V A X／V M S	silence (longer pause or 斜線)
7. frequent words in Internet	T C P／I P	silence (or 斜線)
8. others	中／日／韓文著錄	silence (longer pause)

In addition of the symbol "/", some other non-text symbols can be found within the corpus. We can analyze all the sentences in ASBC, extract the sentences which contain the symbol "/" and then classify these sentence into eight categories (referring to Table 1). Although there are several categories which speech for non-text symbol "/" are silence, its duration for silence in prosodic parameter is till different than other kind of pause.

The paper is organized as follows: in section 2, we will first present related information and previous works. Section 3 addresses the system structure of multiple decision classifiers, including the binary function classifier with decision tree and statistical corpus-based decision classifier. Section 4 displays the testing results of evaluation for each classifier combination. Finally, we will present the conclusion and future works.

2 Related Information and Previous works

2.1 Some Characteristics of Mandarin Characters and Words

A Mandarin Word (詞) is composed of one to several characters(字). The combination of one to several of such characters gives an almost unlimited number of words, in which at least some 10^4 of them are frequently used and can be found in Chinese dictionaries. A nice feature of Chinese language is that all the characters are mono syllabic, and the total number of phonological syllables is about 1345. Another important feature of Mandarin is certainly the existence of tones for syllables. Mandarin is a tonal language; in general, every syllable or character is assigned a tone. In fact, it is well known that four lexical tones (Tone 1,2,3 and 4) are primarily characterized by their pitch contour patterns. There are many works subject to the Chinese features of computational syntax, such as paper [Lee 1987] and [Lee 1993] .

2.2 Mandarin Word Segmentation

Segmentation(斷詞) is usually a difficult work in Mandarin Natural Language Processing (NLP) because of the absence of separation between words, although it is easy for other languages, e.g., English. There are two competing approaches have been used separately: the *rule-base approach* and the *statistical approach* [Fan 1988] and [Liang 1991]. Also some approaches use the *hybrid method* [Nie 1995].

2.3 The Issues of Natural Language Processing

The development of a natural language processing (NLP) system should resolve following issues [Su 1996] :

1) Knowledge representation : the way that can describe and organize linguistic knowledge for the natural language. Some knowledge can be described using a set of features in statistical methods.
2) Knowledge control : the way to apply linguistic knowledge for processing efficiently. A system can use statistical language models based on the maximum likelyhood for most possible estimation.
3) Knowledge integration : the way to use different knowledge source.
4) Knowledge acquisition : the way to acquire knowledge of needed natural language cost-effectively and systematically. The statistical approaches can collect the knowledge automatically.

2.4 Language Models

Currently, probabilistic language modeling (LMs) have been shown effective in many applications (especially, on the various knowledge acquisition of natural language processing). The purpose of LMs is to choose one desired result or category from several candidates for various kinds of linguistic problems, for example, assigning the best part-of-speech (POS) to each word in a sentence or estimate the most possible category to ambiguous symbol. Thus, the LM can be considered as a classifier processing. Basically, the criterions of LMs can be characterized into following types: 1) rule-based methods, 2) purely statistical methods, and 3) corpus-

based statistical methods. Within the rule-base methods, linguists exploit the linguistic theories and experience, on which the rule can be deduced and generated, to express knowledge of natural language. Usually, generating a lot of rules is labor-intensive and heuristic. It is hard to make the rule consistent and scale-up. Within the pure statistical methods, knowledge is expressed in terms of the estimation value of certain events, such as Markov chain and N-gram criterion [Brown 1992], that use the conditional probability on the occurrence of adjacent words and need very large parameter space. On the other hand, the statistical corpus language models can handle non-deterministic situations based on the probability measured from the training corpus. It is apparent that knowledge acquisition will be less expensive and less labor-intensive.

2.5 Previous works

Two frequently criterions for classification: one is *maximum likelihood classifier*, the other is called the *Bayesian classifier*. The problems we will resolve are the semantic ambiguities of non-text symbols for Mandarin text.

There are several methods that resolve the classification problems of linguistic and semantic ambiguity for natural language processing：

1) N-garm taggers: [Merialdo 1990] may be used to tag each word in a sentence with its part-of-speech(POS), thereby resolving those pronunciation ambiguities.
2) Bayesian classifiers: Bayesian have been used for a number of sense disambiguity. An implementation proposed in [Golding 1995]
3) Decision tree: [Brown 1991] can be effectively at handling complex conditional dependencies and noninde-pendence, but often encounter severe difficulties with very large parameter space.
4) Hybrid methods： [Yarowsky 1997] combines the strengths of each of preceding paradigms. It is based on the formal model of decision tree.
5) Multiple Decision classifiers: [Rodova 1997] take interest in speaker identification.

3 The Proposed Approach

3.1 System structure

In contrast to *2-gram*, *3-gram* and *n-gram* Language models, this paper proposes an approach of multiple layer decision classifiers which can resolve the category ambiguity of oral expression for non-text symbols efficiently. The proposed approach is composed of multiple layer decision classifiers (currently, we have constructed two classifiers): the first layer of decision classifier is constructed as decision tree under the linguistic knowledge and plays with a binary function. In this layer, some impossible categories will be excluded and remained categories are all the promising categories. The second layer of the proposed approach is a statistical corpus-based method, in which all the words (lexicons) in sentence play as voter under voting criterion and vote for each category with statistical parameters.

These multiple layer decision classifiers are combined together with *multiply* operation. Like the political mechanism, all voters will give their suffrage to each category with a statistical score. Finally the category with maximum voting score can be predicted as the goal category for non-text symbol. The structure of multiple decision classifier is shown as Figure 1.

3.2 The Binary Classifier based on Decision Tree

The decision tree classifier plays as a binary logical function, which is to deduce all promising categories for the non-text symbol based on Mandarin linguistic knowledge. This classifier will assign probability value 1 to all promising categories. On the other hand, some categories will be excluded and assigned a probability value 0. For example, the substring "3/4"may belongs to several possible categories: *date* (March 4th), *fraction* (three fourth) and *tempo* (three slash four pulses), these categories will be assigned a value 1. But the substring "14/2" and "SUN4/75" could not belong to the category *date* and *tempo*, these category will be assigned a probability value 0. Within the binary classifier, all the promising category with probability value 1 will pass into statistical classifier in the second layer, which will decide the final category of non-text symbols.

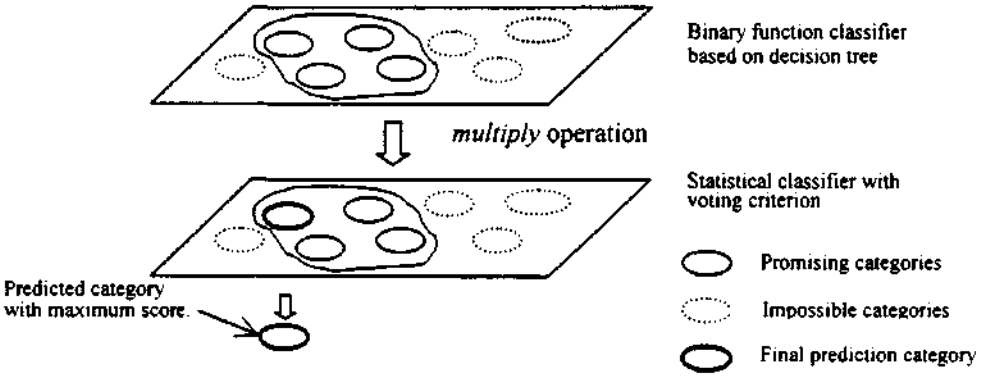

Figure 1: Multiple layer decision classifiers contain two classifiers, which are merged together with *multiply* operation.

The decision tree can be displayed as Figure 2. A successive answers to questions (Q_1, Q_2, ...,Q_n) about the syntax and semantic meaning for left and right neighbor (tokens or lexicons) of non-text symbol in sentence will decide which path should be traced into based on the linguistic knowledge. Finally, one leaf node will be reached and a *set* of categories will be deduced. Within the *set*, all the categories will be assigned a value 1. However, other categories will be assigned a value 0. The key point for constructing an effective decision tree is how to exploit the linguistic knowledge and the skill of decision making. All possible categories should be keep inside the set, otherwise the precision rate will be reduced. In our proposal, the probability value for each category cab be described as follow:

$$P_{ij}(\Phi_j) = \begin{cases} 0 & \text{if } i \in I \text{ and } \Phi_j \notin set. \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

where i=1,2, ...,I. i is labeled as the i^{th} layer decision classifier. I is the number of total decision classifiers (currently, we have developed two decision classifiers, so I=2). j=1,2,...J, and J is the number of categories for non-text symbols. Φ_j is labeled as the category j for non-text symbols. $P_{ij}(\Phi_j)$ is the probability value of category j (Φ_j) for the layer i classifier. *set* is deduced from decision tree classifier and contains all promising categories. Currently in our approach, just first layer classifier plays as a binary function. So, Equation (1) can be explained further as follow: if I = 1 and $\Phi_j \notin set$, $P_{ij}(\Phi_j) = 1$. Otherwise, $P_{ij}(\Phi_j) = 0$.

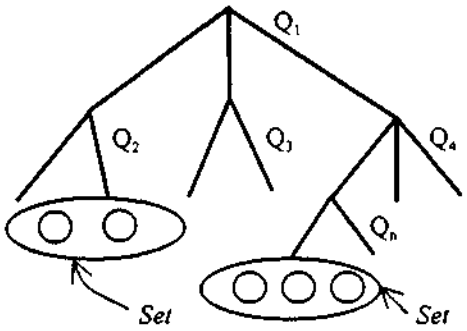

Figure 2: The *set* in leaf node contains all the promising categories deduced on the Decision Tree, which have a probability value 1.

Basically, the decision tree classifier is generated according to linguists' experience and theories. The remained categories are all the possible categories that the non-text symbol may belong to. Thus, the voting approach can predict the only one among all the possible categories. It is so apparent that processing of adopting decision tree can improve the precision rate.

3.3 The Decision Classifier with voting criterion

Figure 3 presents the system structure of our approach, which contains several principal phases:1) Training phase, 2) testing phase and 3) translating phase. The output of translating phase will be sent into the next phase for linguistics analysis further, which could promote the overall intelligibility and performance of TTS system.

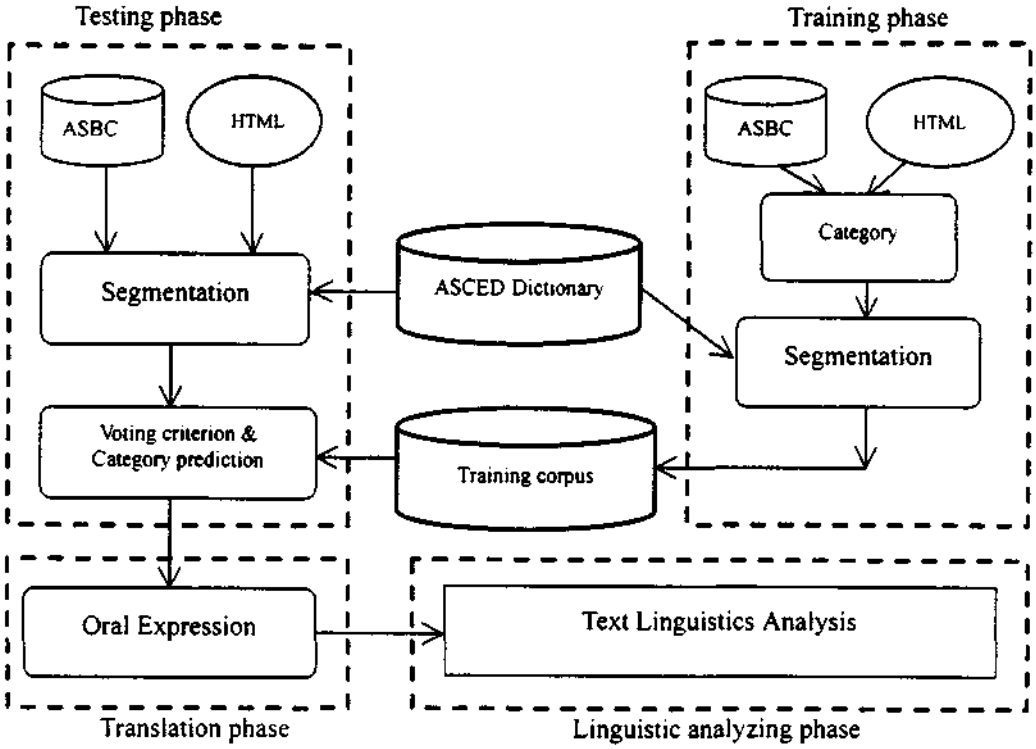

Figure 3: The principal phases of statistical decision classifier.

(C) 故 行 政 院 於 週 四 （ 1／4）復會。
So, the Executive Yuan rehold meeting on Thursday (January 4th)。
(C') 故 行 政 院 於 週 四 （ 1 ／4）復會。

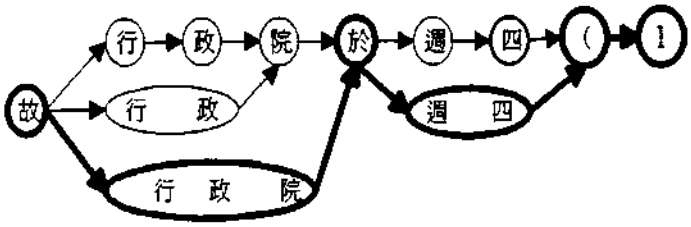

Figure 4: Some possible sequences of segmentation based on the maximum length of words within several possible combinations for chunka (*CHa*).

Training Phase

i) The text source

The Academic Sinica Balance Corpus (ASBC) contains 317 text files and 5.22M words in Chinese Mandarin totally, these files will be classified into each categories. Each sentence in original ASBC is tagged with part-of-speech (POS) and segmented into several words, the tags and white separation (space) between words will be removed during processes. After analyzing the corpus, we find that the nonuniform distribution of sentence appears in some categories for each non-text symbol. For example, the sentences in category of *music tempo* for symbol "/" just appear 3 times totally. The situation also appears in other non-text symbols. We collect further and download the text files from HTML source or BBS posted papers, and then remove all the HTML tags (such as <HTML>, <P>, <A href=" ....", and so on) and unnecessary symbols.

ii) The category classification of each non-text symbol

The text source for training phase can be extracted from ASBC and Internet HTML and BBS files semi-automatically. First, we category the source for each non-text symbol, the extracted sentences will be distributed into one or several categories related with the symbol based on the lexical and semantics knowledge. The categories for non-text symbol "/" are listed in Table 1.

iii) Segmentation

Word segmentation paradigm is based on the Academia Sinica Chinese Electronic Dictionary (ASCED), which contains near 80,000 words. The words in ASCED are composed of one to 10 characters. Our principal rules of segmentation are subject to maximal length of word first and then to least number of words in a segmented pattern based on the **dynamic programming method** (Veterbi searching). The occurrence of word within ASCED will be considered under the situation of same number of words among patterns. The priority is that segmented pattern which contains the maximal length of word will be chosen. If two patterns have same maximum length, we compare further the total number of words in the pattern; then the pattern that is composed of least number of words will be chosen. The same segmentation's criterion will be used within the testing phase.

Any trained sentence will be partitioned into two chunks: chunka (*CHa*) and chunkb (*CHb*), which contain the substring in front of and following the non-text symbol respectively. For example, the sentence (C) , which contains 14 characters (including the parenthesis and punctuation, but exculding the non-text symbol "/" itself) , is composed of two chunks of substring *Cha* and *CHb*: *CHa* contains the substring "故行政院於本週四（1" and *CHb* contains the substring "4）復會。". The word with maximum length in *CHa* is "行政院". After segmentation process, there are six words in *CHa* : 故, 行政院, 於, 週四, （ and 1: five words in *CHb* : 4, ）, 復, 會 and 。. According to the segmentation criterion, the pattern sequence in (C') is the most favorit e one. Figure 4 shows some possible sequences for sentence (C).

iv) The training corpus

After the word segmentation, the word may be appended into the training corpus. Each record contains the evidences: word itself, word category, occurrence and the location subject to the non-text symbol.

The algorithm of training phase is listed as follow:

- collect the source data from ASBC and Internet.
- extract all the sentences from collected data, which contain one or more than one non-text symbols usually.
- classify these source data into categories under linguistic and prosody for TTS system usage.

  - read-in sentence in category. The sentence will be partitioned into two substrings: chunka (*CHa*) and chunkb (*CHb*) which are in front of and following the non-text symbol respectively
  - segmentation processing will generate the words of two substrings
    - if the word never appear in corpus for this category and respective location
      - append a record into the corpus.
      - each record contains four fields:
        - word,
        - occurrence,
        - category id. for the non-text symbol,
        - the location respective to the non-text
    - else
      - word's occurrence for the category will be increased by 1

- the training corpus has been constructed, the phase terminated.

The Testing Phase

The segmentation task of testing phase adopts same criterions as that in training phase shown in section 3.3. A sentence will be divided into substring *CHa* and *CHb*. For each word, the probability of each category can be calculated and summed up based on the evidence (parameters found in training corpus) respectively. It is called the ***voting criterion.***

Based on the *voting criterion,* each word in *CHa* and *CHb* have a probability value, which looks like the voting the suffrage, to every category of the non-text symbol. Like the political voting mechanism, the only category, which gets the tickets in majority (maximum score), will become to be the predicted category. In our voting criterion, two score rules are proposed: one is based on the *preference scoring*, the other is based on the *winner-take-all scoring.*

Voting criterion with *preference scoring* rule

The prediction processing is based on the occurrence of each word inside training corpus for each category. Usually, the sentence *C* is composed of three parts: substring *CHa* ,non-text symbol *N* and substring *CHb*. *C*, *CHa* and *CHb* could be expressed as follow:

$$C = CH_a + N + CH_b \tag{2}$$

$$CH_a = w_{a1} w_{a2} \cdots w_{aj} \cdots w_{a_m} \tag{3}$$

$$CH_b = w_{b1} w_{b2} \cdots w_{bj} \cdots w_{b_n} \tag{4}$$

where a_m and b_n are the total number of words in *CHa* and *CHb* respectively. It is apparent that *CHa* and *CHb* contain one or several non-text symbols. Also, *CHa* and *CHb* may be an empty substring.

For each word in *CHa* and *CHb*, the score *S* of each word voting for category j (Φ_j) of non-text symbol can be computed as follow:

$$S_{ajk_1}(w_{ak_1}) = \frac{C_{aj}(w_{ak_1})}{TC_a(w_{ak_1})} \tag{5-1}$$

$$S_{bjk_2}(w_{bk_2}) = \frac{C_{bj}(w_{bk_2})}{TC_b(w_{bk_2})} \tag{5-2}$$

where $1 \le k_1 \le m$ and $1 \le k_2 \le n$, w_{ak_1} and w_{bk_2} are labeled as the k_1^{th} and k_2^{th} word in *CHa* and *CHb*. $C_{aj}(w_{ak_1})$ and $C_{bj}(w_{bk_2})$ is the occurrence of word w_{ak_1} and w_{bk_2} in category j (Φ_j) . $TC_a(w_{ak_1})$ and $TC_b(w_{bk_2})$ mean the total occurrence of w_{ak_1} and w_{bk_2} in the corpus for the non-text, which can be computed as:

$$TC_a(w_{ak_2}) = \sum_{j=1}^{J} C_{aj}(w_{ak_1}) \tag{6-1}$$

$$TC_b(w_{bk_2}) = \sum_{j=1}^{J} C_{bj}(w_{bk_2}) \tag{6-2}$$

For the second decision classifier, the total score TS_a and TS_b of all words in *Cha* and *CHb* for category j(Φ_j) of non-text symbol can be computed as follow:

$$TS_a(\Phi_j) = \sum_{k_1=1}^{a_m} S_{ajk_1}(w_{ak_1}) , \tag{7-1}$$

$$TS_b(\Phi_j) = \sum_{k_1=1}^{a_m} S_{bjk_2}(w_{ak_2}) \tag{7-2}$$

The second layer total score *TS* of whole sentence for each category j (Φ_j) is displayed as follow:

$$TS(\Phi_j) = (TS_a(\Phi_j) + TS_b(\Phi_j)) \tag{8}$$

where 1<=j<=J. The overall total score TS^* merging the first and second decision classifier for category j is computed with the ***multiply*** (*) operation:

$$TS^*(\Phi_j) = P_{1j}(\Phi_j) * TS(\Phi_j) , \quad \Phi_j \in set \tag{9}$$

where *set* is composed of all the promising categories deduced by first layer decision tree classifier.

$$TS^*(\Phi_{j^*}) = \text{argmax}\,(TS^*(\Phi_j)) \tag{10}$$

where 1<=j<=J, Φ_{j^*} is the final category of our prediction. Under the multiple decision classifiers, the final predicted category should be subject to the category of TS^*, which has the maximum score.

Voting criterion with *winner-take-all* scoring rule

In contract to the *preference scoring rule* above, the Voting criterion with *winner-take-all* adopt the a different scoring rule. For each word in *CHa* and *CHb*, $S_{aj^*k_1}(w_{ak_1})$ and $S_{bj^*k_2}(w_{bk_2})$ will be the winner for category $j^*(\Phi_{j^*})$ for word w_{ak_1} and w_{bk_2} and assigned a probability value 1 as voting score .

Category $j^*(\Phi_{j^*})$ is subject to the category which the have the maximum score among all the promising categories for w_{ak_1} and w_{bk_2} . Equation (5) should be changed. Equation (6) – (10) don't need to make any change at all.

Voting criterion with *winner-take-all* scoring rule looks like the voting scheme in political mechanism, in which everyone just can vote for the favorite person among all the candidates. The suffrage of each voter will be voted for whom to be preferred. In our approach, the favorite category for each word will be assigned a score 1 whereas all other categories will have a score 0. The total scores of each category for non-text symbol can be accumulated for each word in sentence.

3.4 Translation phase

The final phase of second classifier is the translation processing. The non-text symbol can be translated into its Mandarin oral expression of text in which the category has been predicted by testing phase. For instance, sentence (D) contains a non-text symbol "/", which is predicted as the *date* category and "4/10" in (D) will be translated into the Mandarin oral expression "四月十日"

(D) 這本雜誌已於上週六（4／10）出版。
This magazine was published last Saturday (April 10th).
(D') 這本雜誌已於上週六（四月十日）出版。
Je4 ben3 tza2 jr4 yi3 yi2 sang4 jou1 liou4 sz4 yue4 sz4 r chu1 bian3.
(E) ，本ＯＳ適用於ＳＭＰ／ＭＰ等電腦架構，
，This ＯＳ is suited for the computer architecture of ＳＭＰ／ＭＰ，
(E')，本 ＯＳ 適 用 於 ＳＭＰ ＭＰ 等 電 腦 架 構 ，
，ben3 ＯＳ su4 un4 iu2 ＳＭＰ ＭＰ dun1 den4 lau3 chie4 go4 ，

Table 2: precision rate of statistical decision classifier for each category of non-text symbol "/".

second layer classifier only	statistical decision classifier(not merging first layer classifier)															
	preference score rule								*winner-take-all score rule*							
category	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
Inside test (%) ⇨	97	100	92	82	100	100	100	92	97	67	46	64	100	72	97	92
number of *Nt*	134	18	26	44	30	36	68	26	134	18	26	44	30	36	68	26
outside test (%) ⇨	86	80	100	88	100	71	88	100	86	60	25	75	80	71	88	100
number of *Nt*	30	10	8	16	10	14	16	10	30	10	8	16	10	14	16	10

Ps. *Nt* stands for the non-text symbol.

such as the pattern in (D'). Another example in (E), the symbol "/" should be pronounced as silence in Mandarin usually. The output text of this phase will be processed further with text linguistic analysis in TTS system.

4 Evaluations

Our approach has been implemented on a platform of personal computer (PC) with Intel Pentium III. The language package for system development is in C++ environment. Four fifth of text source is used for training phase, the remained source is used for testing phase.

Two layer decision classifiers have been generated. We evaluate the results of inside test and outside test for the second layer *statistical classifier* with two different *voting criterion*, then we combined it with the first layer *decision tree classifier* to compare the performance of precision rate. The precision rate (PR) is defined as follows:

$$\mathbf{PR} = \frac{\#\text{ of correct prediction category}}{\text{total }\#\text{ of non-text symbol}} \qquad (11)$$

4.1 Evaluation results for second layer classifier

The results for second layer classifier are listed in Table 2. Total number of non-text symbol "/" for inside and outside test are 382 and 114 respectively. The overall inside test and outside test for different voting criterion are listed respectively in Table 3.

4.2 Evaluation results merging two layer classifiers

Under the multiple layer decision classifier structure, first and second layer classifier are merged together to improve the overall precision rate. Exploiting the first layer classifier to exclude some impossible categories, the results are attractive and displayed in Figure 5 and Figure 6. As shown, the final results of inside test and outside test is 96.3% and 91.2%, which are obtained by merging the first layer classifier and second layer classifier with voting criterion of preference scoring scheme. L1 and L2 in Figure 5 and Figure 6 represent the first layer and second layer classifier respectively. L1+L2 means the merging of L1and L2.

Table 3: The precision rate of inside test and outside test of second layer(L2) statistical decision classifier

	preference scoring	*Winner-take-all scoring*
Inside test(%)	95.8%	85.7%
Outside test(%)	85.8%	77.1%

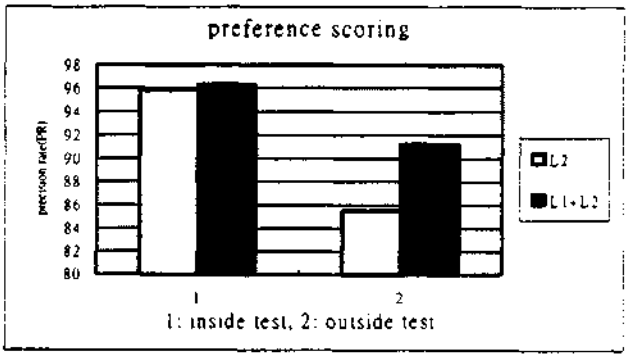

Figure 5: The precision rate (PR) of evaluation, number 1 and 2 in x-axis stand for the inside test and outside test.

[1] Symbol "/" should be a silence speech.

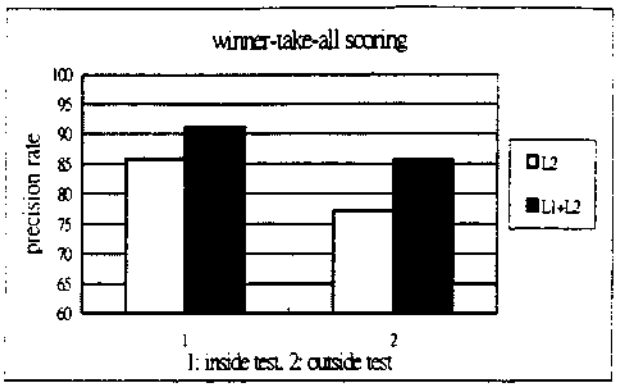

Figure 6: The precision rate (PR) of evaluation, number 1 and 2 in x-axis stand for the inside test and outside test.

5. Conclusion and Future Works

In the paper, we have developed an effective approach, which contains the multiple layer decision classifiers and can resolve the category ambiguity of non-text symbols in Mandarin text. In contract to the *2-gram* and *n-gram* Language Models, our approach just need smaller size of corpus and still can hold the linguistic knowledge for statistical parameters. Currently, there are just two classifiers in our approach: the first classifier is based on the decision tree to deduce promising categories, the second classifier is on the statistical corpus-based classifier with two voting criterion. Final precision rate of inside and outside test reach 96.3% and 91.22% respectively.

In addition to the non-text symbols " /" addressed in the paper, there are some other symbols, such as *, %, [] and so on, in which the oral ambiguity problems will be incurred and should be resolved. We will expand our approach to other related domains. Also, the abbreviation of some company and institute's name in text can be expanded similar to the expansion of non-text symbols in sentences. Basically, the topics which should be researched in the future include :

1) Distance dependency for each lexicon (weighting value and windows)
2) Patterns of special and frequent case for non-text symbols in text files.
3) The training and learning algorithm
4) The smoothing and normalization methods.
5) Expand the current two layer into more layer classifiers to resolve complicated linguistic classification problem.

Acknowledgement

The authors would like to appreciate Academia Sinica for exploiting ASBC Corpus.

References

黃居仁等(1995). "中央研究院平衡語料庫簡介", Proceeding of ROCLLING VII, pp. 81-99.

Brown P., Pietra S. D., Pietra V. D. and Mercer R. (1991). "Word Sense Disambiguation Using Statistical Methods". In Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, pp. 264- 270.

Brown P. F., Pietra V. J., deSouza P. V., Lai J. C. and Mercer R. L.(1992). "Class-based n-gram Models of natural Language", Computational Linguistics, vol. 18, No. 4, pp. 467-479.

Fan C. K. and Tsai W. H. (1988). "Automatic word identification in Chinese sentences by the relaxation technique", Computer Processing of Chinese and Oriental Languages, vol. 4, pp. 33-56.

Golding A. R.(1995). "A Bayesian hybrid method for Context-Sensitive Spelling Correction", In Proceedings of the third workshop on Very Large Corpora, pp. 39-53, Boston, USA.

Lee Lin-Shan, et. al, (August 1987). "A Mandarin Dictation machine Base upon Chinese Natural Language Analysis", The 10th International Joint Conference on Artificial Intelligence, AAAI .,Milano, Italy, pp. 619-621.

Lee Lin-Shan, et. al, (April 1993). "Golden mandarin (I)—A real-time mandarin Speech dictation Machine for Chinese Language with Very large Vocabulary", IEEE Trans. On Speech and Audio Processing, vol.1, No. 2, pp. 158-179.

Liang N. Y. and Zhen Y. B. (1991). "A Chinese word segmentation model and a Chinese word segmentation system PS-CWSS", COLIPS, vol. 1 , pp. 51-55.

Merialdo B.(1990). "Tagging Text with a Probabilistic Model", In Proceeding of the IBM Natural Language ITL, Paris, France, pp. 161-172.

Nie Jian-Yun, Ren Xiaobo, Brisebois Martin. (1995). "A unifying Approach to segmentation of Chinese and its Application to Text Retrieval", Proceeding of ROCLING VII, ROC, pp. 175-190.

Rodova V. and Psutka J. (1997). "An Approach to Speaker Identification Using Multiple Classifiers", ICASSP, pp. 1135-1139

Su K. Y. and Chang J. S. (1992). "Semantic and Syntactic Aspect of Score Function", Proceeding of COLING-88, pp 642-644.

Su K. Y., Chiang T. H., Chang J. S. (August 1996). "A Overview of Corpus-Based Statistical Oriented (CBSO) Techniques for Natural Language Processing", Computational Linguistics and Chinese Language Processing, vol. 1, no. 1, pp.101-157.

Yarowsky David. (1997). "Homograph Disambiguation in Text-to Speech Synthesis".