# Semantic classification for Patterns Containing Non-Text Symbols in Mandarin Text

**Feng-Long Hwang[12], Ming-Shing Yu[1],Ming-Jer Wu[1], Shyh-Yang Hwang[1]**

[1]TTS Lab., Department of Applied Mathematics, National Chung-Hsing University
Taichung, Taiwan, R.O.C. 402, Tel: +886-4-2860133 ext: 609
[2]National Lien-Ho Institute of Technology, Miauli, R.O.C,Tel:+886-37-332997
flhwang@amath.nchu.edu.tw, {msyu,mjwu,syhwang}@dragon.nchu.edu.tw

**ABSTRACT**

In this paper, we address the semantic classification of non-text symbols in Mandarin text using multiple decision classifiers. Some non-text symbols (e.g., "/" and ":") appear frequently within the Mandarin texts (such as newspaper, magazine and files in Internet). Usually, these symbols in sentence may have more than one possible oral expression. In contrast to *2-gram*, *3-gram* and *n-gram* language models, the paper proposes the multiple layer decision classifiers, which can resolve the category ambiguities of oral expression for patterns containing one or several non-text symbols in Mandarin texts efficiently. There are two principal phases in our proposed approach: training phase and classification phase. Currently, classification phase contains two decision classifiers. We can predict the correct category of the non-text symbols then translate the non-text symbols into correct oral expression further. The empirical precision rates for inside and outside test are *97.8% and 93.0%* respectively.

## 1 Introduction

The goal of 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. The phase of speech synthesis, which should transforms the prosodic feature and synthesis units in the acoustic inventory according to the prosody of speech, is to generate the output of Mandarin speech with clear intelligibility and great comprehensibility. The acoustic inventory may contain about 400 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 and non-text symbols, there are several possible different oral expressions based on its contextual information and non-text symbols in sentence. There are some non-text symbols (e.g., "／" and " : ") within the Mandarin texts (such as newspaper, magazine and files in Internet). For example, the pattern

of "2/3" can be translated into "February three " or "two third"; and the pattern of "9:15" may be translated into "nine versus fifteen" or "fifteen minutes past nine". The pattern of "3/5" in (A) is categorized into *date* category(三月五日) while pattern of "3/5" in (B) into *fraction* category(五分之三) (A') and (B') are the oral expression with respect to (A) and (B). Some major types of homographs are listed in [Yarowsky,1997].

(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 de jia4 ge2 pian2 yi2 wu3 fen1 jr1 suan1 tzuo3 you4.

The Academic Sinica Balanced Corpus version 3.0 (ASBC) [黃居仁等,1995] includes 317 text files distributed in different topics, occupying 118MB memory and 5.22 millions of words 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), and each word in the sentence is tagged with its related part-of-speech (POS). There are several kinds of non-text symbols (such as /, %, :, X ,...., and so on). Each non-text symbol may have different meanings subject to the syntax and semantics, such situation (like sentence A & B above) is so-called oral ambiguity. Different semantic category for each non-text symbol should be translated into its related oral expression. On the other hand, there is a one-to-many possible correspondence between a non-text lexical symbol and its possible semantic meanings. 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. Based on the linguistic knowledge and usage of prosody in TTS systems, the possible semantic categories of non-text symbol slash "／" are classified and shown on Appendix A.

The so-called **non-text symbols** are defined as follow: the symbols that are not the Mandarin characters and have several different semantic meanings and oral expressions within a sentence. Such symbols including some punctuation (such as ":", "．" , and so on) will be found in text frequently.

The paper is organized as follow: in section 2, we first present previous works and then addresse the overall structure of proposed approach. Section 3 focuses on the multiple decision classifiers. Section 4 displays the empirical the testing results of evaluation. Finally, we will present the conclusions and future works.

## 2 The Proposed Approach

### 2.1 previous works

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 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 disambiguation. An implementation was proposed in [Golding ,1995].
3) Decision tree: [Brown ,1991] can be effectively at handling complex conditional dependencies and nonindependence, 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.

### 2.2 The Proposed System Structure

The system structure is shown as Figure 1. It contains two principal phases:1) **training phase** and 2) **classification phase**. In the training phase, the feature corpus will be trained using several parameters of linguistic knowledge of the pattern containing non-text symbols. In the classification phase, the patterns containing non-text symbols in sentence will be classified using the multiple decision classifiers, in which the output of predicted category will be sent into the translating phase to translate the pattern to correct oral expression. The output text can be processed for linguistics analysis further, which could promote the overall performance of TTS system. In contrast to *2-gram*, *3-gram* and *n-gram* Language models (LMs), this paper proposes an approach of multiple decision classifiers which can resolve the category ambiguity of oral expression for non-text symbols efficiently. In multiple decision classifiers, currently we have generated two classifiers: the first decision classifier is constructed as decision tree under the linguistic knowledge and plays as a binary function. Within first classifier some impossible categories will be excluded and all remaining categories are the promising categories. The second classifier employes statistical 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 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. Basically, the decision tree classifier is generated according to linguists' experience and knowledge. The remained categories are all the possible categories that the non-text symbol may belong to.

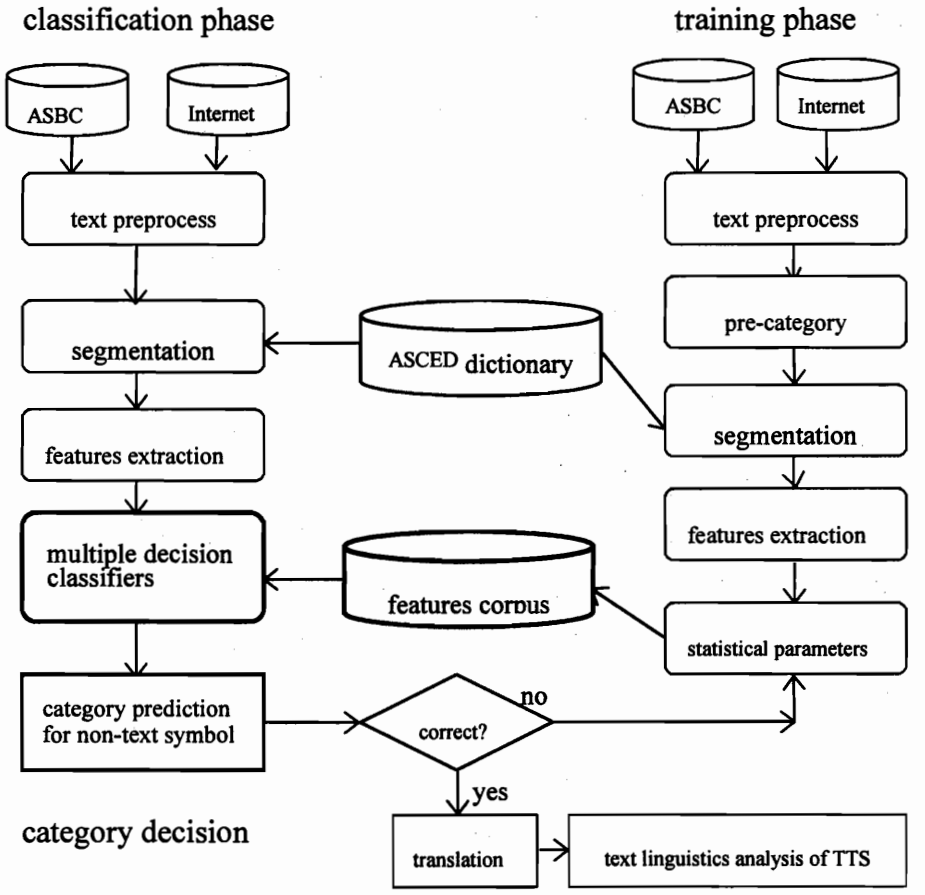


Figure 1:The principal phases of statistical decision classifier with voting criterion.

2.3 Training Phase

**i) The text preprocess**

The Academic Sinica Balance Corpus (ASBC) contains 317 text files and 4.55M characters in Chinese Mandarin [黃居仁等,1995]. 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. In the text preprocesses, we further collect and download the more text from HTML source and BBS posted papers, and then remove all the HTML tags (such as <HTML>, <P>, <A href=" ....", and so on) and other unnecessary symbols in these files.

**ii) The pre-category 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 the symbol based on the lexical and semantics knowledge. The eight possible categories for non-text symbol "／" are listed in Appendix A.

**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** (Viterbi searching). The priority scheme 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 priority will be used within the training phase and testing phase.

**iv) Constructing corpus for statistical parameters**

After the segmentation for *CHa* and *CHb*, the feature of each word will be used as the statistical parameters, all of which will be recorded in the training corpus statistically. Each record contains the four feature evidences explained above.

2.4 Classification Phase

**i)The text preprocess**

Text preprocess in this phase process the same task as that in training phase.

**ii) Segmentation**

segmentation task in classification phase uses same criterions as that used in training phase shown in precious section also. A sentence with non-text symbols 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 parameters found in feature corpus respectively.

**iii) The features extraction**

Feature extraction in this phase does same task as that in training phase.

**iv) Multiple decision classifiers**

The goal of multiple decision classifiers is to predict the correct category, to which the non-text symbols belong. The structure details will be described in next section.

Within the classification phase, some categories output in sentence could be mispredicted. To make the multiple decision classifier more robust, these sentences can be sent back into statistical parameter process in training phase and adapts dynamically the parameters of feature corpus to raise the precision rate. The feedback usually can solve the unseen events (words) in training text, the situation of unseen words often appears in natural language processing.

## 3 The Multiple decision classifiers

3.1 The Structure of Multiple decision classifier

In contrast to *2-gram*, *3-gram* and *n-gram* Language Models, this paper proposes an approach of multiple decision classifiers, which can resolve the category ambiguity of oral expression for non-text symbols efficiently. Within the classification phase, we have

constructed two classifiers: the first decision classifier is generated and shown as decision tree based on the linguistic knowledge. Some impossible categories will be excluded while the remaining categories are all the promising categories. The second classifier employes a corpus statistics-oriented technique to estimate the final category with maximum score. All the words (lexicons) in sentence play as voter under the voting criterion and vote for each category with statistical parameters score.

These multiple decision classifiers are combined together with *multiply* operation. Like the political mechanism, all voters will give their suffrage to each category with a statistical probability score. Finally, the category with maximum statistical parameters score can be predicted as the goal category for non-text symbol. The overall system structure of multiple decision classifiers is shown as Figure 2.

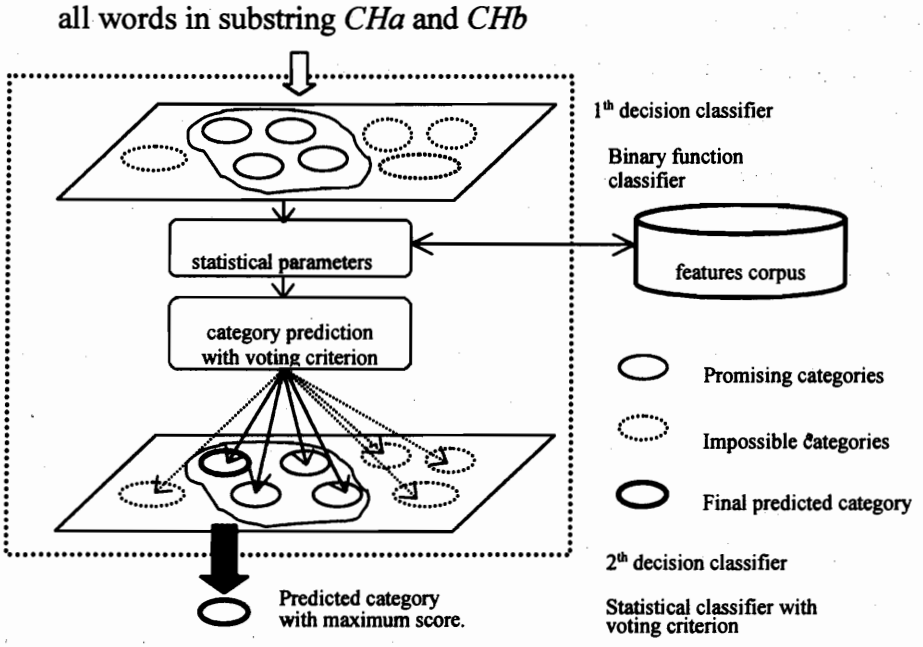


Figure 2: Multiple decision classifiers contain two classifiers, which are merged together with *multiply* operation.

The function of multiple decision classifiers can be described as follow:

Suppose that $C$ denotes the sentence with non-text symbols, $\Phi_1$ and $\Phi_2$ denote the 1st and 2nd classifier respectively. *set* is the set containing all promising categories induced by 1st classifier. $\Phi$ denotes the multiple decision classifiers, which is composed of the 1st decision tree classifier and 2nd statistical decision classifier merging with *multiply* operation. $TS(\bullet)$ will compute the total score for all categories based on the voting criterion and statistical parameters schemes.

$$\Phi_1(C) = set, \tag{1}$$

$$\Phi_2(C) = \mathrm{TS}(\Omega_j), \quad \Omega_j \in set \tag{2}$$

$$\Phi(C) = \Omega_{j^*}, \quad \exists! \Omega_{j^*} \in set \text{ and } \mathrm{TS}(\Omega_{j^*}) = \arg\max_{j=1,2,\ldots J} \mathrm{TS}(\Omega_j) \tag{3}$$

where j is the number of category for non-text symbols.

3.2 The Binary Function Classifier based on Decision Tree

The decision tree classifier plays as a binary logical function, which is to induce all promising categories for the non-text symbol based on Mandarin linguistic knowledge. The classifier will assign probability value 1 to the promising categories. On the other hand, some categories will be excluded and assigned a probability value 0. For example, the pattern of "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 pattern of "14/2" and "SUN4/75" could not belong to the category *date* and *tempo,* all these categories will not be the possible category for non-text symbol and be assigned a probability value 0.

A successive answers to questions: $Q_1, Q_2, \cdots, Q_n$, which are the questions about the syntax and semantic meaning for left and right neighbor (tokens or words) of non-text symbol in sentence, will decide which path should trace into based on the linguistic knowledge. Finally, one leaf node in decision tree will be reached and a *set* of categories will be contained. Within the *set*, all the categories will be assigned a probability value 1 while all 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 making decision tree. 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:

$$\underset{i=1,2}{P_i}(\Omega_j) = \begin{cases} 0 & \text{if i} \in 1 \text{ and } \Omega_j \notin set. \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

where i=1,2, …,I. i is labeled as the ith decision classifier. I is the number of total decision classifiers (currently, we have developed two decision classifiers, so I=2). If we have $J$ categories $\Omega_1\Omega_2 \cdots \Omega_J$ and $\Omega_j$ denotes the category j for non-text symbols. $P_i(\Omega_j)$ is the probability value of category j for the ith classifier . *set* is induced from the decision tree classifier and contains all promising categories. These promising categories will be passed into 2nd decision classifier further, one of which will be the final predicted category. First classifier plays as a binary function in our approach. So, Equation (4) can be explained further as follow: if i=1 and $\Omega_j \notin set$ , then $P_1(\Omega_j) = 0$. Otherwise, $P_i(\Omega_j) = 1$ .

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 possible categories. It is so apparent that processing of adopting decision tree can improve the precision rate.

3.3 The Statistical Decision Classifier with voting criterion

The segmentation task of testing phase adopts same criterions as that in training phase shown in section 2.3. 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 feature corpus) respectively. It is called the **voting criterion**.

Based on the *voting criterion,* each word in $CHa$ and $CHb$ have a statistical probability value, which looks like the voting 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 in our approach) will become to be the predicted category. In our voting criterions, three scoring schemes are proposed: which are the *preference scoring* and the *winner-take-all* criterion. These voting criterions will be implemented and compared with each others to find which one can achieves the best empirical results.

3.3.1 Voting criterion with preference scoring

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:

$$\begin{aligned} C &= CH_a + N + CH_b \\ CH_a &= w_{a_1} w_{a_2} \cdot \cdot \cdot w_{a_j} \cdot \cdot \cdot \mathrm{w}_{a_m} \\ CH_b &= w_{b_1} w_{b_2} \cdot \cdot \cdot w_{b_j} \cdot \cdot \cdot \mathrm{w}_{b_m} \end{aligned} \tag{5}$$

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 different non-text symbols. Also, $CHa$ and $CHb$ may be an empty substring.

For each word in $CHa$ and $CHb$, the word appearance probability appearing in category j of non-text symbol can be computed based on three different statistical parameters scheme: which are word-based, category-based and corpus-based . In this work, the word appearance probability can be considered as the probability the word may appear in certain category for non-text symbol. The appearance probability can be regarded as a score for each word in $CHa$ and $CHb$ to vote for each category of non-text symbol further.

There are three statistical probability schemes, on which the value can be considered as the probability for each word to appear in each category.

**(1) word-based statistical probability**

For all words in $CHa$ and $CHb$, the appearance probability score $S_a$ and $S_b$ of each word voting for category j $(\Omega_j)$of non-text symbol can be computed as:

$$S_a(w_{ak_1} \mid \Omega_j) = \frac{C_a(w_{ak_1} \mid \Omega_j)}{TN_a(w_{ak_1})}, \quad S_b(w_{bk_2} \mid \Omega_j) = \frac{C_b(w_{bk_2} \mid \Omega_j)}{TN_b(w_{bk_2})} \tag{6}$$

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 $k_1^{\text{th}}$ and $k_2^{\text{th}}$ word in $CHa$ and $CHb$. $C_a(w_{ak_1} \mid \Omega_j)$ and $\mathrm{C}_b(w_{bk_2} \mid \Omega_j)$ are the occurrence of $w_{ak_1}$ and $w_{bk_2}$ for category j of non-text symbol. $TN_a(w_{ak_1})$ and $TN_b(w_{bk_2})$ stand for the total frequency of $w_{ak_1}$ and $w_{bk_2}$ within features corpus with respect to the location proceeding and following non-text symbol, which can be computed as follow:

$$TN_a(w_{ak_1}) = \sum_{j=1}^{J} C_a(w_{ak_1} \mid \Omega_j), \quad TN_b(w_{bk_2}) = \sum_{j=1}^{J} C_b(w_{bk_2} \mid \Omega_j) \tag{7}$$

$$\sum_{j=1}^{J} S_a(w_{ak_1} \mid \Omega_j) = 1, \quad \sum_{j=1}^{J} S_b(w_{bk_2} \mid \Omega_j) = 1 \tag{8}$$

Based on the definition above, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ can be considered as the probability value in which the $w_{ak_1}$ and $w_{bk_2}$ will appear in the category j . As the result, our voting criterions are based on this probability value.

In the paper, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ stand for the suffrage for each word (voter) to vote for certain category j $(\Omega_j)$.

**(2) category-based statistical probability**

With respect to Equation (6), the denumerator will be computed based on the total occurrence for the all words which appear in category j$(\Omega_j)$ . Equation (8) can't hold in this scheme.

**(3) corpus-based statistical probability**

With respect to Equation (6), the denumerator will be computed based on the total occurrence for the all words which appear in feature corpus. Equation (8) can't hold in this scheme.

For the 2nd decision classifier, the total score $TS_a$ and $TS_b$ for all words in substring *CHa* and *CHb* to vote for categories j of non-text symbol can be computed.

The overall total score TS of 1st and 2nd decision classifier for category j is computed with the *multiply* operation:

$$\underset{\text{j=1,2,...J}}{TS(\Omega_j)} = P_1(\Omega_j) * P_2(\Omega_j) * TS(\Omega_j), \quad \Omega_j \in set \tag{9}$$

where $P_2(\Omega_j)$ denotes the probability value of category j $(\Omega_{j^*})$ in the 2nd classifier. In our approach, $P_2(\Omega_j) = 1$, j = 1,2,....,J. *set* is composed of all the promising categories induced by 1st decision tree classifier.

$$TS(\Omega_{j^*}) = \underset{j=1,2,...,J}{\arg\max}\,(TS(\Omega_j)) \tag{10}$$

where $TS(\Omega_{j^*})$ will return the maximum score subject to category j* $(\Omega_{j^*})$ based on 1st decision classifier and 2nd statistical decision classifier. $TS(\Omega_{j^*})$ will be used in Equation (3) for the multiple decision classifiers to predict the final category j*$(\Omega_{j^*})$.

3.3.2 Voting criterion with winner-take-all scoring

In construct to the preference scoring criterion above, the Voting with *winner-take-all* adopts a different scoring rule. For each word in *CHa* and *CHb*, $S_a(w_{ak_1} \mid \Omega_j)$ and $S_b(w_{bk_2} \mid \Omega_j)$ will have the total parameter score 1 of category j* for word $w_{ak_1}$ and $w_{bk_2}$ and assigned a score value 1. $S_a$ (similar to $S_b$) in Equation (6) should be changed as follow:

$$S_a(w_{ak_2} \mid \Omega_{j^*}) = \begin{cases} 1 & \exists! \Omega_{j^*} \in \Omega \text{ and } S_a(w_{ak_1} \mid \Omega_{j^*}) = \underset{j=1,2,...J}{\arg\max}(S_a(w_{ak_1} \mid \Omega_j)) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Within the classification phase, some sentences could be mispredicted. To make the statistical decision classifier more robust, these sentences can feedback into category process in training phase and adopt parameters in features corpus. The feedback usually cab solve the unseen events in source, the situation appear often in natural language.

3.4 Unknown word

There are a lot of words in natural language, usually more several ten thousands. New lexicons or tokens will be generated in near future. Within natural language processing (NLP), it is so hard to collect all the words. In our paper, the so-called unknown words can be considered that the words do not appear in our corpus (feature corpus), which have been generated in the training phase. It is so apparent that the distribution and total number of collected word will affect the statistical parameters seriously, especially on the statistical models. Another situation is the data sparsity. The smoothing techniques can resolve the situation.

Based on the ASBC and ASCED corpus, the ASBC source is divided into four groups. we compute the total frequency and number of words in these four groups to derive the relation, in which we can predict the probability unknown words. The fitting regression curve can be employed to estimate the probability for unknown words. $Y(X) = aX^2 + bX + c$. We can find the derivative of $Y(X_j)$. Within classification phase, Value $X_j$ represent the total frequency of collected words in feature corpus for category j ($\Omega_j$). The first order derivative of $Y(X_j)$ can be considered that the probability of unknown word in category j ($\Omega_j$) . Such probability will be used as voting score for unknown words to vote for category j.

3.5 Translate Oral Expansion

The output of multiple decision classifiers is the unique predicted category. Based on the category, the non-text symbol can be translated into its oral expression of text in which the category has been predicted by testing phase. Sentence (C) contains a non-text symbol "/", which is predicted as the *date* category and the pattern of "4/10" in (C) will be translated into the oral expression "四月十日" in sentence (C'). The output text of this phase will be processed further with text linguistic analysis in TTS system.

(C) 這本雜誌已於上週六（4／10）出版。

This magazine was published last Saturday (April tenth).

(C') 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 r1 chu1 bian3.

## 4 Implementation and Evaluation

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. Two decision tree classifiers have been generated. We evaluate the results of inside test and outside test for 2nd *statistical classifier* with two different *voting criterions*, then we combined it with *decision tree classifier* to compare the performance of precision rate. The precision rate is

defined as:

**Precision rate(PRs)** = # of correct prediction categories / total # of non-text symbol (12)

4.1 Evaluation only for statistical decision classifier

The results for 2nd classifier with different voting score criterion and statistical parameters are listed in Table 1. Total number of non-text symbol "/" for inside and outside test are 564 and 202 respectively.

4.2 Evaluation for merging two decision classifiers together.

Under the multiple decision classifier structure, the 1st and 2nd decision classifier are merged together to improve the overall precision rate. exploiting the 1st classifier to exclude some impossible categories first, the results are attractive and listed in Table 1 also. As shown, the final results of inside test and outside test is 97.3% and 92.9%, which are obtained by merging the 1st and 2nd classifier with voting criterion of preference score and category-based statistical parameters in 2nd classifier.

Table 1: The overall precision rate of inside test and outside test of 2nd statistical decision classifier for symbol "/"

| Precision rate(%) | multiple decision classifier, merging or not? | 2nd decision classifier, word-based statistical scheme | | | |
|---|---|---|---|---|---|
| | | voting with preference score | | winner-take-all score | |
| | | inside test | outside test | inside test | outside test |
| word-based statistical scheme | without 1st classifier | 95.4 | 86.3 | 85.9 | 77.3 |
| | with 1st classifier | 96.2 | 91.2 | 90.5 | 85.7 |
| category-based statistical scheme | without 1st classifier | 96.0 | 92.8 | 92.9 | 84.8 |
| | with 1st classifier | 97.3* | 92.9* | 96.1 | 89.4 |
| corpus-based statistical scheme | without 1st classifier | 95.5 | 86.1 | 89.2 | 81.1 |
| | with 1st classifier | 96.3 | 89.9 | 90.1 | 85.5 |

Table 2 is the results for non-text symbol " : ", based on the preference score voting criterion and word-base statistical parameters. The average rate of inside testing and outside testing are 97.8% and 93.0%. Notation of *N* in Table 2 stand for non-text symbol. The total word occurrence for non-text symbol " : " is 14406.

Table 2: The overall precision rate of inside test and outside test for non-text symbol " : ", the 1st and 2nd decision classifier merging.

| multiple decision classifier | voting with preference score , word-base statistical parameters | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | inside testing | | | | | | | outside testing | | | | | | |
| category | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| PRs rate(%) | 99 | 100 | 98 | 95 | 100 | 100 | 97 | 86 | 100 | 100 | 88 | 100 | 78 | 97 |
| Total no. of *N* | 272 | 105 | 126 | 21 | 85 | 35 | 351 | 68 | 31 | 30 | 8 | 22 | 9 | 83 |

## 5 Conclusion and Future Works

In the paper, we have developed an effective approach, which can classify the semantic category of patterns containing non-text symbols and resolve the category ambiguity 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 semantic and linguistic knowledge for statistical parameters and features. Currently, we have developed two decision classifiers: one is based on the decision tree to induce promising categories the other is on the statistical decision classifier with two voting criterion with word-based, category-based and corpus-based statistical parameter schemes. Final precision rate of inside and outside test achieves the performance of 97.8% and 93.0% 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. The topics which should be researched further in the future include:

1) Patterns of special and frequent cases for non-text symbols in text.
2) The extraction training parameters and learning algorithms.
3) The POS of word and smoothing techniques for unknown words.
4) Expand the current two classifiers into more classifiers to resolve complicated linguistic classification problem.

## References


- 黃居仁等, *中央研究院平衡語料庫簡介*, Proceeding of ROCLLING VII, pp. 81-99, 1995.
- P. Brown, S. Della Pietra, V. Della Pietra and R. Mercer. *Word Sense Disambiguation Using Statistical Methods*. In Proceeding of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, pp. 264- 270, 1991.
- A. R. Golding, *A Bayesian hybrid method for Context-Sensitive Spelling Correction*, In Proceedings of the third workshop on Very Large Corpora, pp. 39-53, Boston, USA, 1995.
- B. Merialdo, *Tagging Text with a Probabilistic Model*, In Proceeding of the IBM Natural Language ITL, Paris, France, pp. 161-172, 1990.
- V. Rodova and J. Psutka, An Approach to Speaker Identification Using Multiple Classifiers, ICASSP, 1997,pp. 1135-1139.
- David Yarowsky, *Homograph Disambiguation in Text-to Speech Synthesis*, Book of Progress in Speech Synthesis, 1997, chapter 12,pp 157-172.


Appendix A: 8 categories and its related oral expression for non-text symbol slash "／".

| category | some lexical patterns with non-text symbol "／" | 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 | ／ｄｅｖ／ｎｕｌｌ | 根目錄ｄｅｖ斜線ｎｕｌｌ |
| 5, computer words | Ｉ／Ｏ | silence(or 斜線) |
| 6. production version | ＶＡＸ／ＶＭＳ | silence(longer pause or 斜線) |
| 7. frequent words in | ＴＣＰ／ＩＰ | silence(or 斜線) |
| 8. others | 中／日／韓文著錄 | silence(longer pause) |