# Statistical Models for Deep-structure Disambiguation

TungHui Chiang* and Keh-Yih Su**

*Advanced Technology Center
Computer and Communication Research Laboratories,
Industrial Technology Research Institute
Hsinchu, Taiwan 310, R.O.C.
Email: thchiang@e0sun3.ccl.itri.org.tw

**Department of Electrical Engineering
National TsingHua University
Hsinchu, Taiwan 300, R.O.C.
Email: kysu@bdc.com.tw

## Abstract

In this paper, an integrated score function is proposed to resolve the ambiguity of deep-structure, which includes the cases of constituents and the senses of words. With the integrated score function, different knowledge sources, including part-of-speech, syntax and semantics, are integrated in a uniform formulation. Based on this formulation, different models for case identification and word-sense disambiguation are derived. In the baseline system, the values of parameters are estimated by using the maximum likelihood estimation method. The accuracy rates of 56.3% for parse tree, 77.5% for case and 86.2% for word sense are obtained when the baseline system is tested on a corpus of 800 sentences. Afterwards, to reduce the estimation error caused by the maximum likelihood estimation, the Good-Turing's smoothing method is applied. In addition, a robust discriminative learning algorithm is also derived to minimize the testing set error rate. By applying these algorithms, the accuracy rates of 77% for parse tree, 88.9% for case, and 88.6% for sense are obtained. Compared with the baseline system; 17.4% error reduction rate for sense discrimination, 50.7% for case identification, and 47.4% for parsing accuracy are obtained. These results clearly demonstrate the superiority of the proposed models for deep-structure disambiguation.

## 1 Introduction

For many natural language processing tasks, e.g., machine translation, systems usually require to apply several kinds of knowledge to analyze input sentence and represent the analyzed results in terms of a deep structure which identify the thematic roles (cases) of constituents and the senses of words. However, ambiguity and uncertainty exist at the different levels of analysis. To resolve the ambiguity and uncertainty, the related knowledge sources should be properly represented and integrated. Conventional approaches to case identification usually need a lot of human efforts to encode ad hoc rules [1,2,3]. Such a rule-based system is, in general, very expensive to construct and difficult to maintain. In contrast, a statistics-oriented corpus-based approach achieves disambiguation by using a parameterized model, in which the parameters are estimated and tuned from a training corpus. In such a way, the system can be easily scaled up and well trained based on the well-established theories.

However, statistical approaches reported in the literature [4,5,6,7] usually use only surface level information, e.g., collocations and word associations, without taking structure information, such as syntax and thematic role, into consideration. In general, the structure features that

characterize long-distance dependency, can provide more relevant correlation information between words. Therefore, word association information can be trained and applied more effectively by considering the structural features. In many tasks, such as natural language understanding and machine translation, deep-structure information other than word sense is often required. Nevertheless, few research was reported to provide both thematic role and word sense information with statistical approach.

Motivated by the above concerns, an integrated score function, which encodes lexical, syntactic and semantic information in a uniform formulation is proposed in this paper. Based on the integrated score function, the lexical score function, the syntactic score function, and the semantic score function are derived. Accordingly, several models encoding structure information in the semantic score formulation are proposed for case identification and word-sense discrimination.

To minimize the number of parameters needed to specify the deep-structure, a deep-structure representation form, called **normal form** which adopts "*predicate-argument*" style, is used in our system. By using this normal form representation, the senses of content words and the relationships among constituents in a sentence can be well specified. The normal form used here is quite generalized and flexible; therefore, it is also applicable in other tasks.

When the parameters of the proposed score function are estimated with the maximum likelihood estimation (MLE) method, the baseline system achieves parsing accuracy rate of 56.3%, case identification rate of 77.5%, and 86.2% accuracy rate of word sense discrimination. Furthermore, to reduce the estimation error resulting from the MLE, Good-Turing's smoothing method is applied; significant improvement is obtained with this parameter smoothing method. Finally, a robust discriminative learning algorithm is derived in this paper to minimize the testing set error, and very promising results are obtained with this algorithm. Compared with the baseline system; 17.4% error reduction rate for sense discrimination, 50.7% for case identification, and 47.4% for parsing accuracy are obtained. These results clearly demonstrate the superiority of the proposed models for deep-structure disambiguation.

## 2 The Integrated Score Function

The block diagram of the deep-structure disambiguation system is illustrated in Figure 1. As shown, the input word sequence is first tagged with the possible part-of-speech sequences. A word sequence would, in general, correspond to more than one part-of-speech sequence. The parser analyzes the part-of-speech sequences and then produces corresponding parse trees. Afterwards, the parse trees are analyzed by the semantic interpreter, and various interpretations represented by the normal form are generated. Finally, the proposed integrated score function is adopted to select the most plausible normal form as the output. The formulation of the scoring mechanism is derived as follows.

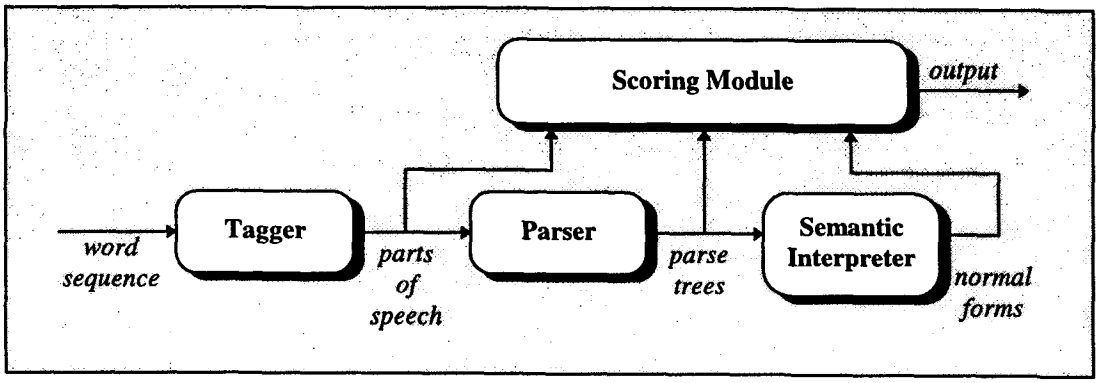


*FIGURE 1. Block diagram of the deep-structure disambiguation system*

For an input sentence, say $W$, of $n$ words $w_1, w_2, \cdots, w_n$, the task of deep-structure

disambiguation is formulated to find the best normal form $\hat{\mathbf{N}}$, parse tree $\hat{\mathbf{L}}$, and parts of speech $\hat{\mathbf{T}}$, such that

$$\left(\hat{\mathbf{N}},\hat{\mathbf{L}},\hat{\mathbf{T}}\right)=\arg\max_{\mathbf{N}_i,\mathbf{L}_j,\mathbf{T}_k} P\left(\mathbf{N}_i,\mathbf{L}_j,\mathbf{T}_k|W\right),$$

where $\mathbf{N}_i$, $\mathbf{L}_j$, $\mathbf{T}_k$ denote the $i$-th normal form, the $j$-th parse tree and the $k$-th part-of-speech sequence, respectively; $P\left(\mathbf{N}_i,\mathbf{L}_j,\mathbf{T}_k|W\right)$ is called the *integrated score function*. For computation, the integrated score function is further decomposed into the following equations.

$$\begin{aligned}P\left(\mathbf{N}_i,\mathbf{L}_j,\mathbf{T}_k|W\right)&=P\left(\mathbf{N}_i|\mathbf{L}_j,\mathbf{T}_k,W\right)\times P\left(\mathbf{L}_j|\mathbf{T}_k,W\right)\times P\left(\mathbf{T}_k|W\right)\\&=S_{sem}\left(\mathbf{N}_i\right)\times S_{syn}\left(\mathbf{L}_j\right)\times S_{lex}\left(\mathbf{T}_k\right),\end{aligned}$$

where $S_{sem}\left(\mathbf{N}_i\right), S_{syn}\left(\mathbf{L}_j\right), S_{lex}\left(\mathbf{T}_k\right)$ stand for the *semantic score* function, *syntactic score* function, and *lexical score* function, respectively; they are defined as follows:

$$S_{sem}\left(\mathbf{N}_i\right)=P\left(\mathbf{N}_i|\mathbf{L}_j,\mathbf{T}_k,W\right)$$

$$S_{syn}\left(\mathbf{L}_j\right)=P\left(\mathbf{L}_j|\mathbf{T}_k,W\right)$$

$$S_{lex}\left(\mathbf{T}_k\right)=P\left(\mathbf{T}_k|W\right).$$

The derivations of these score function are addressed as follows.

## 2.1 The Lexical Score

The lexical score for the $k$-th lexical (part-of-speech) sequence $\mathbf{T}_k$ associated with the input word sequence $W$ is expressed as follows:

$$\begin{aligned}S_{lex}\left(\mathbf{T}_k\right)&=P\left(\mathbf{T}_k|W\right)=P\left(t_{k,1}^{k,n}|w_1^n\right)\\&=\frac{P\left(w_1^n|t_{k,1}^{k,n}\right)\times P\left(t_{k,1}^{k,n}\right)}{P\left(w_1^n\right)},\end{aligned}$$

where $t_{k,i}$, denoting the $i$-th part-of-speech in $\mathbf{T}_k$, stands for the part-of-speech assigned to $w_i$. Since $P\left(w_1^n\right)$ is the same for all possible lexical sequences, this term can be ignored without affecting the final disambiguation results. Therefore, $S_{lex}^*\left(\mathbf{T}_k\right)\left(\equiv P\left(w_1^n|t_{k,1}^{k,n}\right)\times P\left(t_{k,1}^{k,n}\right)\right)$ instead of $S_{lex}\left(\mathbf{T}_k\right)$ is used in our implementation. Like the standard trigram tagging procedures, the lexical score $S_{lex}^*\left(\mathbf{T}_k\right)$ is expressed as follows:

$$\begin{aligned}S_{lex}^*\left(\mathbf{T}_k\right)&=P\left(w_1^n|t_{k,1}^{k,n}\right)\times P\left(t_{k,1}^{k,n}\right)\\&\approx\prod_{i=1}^{n}P\left(t_{k,i}|t_{k,i-1},t_{k,i-2}\right)\times P\left(w_i|t_i\right).\end{aligned}$$

## 2.2 The Syntactic Score

The tree in Figure 2 is used as an example to explain the syntactic score function. The basic derivation of the syntactic score includes the following steps.

●First, the tree is decomposed into a number of phrase levels, such as $L_1, L_2, \cdots, L_8$ in Fig. 2.

- Secondly, the transition between phrase levels is formulated as a *context-sensitive* rewriting process. With the formulation, each transition probability between two phrase levels is calculated by consulting a finite-length window that comprises the symbols to be reduced and their left and right contexts.

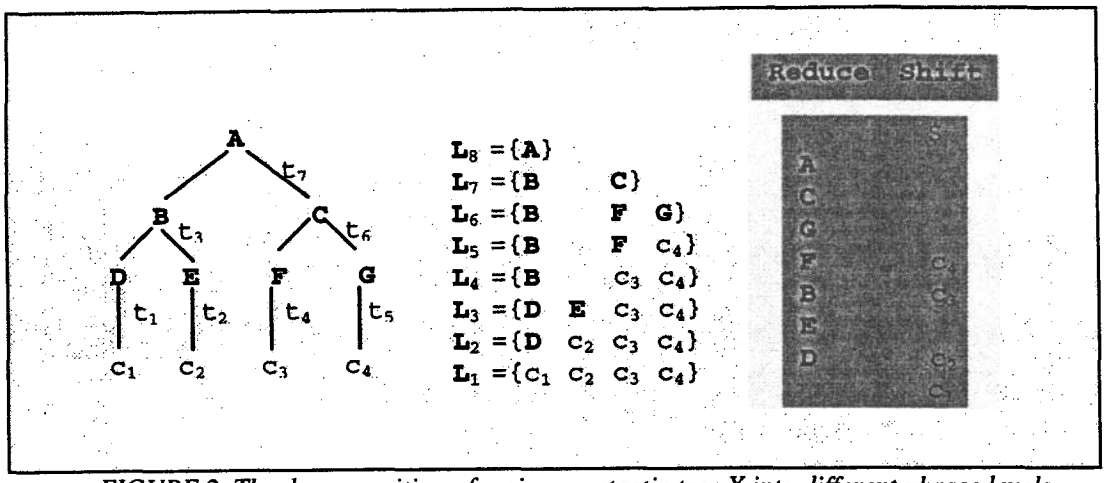


*FIGURE 2. The decomposition of a given syntactic tree X into different phrase levels.*

Let the label $t_i$ in Fig. 2 be the time index for the $i$-th state transition, which corresponds to a reduce action, and $L_i$ be the $i$-th phrase level. Then the syntactic score of the tree $\mathbf{L}_A$ in Figure 2 is defined as follows [8,9]:

$$
\begin{aligned}
&S_{syn}(\mathbf{L}_A) = P(L_2^8 | L_1) \\
&= P(L_8, L_7, L_6 | L_1^5) \times P(L_5 | L_1^4) \times P(L_4, L_3 | L_1^2) \times P(L_2 | L_1) \\
&\approx P(L_8, L_7, L_6 | L_5) \times P(L_5 | L_4) \times P(L_4, L_3 | L_2) \times P(L_2 | L_1) \\
&\approx P(L_8 | L_5) \times P(L_5 | L_4) \times P(L_4 | L_2) \times P(L_2 | L_1) \\
&\approx P(A | \{\phi, B, F\}, c_4 \{\$\}) \times P(F | \{\phi, B\}, c_3, \{c_4, \$\}) \times P(B | \{\phi, D\}, c_2, \{c_3, c_4, \$\}) \times P(D | \{\phi\}, c_1, \{c_2, c_3, c_4, \$\}) \\
&\approx P(A | l_4, c_4, r_4) \times P(F | l_3, c_3, r_3) \times P(B | l_2, c_2, r_2) \times P(D | l_1, c_1, r_1),
\end{aligned}
$$

where $\phi$ and \$ correspond to the begin-of-sentence and the end-of-sentence symbols, respectively; $l_i$ and $r_i$ stand for the left and right contextual symbols to be consulted in the $i$-th phrase level. If **M** number of left contextual symbols and **N** number of right contextual symbols are consulted in computation, the model is said to operate in the $\mathbf{L_M R_N}$ mode.

Note that each pair of phrase levels in the above equation corresponds to a change in the LR parser's stack before and after an input word is consumed by a shift operation. Because the total number of *shift* actions, equal to the number of product terms in the above equation, is always the same for all alternative syntactic trees, the normalization problem is resolved in such a formulation. Moreover, the syntactic score formulation provides a way to consider both *intra-level context-sensitivity* and *inter-level correlation* of the underlying context-free grammar. With such a formulation, the capability of *context-sensitive* parsing (in probabilistic sense) can be achieved with a *context-free* grammar.

## 2.3 The Semantic Score

To simplify the computation of the semantic score, a structure normalization procedure

is taken beforehand by the semantic interpreter to convert a parse tree into an intermediate normal form, called **normal form one (NF1)**, which preserves all relevant information for identification of cases and word senses. The implementation of the normalization procedure includes a syntactic normalization procedure and a semantic normalization. procedure.

In the syntactic normalization procedure, many parse trees that are syntactically equivalent should be normalized first. Such syntactic variants may result from a writing convention, function words, or non-discriminative syntactic information, such as punctuation markers. Excessive nodes for identifying the various bar levels in the phrase structure grammar are also deleted or compacted.

Afterwards, different syntactic structures that are semantically equivalent are normalized to the desired **normal form (NF)** structure. In the NF representation, the tense, modal, voice and type information of a sentence are extracted as features. By taking the sentence "*To meet spectrum-analyzer specification, allow a 30-min warm-up before making any measurement.*" as an example, the parse tree, NF1, and the desired normal form structure are illustrated in Figure 3.

To compute the semantic score, the normal form is first decomposed into a series of production rules in a *top-down* and *leftmost first* manner, where each decomposed production rule corresponds to a "*case subtree*". For instance, the normal form in Figure 3(c) is decomposed into a series of case subtrees , where

$\Gamma_1$: PROP $\rightarrow$ PURP VACTN GOAL TIME
$\Gamma_2$: PURP $\rightarrow$ VSTAT GOAL
$\Gamma_3$: GOAL $\rightarrow$ HEAD HEAD
$\Gamma_4$: GOAL $\rightarrow$ HEAD HEAD
$\Gamma_5$: TIME $\rightarrow$ VACTN THEME
$\Gamma_6$: GOAL $\rightarrow$ QUAN HEAD.

Similarly, the NF1 structure is also decomposed into another set of production rules, each of which corresponds to a *Normal Form One (NF1) subtree*. For example, the NF1 structure in Figure 3(b) is decomposed into the following NF1 subtrees:

$E_1$: S $\rightarrow$ SS* v NP SS**
$E_2$: SS* $\rightarrow$ v NP
$E_3$: NP $\rightarrow$ n n
$E_4$: NP $\rightarrow$ n n
$E_5$: SS** $\rightarrow$ v NP
$E_6$: NP $\rightarrow$ quan n.

In such a way, the semantic score can be defined in terms of the case subtrees and the NF1 subtrees.

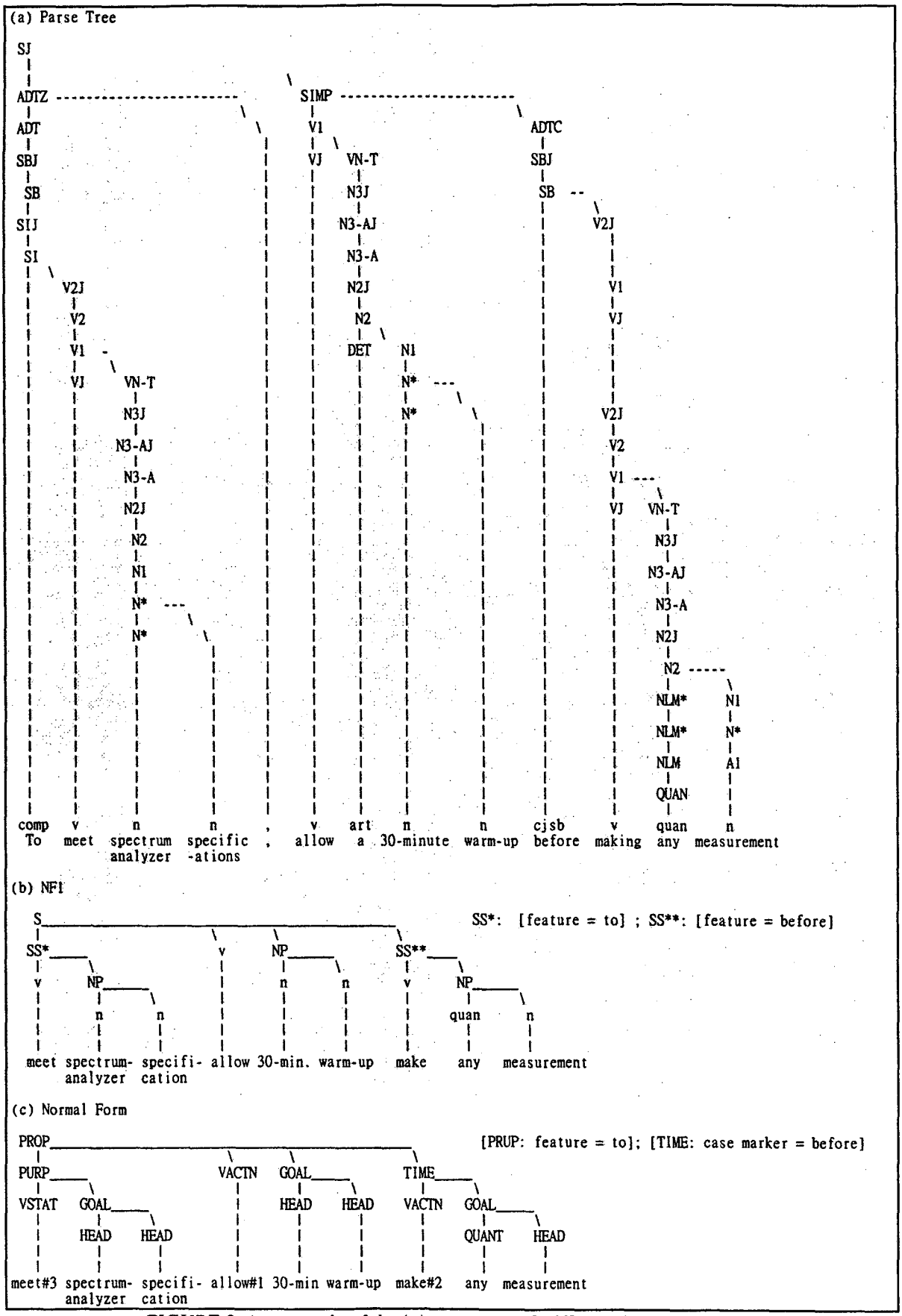


*FIGURE 3. An example of the (a) parse tree, (b) NF1 and (c) normal form.*

Formally, regarding the NF1 alternatives, the semantic score $S_{sem}(\mathbf{N}_i)$ can be expressed as follows:

$$\begin{aligned} S_{sem}(\mathbf{N}_i) &= P(\mathbf{N}_i|\mathbf{L}_j,\mathbf{T}_k,W) \\ &= \sum_{\Phi} P(\mathbf{N}_i,\Phi|\mathbf{L}_j,\mathbf{T}_k,W) \\ &= \sum_{\Phi} P(\mathbf{N}_i|\Phi,\mathbf{L}_j,\mathbf{T}_k,W)\times P(\Phi|\mathbf{L}_j,\mathbf{T}_k,W) \end{aligned}$$

where $\Phi$ denotes the possible NF1 structures with respect to $\mathbf{N}_i$ and $\mathbf{L}_j$. Theoretically, a parse trees may be normalized into more than one NF1 structure; however, this happens seldom in our case. That is, it is almost true that the normalization procedure can be considered as a one-to-one mapping, which indicates $P(\Phi|\mathbf{L}_j,\mathbf{T}_k,W)\approx 1$ in our task. Under this assumption, the semantic score can be simplified as:

$S_{sem}(\mathbf{N}_i)\approx P(\mathbf{N}_i|\Phi_j,\mathbf{L}_j,\mathbf{T}_k,W)$.

Since the normal form comprises the cases of constituents and the senses of content words, the representation of the normal form can be thus rewritten as $\mathbf{N}_i=\{s_{i,1}^{i,n},\Gamma_{i,1}^{i,M_i}\}$, where $s_{i,1}^{i,n}$ is the word senses corresponding to $W(=w_1^n)$; $\Gamma_{j,1}^{j,M_i}=\{\Gamma_1,\Gamma_2,\cdots,\Gamma_{M_i}\}$ is the $M_i$ case subtrees which define the structure of the normal form $\mathbf{N}_j$. In such a way, the semantic score is rewritten as follows:

$$\begin{aligned} S_{sem}(\mathbf{N}_i) &\approx P(\mathbf{N}_i|\Phi_j,\mathbf{L}_j,\mathbf{T}_k,W) \\ &= P(s_{i,1}^{i,n},\Gamma_{i,1}^{i,M_i}|\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n) \\ &= P(s_{i,1}^{i,n}|\Gamma_{i,1}^{i,M_i},\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n) \\ &\quad\times P(\Gamma_{i,1}^{i,M_i}|\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n) \\ &= S_{sense}(s_{i,1}^{i,n})\times S_{case}(\Gamma_{i,1}^{i,M_i}), \end{aligned}$$

where $L_{j,1}^{j,N_j}=\{L_1,L_2,\cdots,L_{N_j}\}$ corresponds to the $N_j$ sentential forms (phrase levels) with respect to the parse tree $\mathbf{L}_j$. $\Phi_{j,1}^{j,M_i}=\{\Phi_1,\Phi_2,\cdots,\Phi_{M_i}\}$ stands for the NF1 subtrees transformed from $L_{j,1}^{j,N_j}$ respectively. $S_{sense}(s_{i,1}^{i,n})=P(s_{i,1}^{i,n}|\Gamma_{i,1}^{i,M_i},\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n)$ is the *word-sense score* and $S_{case}(\Gamma_{i,1}^{i,M_i})=P(\Gamma_{i,1}^{i,M_i}|\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n)$ is the *case score*.

Different models for case identification and word-sense disambiguation are further derived below.

- **Case Identification Models**

To derive the case identification model, it is assumed that the required information for case identification from the parse tree $L_{j,1}^{j,N_j}$, parts-of-speech $t_{k,1}^{k,n}$, and the word $w_1^n$ has be represented by the NF1. Based on this assumption, the case score, $S_{case}(\Gamma_{i,1}^{i,M_i})$, is thus approximated as follows:

$$S_{case}\left(\Gamma_{i,1}^{i,M_i}\right) = P\left(\Gamma_{i,1}^{i,M_i}\middle|\Phi_{j,1}^{j,M_i}, L_{j,1}^{j,N_j}, t_{k,1}^{k,n}, w_1^n\right)$$
$$\approx P\left(\Gamma_{i,1}^{i,M_i}\middle|\Phi_{j,1}^{j,M_i}\right)$$
$$= \prod_{m=1}^{M_i} P\left(\Gamma_{i,m}\middle|\Gamma_{i,1}^{i,m-1}, \Phi_{j,1}^{j,M_i}\right).$$

Again, the number of parameters required to model such a formulation is still too many to afford, unless more assumptions are made.

Since the decomposition of the normal form structures has been carried out in the *top-down* and *leftmost-first* manner, the case subtree $\Gamma_{i,m}$ depends on its previously decomposed case subtrees, which are either the siblings or the ancestors of the subtree $\Gamma_{i,m}$. Therefore, in addition to the NF1 representation $\Phi_{i,1}^{i,M_i}$, the determination of cases in the case subtree $\Gamma_{i,m}$ is assumed to be highly dependent on its ancestors and siblings. In computation, if **N** number of ancestors and **R** number of siblings of $\Gamma_{i,m}$ have been consulted, the case score function is approximated as:

$$S_{case}\left(\Gamma_{i,1}^{i,M_i}\right) = \prod_{m=1}^{M_i} P\left(\Gamma_{i,m}\middle|\Gamma_{i,1}^{i,m-1}, \Phi_{j,1}^{j,M_i}\right)$$
$$\approx \prod_{m=1}^{M_i} P\left(\Gamma_{i,m}\middle|\left\{\Gamma_{A_1}, \Gamma_{A_2}, \cdots, \Gamma_{A_N}\right\}, \left\{\Gamma_{S_1}, \Gamma_{S_2}, \cdots, \Gamma_{S_R}\right\}, \Phi_{j,1}^{j,M_i}\right),$$

where $\Gamma_{A_i}$ and $\Gamma_{S_j}$ denote the *i*-th ancestor and the *j*-th sibling of $\Gamma_{i,m}$, respectively. A model using this case score function is hereby said to operate in an $\mathbf{A_N S_R}$ mode. For example, when the model is operated in $A_1S_0$ mode, the case score of the normal form in the previous example is expressed as:

$$S_{case}^{(A_1S_0)}\left(\Gamma_{i,1}^{i,M_i}\right) = P\left(\Gamma_1|\Phi_1\right) \times P\left(\Gamma_2|,\Gamma_1,\Phi_2\right) \times P\left(\Gamma_3|,\Gamma_2,\Phi_3\right)$$
$$\times P\left(\Gamma_4|,\Gamma_1,\Phi_4\right) \times P\left(\Gamma_5|,\Gamma_1,\Phi_5\right) \times P\left(\Gamma_6|,\Gamma_5,\Phi_6\right)$$
$$= P\left(\text{PROP} \rightarrow \text{PURP VACTN GOAL TIME} \middle| S \rightarrow SS^* \, v \, NP \, SS^{**}\right)$$
$$\times P\left(\text{PURP} \rightarrow \text{VSTAT GOAL} \middle| \begin{matrix}\text{PROP} \rightarrow \text{PURP VACTN GOAL TIME} \\ SS^* \rightarrow v \, NP\end{matrix}\right)$$
$$\times P\left(\text{GOAL} \rightarrow \text{HEAD HEAD} \middle| \begin{matrix}\text{PURP} \rightarrow \text{VSTAT GOAL} \\ NP \rightarrow nn\end{matrix}\right)$$
$$\cdots$$
$$\times P\left(\text{THEME} \rightarrow \text{QUAN HEAD} \middle| \begin{matrix}\text{TIME} \rightarrow \text{VACTN THEME} \\ NP \rightarrow nn\end{matrix}\right)$$

- **Word-sense Disambiguation Model**

To make the word-sense score function feasible for implementation, we further assume that the senses of words depend only on the case assigned to the words, the parts-of-speech, and the words themselves only. Therefore, the word sense score function is approximated as follows.

$$
\begin{aligned}
S_{sense}\left(s_{i,1}^{i,n}\right) &= P\left(s_{i,1}^{i,n}\middle|\Gamma_{i,1}^{i,M_i},\Phi_{j,1}^{j,M_i},L_{j,1}^{j,N_j},t_{k,1}^{k,n},w_1^n\right) \\
&\approx P\left(s_{i,1}^{i,n}\middle|\Gamma_{i,1}^{i,M_i},t_{k,1}^{k,n},w_1^n\right) \\
&\approx P\left(s_{i,1}^{i,n}\middle|c_{i,1}^{i,n},t_{k,1}^{k,n},w_1^n\right) \\
&\approx \prod_{m=1}^{n} P\left(s_{i,m}\middle|s_{i,1}^{i,m-1},c_{i,1}^{i,n},t_{k,1}^{k,n},w_1^n\right),
\end{aligned}
$$

where $c_{i,m}$ denotes the case of $w_{i,m}$ which is specified by the case subtrees $\Gamma_{i,1}^{i,M_i}$. Currently, a simplified model, called case dependent (CD) model, is implemented in this paper. In the case-dependent model, the sense of a word is assumed to depend on its case role, part-of-speech and the word itself. Thus, the word sense score in this model, denoted by $S_{sense}^{(CD)}$, is approximated as follows:

$$
S_{sense}^{(CD)}\left(s_{i,1}^{i,n}\right) \approx \prod_{m=1}^{n} P\left(s_{i,m}\middle|c_{i,m},t_{k,m},w_m\right).
$$

## 3. The Baseline System

### 3.1 Experimental Setup

**A. Corpora:** 3,000 sentences in English, extracted from computer manuals and related documents, are collected and are parsed by the **BehaviorTran** system [10], which is a commercialized English-to-Chinese machine translation system developed by Behavior Design Corporation (BDC). The correct part-of-speech, parse trees and normal forms for the collected sentences are verified by linguistic experts. The corpus is then randomly partitioned into the training set of 2,200 sentences and the testing set of the remaining 8,00 sentences to eliminate possible systematic biases. The average number of words per sentence for the training set and the testing set are 13.9 and 13.8, respectively. On the average, there are 34.2 alternative parse trees per sentence for the training set, and 31.2 for the testing set.

**B. Lexicon:** In the lexicon, there are 4,522 distinct words extracted from the corpus. Different sense definitions of these words are extracted from the Longman English-Chinese Dictionary of Contemporary English. For those words which are not included in the Longman dictionary, their sense are defined according to the system dictionary of the BehaviorTran system. In total, there are 12,627 distinct senses for those 4,522 words.

**C. Phrase Structure Rules:** The grammar is composed of 1,088 phrase structure rules, expressed in terms of 35 terminal symbols (parts of speech) and 95 nonterminal symbols.

**D. Case Set:** In the current system, the case set includes a total number of 50 cases, which are designed for the next generation BehaviorTran MT system. Please refer to [11] for the details of the case set.

To evaluate the performance of the proposed case identification models, the recall rate and the precision rate of case assignment, defined in the following equations, are used.

$$\text{recall} \equiv \frac{\text{No of matched case trees specified by the model}}{\text{Total no of case trees specified by the linguistic experts}}$$

$$\text{precision} \equiv \frac{\text{No of matched case trees specified by the model}}{\text{Total no of case trees specified by the model}},$$

where a case tree specified by the model is said to *match* with the correct one if the corresponding cases of the case tree are fully identical to those of the correct case tree.

### 3.2 Results and Discussions

In the baseline system, the parameters are estimated by using the maximum likelihood estimation (MLE) method. The results of the deep-structure disambiguation system with the $A_1S_0$+CD model is summarized in Table 1. For comparison, the performance of the parser, without combined with the semantic interpreter, is also listed in this table. As expected, the accuracy of parse tree selection is improved as the semantic interpreter is integrated.

| | Parser | Baseline | +Smoothing | +Smoothing +Learning |
|---|---|---|---|---|
| Parse Tree | 50.1 | 56.3 | 61.4 | 77.0 |
| Case | | 77.5 | 84.2 | 88.9 |
| Recall/Precision | | 76.9 | 83.4 | 88.3 |
| Sense | | 86.2 | 87.2 | 88.6 |

*TABLE 1. Summary of the performance for the deep-structure disambiguation system.*

When the error of the baseline system was examined, we found that a lot of errors occur because many events were assigned with zero probability. To eliminate this kind of estimation error, the parameter smoothing method, Good-Turing's formula [12], is adopted to improve the baseline system. The corresponding results are listed in the third column of Table 1, which show that parameter smoothing improves the performance significantly.

In addition, a robust learning algorithm, which has been shown to perform well in our previous work [9], is also applied to the system to minimize the error rate of the testing set. The basic idea for the robust learning algorithm to achieve robustness is to adjust parameters until the score differences between the correct candidate and the competitors exceed a preset margin. The parameters trained in such a way, therefore, provide a tolerance zone for the mismatch between the training and the testing sets. Readers who are interested in details of the learning algorithm are referred to [11]. When the robust learning algorithm is applied, very encouraging result is obtained. Compared with the baseline system, the error reduction rate is 50.7% for case and 17.4% for sense discrimination, and 47.4% for parsing accuracy. As the parser, before coupling with the semantic interpreter, is considered, the performance is improved from 50.1% to 77.0%, which corresponds to 53.9% error reduction.

## 4 Error Analysis

To explore the areas for further improving the deep-structure disambiguation system, the errors for 200 sentences extracted randomly from the training corpus have been examined. It is found that a very large portion of error come from the syntactic ambiguity. More precisely, most syntactic errors result from attachment problems, including prepositional phrase attachment and modification scope for adverbial phrases, adjective phrases and relative clauses. Only less than 10% of errors arise due to incorrect parts-of-speech. Since the normal form cannot be correctly constructed without selecting the correct parse tree, errors of this type deteriorate system performance most seriously.

In addition, errors for case identification is one of the problems that make the deep-

structure disambiguation system unable to achieve a high accuracy rate of normal form. Excluding the effect of syntactic ambiguity, we checked out the errors of the semantic interpreter and found that 44.9% of normal form errors occur in identifying case. As these errors are examined, it is found that more than 30% of the incorrect normal forms have only one erroneous case. Among them, a lot of errors occur in assigning the case for the first noun of a compound noun. Taking the compound noun "*shipping materials*" as an example, the corresponding cases for the words "*shipping*" and "*materials*" are both annotated as the "HEAD" case in the corpus, as shown in Figure . However, they are assigned the cases "MODIFIER" and "HEAD", respectively. Error of this kind is usually tolerable for most applications.

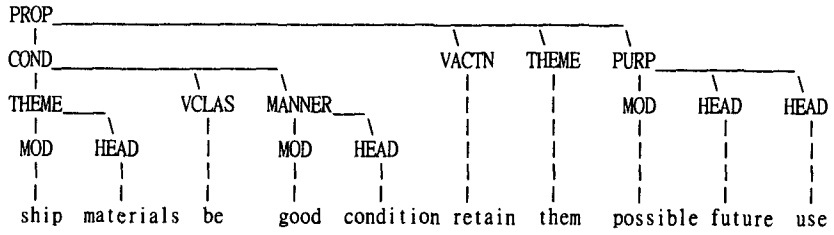


*FIGURE 4. Example of error type 1.*

Another important type of case error is to determine the class of a verb. A constituent with an action verb tends to prefer the case frame in the form of [VACTN AGENT (INSTR, ...), THEME], where AGENT, INSTR, and THEME are the arguments of the action verb, assigned by the VACTN case. On the contrary, a constituent with a stative verb would have the case frame in the form of [VSTAT THEME GOAL]. Therefore, once the class of a verb is recognized incorrectly, the cases for the verb's arguments and adjuncts will not be identified correctly. Therefore, the errors of this kind would have more serious effects on the case recall rate and the precision rate than the case structure accuracy rate.

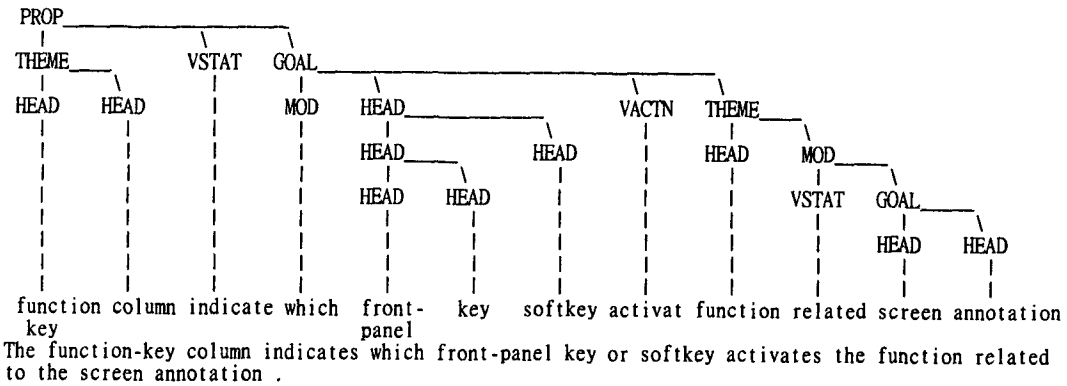


*FIGURE 5. Example of error type 2.*

## 6. Conclusions

In this paper, a deep-structure disambiguation system, integrating a semantic interpreter, a parser and a part-of-speech tagger, is developed. In this system, deep-structure ambiguity is resolved with the proposed integrated score function. This integrated score function incorporates the various knowledge sources, including parts-of-speech, syntax and semantics, in a uniform formulation to resolve the ambiguities at the various levels. Based on the integrated score function, the lexical score function, the syntactic score function, the case score function and the sense score function are derived accordingly. In addition, different models are derived in this paper to carry out case identification and word-sense discrimination.

To reduce the estimation error from maximum likelihood estimation, the Good-Turing's

smoothing method is also applied. Parameter smoothing is shown to improve the performance significantly. Finally, the parameters are adapted by using the robust discriminative learning algorithm. With this learning algorithm, 17.4% error reduction rate for sense discrimination, 50.7% for case and 47.4% for parsing accuracy are obtained compared with the baseline system. These results clearly demonstrate the superiority of the proposed models for deep-structure disambiguation.

## Reference


[1] F. C. N. Pereira and D. G. D. Warren. "Definite clause grammar for language analysis - a survey of the formalism and a comparison with augmented transition networks." Artificial Intelligence, 13(3): 231-278, 1980.

[2] A. E. Robinson. "Determining verb phrase referents in dialogue." American Journal of Computational Linguistics, 7(1):1-16, 1981.

[3] M. Kay. "Parsing in functional unification grammar." In D.R. Dowty, L. Karttunen, and A. Zwicky, editors, Natural Language Parsing, page 251-278. Cambridge University Press, 1985.

[4] Peter F. Brown, Stephen A. Della Pietra, Vicent J. Della Pietra, and Robert L. Mercer. "Word-sense disambiguation using statistical methods." In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pages 264-270, 1992.

[5] I. Dagan, A. Itai, and U. Schwall. "Two languages are more than one." In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, pages 130-137.

[6] Willian A. Gale, Kenneth W. Church, and David Yarowsky. "Using bilingual materials to develop word sense disambiguation methods." In Proceedings of the 4th International Conference on Theoretical Methodological Issues in Machine Translation, pages 101-112, Montreal, Canada, 25-27, June 1992.

[7] David Yarowsky, "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora." In Proceedings of the 14th International conference on Computational Linguistics, pages 454-460, Nates, France, August 1992.

[8] T. H. Chiang, Y. C. Lin, and K. Y. Su, "Syntactic ambiguity resolution using a discrimination and robustness oriented adaptive learning algorithm." In Proceedings of the 14th International Conference on Computational Linguistics, pages 352-358, Netes, France, August, 1992.

[9] T. H. Chiang, Y. C. Lin, and K. Y. Su. "Robust learning, smoothing, and parameter tying on syntactic ambiguity resolution." Computational Linguistics, pages 321-329, 1995.

[10] S. C. Chen, J. S. Chang, J. N. Wang, and K. Y. Su. "ArchTran: A corpus-based statistics-oriented English-Chinese machine translation system." In Proceedings of Machine Translation Summit III, pages 33-40, 1991.

[11] T. H. Chiang. "Statistical Models for Deep-structure Disambiguation." PhD dissertation, National Tsinghua University. Taiwan, R.O.C., 1996.

[12] I. J. Good. "The population frequency of species and the estimation of population parameters." Biometrika, 40:237-364, 1953.