Developing a Chinese Module in UNITRAN *

Zhibiao Wu, Loke Soo Hsu, Martha Palmer, Chew Lim Tan
Department of Information System & Computer Science
National University of Singapore
Republic of Singapore, 0511
E-mail: wuzhibia,hsuls,mpalmer,tancl@iscs.nus.sg

Abstract

This paper will share with the readers our experiences gained in a project of translating Chinese to other languages with Principle Based Machine Translation (PBMT). UNITRAN is a prototype system developed in MIT which translates simple sentences among English, Spanish and German. Based on Government Binding (GB) theory and Lexical Conceptual Structure (LCS) theory, UNITRAN serves a good model for applying GB and LCS to achieve the principle based machine translation. We have tried to put Chinese into the system. Now the system can translate among the four languages properly. In the following sections, we will first introduce the basic idea of PBMT. Then we briefly explain how UNITRAN translates Chinese to English. Our major focus will be the Chinese language parameter setting. Some GB parameters will be discussed in certain detail. And finally, the last section will discuss the merit of the PBMT and problems arise from this approach.

1. Introduction

With the development of Government Binding (GB) theory (Sells, 1985) and Lexical Conceptual Structure (LCS) theory (Jackendoff, 1991), the principle based machine translation (PBMT) has drawn some attention. The basic idea of PBMT is that by highly abstracting the regularities existing in human languages, a large part of the language grammar and lexical semantics can be covered by a small number of principles. Different languages get their particular expressions by setting parameters for these principles. The idea is based on the assumption that human has an innate Universal Grammar which enables one to compose new lexical concepts based on a set of semantic primitives.

UNITRAN is a PBMT machine translation prototype developed in MIT by Bonnie Dorr (Dorr, 1990). The system can freely translate single sentences among English, Spanish and German. UNITRAN elaborates the idea of PBMT to its full strength. At each level of morphological, syntactic and semantic processing, the system is designed based on a small

*Special thanks to Bonnie Dorr for her kind permission to use some of the materials in her PhD thesis.

set of principles. Particular languages realized themselves in the system by a set of parameter setting files in the system. This approach brings a lot of merits. First, it is easy to extend the system's ability to handle a new language. By specifying the parameter files, the major parts of the system remind unchanged. Secondly, Different languages have divergences in syntax, semantics and pragmatics level. Since the divergences of the languages can be represented by different parameters for the same principle, language divergences can be easily resolved.

In order to investigate the strength of UNITRAN and the idea of PBMT, and to see whether Chinese is suitable for a PBMT treatment, we have tried to put Chinese in UNITRAN. In this paper, we will focus on the Chinese GB setting. In Section 2, we will briefly introduce the GB and LCS theory. In Section 3, we will present examples to show how UNITRAN handle the Chinese sentence. The Chinese GB parameter setting will be discussed in Section 4. And finally, the last section will discuss the merit of the PBMT and problems may arise from this approach.

2. GB and LCS

This section will briefly introduce the GB and LCS theory and how they are realized in UNITRAN system.

2.1. Government Binding theory

A detail introduction of GB theory can be found in (Sells, 1985). The Chinese GB theory has been developed by (Huang, 1982), (Li, 1990), and recently, Professor Tang T. C. has done a lot of work on this subject (Tang, 1989; Tang, 1992). Followingly, I briefly summarize the theory.

The basic idea of the theory can be shown in the following figure.

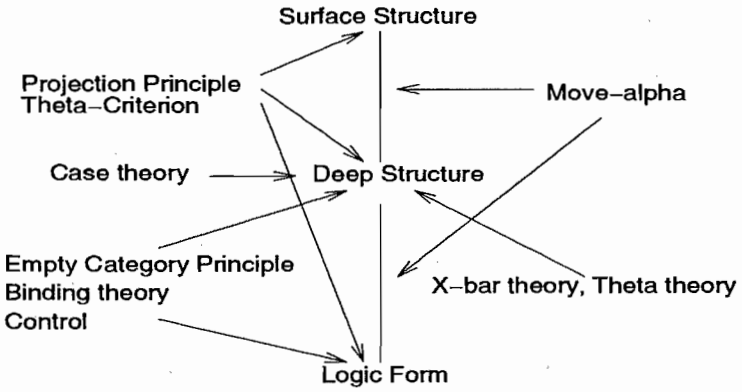


Figure 1: Government-Binding Theory

The language expression is a projection from the lexical semantics. By θ criterion and projection principle, the arguments of a LCS expression can be projected to deep structure to

form word phrase. By 'Moving anything anywhere' (Move-α) with some constraint principles, the surface structure can be derived. Some of the important concepts is explained below:

X-bar theory says that language expressions can be classified into several phrase categories. The syntactic behaviors of each category are similar. For example, in UNITRAN, English is been classified into six categories V, N, A, P, C and I standing for Verb, Noun, Adjective, Preposition, Complementizer and Inflection. Each category follows the rule schemata: For word belongs to a X category, it can form the intermediate phrase X′ with the complement of the X category. The X′ can recursively form a new X′ with the adjunct of the X category. And the X″ i.e. the maximum projection of X is finally formed with specifier and X′. The word phrase is then projected to the surface by Move-α with some constraints like Empty Category Principle, Binding Theory, Control and Case Theory.

θ theory is for the linking from the deep structure to the logic form. θ role is the thematic roles of predicate's arguments. θ criterion says that each argument bears one and only one θ role, and each role is assigned to one and only one argument. This derived the linking rules in UNITRAN.

Projection principle says that representations at each level are projected from the lexicon in that they observe the subcategorization properties of lexical items.

Move-α is an operation from deep structure to surface structure. It means 'Move anything anywhere'. But the moving must obey some constraints. Following are the example of NP-movement and Wh-Movement.

```
1) NP-movement:
   d:    [NP ]  INFL kiss-en Bill
   s:    Bill_i INFL kiss-en e_i

2) Wh-movement:
   d:  [COMP      ] Bill INFL see who
   s:  [COMP who_i] Bill INFL see e_i
```

Bounding says that any application of Move-α may not cross more than one bounding node.

Abstract case: is a notion of NP to show its relation to verbs. NP have cases, verb assigns Accusative Case to its object. Four cases are used in UNITRAN. They are Nominative, Accusative, Dative, and Genitive.

Trace is concerned with the empty position left behind when a constituent has moved by Move-α.

Binding is concerned with the coreference relations among noun phrases. Following sentence show that illegal bindings with the mark *.

```
  John_i sees himself_i
* John_i sees himself_j
* John_i sees him_i
  John_i sees him_j
```

2.2. Lexical Conceptual Structure

Lexical conceptual structure is a compositive representation method for lexical semantics. The building blocks can be classified into several types as: EVENT, STATE, POSITION, PATH, THING, PROPERTY, LOCATION, TIME and MANNER. For each type, there is a set of semantic primitives. For example, we have HERE, THERE, LEFT, RIGHT, UP, DOWN ... in the type of LOCATION. Some of the primitives are predicates which takes arguments. According to Jackendoff's theory (Jackendoff, 1991), every sentence meaning or word meaning can be represented by primitives or the composition of primitives. The composition is done by observing the θ theory. Following is an example of Chinese "划伤 (HuaShang)" event.

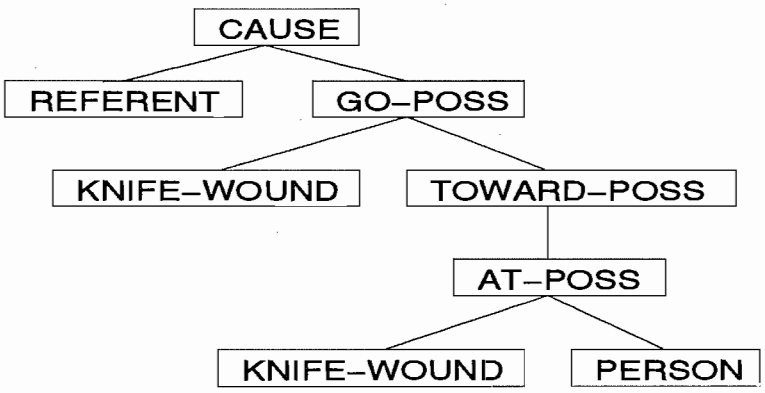


Figure 2: Underlying LCS for Chinese verb HuaShang

This is the interlingua that is used as the pivot from source to target language. The underlying form conveys the meaning that a *referent* cause a *person* to possess a *knife-wound.*

2.3. A Chinese LCS in UNITRAN

Taking UNITRAN as an example. UNITRAN used LCS (Jackendoff, 1991) and Pinker's verb representation with manner (Pinker, 1991) to represent verb semantics. For the Chinese sentence "小明 跑步 上学 (XiaoMing PaoBu Shang Xue)", the logic form of the whole sentence is derived from the verb semantic representation of "跑步 (PaoBu)". The LCS representation i.e. the argument structure of "跑步 (PaoBu)" is:

(DEF-ROOT-WORDS (GO-LOC Y (FROM-LOC (AT-LOC Y Z1)) (TO-LOC (AT-LOC Y Z2)))
:ROOTS ((跑步 (Y (* Y))
(Z1 :OPTIONAL ((* FROM-LOC) (AT-LOC (Y) (Z1))))
(Z2 (UC (CASE ACC)) ((* TO-LOC) (AT-LOC (Y) (Z2))))
(MODIFIER JOGGINGLY))

This representation defines "跑步 (PaoBu)" falling into the class of GO-LOC. GO-LOC is a three place predicate which represents "motion with manner". The definition of "跑步 (PaoBu)" can be read as "Y is in a motion from location Z1 to a location Z2 with a

'JOGGINGLY' manner". By not arguing on the semantic representation schemes, let's see how the surface structure "小明 跑步 上 学 (XiaoMing PaoBu Shang Xue)" can be analyzed to form a Logic form LCS representation. According to the X-bar theory and the principles like:

$$\text{I-MAX} -> \text{V-MAX N-MAX}$$
$$\text{V-MAX} -> \text{V P-MAX}$$
$$\text{P-MAX} -> \text{P N-MAX}$$

The GB parse tree of the sentence is shown in Figure 3, it is derived by X-bar theory and the other constraints. The composed LCS of the sentence is shown in Figure 4.

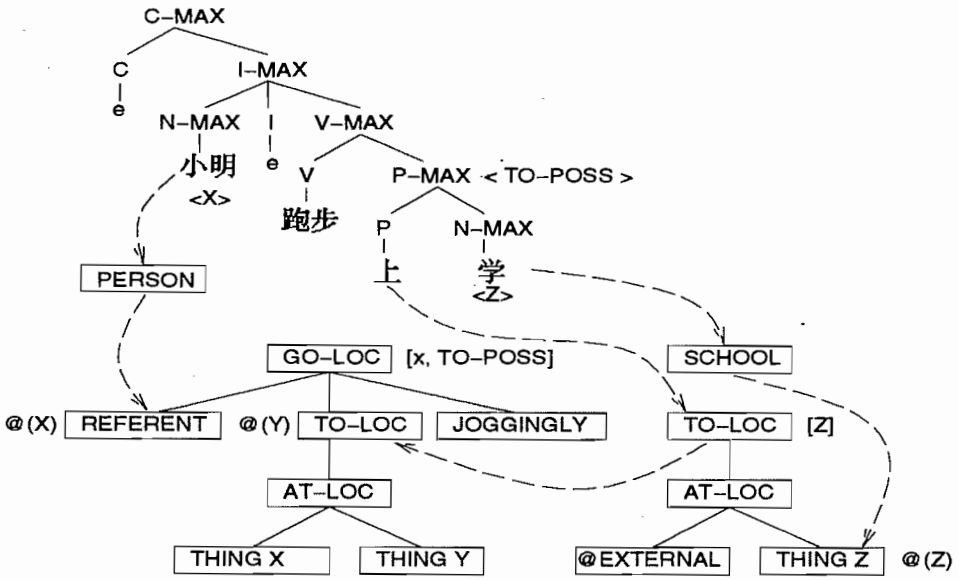
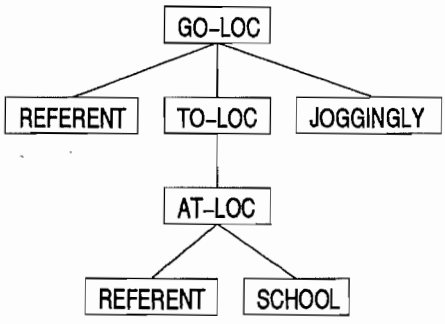


Figure 3: Parse tree and θ role assignment



GO–LOC(REFERENT TO–LOC (AT–LOC (REFERENT SCHOOL)) JOGGINGLY)

Figure 4: Logic form i.e. LCS representation

3. Handling Chinese in UNITRAN

Based on GB and LCS, UNITRAN separates the data and program quite well. Followingly, by presenting an example of how UNITRAN processes Chinese sentence, we will discuss some of the good features of UNITRAN. Obviously, we cannot go into every detail of the system. Interested reader please refer to (Dorr, 1990). The overall design is shown in Figure 5.

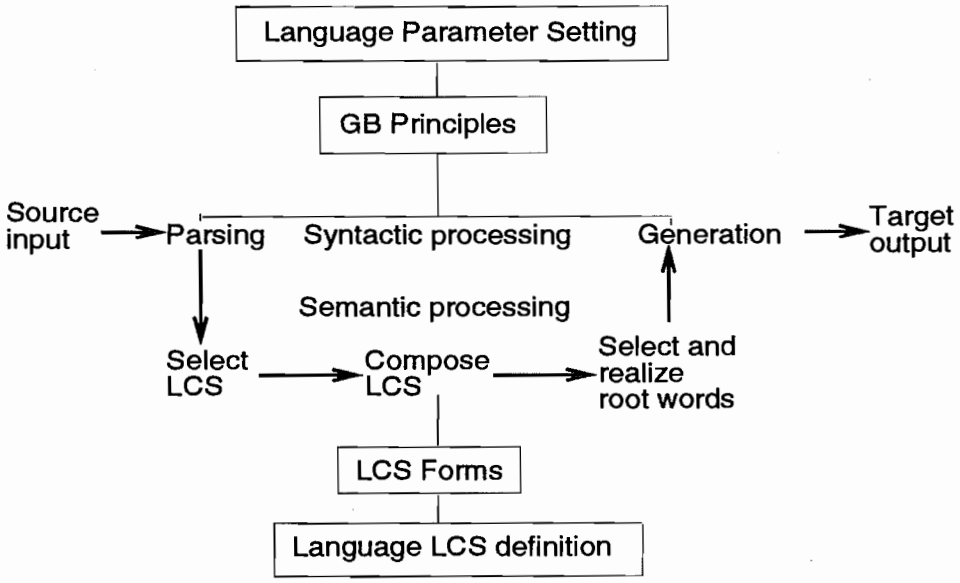


Figure 5: Overall design of UNITRAN

Following is a translation output from English "I stab him" to Chinese.
After Projection and assign X-bar structure, 14 structures left.
Running Translation Examples ...
Parsing (I STAB HIM) ... Done (14 trees in 0.38 seconds).
Assigning X-bar structure to (I STAB HIM) ... Done (14 structures in 0.75 seconds.)

After applying constraints, only one structure left.
Applying Bounding (Trace Linking) ... Done (3 structures in 0.13 seconds.)
Applying X-bar (Feature Matching) ... Done (6 structures in 0.17 seconds.)
Applying Case ... Done (2 structures in 0.25 seconds.)
Applying Binding ... Done (2 structures in 00.20 seconds.)
Applying Theta ... Done (1 structures in 00.20 seconds.)

Following is the parse tree for the sentence.
(0
((C-MAX (C "e")
(I-MAX (N-MAX (N "i")) (I "e")
(V-MAX (V "stab") (N-MAX (N "him")))))))

Lexical conceptual structure is composed.
Composing LCS ... Done (1 structures in 3.18 seconds.)

Following is the LCS structure for the sentence.
(0
(CAUSE REFERENT
(GO-POSS KNIFE-WOUND
(TOWARD-POSS (AT-POSS KNIFE-WOUND REFERENT)))
WITH-INSTR))

Generation begins.
Generating ...

After lexical selection, assign X-bar structure to the generated sentence.
Assigning X-bar structure to 我自己 划伤 他 ...
Assigning X-bar structure to 我的 划伤 他 ...
Assigning X-bar structure to 我 划伤 他 ...
Done (3 structures in 0.50 seconds.)

Apply constraints to the generated structures.
Applying X-bar (Feature Matching) ... Done (6 structures in 00.80 seconds.)
Applying Case ... Done (1 structures in 0.13 seconds.)
Applying Binding ... Done (1 structures in 0.00 seconds.)
Applying Theta ... Done (1 structures in 00.20 seconds.)
Applying Bounding (Trace Linking) ... Done (1 structures in 00.20 seconds.)

Only one legal structures left. Following shows the parse tree for translated Chinese sentence.
(0
((C-MAX (I-MAX (N-MAX (N "我")) (I "e")
(V-MAX (V "划伤") (N-MAX (N "他"))))
(C "e"))))

The feature matching is shown the the following figure:

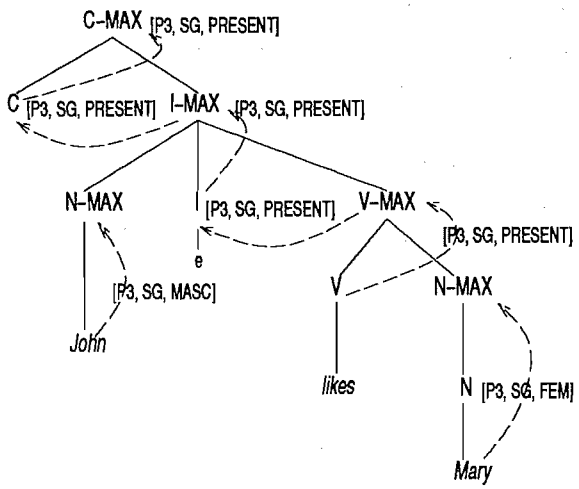


Figure 6: Feature matching in X-bar module

The target language generation is divided into two steps. One is the lexical selection. The

other one is the syntactic realization. This is done by matching the underlying LCS to the appropriate root word from the target language possible set. The lexical selection of Spanish root word for the *stab* event is shown in Figure 7.

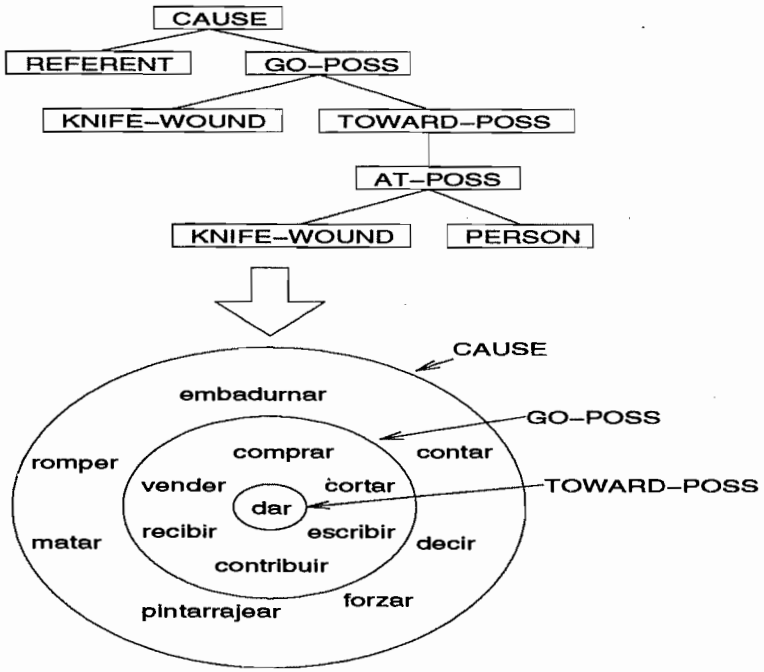


Figure 7: Lexical selection of Spanish root word for English stab event

4. Chinese Parameter Setting

We now come to the main part of the paper: setting those parameter files for Chinese. UNITRAN uses PC-KIMMO as its morphological processor. PC-KIMMO is known as a good tool for morphological processing on languages which have inflections. We let UNITRAN skip the KIMMO processing for Chinese, since Chinese don't have inflections. Since we follow the examples of English verb definition to define Chinese verbs. The main focus will be on Chinese GB parameter setting. In UNITRAN, there is already a set of modules for the GB theory with English, Spanish and German. What we have done is to set those parameters related to Chinese GB theory.

4.1. X-bar module

4.1.1. Basic categories

The choice of X are determined by the basic categories. For Chinese, we employ the view of (Tang, 1989) to classify Chinese language into eight categories. We put Chinese Adverb and Adjective together. Noun phrase is further divided into Determiner phrase and Qualifier phrase. Tense and Aspect is the head for I phrase. The sentence particle is the head for C. The basic categories parameter is set as follows:

Language	Basic categories
Chinese	C, I, V, N, P, A, Q, D

Table 1: Basic categories for Chinese

4.1.2. Constituent Order

The constituent order parameter accounts for the word order distinctions among different languages. According to Tang T. C. , for Chinese, the noun phrase is head final. Adjective phrase is head initial for transitive adjective, head final for intransitive adjective. Proposition phrase is head initial. Complementizer is head final. Determiner is head initial, specifier initial. Qualifier phrase is head initial and specifier initial. We differ with Tang T. C. in viewing that verb phrase is head initial. The constituent order parameter is set as follows:

Category	Chinese
I	SPEC-INITIAL, HEAD-INITIAL
N	SPEC-INITIAL, HEAD-FINAL
C	SPEC-INITIAL, HEAD-FINAL
A	SPEC-INITIAL, HEAD-INITIAL
P	SPEC-INITIAL, HEAD-INITIAL
D	SPEC-INITIAL, HEAD-INITIAL
Q	SPEC-INITIAL, HEAD-INITIAL
V	SPEC-INITIAL, HEAD-INITIAL

Table 2: Constituent order for Chinese

4.1.3. Base-Generated Specifiers

There are two types of specifiers: ones that are base generated in θ position and ones that are moved to a θ-bar position. The base-generated specifiers are assumed to be optional unless the :OBLIGATORY marker is included in the parameter setting. For noun phrase, the specifiers can be a noun phrase or the determiner phrase. For qualifier phrase, the specifiers must be a number. For determiner, the specifier can be the noun phrase. The base-generated specifiers parameter is set as follows:

Category	Chinese
I	N-MAX
N	DET, N-MAX
C	N-MAX
D	N-MAX :OBLIGATORY
Q	NUM :OBLIGATORY

Table 3: Base generated specifier for Chinese

4.1.4. Base-generated Adjuncts

The base adjuncts parameter specifies the position (left, right or free) and the level (minimal or maximal) of each adjunct with respect to the category to which it is adjoined. The base-generated specifiers parameter is set as follows:

Category	Position	Chinese Adjuncts
N	LEFT-MAX	Q-MAX, A-MAX
N	RIGHT-MAX	C-MAX
V	LEFT-MAX	ADV
V	FREE-MAX	P-MAX
A	LEFT-MAX	ADV, P-MAXD
A	RIGHT-MAX	C-MAX
I	LEFT-MAX	ADV, N-MAX, P-MAX
C	LEFT-MAX	ADV, N-MAX, P-MAX

Table 4: Base-generated adjuncts for Chinese

4.1.5. Complements

The Chinese complements parameter is is set as follows:

Category	Chinese Complements
V	(N-MAX), (P-MAX) (P-MAX P-MAX) (N-MAX P-MAX) (C-MAX) (P-MAX C-MAX) (C-MAX P-MAX) (A-MAX) (ADV) (V-MAX)
P N	(N-MAX), (Q-MAX) (P-MAX), (C-MAX) (N-MAX) (D-MAX)
Q	(N-MAX)
D	(N-MAX), (D-MAX)
A	(C-MAX), (N-MAX)
I	(V-MAX), (A-MAX)
C	(I-MAX)

Table 5: Complements for Chinese

4.2. Government Parameter

Government parameter is a key point to those constraints.

4.2.1. Governors

The governors for each Chinese categories are:

Language	Governors
Chinese	V, N, A, P, Q, D, ASP, PAR

Table 6: Governors for Chinese

Here, ASP is for aspect and tense, PAR is for sentence particles.

4.2.2. Proper Governors

The proper governors for Chinese are:

Language	Governors
Chinese	V, P, ASP

Table 7: Proper Governors for Chinese

4.3. Bounding

4.3.1. Bounding node

The bounding node in Chinese are:

language	Bounding nodes
Chinese	I, N

Table 8: Bounding node for Chinese

4.3.2. Moved Specifiers

Category	Chinese Moved Specifiers
I	N-MAX
C	N-MAX, P-MAX

Table 9: Moved Specifiers for Chinese

4.3.3. Moved adjuncts

Category	Position	Chinese Moved Adjuncts
I-MAX	LEFT-MAX	ADV, P-MAX

Table 10: Moved Adjuncts for Chinese

4.4. Trace

The parameters for Trace in Chinese is set as follows:

Trace parameter	Chinese
Traces	N-MAX, P-MAX
Empties	N-MAX in Specifier of I

Table 11: Trace for Chinese

5. Discussion

In the last section, we have shown some of the parameters set for Chinese GB grammar. Several sample Chinese sentences have been successfully run on UNITRAN. This shows that a new language can be easily added into the system just by setting parameters for those principles. However, the merit comes together with the deficits. The requirement of highly abstracted principles for all human languages is very difficult to meet. The Chinese grammar we set in UNITRAN is by no mean a complete one. Although there is some universal rules for human languages to form a core grammar, each particular language has its own idiosyncrasy. These 'periphery' phenomena need the system to handle them piece by piece(Tang, 1989). Unfortunately, the number of irregularities is very larger than the number of principles. Therefore more effort is needed to show that the PBMT style of UNITRAN is suitable to scale up for unrestricted text.

REFERENCES

CHOMSKY, N. (1956). *Syntactic Strcutures.* Mouton.

CHOMSKY, N. (1965). *Aspects of the theory of syntax.* MIT Press.

DORR, B. J. (1990). *Lexical Conceptual Structure and machine Translation.* PhD thesis, MIT.

DOWTY, D. (1991). Thematic Proto-roles and Argument Selection. *Language*, 67(3).

DOWTY, D. R. (1979). *Word Meaning and Montague Grammar.* D. Reidel Publishing Company.

HUANG, J. (1982). *Logic Relations in Chinese and the Theory of Grammar.* PhD thesis, MIT.

HUDSON, R. (1984). *Word Grammar.* Blackwell.

HUTCHINS, W. J. & SOMERS, H. L. (1992). *An Introduction to Machine Translation.* Academic Press, London.

JACKENDOFF, R. (1991). *Semantic Structures.* MIT Press.

LEVIN, B. (1987). Approaches to Lexical Semantic representation. In LEVIN, B., editor, *Readings for Lexical Semantics.* Northwestern University.

LEVIN, B. (1992). English Verb Classes and Alternations: A Preliminary Investigation. Technical report, Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208.

LI, Y. H. A. (1990). *Order and Constituency in Mandarin Chinese.* Kluwer Academic Publishers.

NIRENBURG, S., CARBONELL, J., TOMITA, M., & GOODMAN, K. (1992). *Machine Translation: A Knowledge-Based Approach.* Morgan Kaufmann Publishers.

NIRENBURG, S. & NIRENBURG, I. (1988). A Framework for Lexical Selection in Natural Language Generation. In *COLING88.*

PALMER, M. (1990a). Customizing verb definitions for specific semantic domains. *machine Translation*, 5(30).

PALMER, M. & POLGUÈRE, A. (1992). A Computational Perspective on the Lexical Analysis of Break. In *Proceedings of the Workshop on Computational Lexical Semantics*, Touluse, France.

PALMER, M. S. (1990b). *Semantic Processing for Finite Domains.* Cambridge University Press.

PINKER, S. (1991). *Learnability and Cognition.* The MIT Press.

PUSTEJOVSKY, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17(4).

SELLS, P. (1985). *Lectures on Contemporary Syntactic Theories.* CSLI.

SOWA, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley.

TAN, C. L., HSU, L. S., & WU, Z. (1992). On Self-organized approaches to NLP. Technical report, Department of Information System & Computer Science.

TANG, T. (1989). Principle and Parameter Grammar and the comparative analysis between Chinese and Chinese. In *Singapore Symposium on the World Chinese Teaching.*

TANG, T. C. (1992). Grammar theory and Machine Translation: Principle and parameter Grammar. In *ROCLING V R.O.C. Computational Linguistics Conference V.*

TOMITA, M., editor (1991). *Current Issues in Parsing Technology.* Kluwer Academic.

WU, Z., HSU, L. S., & TAN, C. L. (1992). A Survey on Statistical Approaches to Natural Language Processing. Technical Report TRA4/92, Department of information system and computer science, National Universoty of Singapore. Submitted to *computational lingistics.*

WU, Z., PALMER, M., HSU, L. S., & TAN, C. L. (1993). Toward the Similarity-based Fuzzy Lexicon. Technical Report TRE4/93, Department of Information System and Computer Science, National Universoty of Singapore.