Lexicon-Driven Transfer In English-Chinese Machine Translation

Chung-Teng Sun and Jyun-Sheng Chang

Department of Computer Science National Tsing Hua University Hsinchu, Taiwan 30043

Abstract

This paper identifies the major differences between English and Chinese due to lexical idiosyncrasy and describes a proposed mechanism for bridging the differences in the generation phase of an English-Chinese machine translator. The method involves uniform intermediate representation for clauses and noun phrases, a minimal set of transfer operations and an active bilingual lexicon that encodes the needed transfer. Using the method, we are able to deal with lexical idiosyncrasy in both languages in a modular and efficient manner.

1. Introduction

Much effort has been devoted to research and development of machine translation since 1950s (Slocum 1985). However, the quality of the output produced by most machine translation systems is not high enough to have any marked effect on translation productivity.

A machine translation system produces a variety of expressions in the target language, including *good*, *fair*, and *poor* expressions [Tsutsumi, 1990]. To improve the quality of translation and get the *good* sentences which can be easily understood or postedited, the following major functions should be implemented with great care.

- 1. Selection of equivalents for words;
- 2. Reordering of words; and
- 3. Improvement of sentences styles.

In this paper, we will propose a practical way of designing a generator in machine translation system that bridges the differences between English and Chinese languages. Using this generator we hope to achieve the above functions and improve translation quality.

1.1 Rule-Based Machine Translation

Modern rule-based machine translation systems use either *transfer* approach or *interlingual* approach.

Transfer approach is characteristic of a system(e.g., TAUM) in which the internal representations of a grammatical unit (e.g., sentence) in analysis and synthesis are different depending on the source and target languages. This implies the existence of a third translation stage which maps one language-specific representation into another : this stage is called *Transfer*. Thus, the overall transfer translation process is *Analysis* followed by *Transfer* and then *Synthesis*. **Interlingual approach** is characteristic of a system (e.g., CETA) in which the internal representation of the source language input is intended to be independent of any language, and the same representation is used to analysis the source language and to synthesize the target language output [Slocum 1985].

The differences between two languages can be classified into two kinds

- (1) syntactical differences: the general differences in word order.
- (2) differences that are caused by the idiosyncrasy of individual words in the two languages.

As for the syntactical difference, various systems use different approaches to deal with them. The transfer approach uses structure transfer rules to express the differences. The interlingual (or pivot) approach use a non-syntactical representation, and provide mapping mechanisms between syntax and the representation, so the differences can be resolved via a language independent representation.

We are currently developing a machine translation system that uses a mixed approach. On the syntactical and semantic levels, it is interlingual and on the lexical level it takes the transfer approach.

1.2 System Model

In this paper, we concentrate on the generation process of our MT system. The generator consists of two phases: the **lexicon-driven transfer phase** and the **surface generation phase**. The first phase deals specially with global reorganization of intermediate representations in order to bridge the differences caused by lexical idiosyncrasy between source and target languages. The reorganization includes structural and lexical transfer. Currently, transfer is done sentence by sentence, using only information from sentential analysis. No analysis and reorganization on the discourse level is performed. The system overview is shown in Figure 1.



I.R. : Intermediate Representation



This paper focuses on the problem lexical idiosyncrasy. Our approach emphasizes the role of lexicon, and in particular, it shares the idea with [Tsujii, 1990] which proposes that the bilingual lexicon play the central role in the transfer phase.

1.3 Paper Outline

This paper describes research from continuing previous works on Chinese sentence generation [Kuo 1989, Chen 1990 and Liao 1990] and focuses on the following:

- 1. Adding lexical transfer phase before sentence generation.
- 2. Designing a bilingual lexicon which controls the transfer by means of necessary tests and actions.

3. Rewriting the driver in Prolog, making various changes in the input format and extending the Chinese systemic grammar

In the following, we will concentrate on the first two areas. The rest of the paper is organized as follows: Section 2 identifies the most prominent lexical idiosyncrasy between English and Chinese. Section 3 describes our proposal for resolving these differences in order to obtain fluent target Chinese text. Section 4 compares our approach with previous works and summarizes the paper.

2. Differences between English and Chinese

In discussions on the differences caused by lexical idiosyncrasy between English and Chinese, it is necessary to consider the differences in the ways people recognize things and express their ideas about them. It is observed that there is a difference in viewpoint [Wu, 1990] between English and Chinese. So we sometimes have to restructuring them to give an appropriate translation. Each language also has its own specific word constructions, which are used to express specific meanings. These specific constructions can not be directly translated into other languages. We will describes the differences caused by *lexical idiosyncrasy* and specific constructions caused by target words in following two subsections.

2.1 Structure Transfer Caused by Lexical Idiosyncrasy

Most contrastive linguistic analysis of English and Chinese are quick in pointing out that the most prominent difference is in the way that nouns and verbs are used. In English, every sentence has at most one finite verb, so for complex information to come across, most of the information has to be expressed in terms of nouns. This explains the abundance of English nouns. There are more nouns than verbs and a verb can turn into a nominal counterpart through inflexion transformation. On the contrary, Chinese sentences are common to have more than one verb, and for this reason, verbs abounds in Chinese [Chen, 1988]. Consequently, the most appropriate translation of an English noun often turns out to be a verb in Chinese. Some English verbs turn into other parts of speech other than a noun, such as adjective or adverbs in English. However, because of heavy reliance on verbs in Chinese, the most suitable translation is again a verbal counterpart in Chinese.

The following are some examples of structural transfer due to translating an English noun into a Chinese verb. In each example, the first sentence (a) is the source English, the second one (b) is

the direct translated Chinese sentence without any structure transfer, and the last one (c) is the Chinese sentence generated with structural transfer.

(1-a): He pretended illness yesterday. (2-a): He is a good speaker of English. (1-b):昨天他假裝病。 (2-b): 他是個好的英語講者。 (1-c): 昨天他假裝生病。 (2-c):他講英語講得很好。 (3-a): New Lab animals reduce testing of drugs on humans. (4-a): I have a severe headache. (3-b): 新的實驗動物減少在人身上藥物的試驗。 (4-b) 我有**厲害的頭痛**。 (3-c): 新的實驗動物減少在人身上試驗藥物。 (4-c):我頭痛得厲害。 (5-a): The arrival of a train at the station. (6-a): He quits the job surprisingly. (5-b): 火車的抵達車站 (6-b): 他令人吃驚地辭掉工作。 (5-c): 火車抵達車站 (6-c):他辭掉工作令人吃驚。 (8-a): This sentence is <u>untranslatable</u>. (7-a): He is talkative. (7-b): 他是愛説話的。 (8-b):這個句子是無法翻譯的。 (8-c): 這個句子無法翻譯。 (7-c): 他很愛説話。

English prepositions sometimes are best translated into verbs in Chinese. For instance,

| (9-a): <u>in</u> uniform | (10-a): <u>in</u> hat |
|----------------------------------|-----------------------------|
| (9-b): | (10-b):_ |
| (9-c): 穿制服 | (10-c): 戴帽子 |
| (11-a): <u>by</u> train | (12-a): a path by the river |
| (11-b): | (12-b): _ |
| (11-c): 坐火車 | (12-c): 沿河道路 |
| (13-a): a telegram with bad news | (14-a): a man with glasses |
| (13-b):_ | (14-b): _ |
| (13-c): 帶著壞消息的電報 | (14-c): 戴眼鏡的人 |

Because the different way that verbs are used in the two languages, even when a verb is translated into a verb in the target language, there could be incompatibility in their argument structure and that calls for some kinds of structural transfer too. For instance,

(15-a): we shall give <u>special consideration</u> to your opinion.
(15-b): 我們將給你的意見特別考慮。
(15-c): 我們將特別考慮你的意見。

(17-a): he <u>skied twice</u> this year. (17-b): 他今年滑雪兩次。 (17-c): 他今年滑兩次雪。 (16-a): this <u>surprised</u> everybody°.
(16-b): 這**震驚**大家。
(16-c): 這使大家震驚。

(18-a): they <u>lived a happy life.</u>
(18-b): 他們過著快樂的生活。
(18-c): 他們生活得很快樂。

We have described some types of structure differences caused by lexical idiosyncrasy. Generally, the major structural changes are caused by the differences in part of speech of a equivalent concept in the two languages. It is clear that if we can deal adequately with these structure differences, we can improve translation quality considerably.

2.2 Lexical Influences on Construction of Target Sentence

Individual words can sometimes influence the selection of structure in construction of a target sentence. The lexical influence on constructions is complex, including wide-range and local-range restructuring. The following are some examples of lexical influence on sentence construction.

2.2.1 Lexical Influence on Order

The selection of a target word may influence the order of phrases in a sentence. For example,

| (19-a) She sang sadly. | (20-a) She sang that song well. |
|------------------------|---------------------------------|
| (19-b)她憂傷地唱歌 o | (20-b)她很好地唱那首歌。 |
| (19-b) 她唱歌唱得很憂傷 o | (20-c) 她唱那首歌唱得很好。 |

Both (19-b) and (19-c) are appropriate translation for (19-a). But for (20-a), (20-c) is appropriate while (20-b) is not a good translation. This is because there is the adverb 很好 unlike 憂傷地 must locate after the verb in Chinese.

2.2.2 Lexical Influence on Selection of Sentence Construction

The conceptual content of a lexical item may determine the structure of the target sentence. For example,

(21-a) That book has been stolen (21-b) 那本書被偷了。 (23-a) That book has been stolen by him.(23-b) 那本書被他偷了。

(22-a) That book has been published. (24-a) That book has been published by Tsin-Hua bookstore.
(22-b) 那本書出版了。
(24-b) 那本書是清華書局出版的。

The bei-construction (被字句) in Chinese is used essentially to express an adverse situation, one in which something unfortunate has happened, and also express disposal. That is, the bei sentence describes an event in which an entity or person is dealt with, handled, or manipulated in some way [Li-Thompson, 1982]. So we must check adversity and disposal features of each verb to decide whether to use the bei-construction. For example, since the verb publish in Chinese is not a adverse verb, (24-b) is the shi-de-construction (是一的一句), not the bei-construction.

2.2.3 Lexical Influence on Lexical Selection

The selection of target words may influence each other. For example,

(25-a) That book is not valuable. (25-b) 那本書沒有價值。

There are four negative forms in common use in Chinese: 不, 別, 沒, 沒有. The scope position and form of negative particles in Chinese are decided mostly according to the features collected from the analysis phase. But there are words such as *valuable* in (25-a), Chinese counterpart "有價值" carrying the head "有". In this case, the negative form is "沒有". Other example is interesting and "有意思".

These cases can only be handled appropriately using a bilingual lexicon.

3. Lexicon-Driven Transfer

This section discusses the main idea of bilingual lexicon-driven transfer, and show how our framework treats structural changes caused by lexical idiosyncrasy described in Section 2.

3.1 Main Idea about Lexicon-Driven Transfer

We propose to resolve the structure difference in translation due to lexical idiosyncrasy according to the following considerations:

1. The transfer needed should be captured in the lexicon.

2. The intermediate representation should encode a clause, a verb phrase and a noun phrase as similar as possible so that easily interchangeable. This idea is similar to [Allen 1987] which propose that the logic form for NPs that describe events should be virtually identical to the representation of sentences that describe event.

3. The set of transfer operations should be kept minimal for simplicity and efficiency.

The intermediate representation has three layers: *event*, *entity*, and *lexeme* corresponding to the syntactical structures of clause, noun phrase and word. The lexeme layer is atomic containing only the target word while the event and entity layers share the same structure with a head and various thematic cases and modifiers. The head of an event is of course the main verb and that of an entity is the head noun. And to operate on this representation, we proposed four basic operations for encoding transfer:

1. **Raise** : a constituent in the intermediate representation can be raised one step up the constituency structure without changing the slot names (functional role) of its subconstituents.

2. Modify : the slot name of a constituent can be changed into another.

3. **Insert** : a constituent of any slot name can be inserted on the same level of the lexical item being considered.

4. Delete : a constituent can be deleted from the intermediate representation.

3.2 How to Deal with Structure Differences

Following are some examples which show what structure changes caused by lexical idiosyncrasy between English and Chinese languages have been done using these four operations.

3.2.1. Raise operation

Example (2-a): He is a good speaker of English.

The source intermediate representation using Direct Acyclic Graph (DAG)notation is shown in Figure 2.:



Figure 2. DAG of He is a good speaker of English

The linear format of representation is the following:

[cat : event, features : [transitive], agent : [cat : pro, lex: he], head: [cat: bv, lex: is], theme:[cat:entity, features:[indefinite], mod:[cat: adj, lex: good], head:[cat: n,lex: speaker], theme: [cat: n, lex: English]]].

If we generate a Chinese sentence, using this intermediate representation directly, we will get "他是個好的英語講者". That is an inappropriate sentence in Chinese. To get the appropriate sentence, we need to raise the noun phrase to the verb phrase position, resulting the DAG in Figure 3 as the target intermediate representation, with Chinese lexical item inserted in the DAG.



Figure 3. DAG of 他講英語講得很好

The corresponding linear form is

[cat : event, features : [transitive, post_mod], agent : [cat : pro, lex: 他)], mod:[cat: adj, lex 很好], head: [cat: v, lex: 講], theme:[cat: n, lex: 英語]].

Using this target representation, we can generate a fluent Chinese sentence. The crucial point is that how and when we can raise Figure 2 to Figure 3 to get suitable target intermediate representation and then generate a appropriate Chinese sentence. Let us consider, for example, some possible lexical entries for the noun "speaker" :

- lex(speaker, hn, [human], [head_of(theme), is(head_of(sentence):cat, bv))], [delete(head), raise(*)],講).
- lex(speaker, hn, [human], [head_of(theme_mod)], [raise(*),insert_f(theme,transitive), modify(theme:cat:entity, theme:cat:event)],講).

lex(speaker, hn, [inanimate], [], [], 喇叭)

Six arguments in each lexical entry are (1)English word, (2)part of speech, (3)semantic attribute list, (4)condition test list, (5)transfer action list and (6)Chinese word. When we unify the first lexical entry of "speaker", the condition list, [head_of(theme), is(head_of(sentence):cat, bv))],will be instantiated. If the logic form satisfies this condition list, that is, "speaker" is the head of theme

and the *cat* of the head of theme is bv, we then execute the action described in transfer rule list, raising theme to upper level of DAG and automatically replacing the original head of sentence with " \ddagger ".

Example (2'): We consider him a good speaker of English.

The source DAG is shown in Figure 4.



Figure 4. DAG of We consider him a good speaker of English

It is like above example (a), we execute the *raise* operation. But in this case, when we raise "講" up to clause level, it can not replace the original sentence head "認為", because "認為" is a verb with substance unlike "is". So the right thing to do seems to be raise the np containing this lexical item to a clausal level, by changing the *cat* slot from entity to event and inserting transitivity in the features slot.



Fig 5. DAG of 我們認為他講英語講得很好

3.2.2. Modify and Insert operations

Example (16-a): This surprised everybody.

The source DAG is shown in Figure 6.



Figure 6. DAG of This surprised everybody

The equivalent Chinese lexical item for "surprise" is "使 震驚", so the "使" need to occur before the *experiencer* role. Let us consider the lexical entry of "surprise".

lex(surprise,v,_,[],[modify(head,head'),insert(head,"使")],震驚

Executing the transfer operations described in Transfer-list, we will get new DAG shown in Figure 7.



Figure 7. DAG of 這使大家震驚

3.2.3. Delete operation

Example (18-a): They lived a happy life.

The source DAG is shown in Figure 8.



Figure 8 DAG of They lived a happy life

Live in English is transitive while "<u>生</u>活" in Chinese is intransitive, so the "life" must be deleted in target DAG. So the lexical entry of "live" is as follows:

lex(live,v,_,[is(head_of(theme):lex, "life")],
 [raise(theme:mod),delete(theme),modify(f:transitive,f:intransitive)]
 ,生活).

Executing the *delete* operation, we will get the new DAG as Figure 9.



Figure 9. DAG of 他們生活得很快樂

3.2.4 Relative Clauses

Many relative clauses begin with a **relative pronoun**. The relative pronoun can acts as the subject or object in the relative clause, which always influences the translation of a sentence. In following, we will discuss the translation of relative clauses using the proposed mechanism. We will consider the two kinds of relative clauses, *defining* and *non-defining* relative clauses.

3.2.4.1 The translation of Defining Relative Clauses

Defining relative clauses explain which person or thing you are talking about. For example, if you say 'the teacher', it might not be clear who you mean, so you might say, 'The teacher who taught me English married yesterday'. In this sentence, 'who taught me English' is a defining relative

clause. Defining relative clause is a kind of quantifier, so the whole sentence should be translated as '教我英文的那個老師,昨天結婚了'

3.2.4.2 The Translation for Non-defining Relative Clause

Non-defining relative clauses give further information which is not needed to identify the person, thing, or group you are talking about. For example, 'The teacher, who taught me English, married yesterday' should be translated as '那個老師, 教我英文,昨天結婚了' Here 'who taught me English' is only the added information, so we can take advantage of the topic-comment construction of Chinese, treat it as a comment in Chinese sentence and place it after the topic/subject.

Another important issue is the translation of the relative pronoun itself in non-defining relative clauses. For example, 'I like English, which is an interesting language' should be translated as '我喜歡英文,英文是種有趣的語言'. The relative pronoun 'which' refers to the preceeding object 'English' and often appears in non-anaphoric form in Chinese.

3.2.4.3 The Informations in Lexicon

We can then encode the above informations about the translation of relative clauses in lexicon as following.

lex(who, ip, _,[],[],誰)

The ' **†** ' sign means the upper level (parent node) of current node.

By these lexical transfer rules, the following English intermediate representation can be transferred to appropriate Chinese intermediate representation and then turn into suitable translation. For example:

[cat:event, f:[transitive],

time:[cat:n,lex:yesterday]]



Figure 10. The DAG of English relative clause



Figure 11. The DAG of Chinese defining relative clause



Figure 12. The DAG of Chinese non-defining relative clause

3.3 How to Deal with Special Constructions

In addition to processing structure mismatches problem, our lexical transfer rules in lexicon can also deal with lexical influences on constructions that we described in Section 2. To consider the sentences (19-a), *She sang sadly*, and (20-a), *She sang that song well*, the default structure building rule in our Chinese systemic grammar specify that the adverb should be located before verb, such as sentence (19-b),她爱傷地唱歌. But this default rule is not suitable for the case of

(20-a), because (20-b), 她很好地唱歌, is inappropriate. Let us consider the lexical entries for the adverb "well":

lex(well,adv,_,[], [insert_f(post_mod)],很好).

The condition list is empty, so the transfer action in transfer rule list will be fired, inserting feature *post_mod* in current level in DAG. Using this new DAG to generate Chinese sentence, we will get modifier into the right position, such as (20-b), 她唱歌唱得很好.

Similarly, Let us examine the sentences (21)-(24). Whether an English sentence with passive voice should be translated into *bei_construction* in Chinese or not is determined by the attributes of the Chinese head verb. So the lexical entries for the verbs "steal" and "publish" are as follows :

lex(steal,v,_,[with(passive)],[insert_f(bei)],倚). lex(publish,v,_,[with(passive),has(agent)], insert_f(shi_der), 出版). lex(publish,v_,[],[],出版).

As we mentioned in the previous section, since the verb "偷" has the attributes of *adverse* and *disposal*, we will generate bei-construction (被字句) for the passive sentence.

Besides structure changes caused by lexical items, selection of suitable target words has been problematic in MT. For example, the negative forms in English can be translated into different target expressions, sometimes depending on the target word of what item be negated. So we need a lexical transfer rule to decide what negative form to use in surface generation phase. Consider the lexical entry of adjective "valuable" :

```
lex(valuable, a, _, [with(negative)],[insert_f(mei)],有價值).
```

In the surface generation phase we will generated "不" as default negative form based on the syntactic and semantic analysis on the word "valuable". But the equivalent Chinese word containing the head "有", so it is necessary to replace "不" with "沒" in lexical transfer phase, otherwise, the negative form is inappropriate in Chinese.

4. Conclusions

4.1 Summary

Existing transfer-based MT systems, deal with all the differences of the two languages in one complex transfer phase [Chen 1988 b.], In this paper we present an approach that only deal with lexical idiosyncrasy using an active bilingual lexicon in transfer, to minimize the complexity of the transfer unit and ease the task of retargeting of a translator. It is a mixture of the interlingua and transfer approach [Tsutsumi 1991]. Let us discuss the advantage of this approach over the conventional transfer method in the following.

Our idea is similar to [Alonso, 1990], in making the intermediate representation as universal as possible based on case grammar [Fillmore, 1971],[Tang, 1975],[Huang and Wang 1988]. But this representation deviates from the interlingual approach in that it does not include a universal representation for lexical items [Nirenburg 1990]. For the characteristic of Interlingua, our system guarantees the independence of analysis and generation grammars, which is a basic requirement for practical multilingual MT systems, and at the same time, minimizes the size and complexity of the transfer modules, by using a bilingual lexicon.

Using this approach, we intend to achieve the following goals:

(1) The transfer module for a language-pair is reduced to the bilingual lexicon. The global syntactical reorganization is dealt with using a generator with an explicit grammar of the target language [Kuo 1989, Chen 1990 and Liao 1990].

(2) The source language analysis module is target-language independent. The analysis module produces an intermediate representation as output which is as interlingua as possible.

(3) The target language generation module is based on an explicit grammar which is completely source-language independent.

4.2 Future Work

Our system can be improved in the following respect:

(1) Extending the scope of the grammar : The grammar used in our system does not have a very large scope. We feel that the inclusion of the following is most urgent in improving translation quality:

1. Interogative sentences: Question word questions.

A-not-A question.

Particle question.

2. Serial verb construction.

3. Nominalization.

Besides, some existing parts should also be extended, such as arrangement of various cases in different type of sentences and selection of conjunction.

(2) **Implementing macros in transfer rules :** we can further define macros to represent the relative tests and actions for transfer. It can minimize the size and ease the maintainence of the transfer modules.

(3) Using corpus to train transfer rules : we plan to use a large English-Chinese bilingual corpus to train transfer rules stochastically. In this way, we hope to ease the work of analyzing and formulating transfer rules between these two languages.

5. Acknowledgement

This research was supported by the National Science Council under Contracts NSC79-0408-E007-17 and NSC80-0408-E007-13 and by the Telecommunication Laboratories under Contract TL-NSC-79-016.

References

ζ

[1] James Allen, Natural Language Understanding, The Benjamin/Cummings Publishing Company, Inc. pp 215, 1987.

[2] Juan Alberto Alonso, Transfer InterStructure: Designing an "Interlingua" for Transfer-based MT Systems, COLING 1990.

[3] D.A. Chen, English and Chinese-A Comparative Study,新學識文教出版中心, 1988.

[4] J.R. Chen, And I.P.Lin, 利用ATN製作英中機器自動翻譯轉換生成子系統, Master Paper, Department of Computer Engineering, National Taiwan University, Taipei, Taiwan, 1988.

[5] C.C. Chen and J.S.Chang, A Multi-Lingual Sentence Generator Based on Systemic Grammar, Master Paper, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, 1990.

[6] C.J. Fillmore, *Some Problems for Case Grammar*, Gerogetown University Monograph Series of Languages and Linguistics 22. PP 35-36, 1971.

[7] C.R Huang and L.J Wang, 格 框 在 機 器 翻 譯 的 應 用, ROCLING I, pp 99-125, 1988.

[8] H.W. Kuo and J.S Chang, Systemic Generation of Chinese Sentences, ROCLING II, pp. 187-212, 1989.

[9] J.W.Liao and J.S.Chang, Computer Generation of Chinese Commentary on Othello Games, ROCLING III, pp. 393-415, 1990.

[10] C.N. Li, and S.A. Thompson, *Mandarin Chinese - A Functional Reference Grammar*, University of California Press, California.

[11] Sergei Nirenburg, Lexical and Conceptual Structure for Knowledge-Based Machine Translation, ROCLING III, pp 105-130, 1990.

[12] Slocum, J. 1985. A survey of Machine Translation: Its History, Current Status, and Future Prospects. AJCL 11(1):1-17.

[13] Taijiro Tsutsumi, Wide-Range Restructuring of Intermediate Representations in Machine Translation, AJCL 16(2):71-78.

[14] T.C.C. Tang, A Case Grammar Classification of Chinese Verbs, Hai-Guo Book Company, Taipei, Twiwan 1975.

[15] Gr. Thurmair, Complex Lexical Transfer in METEL, COLING 1990.

[16] Jun-ichi Tsujii and Kimikazu Fujita, Lexical Transfer based on Bilingual Signs -Towards Interaction during Transfer-, ROCLING III, pp 141-157, 1990.

[17] C.C. Wu, Chinese-English Translation through Contrastive analysis, 文 鹤 出 版 有 限 公 司, 1990 2nd Edition.