Transfer-Based Japanese-Chinese Translation Implemented on an e-mail System

MATSUDA Junichi, KUMAI Hiroyuki
Hitachi, Ltd.,
Central Research Laboratory
Japan

Abstract

We have developed a Japanese-Chinese translation engine that uses a transfer method. This system directly transfers Japanese structure into the equivalent Chinese one (including idiomatic phrases) and can select Chinese equivalent words by using three kinds of co-occurrence data. We have implemented the translation engine on an e-mail system. And the system can send and receive e-mail written in CJK languages, so the sender and receiver can communicate with each other in their own native languages.

1 Introduction

Personal Computers and the Internet are becoming ever more popular around the world. Especially, e-mail has a major communication tool between countries. English is usually the common language used in this communication. However, to communicate more smoothly, it would be better to use each country's native language instead of English[1,2]. Nowadays, in Japan, communication with other Asian countries has become important. We have therefore developed a Japanese-Chinese translation method and implemented it on an e-mail system.

2 Input and Output method

At present, PCs and workstations used in non-English-speaking regions can only handle their native language and English, so it is difficult to display Japanese and, for example Chinese on the same screen.

Recently, to solve this problem, the Universal Multiple-Octet coded character set ISO/IEC10646-1 was standardized according to Unicode. An operating system using Unicode has been developed so the basis for multilingual processing will be ready soon[3].

3 Japanese-Chinese Translation System

3.1 Language Structures of Japanese and Chinese

Japanese and Chinese belong to different language groups, so They have some structural differences as listed below[4].

	Chinese	Japanese
Object location	after verb	before verb
Preposition /postposition	preposition	postposition
Adverb location	before verb /after verb*1	before verb

*1: usually locates before the verb, but frequency adverbs locate after the verb.

However, Japanese and Chinese both use Kanji characters and there are some similarities as follows.

(a) Phrases modifying a noun locate before the noun. Examples:

Japanese: 私が読んだ本
Chinese: 我读的书

(b) Similar quantifiers are used. Examples:

Japanese:5枚の紙
Chinese:五张纸

(c) Similar nouns that indicates direction. Examples:

Japanese:上、前、時
Chinese ：上，前，时候

(d) Sequences of verb phrases have various meanings, but the structure is the same and it is not necessary to analyze the relation between verbs. Examples:

Japanese: 彼は座って本を読んでいる
Chinese: 他坐着看书

(e) Auxiliary verbs have various meanings, but there is sometimes one-to-one correspondence. Examples:

Japanese: 持ってくる
Chinese: 拿来

(f) Word order is not fixed. Examples:

The following two sentences mean "I will go soon".

Japanese: 私はすぐに行きます。

Japanese: すぐに私は行きます。

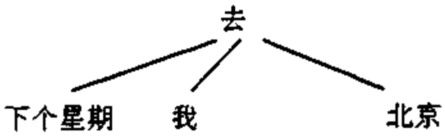


Fig. 4 Chinese Dependency Structure

The following two sentences means "I met him in Beijing".

Chinese: 在北京我见过他.

Chinese: 我在北京见过他.

3.2 Translation Method

Since the structure of Chinese and Japanese is partly similar, we use the transfer method. If the structure is similar, the analysis level is low and the system directly transfers Japanese syntactic structure into the Chinese one. Moreover, Chinese has rich idiomatic expressions. In order to translate idiomatic phrases, it is effective to transfer the Japanese word sequence into the equivalent Chinese one. That is, this transfer is quick and involves simple translation rules. The translation system has three levels of transfer and selects the suitable level according to the sentence. Figure 1 shows the translation flow.

The transfer stages are explained as follows.

(1) Word-level transfer

This level is used to transfer a Japanese word sequence into the Chinese one. Some idiomatic phrases are transferred at this level.

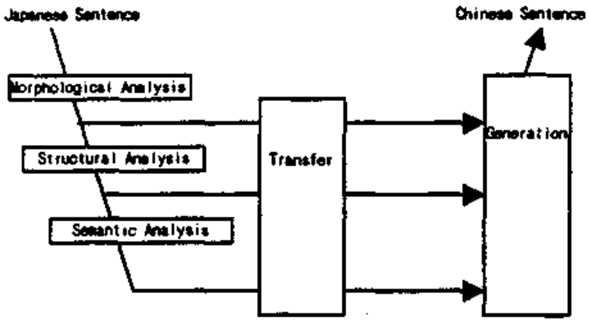


Fig.1 Japanese-Chinese Translation Method

(2) Syntax-level transfer

This level is used to translate some Japanese phrases that are similar to Chinese equivalent phrases.

(3) Semantic-level transfer

This level is used to translate Japanese phrases that are needed to analyze case relations.

3.3 Chinese Sentence Generation

The result of syntactic and semantic analysis is represented as an arrow graph. Nodes represent words and arcs represent relations between nodes. Figure 2 shows the result of syntactic analysis of the Japanese sentence "彼は座って本を読んでいる。". Figure 3 shows the result of semantic analysis of the Japanese sentence "来週中国に私は行きます。".

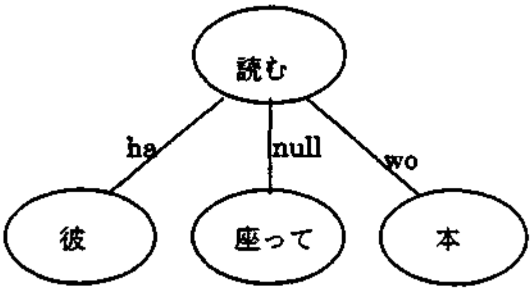


Fig. 2 Result of syntactic analysis

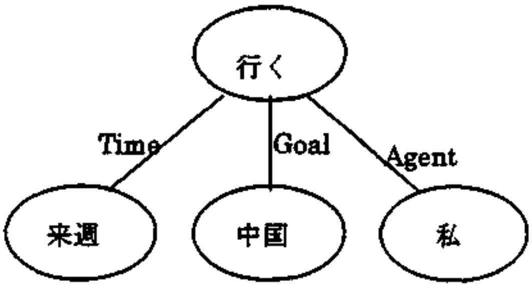


Fig. 3 Result of semantic analysis

In the analysis process, the system decides whether semantic analysis is necessary or not.

From the syntactic or semantic process, a Chinese dependency structure is generated. Figure 4 shows an example of a Chinese dependency structure. The structure means that "下个星期" and "我" modify "去" from the front and that "北京" modifies "去" from the back. This structure corresponds to the sentence "下个星期我去北京".

In the generation process, the system uses generation rules that correspond to the Chinese dependency structure. Examples of generation rules that generate the sentence in Fig. 4 are as follows.

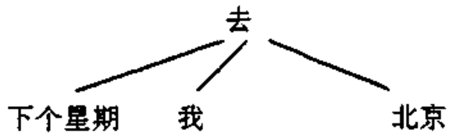


Fig. 4 Chinese Dependency Structure

Rule 1: agent main goal
Rule 2: time main

Here, "main" means main words. Rule 1 means agent locates before the main verb and the goal locates after the verb. Rule 2 means time locates before the main verb.

3.4 Translation of Idiomatic Phrases

Chinese has rich idiomatic phrases. We use a word-level transfer to translate these idiomatic phrases. The reason is as follows.

(a) Idiomatic structures of Chinese and Japanese are sometimes similar. Syntax is therefore not necessary to transfer the idiomatic structure.

(b) There are many idiomatic rules. Transfer rules must therefore be simple so that it is easy to implement a large amount of rules.

An example of these rules is as follows:

x-ば+x-ほど+*y/越-x-越-
*y/x:ps=adjective,y:ps=adjective

where variables "x" and "y" are phrases and "ps" means part of speech. The rule consists of three parts divided by "/". The first part is a Japanese word sequence. The second part is the Chinese equivalent word sequence. The third part is a condition.

If the result of morphological analysis matches the first part and satisfies the condition part, the word sequence is directly transfered into the Chinese word sequence. Only variables "x" or "y" go through the syntactic and semantic analysis paths.

Examples:

(a) 「データの送付は速ければ速いほどよい。」
x-ば+x-ほど+*y/越-x-越-*y
/x:ps=adjective,y:ps=verb
「数据 的 传送 越 快 越 好 。」

(b) 「どうも風邪をひいたようだ。」
どうも+x-た-*ようだ/好象-x-了/x:ps=verb
「好象 感冒 了 。」

3.5 Selection of Equivalent Words

In order to select the proper Chinese equivalence, we use three kinds of co-occurrence data.

(1) syntactic co-occurrence
This is co-occurrence of syntactically related words.
Examples:
(a)市場で野菜を求めた。
在市场买蔬菜.
(b)社長に面会を求めた。
向总经理要求面会.

In the above example, to select equivalence of the verb "求める", co-occurrence of the verb and its object is used. The co-occurrence is described as follows.

买-object-noun(attribute:goods)
要求-object-noun(spelling:面会)

(2) co-occurrence of neighboring words (before)
This is co-occurrence of neighboring words located before the target word.
Examples:
(a)講演プログラム
讲演节目
(b)計算機プログラム
计算机程序

In the above example, to select equivalence of the noun "プログラム", co-occurrence of a noun located before "プログラム" is used. The co-occurrence is described as follows.

noun(spelling:讲演)-节目
noun(spelling:计算机)-程序

(3) co-occurrence of neighboring words (after)
This is co-occurrence data with the neighboring word located after the target word.
Examples:
(a)プログラム言語
程序语言

In the above example, to select equivalence of the noun "プログラム", co-occurrence data with the noun located after "プログラム" is used. The co-occurrence is given below.

节目-noun(spelling: 语言)

4 Experiment

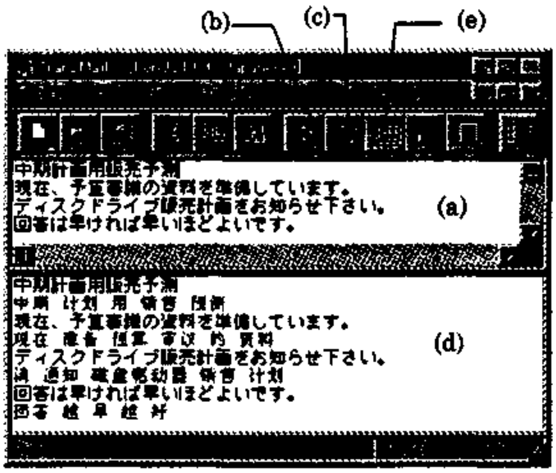
For our experiment, we use 2300 sentences including technical document, newspaper and conversation. 23% of the sentences are exactly correct but other sentences have some errors to be rewritten. The total number of errors are 4473 and the content of errors are as follows.

#	errors	number	%
1	Selection of Chinese equivalent word is not suitable	1528	34.1
2	Word order is not suitable	926	20.7
3	Japanese word is Unknown	367	8.2
4	Japanese postposition is not translated correctly	521	11.6
5	Japanese idiomatic phrase is not translated correctly	208	4.7
6	Compound noun is not translated correctly	190	4.2
7	Chinese auxiliary verb that means tense or aspect is not translated correctly	345	7.7
8	Others	388	8.7

Some types of errors can be solved by adding co-occurrence data or rules for idiomatic phrases as described below.

#	Solution method
1	To add syntactic co-occurrence data
5	To add rules for idiomatic phrases
6	To add co-occurrence data of neighboring words

43% of the errors can be solved by adding some data or rules without changing translation algorithm. These data or rules can be easily added without the knowledge of machine translation.



中期計画用販売予測
現在、予算審議の資料を準備しています。
ディスクドライブ販売計画をお知らせ下さい。
回答は早ければ早いほどよいです。

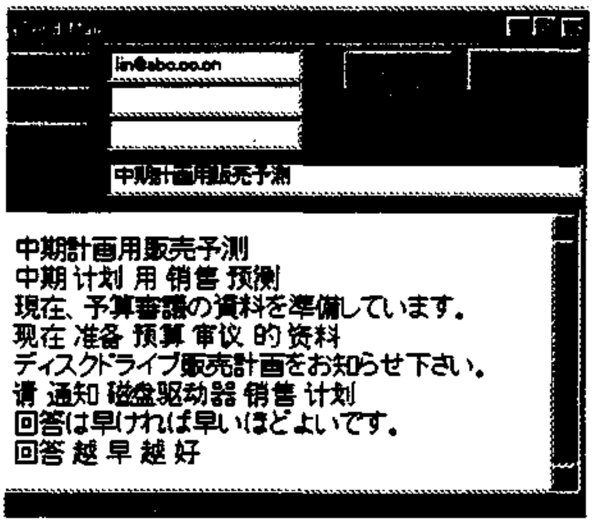
中期計画用販売予測
中期 计划 用 销售 预测
現在、予算審議の資料を準備しています。
现在 准备 预算 审议 的 资料
ディスクドライブ販売計画をお知らせ下さい。
请 通知 磁盘驱动器 销售 计划
回答は早ければ早いほどよいです。
回答 越 早 越 好

Fig. 5 Screen of the e-mail system

5 Implementation on e-mail system

We have implemented the multilingual translation on e-mail system. Figure 5 shows a screen of the e-mail system.

Original sentences are inputted in area (a). The target language can be is selected by button(b). By putting the translation button(c), original and translated sentences are displayed in area (d). Then, by putting the button (e), mail documents are generated and the send-mail screen is displayed. Figure 6 shows a send-mail screen.



lin@abc.co.cn

中期計画用販売予測

中期計画用販売予測
中期 计划 用 销售 预测
現在、予算審議の資料を準備しています。
现在 准备 预算 审议 的 资料
ディスクドライブ販売計画をお知らせ下さい。
请 通知 磁盘驱动器 销售 计划
回答は早ければ早いほどよいです。
回答 越 早 越 好

Fig. 6 Send-mail Screen

The system send both original and translated documents. If the receiver can not understand the exact meaning of the translated sentences, he or she can refer the original documents. Since the system uses Unicode, e-mail in any language can be read.

Also, by adding a translation engine, the system can easily be expanded to increase the number of languages. Figure 7 shows the system configuration.

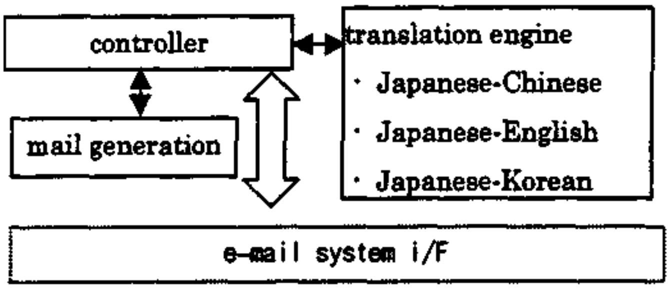


Fig. 7 System Configuration

6 Conclusion

We have developed a Japanese-Chinese translation engine that uses a transfer method. This system directly transfers Japanese structure into the equivalent Chinese one (including idiomatic phrases) and can select Chinese equivalent words by using three kinds of co-occurrence data. We had experiment on 2300 sentences. As a result, 43% of the errors can be solved by adding co-occurrence data or rules for idiomatic phrases. We have implemented the translation engine on an e-mail system. And the system can send and receive e-mail written in CJK languages, so the sender and receiver can communicate with each other in their own native languages.

References

[1] Nishigaki, T., Internet for Multilingual Era Sekai No. 640 Iwanami Shoten, 1997

[2] Miura, What is Multilingualism? Fujiwara Soten, 1997

[3] CICC, Joint Development Research on International Standardization: Multilingual Information Technology, 1999

[4] Mino., Basic Chinese Grammar. Sanshusha