ALT-J/C
A Prototype Japanese-to-Chinese Automatic Language Translation System

Minoru Hayashi, Setsuo Yamada, Akira Kataoka and Akio Yokoo

NTT Communication Science Laboratories
2-4, Hikaridai, Seika-chou, Soraku-gun, Kyoto 619-0237 Japan
hayashi@cslab.kecl.ntt.co.jp

Abstract

This paper describes a prototype Japanese-to-Chinese automatic language translation system. ALT-J/C (Automatic Language Translator - Japanese-to-Chinese) is a semantic transfer based system, which is based on ALT-J/E (a Japanese-to-English system), but written to cope with Unicode. It is also designed to cope with constructions specific to Chinese. This system has the potential to become a framework for multilingual translation systems.

1 Introduction

The history of telecommunications is the history of the expansion of information, and efforts to overcome the barrier of distance, overcome the barrier of time and overcome the barrier of language. To remove the barrier of language several automatic language translation systems have been developed. There are especially a lot of Native-to-English and English-to-Native translation systems that have been developed. Some multilingual translation systems also have been developed (Mitamura 1999, Fu 1999). But even now there are not so many translation systems that have been developed between Asian languages. There are a small number of Japanese-to-Chinese translation systems and Chinese-to-Japanese translation systems, but they are not of good quality.

We began to develop a Japanese-to-Chinese automatic language translation system to contribute to a framework of multilingual translation systems. The Chinese language originates from the Han language incorporated with external linguistic factors. Chinese has more than 50,000 characters and has a complicated grammatical structure. The main reason for developing a Japanese-to-Chinese translation system is that we want to develop a system that has the potential to become a framework for multilingual translation systems developed from a Japanese-to-Chinese translation system. In addition, recently the number of potential users for such a Japanese-to-Chinese translation system has increased. In this study we handle Mandarin, a modern version of standard Chinese.

In this paper we introduce ALT-J/C, a prototype Japanese-to-Chinese automatic language translation system that is a semantic transfer based system. The system implementation is rewritten to cope with Unicode based on the translation engine ALT-J/E (Ikehara et al. 1987, Ikehara et al. 1991), a Japanese-to-English system. ALT-J/E once has been extended to ALT-J/M (Ogura et al. 1999), a Japanese-to-Malay system. In this paper we further extend the system to handle Chinese. To do this we rewrote the system to cope with Unicode, as well as extended the translation system to cope with constructions specific to Chinese. In the next section we describe the translation method of ALT-J/C. Then we look at semantic dictionaries and structure dictionaries in sections 3 and 4. In section 5 we look at some examples of how the system translates. And then we discuss and summarize in sections 6 and 7.

2 Translation Method

ALT-J/C uses a translation method based on the constructive process theory, in which grammar focuses attention on the speaker's intention and recognition. This hierarchical translation method is based on semantic analysis and multi-level translation at appropriate granularity. The translation method is based on two ideas. The first is the separation and synthesis of subjective expressions and objective expressions. Subjective expressions are unconceptualized expressions that express the speaker's emotion, intention and will. Objective expressions are conceptualized expressions that describe the objective world. The second idea is the use of multi-level transfer. This overcomes the limitations of conventional compositional semantics.

As shown in Figure 1, the Japanese input sentence is separated into subjective expressions, such as tense and modal information, and objective expressions, the kernel sentence. The objective part is translated using the multi-level transfer; direct parse tree transfer, idiomatic expression transfer, semantic valency transfer and general pattern transfer processes are applied.

The processing of the automatic language translator, Japanese-to-Chinese system, is as follows. First, the system splits input Japanese text into morphemes. Second, the system analyzes the sentence syntactically. Next, it rewrites complicated Japanese expressions into more easily translatable ones and it semantically evaluates the various interpretations. Then, syntactic and semantic criteria are used to select the best interpretation. This last interpretation is used as input to generate Chinese. Finally, the Chinese sentence is adjusted to give the correct tense and aspect.

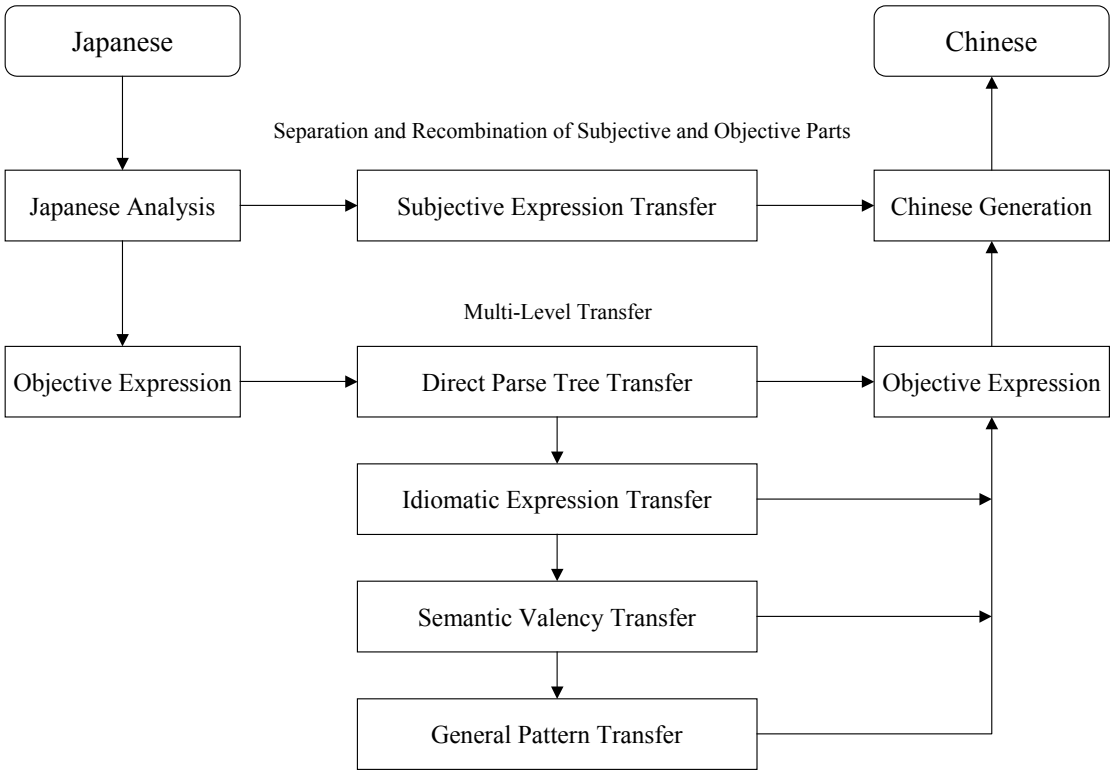

Figure 1: The Multi-Level Translation Method

ALT-J/C has been constructed as a research experimental system and has not been optimised for speed. ALT-J/C uses CommonLisp and runs under Unix OS on both Solaris and Linux. The system implementation is rewritten to cope with Unicode, because Japanese and Chinese use different character sets.

In order to realize semantic analysis we use semantic dictionaries and structure dictionaries, dictionaries based on a precise and detailed semantic attribute system.

3 Semantic Dictionaries

Our lexicons provide detailed information about the use and meanings of words. Most words can express several concepts. Which concept a word expresses in a text is determined by its relationship with the other words in the sentence. For example the noun "hotel" is used as "a common noun - concrete - an agent - an organization - a corporation " and as "a common noun - concrete - a place - a facility - a lodging facility" as shown in Figure 2.

This ontology classifies concepts to use in expressing relationships between words. It represents a kind of common knowledge. Relationships between concepts and words are described in our semantic dictionaries. In ALT-J/C, the ontology has several hierarchies of concepts: 3,000 attributes with a 12-level tree structure for common nouns, 200 attributes with a 9-level tree structure for proper nouns and 108 attributes for predicates.

To transfer Japanese words to Chinese words we provided a Japanese-to-Chinese transfer word semantic dictionary. Japanese and Chinese have different character sets. So in order to enter or display Japanese words and Chinese words at the same time, we implemented a web-based semantic dictionary maintenance tool that is written to cope with Unicode.

Figure 3 shows an example of the web-based semantic dictionary maintenance tool. Each record has a Japanese index form, a Chinese index form, Chinese syntactic information and Chinese semantic attributes. Each word can have up to five common noun attributes and five proper noun attributes. In the case of ホテル *hoteru* "hotel", there are two common noun attributes, 437 (lodging facility) and 374 (corporation) as shown in Fig. 3, and no proper noun attributes. The dictionary created by the use of this tool is used when the system generates Chinese sentences.

In addition, this semantic dictionary also has information about numeral classifiers. There is a great deal of variation, as a classifier can select not just for the object denoted by its target but also for a sub-part of it (Bond et al. 1998, Bond et al. 2000). This semantic dictionary is also used when the system generates Chinese sentences.

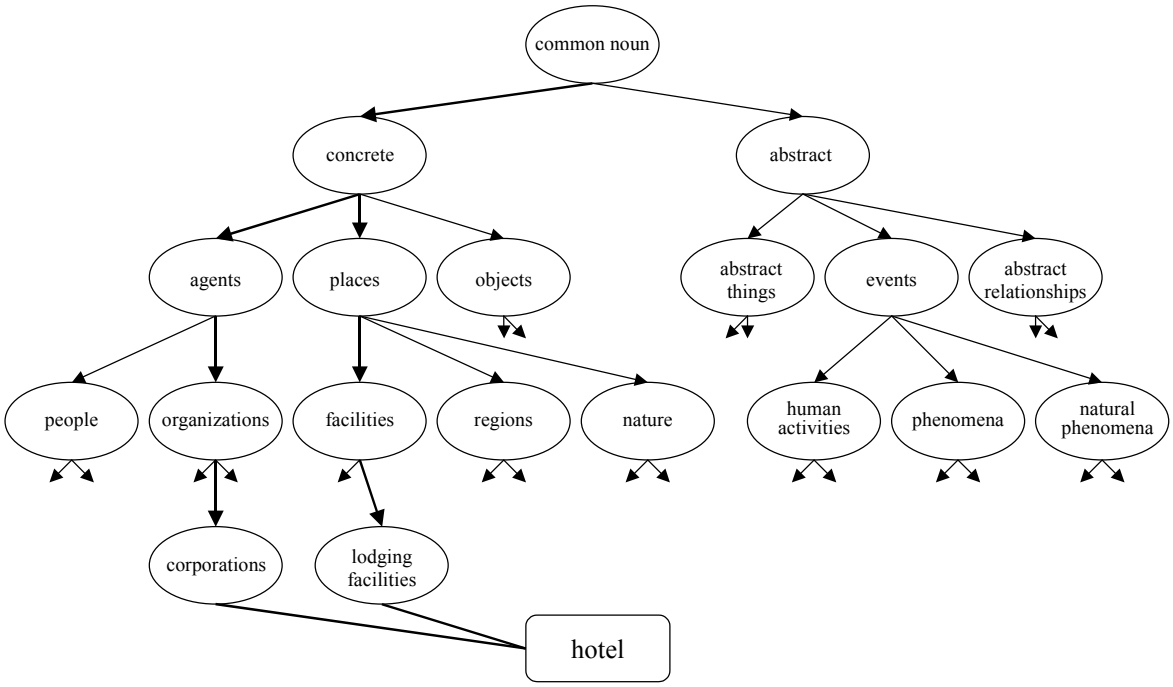

Figure 2: A Semantic Attribute System for Common Nouns

Figure 3: A Semantic Dictionary Maintenance Tool Image.

4 Structure Dictionaries

In the structure dictionary, Japanese-to-Chinese pattern pairs are defined for Japanese verbs using the words defined in the common noun attribute system. A verbal semantic attribute is also given to every pattern pair. This attribute denotes the meaning and function of the pattern in actual sentences. The basic structure of a sentence comes from the relationship between the main predicate and the nouns. A structure transfer dictionary is provided for producing basic sentence structures.

Figure 4 shows an example of the structure dictionary. The structure dictionary describes under which conditions a Japanese verb should translate into which Chinese verb. The Japanese verb 取る *toru* has many meanings and can be translated into various Chinese words. Three structures are shown in Fig. 4: "N1<subject> が *ga* N2<permission, agreement, score> を取る *o-toru*" --- N1 得到 **dédào** N2 "N1 get N2"; "N1<subject> が *ga* N2<seat, lodging facilities, tickets> を取る *o-toru*" --- N1 预约 **yùyuē** N2 "N1 reserve N2"; "N1<subject> が *ga* N2<vacation, leisure> を取る *o-toru*" --- N1 请 **qǐng** N2 "N1 take N2". Each predicate has one or more case-slots associated with it. Each case-slot has information such as grammatical function, case-marker, case-role and semantic restraints on the filler and default order.

N1<主体>が N2<許可、賛成、成績>を 取る ⇨ N1 得到 N2

N1<主体>が N2<席、宿泊施設、券>を 取る ⇨ N1 预约 N2

N1<主体>が N2<休暇、余暇>を 取る ⇨ N1 请 N2

Figure 4: An Example of the Structure Dictionary

5 Translation Examples

In this section we describe examples of Japanese-to-Chinese automatic language translation results, especially translation of verbs and numeral classifiers.

5.1 Correct translation of verbs

For example, when the Japanese input sentence is

"私は　休暇を 取る 許可を　取り、

"*watashi-wa kyuka-o toru kyoka-o tori,*

"I　vacation take permission take,

妻は　ホテルを 取った。"

tsuma-wa hoteru-o totta."

wife　hotel　took."

the ALT-J/C system splits the Japanese sentence into morphemes, and analyzes the sentence syntactically. Then, syntactic and semantic criteria are used to select the best interpretation. In the case of the verb 取る *toru* "take", the structure dictionary, described in section 4, used semantic categories to choose the translation as follows: "take permission" --- 得到 **dédào** "get"; "take vacation" --- 请 **qǐng** "take"; "take hotel" --- 预约 **yùyuē** "reserve". Then, the Chinese sentence is adjusted to give the correct aspect as follows:

"我 得到 了　请 假　的　许可，

"**wǒ dédào le　qǐng jiá　de　xǔkě,**

"I　get　PAST take vacation MOD permission,

妻子 预约　了　宾馆。"

qīzi yùyuē le　bīnguǎn."

wife　reserve　PAST hotel."

The translated result is the upper right sentence in the bottom table in Figure 5. An English translation is

"I got permission to take a vacation and my wife reserved a hotel."

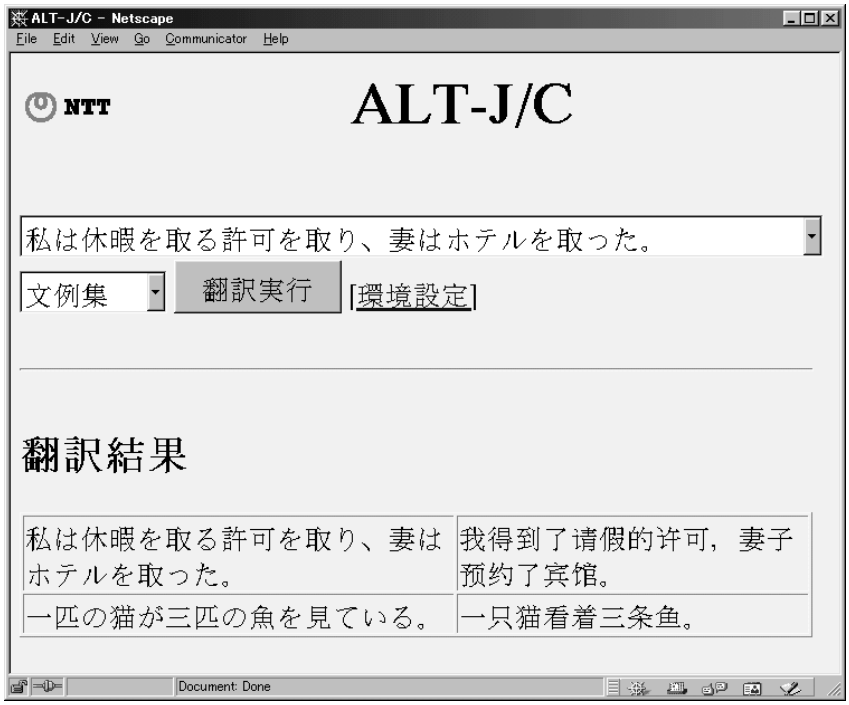

Figure 5: Two Examples of Translated Results.

5.2 Correct translation of numeral classifiers

As another example, when the Japanese input sentence is

"一匹の　猫が　三匹の　魚を　見ている。"

"*ip-piki-no neko-ga san-biki-no sakana-o miteiru.*"

"one CL　cat　three CL　fish　looking at"

the ALT-J/C system splits the sentence into morphemes, and analyzes the sentence syntactically. Then, syntactic and semantic criteria are used to select the best interpretation. The semantic dictionary has numeral classifier combination information. For the Japanese numeral classifier 匹 *hiki (piki, biki)* "animal", the semantic dictionary uses semantic categories to choose the translation as follows: "cat" --- 只 **zhī** "small animal"; "fish" --- 条 **tíao** "long animal". And then the Chinese sentence is adjusted to give the correct aspect as follows:

"一　只　猫　看　着　三　条　鱼。"

"**yì　zhī māo kàn　zhe　sān　tíao yú.**"

"one CL　cat　look at PROGRESSIVE three CL　fish."

The translated result is the lower right sentence in Figure 5. An English translation is

"A cat is looking at three fish."

6 Discussion

We have just started to develop a prototype Japanese-to-Chinese automatic language translation system. We aim to increase the size of the lexicons, and test the system with more grammar structures.

The Chinese language has a number of unique grammatical features, for example, measures for nouns, pronouns, the verb 是 **shì**; nominal predicates, the verb 有 **yǒu**; comparisons, motion verbs and direction indicators, verb and location, modal and similar verbs, preposition and coverbs, and so on. Particularly, preposition and coverb constructions are different from Japanese and from English.

One challenging subject is the 把 **bǎ** construction. Usually the Chinese sentence pattern is "**subject + verb + object**", similar to English. The 把 **bǎ** construction is a grammatical feature unique to the Chinese language. In this construction the coverb 把 **bǎ** has the function of shifting the object of the verb to a pre-verbal position in the pattern of "**subject + bǎ + object + verb**". The pattern "**subject + object + verb**" is similar to Japanese. So to analyze and translate the 把 **bǎ** construction is a very challenging subject for Japanese-to-Chinese automatic language translation systems. Our next subject of research is to incorporate the above grammatical structure, and others, into the system.

Almost all Asian languages have different character sets. Our prototype system can handle a variety of characters, because it is written to cope with Unicode. We hope this prototype system will become the basis of a framework for multilingual translation systems.

7 Conclusion

In this paper we described a prototype Japanese-to-Chinese automatic language translation system ALT-J/C (Automatic Language Translator - Japanese-to-Chinese). The system implementation is written to cope with Unicode. So this system has the potential to become a framework for multilingual translation systems.

Acknowledgements

We wish to thank Yoshifumi Ooyama, Kentaro Ogura, Yoshihiro Matsuo, Francis Bond, Naruhiro Ikeda and the members of the Machine Translation Research Group for their valuable discussions and advice. We also wish to thank the members of the NTT Software Corporation who helped to implement ALT-J/C.

References

Bond, F., Daniela, K., & Shirai, S. (1998). 'Anchoring floating quantifiers in Japanese-to-English machine translation', 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98, Montreal, pp.152--159.

Bond, F, & Paik, K. (2000). 'Reusing an ontology to generate numeral classifiers', 19th International Conference on Computational Linguistics: COLING-2000, Saarbrucken, pp. 90--96.

Fu, A. (1999). 'The research and development of machine translation in China', MT Summit VII, Singapore, pp.86--90.

Ikehara, S., Miyazaki, M., Shirai, S., & Hayashi, Y. (1987). 'Speaker's recognition and multi-level-translation method based on it', Transactions of the Information Processing Society of Japan, 28 (12), pp. 1269--1279.

Ikehara, S., Shirai, S., Yokoo, A., & Nakaiwa, H. (1991). 'Toward an MT system without pre-editing-effects of new methods in ALT-J/E', Third Machine Translation Summit , Washinton DC, cmp-lg/9510008, pp. 101--106.

Mitamura, T. (1999). 'Controlled language for multilingual machine translation', MT Summit VII, Singapore, pp.46--52.

Ogura, K., Bond, F. & Ooyama, Y. (1999). 'ALT-J/M: A prototype Japanese-to-Malay Translation System', MT Summit VII, Singapore, pp.444--448.