Prèmieres journées franco-japonaises sur la traduction assistée par ordinateur, pp.117-122, 1993.

Construction of a Bilingual Dictionary for DUET-E/J — Toward High Performance MT

Ichiko SATA

Assistant Supervisor
Development Department II
Information Systems Research and Development Center
Corporate Research and Development Group
Sharp Corporation
492 Minosho-cho, Yamatokoriyama-city, Nara, Japan 639-11
E-mail address: sata@islix.sharp.co.jp

Abstract

This paper outlines the main features of a bilingual general dictionary for the commercial MT system "DUET-E/J", which contains about 88,000 lexical entries. Recently, the structure of the dictionary has been changed in order to include new kinds of information, such as lexical preference and collocations. Moreover the semantic information, which has been enriched, is now carefully encoded, by referring to a large-scale corpus and to a series of paper dictionaries and by extracting discriminant primitives from these and then describing the primitives as semantic constrains in natural language (Japanese) to be later converted into semantic codes by a computer. A preferential score is given to each of the case frames which carry semantically constrained obligatory cases. Thus the lexical preference information contributes both to avoiding a combinatorial explosion of interpretations and to selecting proper word senses.

One of the great advantages of using a bilingual dictionary for MT systems lies in the fact that it is possible to describe exceptional detailed transfer rules in each lexical entry, which enables the DUET-E/J to generate natural-sounding translations.

1 Introduction

DUET-E/J is a transfer-based English-Japanese machine translation system which is implemented on Sharp's desktop and laptop workstation. The prototype system was released in 1988 by Sharp Corporation in Japan. The new version is under development at the Information Systems Research and Development Center at Sharp. Recent changes in the system design have focused on the enhancement of both the translation part and the user interface. The translation part of the new version is characterized by breadth-first parsing and preferential scoring which is based on sophisticated grammar rules and lexical preference information.

This paper especially outlines the main features of DUET-E/J's new general dictionary. This bilingual dictionary has about 88,000 lexical entries, and each entry contains morphological, syntactic and semantic information, subject field codes, exceptional transfer rules and so on. Those pieces of information are required for the syntactic or semantic analysis or the generation, and sometimes for the transfer. To improve the performance of translation, the structure of the dictionary has been changed, its information enriched, and new kinds of information have been included.

2 Lexical Preference

In the grammar rules, there are a lot of precise definitions to determine the most preferred interpretation. In addition to the grammar rules, it is useful to use the lexical preference information, such as the frequencies and the semantic information, to resolve the syntactic and/or the semantic ambiguities.

2.1 Frequencies

Duet-E/J adopts the original 2-pass method in the syntactic analysis to reduce the syntactic ambiguities based both on the grammar rules and on the information about frequencies described in a lexical entry. There are two types of frequency information: frequencies of the parts of speech and frequencies of the verb patterns.

2.1.1 Frequencies of Parts of Speech

One of the major problems for MT arises from the fact that one word has several parts of speech in the source language, which inevitably increases the syntactic ambiguities. To cope with this problem, the syntactic analyzer uses the frequency information about the parts of speech contained in each lexical entry. Currently, every lexical entry is classified into two categories according to the frequency of the parts of speech which the particular word carries, therefore the frequency here is not absolute but relative. One category is “MAJOR”, and the other is “MINOR”. The syntactic analyzer first attempts to use only parts of speech classified as MAJOR, and when the analysis fails, it starts analyzing the same sentence again using MINORs as well as MAJORs. Suppose the input is:

“baby foods”

and there arise two interpretations as follows:

(a) “special foods for babies”

(b) “to treat foods like a baby”.

Of course, (b) is totally nonsensical, but it is a syntactically possible interpretation. To avoid such nonsensical interpretations, like (b), the verb “baby” is classified as “MINOR”, thus the syntactic analyzer successfully interprets the word “baby” as a noun in the source text.

If a lexical entry consists of more than two words and is classified as MINOR, it is treated in a different but more complicated way according to the grammar rules.

2.1.2 Frequencies of Verb Patterns

Predicate verbs which carry obligatory cases are classified into about 120 verb patterns according to Sharp’s original classification, which is based mainly on A. S. Hornby’s Verb Patterns. All the verb patterns are given syntactic scores in the grammar rules. The syntactic scores are not always applicable to every word. In such case, it is possible to describe the exceptional syntactic scores in the lexical entry.

In the following example,

“The economy continues to show signs of both strength and weakness.”

the verb “show” could be interpreted in two different verb patterns, as follows:

(a) “⟨The economy⟩ ⟨continues to show⟩ ⟨signs of both strength and weakness⟩”
(SUBJECT + VERB + NP)

(b) "⟨The economy⟩ ⟨continues to show⟩ ⟨signs of both⟩ ⟨strength and weakness⟩"

(SUBJECT + VERB + NP + (to be) + NP)

Although in the source sentence, the infinitive "to be" does not exist, it is possible to interpret "strength and weakness" as a complement and get (b). To avoid an interpretation like (b), the verb "show" in the verb pattern "SUBJECT + VERB + NP + NP" is given a low syntactic score if the infinitive "to be" is missing.

2.2 Semantic Information

In this bilingual dictionary, each word sense carries its semantic features in the lexical entries of every noun, pronoun, verb, and in some of the entries of adjectives, adverbs etc., while those semantic features are specified by grammatical governors and modifiers such as verbs, adjectives, prepositions, nouns, adverbs, as semantic constraints. The semantic analyzer performs the semantic co-occurrence check by comparing the described semantic constraints and the semantic features of a governed or modified syntactic category.

There are two sorts of semantic features: the set of semantic categories and that of semantic codes. The former consists of about 50 hierarchical labels, and are sometimes used as semantic constraints to reduce syntactic ambiguities. The latter consists of about 3,000 partially hierarchical 3-to-6-digit numbers and are mainly used as constraints for the word sense selection. Since it is impossible for a human to memorize some 3,000 concepts presented by numbers and make good use of them as semantic constraints, each word sense is mapped into its semantic code(s) and stored in the computer, which enables semantic constraints to be described in natural language (Japanese), then converted into semantic codes, resulting in saving time for encoding semantic constraints and making the work more accurate and effective. All the semantic constraints are carefully encoded by referring to a large-scale corpus and to a series of paper dictionaries and by extracting discriminant primitives from these and then describing the primitives in natural language. A preferential score is given to each of the case frames which carry semantically constrained obligatory cases.

2.2.1 Syntactic Preference

The syntactic analysis is more accurate when the semantic information described in the lexical entries is used in addition to the grammar rules. For example, the verb "hold" can be used in the following 2 verb patterns, which look syntactically the same.

(a) SUBJECT + VERB (intransitive verb) + NP (quantity: duration)

something lasts for a certain period.

(b) SUBJECT + VERB (transitive verb) + NP (quantity: amount)

something can contain a certain amount.

The verb pattern (a) has a higher syntactic score than pattern (b). Giving a strict constraint to the case slot "NP" of pattern (a) makes it possible to ensure that pattern (b) is preferred to pattern (a) if the described constraint does not match the semantic feature of the word in the input sentence. Therefore, each of the following 2 sentences can be analyzed properly, using the 2 different verb patterns above.

(a) "This good weather will hold 3 days."

(This good weather will last for 3 days.)

(b) "This tub holds 3 gallons."

(This tub can contain 3 gallons.)

2.2.2 Semantic Preference

There might be no stable and consistent method for resolving semantic ambiguities, but the semantic preference approach adopted by DUET-E/J could be one of the effective methods for this purpose, although it seems to be rather conventional. To select word sense, semantic restrictions are given to governed case slots or modified heads with preference information.

There are, however, a large number of words which are not contained in DUET-E/J's general dictionary, and more and more new words and technical terms are being produced each day. Those unknown words have several types of rather rough default semantic features which are decided by the system according to their morphological features, although it is of course possible for the users to add lexical entries of some of those unknown words to their own user-built dictionaries and give semantic features to them, if they want to. But those semantic features given by the users would still be very rough, and the semantic constraints described in the lexical entries might accidentally match the default or user-determined semantic features, which could possibly damage the selection of the word senses. Besides this, there are some cases where the semantic analyzer is not able to do the semantic co-occurrence check sufficiently; that is, when case slots which have semantic constraints are missing, and when the sentence structure for the checking mechanism is exceptional, etc.

To eliminate this kind of risk, each case slot of a minor word sense is given as strict a semantic constraint as possible with an obligatory flag, which does not allow the semantic analyzer to pass the case pattern, if the semantically obligatory case is missing in a sentence or if it is impossible to check its semantic feature because of the structure of the sentence. Another effort is also made where one of the most general word senses is given top priority among the candidates whose constraints are not so strict and whose constraints might match the semantic features of unknown words. As a result of these efforts, DUET-E/J offers comparatively safe translations when handling sentences which contain unknown words, or whose structure makes the semantic co-occurrence check insufficient.

DUET-E/J also adopts the semantic fail-safe mechanism which allows the semantic analyzer to select the system's default word sense if the given constraints are too strict in the lexical entries and none of the described case frames is passed in the feature co-occurrence check.

3 Subject Field Code

After the syntactic and semantic ambiguities are eliminated, there still remain several candidates, if the lexical entry has a lot of word senses. To solve this problem, the idea of subject field codes was introduced, because it is obvious that the word senses vary with the domain of use. When a user chooses a specific subject field, DUET-E/J selects appropriate word senses for the subject according to the subject fields codes in lexical entries. Those subject field codes are carefully given by referring to a lot of technical term dictionaries, encyclopedias, etc. and by the advice of cooperative professional translators. There are currently seven subject fields offered by the system.

4 Exceptional Transfer Rules

DUET-E/J generates natural-sounding Japanese, the target language, according to the exceptional detailed transfer rules described in the lexical entries, as well as the general transfer rules. This offers one of the great advantages of using a bilingual dictionary for MT systems.

4.1 Paraphrased Translation

There are many cases where the ordinary passive form generation (-*reru* or -*rareru*) does not sound natural in Japanese. In such cases, a paraphrased word sense is defined in a lexical entry as passive-specific. For example, in the sentence,

> "The subsidiaries' creditors will receive virtually all the $120 million they are owed."

the verb pattern of the verb “owe” is “SUBJECT + VERB + NP + NP”, and the translation of the phrase “$120 million they are owed” should be paraphrased like:

“$120 million they have lent”,

because “*kari-rareru* (to be owed)”, the passive form of “*kari-ru* (to owe)”, sounds quite strange in Japanese. Thus the verb “owe” is given a new passive-specific word sense “*ka-shi-teiru* (to have lent)”.

To make the generation more natural, even in the active form, sometimes a paraphrased word sense is defined in a lexical entry as active-specific. For example, the verb “cause”, when used in the verb pattern “SUBJECT + VERB + NP + to-infinitive”, is given both the active-specific and the passive-specific word senses. In the following example,

“The steam is caused to transmit some of its heat to the liquid.”

the verb “cause” is used in the passive form, and the given passive-specific translation is paraphrased like:

“As a result, the steam transmits some of its heat to the liquid.”

(“*kekkateki-ni*, ‘the steam’ *ha* ‘some of its heat’ *wo* ‘transmit’ *suru*.”)

If it is used in the active form, the sentence would be:

“Something causes the steam to transmit some of its heat to the liquid.”

and the given active-specific translation is paraphrased like:

“By the influence of something, the steam transmits some of its heat to the liquid.”

(“‘something’ *niyotte*, ‘the steam’ *ha* ‘some of its heat’ *wo* ‘transmit’ *suru*.”)

4.2 Word Order

Word order is usually determined by the transfer rules and the deep case descriptions in each lexical entry.

As shown in the above item, the paraphrased translation includes the word sense of the predicate verb and the word order of its case slots. It is possible to specify the word order in a lexical entry, using the case slot numbers which are counted from left to right. In the above example of “cause”, in the passive form, the word sense and the word order is defined as:

“*kekkateki-ni* (1) (2)”

(1) = SUBJECT (NP in the active form, particle: *ha*)
(2) = to-infinitive (predicate: *suru*)

and in the active form:

“(1) (2) (3)”

(1) = SUBJECT (particle: *niyotte*)
(2) = NP (particle: *ha*)
(3) = to-infinitive (predicate: *suru*)

Another way of deciding the word order is to give the information in a lexical entry, which prohibits the free case from being generated in between the obligatory cases. In the ordinary transfer rules, an adverb is usually generated directly in front of a predicate verb, though there are some exceptional rules which determine the position of the adverb to be generated.

In the following example,

“They eventually become red.”

the exceptional transfer rule is defined in the lexical entry of the verb “become”, because it is not natural-sounding if the translation of adverb “eventually” is generated in between the translations of the verb “become” and its obligatory case “red”.