Abstract

IMT/EC is an English-Chinese machine translation system which integrates some outstanding features of the case grammar and semantic grammar into a uniform frame, uses various knowledge in the disambiguation, and tries to modify the object language by itself. In this paper, we first introduce IMT/EC's design motivation and overall architecture, then describe the design philosophy of its translation mechanisms and their processing algorithms.

1. The design motivation

The design of the IMT/EC system is motivated to develop new approaches to the English-Chinese machine translation, such as, to provide the system with powerful analysis mechanisms and MT knowledge base management system, as well as some exceptional processing and learning mechanisms, that is, to make the system be intelligent. In addition, it also tries to integrate as many advantages of conventional machine translation systems into a single system as possible, such as, to provide the system with powerful mechanisms for the processing of various ambiguities and contextual relations. The design of the MT's translation mechanisms are based on the following considerations:

(1) SC-analysis

In the development of machine translation system, in order to disambiguate the source language, we have to analyze the input deeply to get the internal meaning representation of the source language. However, the deeper we analyze the input, the more we lose the clues about how to express the translation, also, that it results in extremely poor or no translations of sentences for which complete analyses can not be derived [Lozoum 85]. To find a suitable analysis depth so as to get both clues and express the translation of the input and to disambiguate the input completely is almost impossible. In the IMT/EC, we try to design a simple grammar analysis mechanism - SC-grammar analysis mechanism to inherit both the outstanding features of case grammar analysis and semantic grammar analysis so as to produce a high quality translation.

(2) Multi-language translations oriented

In present technical conditions, it is impossible to design a general internal meaning representation for all natural languages. Thus, the knowledge based multi-language oriented machine translation system is difficult to be marketed in the near future. A feasible way for designing multi-language oriented machine translation systems might be to separate the processing mechanisms from the language specific rules [as King et al. 85], that is, to apply the same processing mechanism with different language specific rules for different natural language pair translations. In the IMT/EC, we develop a general rule representation form for the representation of various knowledge used in the translation. Knowledge for different language pair translations are stored in the different packages of the knowledge base IMT-KB. The knowledge base is organized in multi-package and multi-level way so as to store rules for the translation of different language pairs and different phases of the processing. Thus, the system can be easily extended for multi-language translation purposes.

(3) Diversity processing

As the disambiguation rules are rather word specific, it is difficult to manage them in the same way. To deal with this problem, we store these rules in their respective word entries and classify them as several categories in the IMT/EC. Each category corresponds to a general subroutine application mechanism, which apply the word specific rules and subroutines in the processing of translation. The subroutines are stored in a natural language specific subroutine package. Some word specific subroutines are directly stored in the respected word entry.

(4) Powerful exceptional processing

Since the natural language phenomena are so abundant that any existed machine translation system can not process all the phenomena, it is essential to provide an exceptional processing mechanism in the system to deal with exceptional phenomena. As IMT/EC incorporates some learning mechanisms, thus, it is more powerful in dealing with the exceptions than others.

(5) Automatic modification of the translation

Generally speaking, machine translation system can only produce rigid translations, it is a desire that MT systems be able to modify the output by itself so as to produce more fluent translations. IMT/EC tries to apply some common sense knowledge and linguistic knowledge of object language to disambiguate the input and modify the translations, thus, to improve the translation quality.

In the following paragraph, we focus on the translation procedure of the system and the algorithms related to it ignoring the knowledge base organization and management mechanisms.

2. The overall architecture of the system

The architecture of the IMT/EC system is as follow.

Fig. The architecture of the IMT/EC
As the rule base and dictionary in a machine translation system are so vast that it is impossible for human beings to find the conflicts and implications among the rules. To modify a rule in the knowledge base often results in many side effects on other rules. Thus, it is necessary to provide a self-reorganization and refinement mechanisms in the knowledge base.

In the IMT/EC, we design a special knowledge base management system IMT-KB to manage all the knowledge used in various processing phases of the translation. In addition, IMT/EC also provides a knowledge base augmentation and knowledge acquisition environment for the system to augment system performance by itself and for the users to improve the knowledge base.

The cell relations connected by dotted lines in the figure above are executed only when the user sets the learning mechanisms in working status. These mechanisms can acquire new knowledge in the dynamic interactive, static interactive, or disconnected ways. They are primarily used to resolve the exceptional phenomena in the translation.

Dynamic Interactive Learning (DIL): Whenever the system encounters a sentence out of its processing range, it produces various possible translations for each segment of the sentence and interacts with human beings when necessary to select an appropriate translation of the segment and combine them to get a correct translation of the sentence. At the same time, it also creates some new rules to reflect the selections. That is, it learns some new knowledge.

Static Interactive Learning (SIL): Whenever the system encounters a sentence out of its processing range, it records down the sentence and its appearance context in a file. After the text has been translated, it begins to analyze the sentence in detail to get various possible translations for each segment of the sentence and interacts with human beings when necessary to get appropriate translations of the segments and combines them to get a correct translation of the sentence. At the same time, it also creates some new rules to reflect the selections, thus, to learn new knowledge.

Disconnected Learning (DL): Whenever the system encounters a sentence out of its processing range, it analyzes the sentence in detail to get all the possible translations, and then evaluates these translations according to the preference rules stored in the IMT-KB to select an appropriate translation and modify the related rules used in the analysis to reflect the selections. It skips over sentences which the translation can not be determined by the preference rules instead of interacting with human beings.

The translation procedure

IMT/EC's translation procedure is divided into several phases, i.e., morphology analysis and dictionary retrieval, SO-grammar analysis, disambiguation and transfer, modification of the translation etc.

The communications between translation mechanisms and the knowledge base are performed by the knowledge base management system IMT-KB, these operations include getting a set of related rules and returning some information for the modification as well as augmentation of the MT knowledge base.

3. The translation procedure

IMT/EC's translation procedure is divided into several phases, i.e., morphology analysis and dictionary retrieval, SO-grammar analysis, disambiguation and transfer, modification of the translation etc.

The communications between translation mechanisms and the knowledge base are performed by the knowledge base management system IMT-KB, these operations include getting a set of related rules and returning some information for the modification as well as augmentation of the MT knowledge base.

3.1. Morphology analysis and dictionary retrieval

In the IMT/EC, words in most common uses can be retrieved by either their base forms or their surface forms, while most of the other words can only be retrieved by their base forms. The tasks of the morphology analysis is to process the prefix, suffix, and compound words. Since these processings are completely natural language specific, in order for the processing mechanisms to be language independent, we develop a language independent morphology analysis mechanism to apply the language specific morphology rules in the morphology analysis.

The morphology analysis rule form is

\[ \text{<surface pattern> } \rightarrow \text{<conditions> } \rightarrow \text{<result>} \]

Here,

- \text{<surface pattern>} is the surface form of the word to be analyzed.
- \text{<conditions>} is the application conditions of the rule.
- \text{<result>} is the definition of the word base form analyzed.

For example,

\[ \{ \begin{align*} & (\text{*} \rightarrow \text{verb}) : \text{def}(*), \text{SV} \nonumber \\text{SP} <- \text{search} & (\text{\texttt{X}}) \nonumber \\text{X} <- \#\text{word} \nonumber \text{for} \text{\texttt{rule}} \text{\texttt{X}} \text{\texttt{SP}} \text{\texttt{do}} \nonumber \text{MATCH} \\text{\texttt{PAT}} <- > \text{\texttt{rule}} \nonumber \text{\texttt{COND}} <- > > \text{\texttt{Brule}} \nonumber \text{\texttt{RES}} <- < < \text{\texttt{Brule}} \nonumber \text{Loop} \text{\texttt{pair}} <- \text{\texttt{f(PAT, X)}} \nonumber \text{if} \text{\texttt{(null(pair)) goto TEST}} \nonumber \text{if} \text{\texttt{(not(check(pair))}} \text{\texttt{break}} \nonumber \text{\texttt{PAT}} <- \text{\texttt{PAT} = pair}} \end{align*} \]

Here, \* and \* are variables indicating that it can be bounded to any sub-character string of the word to be analyzed. \text{def}(X) is the definition of \text{X} in the IMT-KB, SV, PN, COM are surface features of the word.

Rule (1) indicates that when the last character of the surface form of a word is 's' and the remained character string * in the word is a noun, then its surface feature is the singular verb form (SV) of the verb *.

Rule (2) indicates that when the last character of the surface form of a word is 's' and the remained character string * in the word is a noun, then its surface feature is the plural noun form (PN) of the noun *.

Rule (3) indicates that when the character string of a word comprises a character 's', the left part \#1 and the right part \#2 of 's' are both words, then it is a compound word of \#1 and \#2. Thus, it applies morphology rules to analyze the word \#1 and \#2, and returns the value of \text{f(morphology \*1),f(morphology \*2)} as result.

Suppose that,

- \$X indicates that \text{X} is a variable.
- \#X returns the character list of \text{X}.
- \%X returns the last character of \text{X}.
- \%X returns the first part of rule \text{X} or the first element of a list.
- \text{def}(X) is the definition of the word \text{X} which \text{X} needs further analysis and performs recursive analysis when necessary.

Here, \text{null}() is a function which returns true if its argument is empty.

The algorithm for morphology analysis and dictionary retrieval is as follow.

\text{INITIALIZE}

\begin{align*} & \text{\$X} <- \#\text{word} \nonumber \text{\$P} <- \text{search} & (\text{\texttt{X}}) \nonumber \text{\$P} <- \text{\$P} \text{\texttt{u search}} (\text{\texttt{X}}) \nonumber \text{\$result} <- (\; ) \nonumber \text{for} \text{\texttt{rule}} \text{\texttt{X}} \text{\texttt{P}} \text{\texttt{do}} \nonumber \text{\texttt{MATCH}} \text{\texttt{\$PAT} <- > \text{\texttt{rule}} \nonumber \text{\texttt{\$COND} <- > > \text{\texttt{Brule}} \nonumber \text{\texttt{\$RES} <- < < \text{\texttt{Brule}} \nonumber \text{\texttt{Loop}} \text{\texttt{\$pair} <- f(\text{\texttt{\$PAT}, \text{\texttt{X}}}) \nonumber \text{\texttt{if}} \text{\texttt{(null(pair)) goto TEST}} \nonumber \text{\texttt{if}} \text{\texttt{(not(check(pair))}} \text{\texttt{break}} \nonumber \text{\texttt{\$PAT} <- \text{\texttt{\$PAT} = pair}}} \end{align*}
3.2. SD-Grammar Analysis

The SD-grammar analysis mechanism of IMT/EC applies the SD-rules stored in the IMT-KB to disambiguate the structural meanings of the input sentence and produces the structural description for them. The grammar has some outstanding features of the formal grammar and semantic grammar. The rule form is as follows:

\[
\langle \text{S-STRUCTURE} \rangle \rightarrow \langle \text{S-ENVIRONMENT} \rangle \rightarrow \langle \text{R-STRUCTURE} \rangle \rightarrow \langle \text{R-ENVIRONMENT} \rangle \rightarrow \langle \text{TRANSFER} \rangle.
\]

Here, \(\langle \text{S-STRUCTURE} \rangle\) and \(\langle \text{S-ENVIRONMENT} \rangle\) are rule conditions which define the current structural form and contextual features of the input, \(\langle \text{R-STRUCTURE} \rangle\) and \(\langle \text{R-ENVIRONMENT} \rangle\) are result structural form and contextual features of the input, \(\langle \text{TRANSFER} \rangle\) are the transformations related to the rule.

The structural forms, \(\langle \text{S-STRUCTURE} \rangle\) and \(\langle \text{R-STRUCTURE} \rangle\) are represented as strings of syntagmas and words. The contextual environments, \(\langle \text{S-ENVIRONMENT} \rangle\) and \(\langle \text{R-ENVIRONMENT} \rangle\) are represented as vectors, of which each element corresponds to an inter-sentential relation or a special case, their values are used to resolve the ellipsis, anaphora, tense and aspects etc.

It is the principal contextual processing mechanism in the IMT/EC.

Since the contextual vector is used only as a supplement to the pure semantic grammar analysis, especially in the processing of contextual relations, it is not necessary to analyze the input to the extent that one can get all the semantic relations of the input. Thus, the vector processing formalism is completely acceptable.

Two example rules are as follow.

\[
\text{NP} \rightarrow \text{A} \mid \text{B}, \text{change(B1, X)}, \text{NP \ then NP}. \\
\text{change(H2, Z)}, \text{NP \ null}. \\
\text{SC-grammar analysis mechanisms receive the results of morphology analysis or previous SC-reduction, send the messages to the IMT-KB to get related rules, and apply these rules to reduce the input until a non-terminal symbol is reduced}, \text{thus, to produce the structural description of the input.}
\]

The SC-grammar analysis algorithm of the system is:

1. In the entries of the IMT-KB dictionary, we stored not only the word meanings and their disambiguation conditions, but also SC-phrase and semantic rules specific to the entry word. When analyzing a sentence, the system first retrieves the SC-phrase rules specific to the words appeared in the sentence, and applies these rules to find a list of possible phrases of the sentence from the context of the words in the sentence.

The phrase list returned is as follow:

\[
X_1(l_1, j_1), X_2(l_2, j_2), \ldots, X_m(l_m, j_m).
\]

Here, \(X_1, X_2, \ldots, X_m\) are phrase syntagma identifiers, \(l_1, j_1, \ldots, l_m\) and \(j_1, j_2, \ldots, j_m\) are ending positions of the phrases in the input sentence.

2. Find a list of expectation paths from the phrase list as follow:

\[
X_1^p(l_1^p, j_1^p), X_2^p(l_2^p, j_2^p), \ldots, X_m^p(l_m^p, j_m^p).
\]

3. For each complex phrase \(X_i^p\) in the phrase list, we find the SC-rules which head patterns contain the sub-string from \(X_i^p\), and the current expectation path to the IMT-KB to retrieve the SC-rules which head patterns contain a sub-string of \(X_i^p\) and the current analysis position to the current form of the sentence. We try one new path at one backtracking.

4. Send the analyzed component \(M = \{s_1(\ldots), s_2(\ldots), \ldots, s_n(\ldots)\}\) and the current expectation path to the IMT-KB to retrieve the SC-rules which head patterns contain a sub-string of \(M\) and the current analysis position to the current form of the sentence.

5. First, add some newly formed phrases into the phrase list in order for the backtracking of the analysis, then call the case analysis mechanism to check the current analysis results and the current form of the sentence to fill in the case frame A, B in the rule and the context vector. The case analysis algorithm is described in the following paragraph.

6. Check A and the context vector to see whether their values are unifiable. If they are unifiable, then go to (7), else get the next rule from (3) and returns to (4).

7. Store the backtracking information into the temporary stack, substitute the reducing part of the current sentence form with the reduced form, change the current analysis position to the last word of the newly reduced syntax, and change the related element values of the context vector according to the element values of B.

8. Call the semantic processing mechanism to check the result of the analysis to see whether it violates the English collocation rules. If the result violates the collocation rules, the system recovers to the status before the last reduction and gets the next rule from (3) to re-analyze the input. Otherwise, there will be two cases:

   a. If the user only needs the most adequate translation, the system proceeds to analyze the next sentence.

   b. If the user needs all possible translations, the system records down the current result and recovers to the status before the last reduction and gets

119
the next rule from (3) to re-analyze the input in order to get other analysis results.

As we have mentioned before, the case analysis in the SC-analysis is only a complement to the semantic analysis. It is mainly used to deal with the context relation and aspect, tense, modal etc. Thus, the system only needs to analyze those cases which can be used in those purposes. It is much simpler than the case analysis in the case grammar analysis.

The case analysis in the SC-analysis is performed by the following algorithm:

1. Get the value of the case identifier, such as time adverbial, place adverbial, anaphora etc.

When a specific semantic identifier is concerned, the semantic processing mechanisms first finds the word which can match the semantic identifier from the input, and return the phrase which comprises the word and the value of the identifier.

Only simple anaphora phenomena are considered in the SC-analysis. They are processed in two different ways. First, finds the word which can match the semantic identifier from the input, and return the phrase which comprises the word and the value of the identifier.

When a specific semantic identifier is concerned, the semantic processing mechanisms first finds the word which can match the semantic identifier from the input, and return the phrase which comprises the word and the value of the identifier.

(1) Get the value of a specific semantic identifier, such as time adverbial, place adverbial, anaphora etc.

When a specific semantic identifier is concerned, the semantic processing mechanisms first finds the word which can match the semantic identifier from the sentence, such as word with time, place properties, and then get the phrase which comprises the word and the value of the identifier.

(2) The case analysis in the SC-analysis is performed by the following algorithm:

1. Get the value of the case identifier, such as time adverbial, place adverbial, anaphora etc.

When a specific semantic identifier is concerned, the semantic processing mechanisms first finds the word which can match the semantic identifier from the sentence, such as word with time, place properties, and then get the phrase which comprises the word and the value of the identifier.

Only simple anaphora phenomena are considered in the SC-analysis. They are processed in two different ways. First, finds the word which can match the semantic identifier from the input, and return the phrase which comprises the word and the value of the identifier.

(1) Get the value of a specific semantic identifier, such as time adverbial, place adverbial, anaphora etc.

When a specific semantic identifier is concerned, the semantic processing mechanisms first finds the word which can match the semantic identifier from the sentence, such as word with time, place properties, and then get the phrase which comprises the word and the value of the identifier.

Only simple anaphora phenomena are considered in the SC-analysis. They are processed in two different ways. First, finds the word which can match the semantic identifier from the input, and return the phrase which comprises the word and the value of the identifier.
Aknowledgment: We would like to thank Hong Xiong, Zhai Hamo, Ye Yufan, Tong Jiaoxion, Zong Lingyi, Zhang Zifa, Chen Zizong, Chen Zizong and Fu Wei for their cooperation in the implementation of IMT/EC.

5. Reference

4. Summary
In conclusion, we have introduced the translation processing procedure of the English-Chinese machine translation system IMT/EC, and describe its principal processing algorithms.

The main tasks consists:
(a) Change the order of the phrases and words of the translation.
(b) Substitute some words which collocation is not commonly used in the Chinese utterance for the synonymous words.
(c) Insert some conjunctive words when necessary.
(d) Eliminate some redundant words.

The algorithm for these processing is,

(1) According to the Chinese collocation rules defined in the IMT-KB, changes the words and phrases order of the translation which are not in accord with the collocation conventions in Chinese, such as

   Budan ..., Erchia ....

(2) According to the co-occurrence rules of the Chinese words defined in the IMT-KB, check the uses of the Chinese words in the translation. If they are not in accord with the co-occurrence rules, then replaces these words with the Chinese synonymous words until they are accord to the rules. If there is no suitable synonyms, then tries to extend the meaning of some words. The meaning extending rules are defined in the word entries. Its form is as follow,

   \( \text{word} \) :- \( \text{condition} \) 1  \( \text{extension} \) 1 \( \text{condition} \) 2  \( \text{extension} \) 2  \( \text{condition} \) 3  \( \text{extension} \) 3

Here, \( \text{word} \) indicates the word appeared in the sentence, \( \text{condition} \) defines the extending conditions, \( \text{extension} \) is the utterance extended.

If the word can not be replaced or extended, then just returns the source translation.
(3) Check the translation to find the redundant words and eliminates them. The form of deletion rule is

   X Y X Z \( \rightarrow \) p (X), p (Y) I X Y Z

such as, 'NP de NP de \( \rightarrow \) NP NP de'.

Since the modification has no absolute standard and requires a large amount of world knowledge, it is rather difficult to solve this problem in one day. In the IMT/EC, we only deal with the most simple cases. More complex situations can be solved with the application and improvement of the system. Thus, the system is designed to be easily extended with the application.

If the user needs high quality translation, he may call the post editing subroutine to modify the translation by human beings or with the aid of human beings. At the same time, we can also set the learning mechanisms in working status to trace the modification procedure of human beings and produce some useful rules for the system.

3.4. The modification of the translation
The objective of the automatic modification of the translation is to improve the readability of the translation, but this sacrifices part of the accuracy. It is more suitable for the non-scientific literature translation.

The main tasks comprises:
(a) Change the order of the phrases and words of the translation.
(b) Substitute some words which collocation is not commonly used in the Chinese utterance for the synonymous words.
(c) Insert some conjunctive words when necessary.
(d) Eliminate some redundant words.

The algorithm for these processing is,

(1) According to the Chinese collocation rules defined in the IMT-KB, changes the words and phrases order of the translation which are not in accord with the collocation conventions in Chinese, such as

   Budan ..., Erchia ....

(2) According to the co-occurrence rules of the Chinese words defined in the IMT-KB, check the uses of the Chinese words in the translation. If they are not in accord with the co-occurrence rules, then replaces these words with the Chinese synonymous words until they are accord to the rules. If there is no suitable synonyms, then tries to extend the meaning of some words. The meaning extending rules are defined in the word entries. Its form is as follow,

   \( \text{word} \) :- \( \text{condition} \) 1  \( \text{extension} \) 1 \( \text{condition} \) 2  \( \text{extension} \) 2  \( \text{condition} \) 3  \( \text{extension} \) 3

Here, \( \text{word} \) indicates the word appeared in the sentence, \( \text{condition} \) defines the extending conditions, \( \text{extension} \) is the utterance extended.

If the word can not be replaced or extended, then just returns the source translation.
(3) Check the translation to find the redundant words and eliminates them. The form of deletion rule is

   X Y X Z \( \rightarrow \) p (X), p (Y) I X Y Z

such as, 'NP de NP de \( \rightarrow \) NP NP de'.

Since the modification has no absolute standard and requires a large amount of world knowledge, it is rather difficult to solve this problem in one day. In the IMT/EC, we only deal with the most simple cases. More complex situations can be solved with the application and improvement of the system. Thus, the system is designed to be easily extended with the application.

If the user needs high quality translation, he may call the post editing subroutine to modify the translation by human beings or with the aid of human beings. At the same time, we can also set the learning mechanisms in working status to trace the modification procedure of human beings and produce some useful rules for the system.

4. Summary
In conclusion, we have introduced the translation processing procedure of the English-Chinese machine translation system IMT/EC, and describe its principal processing algorithms.

Aknowledgment: We would like to thank Hong Xiong, Zhai Hamo, Ye Yufan, Tong Jiaoxion, Zong Lingyi, Zhang Zifa, Chen Zizong, Chen Zizong and Fu Wei for their cooperation in the implementation of IMT/EC.

5. Reference

4. Summary
In conclusion, we have introduced the translation processing procedure of the English-Chinese machine translation system IMT/EC, and describe its principal processing algorithms.

Aknowledgment: We would like to thank Hong Xiong, Zhai Hamo, Ye Yufan, Tong Jiaoxion, Zong Lingyi, Zhang Zifa, Chen Zizong, Chen Zizong and Fu Wei for their cooperation in the implementation of IMT/EC.

5. Reference

4. Summary
In conclusion, we have introduced the translation processing procedure of the English-Chinese machine translation system IMT/EC, and describe its principal processing algorithms.

Aknowledgment: We would like to thank Hong Xiong, Zhai Hamo, Ye Yufan, Tong Jiaoxion, Zong Lingyi, Zhang Zifa, Chen Zizong, Chen Zizong and Fu Wei for their cooperation in the implementation of IMT/EC.

5. Reference