AUTOTEC: an English to Chinese Machine Translation System

Kefeng Chen
Center for Information Research
University of Florida
Gainesville, FL 32611

Abstract
This paper describes an English to Chinese machine translation system--AUTOTEC. The AUTOTEC is based on a multilevel reconstruction approach, which can greatly improve translation quality of MT system.

1 Introduction
Machine translation from English to Chinese has been under research for more than three decades. Many systems have been developed. However their translation quality compared with human translation is still not high enough to have any marked effect on translation productivity. The large linguistic difference and the shortcomings of the translation approaches used in the systems are two main reasons.

In this paper, we present an English to Chinese MT system--AUTOTEC, which uses a multilevel reconstruction approach [1]. The translation in the system is performed in a sentential, phrasal, and lexical order. Through the hierarchical processing, the large linguistic differences between English and Chinese can be well bridged.

2 The Multilevel Reconstruction Approach
The multilevel reconstruction approach focuses on bridging linguistic gaps and improving translation quality. It follows language sentential, phrasal, lexical hierarchical structure. The reconstruction takes place after analysis stage. It is equivalent to the transfer and generation stage of the traditional approach.

A sentence structure is the foundation of a sentence. English and Chinese belong to different language groups. From word order point of view, English is a predominantly SVO (Subject Verb Object) type of language. Chinese is an SOV and SOV (Subject Object Verb) mixed type of language [2]. Their sentence structures range from similar to quite different. The task of sentential reconstruction is to establish a most appropriate Chinese sentence structure for the translation.

The phrasal reconstruction takes sentence constituents as its working elements. It has two major tasks: to map English sentence constituents to the established Chinese sentence structure and to resolve the phrasal level differences between English and Chinese.

The lexical reconstruction substitutes all English words with Chinese words and perform necessary processing to meet Chinese word specific requirements. It finalizes the translation.

3 The AUTOTEC System
The AUTOTEC system consists of two major components which perform English sentence analysis and Chinese sentence synthesis.

The analysis component analyzes an English sentence into two related structures--constituent structure (CS) and feature representation structure (FRS) [1]. The CS is a standard context-free surface parse of an English sentence. It provides the sentence surface structure information. The FRS gives the sentence syntactic, semantic and functional information in an attribute-value pair form. It provides an effective interface between the CS and reconstruction knowledge bases.

The synthesis component includes three modules: sentential reconstruction module, phrasal reconstruction module and lexical reconstruction module. Each module has its own knowledge base to guide its reconstruction.

In the sentential reconstruction module, the determination of a Chinese sentence structure is based on three factors: sentence patterns, linguistic features, and grammar rules.

In translation, sentence patterns play an important role. Many English sentence patterns traditionally have fixed translations. Their translation must follow the convention. For example, English "there-be" sentence, as a pattern, must be translated into Chinese "(place) + YOU + (entity)." Any other structure would not be appropriate. The first phase of the sentential reconstruction checks whether the English sentence is a special pattern and processes it accordingly.

Linguistic features also affect translation. The features include verb feature, features of post verb elements, sentence voice etc. For example, the passive voice in Chinese is used almost exclusively for expressing adverse situation. To translate an English "patient be v-en by agent" passive sentence into Chinese, the verb feature has to be considered. If the verb includes a unfortunate feature, the Chinese "patient BEI agent v" structure can be used, otherwise, an active sentence should be used instead.

The consideration of linguistic features in selecting sentence structure will make it easier to determine when to translate an English sentence into Chinese BA sentence, which has been a puzzling problem. In the second phase of sentential reconstruction, the linguistic features are checked to determine an appropriate sentence structure.

For the English sentences which are not considered
special pattern or they have no structural affecting features, the SVO structure (NP V NP) will be transformed into Chinese SVO (NP V NP) structure.

The phrasal reconstruction module first maps the English sentence constituents to the Chinese sentence structure created in the sentential reconstruction module. The mapping is based on the sentential structure requirements. For example, the first slot of BA structure 

\[<\text{agent}> \text{ BA} <\text{patient}> <\text{verb}> <\text{complement}>\] 

requires a constituent to be a noun phrase and its case role be an agent. If a noun element meets this requirement, it will be mapped to the first slot. The result of the mapping is a meta-sentence which has a Chinese sentence structure filled with English constituents.

When adverbials exist, they need special treatment. If several adverbials occur in a sentence, they need to be reordered. If their function roles are different, the reordering can be based on their function roles. For example, when manner, time, and place adverbials occur in a sentence, their English order is usually manner-place-time, such as "She ate quietly in her room last night." The Chinese order is time-place-manner, "She last night in her room quietly ate." The reordering rule is: manner-place-time \rightarrow time-place-manner.

If several adverbials have the same function role, English tends to organize them to be specific to general, from small to large and from individual to group. Chinese order is in the opposite, from general to specific, from large to small and from group to individual. For example, the English expression "7:30 pm, Wednesday, January 30, 1990" in Chinese order is "1990, January 30, Wednesday, pm, 7:30." "Mary lived in a hut near a river in a remote area" in Chinese order is "Mary in a remote area near a river in a hut."

To obtain correct Chinese adverbial word order, two steps are taken in AUTOTEC. One is to incorporate semantic analysis in the parsing stage to obtain the function role of constituents. The other is to use an inference engine based on the larger-concept-first principle. The larger-concept-first principle states that a word (or a phrase) which expresses a larger concept than the other words (or phrases) should precede the other words (or phrases). That is, if English words (or phrases) A, B, and C are in the same category and have the same function, their Chinese equivalent D E F should satisfy the condition D > E > F in conceptual level.

The second task of phrasal reconstruction is to obtain Chinese phrase order. It is done through grammatical transformation. For example, the phrase "a small computer with large memory" is transformed into Chinese order "a computer with large memory small computer" through transformational rule:

\[
\text{DET} (\text{[ADJ]*} \text{ N} \text{ [PP]} \rightarrow \text{DET} \text{ [PP]} [\text{[ADJ]*} \text{ N}].
\]

The lexical reconstruction module substitutes all English words with equivalent Chinese words. Chinese words are chosen from a bilingual dictionary. The main principles for choosing a word are: (1) domain restriction; (2) semantic primitive information; (3) context information and pragmatic information.

The domain restriction limits word ambiguity. In most cases, there are several possible translations of a given word. However, if the translation domain is known, its correct meaning can usually be quickly determined. In general, the narrower the specialized domain of the original text, the more likely it is that individual word will have unique translation. In AUTOTEC, every meaning of a word in the dictionary is marked with suitable field.

After domain restriction, the semantic marker and context information can help further narrowing down the selection of correct word meaning.

The word specific processing, which includes realizing tense and aspect, adding measure word etc., will finalize the translation.

4 Processing Example

A complete sentence translation process by AUTOTEC using multilevel reconstruction approach is demonstrated in this section. The sample sentence is:

Bill moved the computer with a color monitor to his office.

In the sentential reconstruction level, the system determines the Chinese sentence structure is 

\[<\text{agent}> \text{ BA} <\text{patient}> <\text{verb}> <\text{complement}>\] 

based on the sentence CS and FRS information.

In the phrasal reconstruction level, the CS constituents are first mapped to the structure slots. The sentence becomes:

\[<\text{Bill}> \text{ BA} <\text{the computer with a color monitor}> <\text{moved}> <\to \text{his office}>.\]

After filling out the sentence structure, the phrasal reconstruction will generate correct Chinese phrase expressions. The sentence becomes:

Bill BA the with a color monitor computer moved to his office.

In the lexical level, the substitution of English with Chinese words yields:

Bir ba na dai yi caise xianshiqi jisuanji ban dao tade bangongshi.

After word specific processing, the final Chinese translation becomes:

Bir ba na tai dai yi ge caise xianshiqi de jisuanji ban dao tade bangongshi qu le.

5 Conclusion

The AUTOTEC uses the multilevel reconstruction approach and achieves high translation quality. The multilevel reconstruction approach is based on the language hierarchical concept. From computational perspective, the use of multilevel reconstruction approach breaks the translation process into subprocesses. Thus it will simplify the knowledge-base organization, translation-rule writing and programming.

6 References