A PRAGMATIC APPROACH TO MACHINE TRANSLATION FROM CHINESE TO ENGLISH

- Ching-Yi Dougherty -
Merill College
University of California
California 95 060 USA

Chinese machine translation can be achieved by organizing all the necessary linguistic data in such a way that the computer can compare and retrieve them in the most economical way. We are constantly reminded that the storage space in the computer is limited, and the processing time is expensive. We must aim at the efficiency of the system without sacrificing accuracy.

Five types of linguistic data have to be stored in the computer before a translation can be rendered: (1) a Chinese to English dictionary, (2) Chinese syntactic rules, (3) syntactic conversion rules from Chinese structure to English structure, (4) English morphological rules, and (5) the text to be translated.

(1) The dictionary must have the capability to distinguish automatically the different meanings which a Chinese lexeme represents. Human readers can do this and so can the machine if enough linguistic information is stored in the computer. To a certain extent the meanings of a given lexeme can be differentiated by its different syntactic functions. For example the word i means 'one' when it is used as a numeral. A numeral is defined as the class of words which is followed by a classifier, a measure, a collective and a

partitive noun and other numerals. I means 'once' when it is used as an adverb which is always followed by a verb. Anything else which is not covered by these two rules will have to be listed in the dictionary either as an idiom or as a lexeme with its immediate constituent. Sometimes a more refined syntactic codes are needed to distinguish the different meanings of a given verb, as in the case of tzuoh 作. When it is followed by an inanimate noun, it is translated to 'make', by an abstract noun, it is translated to 'do',and by a human noun, it is translated to 'be'. Any other constitutes which are exceptions to the rules will have to be listed in the dictionary, such as tzuoh ren 'behave', tzuoh fann 'cook', and tzuoh show 'celebrate a birthday'.

For the convenience of programming, each dictionary entry, which may be a lexeme, an idiom or a constitute, has only one syntactic code and its respective meaning. After a sentence has been analyzed by the automatic parser, the syntactic function of each lexeme is determined. Together with the method of longest match in dictionary lookup, the correct meaning of a given lexeme in a given context can be chosen by the machine.

(2) The Chinese syntactic rules are formulated in such a way that they may reduce the ambiguities in parsing to a minimum. It is assumed that in a good scientific writing, the sentences are not ambiguous and there is only one correct way of parsing which can carry the process to the end of the sentence. Anybody who has had any experience with the automatic parser will know this goal is

hard to achieve. Several ways have been found to reduce the ambiguities, and probably there are others.

One of the methods is to add more refinements to the syntactic codes. Semantic elements were introduced in the formulation of the syntactic codes, so that the constitutes must be meaningful as well grammatical. For example: by adding the human element to the codes, only the human nouns can be the agents of the human verbs. If an inanimate noun occurs before a human verb, it is likely to be the goal rather than the agent of the verb. For further refinement, the element of plurality is also added to the codes, so that only the plural nouns can be the agents of plural verbs. There is a class of adverbs, such as i torng 'together', huh shiang 'mutually' and bii tsyy 'to each other' usually pluralize any verb that follows them.

The second method of reducing ambiguities is to introduce higher levels to the noun and verb phrases. There are five levels of noun phrases for example. The terminal code is called level 1. When the noun is modified by an adjective, a noun or an adjective phrase, the noun phrase is called level 2. Level 1 or 2 modified by a number and a classifier is called level 3. Modified by a determiner, it is called level 4. When the latter is modified by a pronoun, a relative clause or an apposition, it is called level 5. The noun phrase of level 5 is a closed noun phrase; nothing else can be added to it. The constitute IND (indicative expression) is formed by a closed noun phrase and a closed verb phrase. Even if the noun phrase consists of only the terminal code, this rule also applies linearly.

Please see Diagram I.

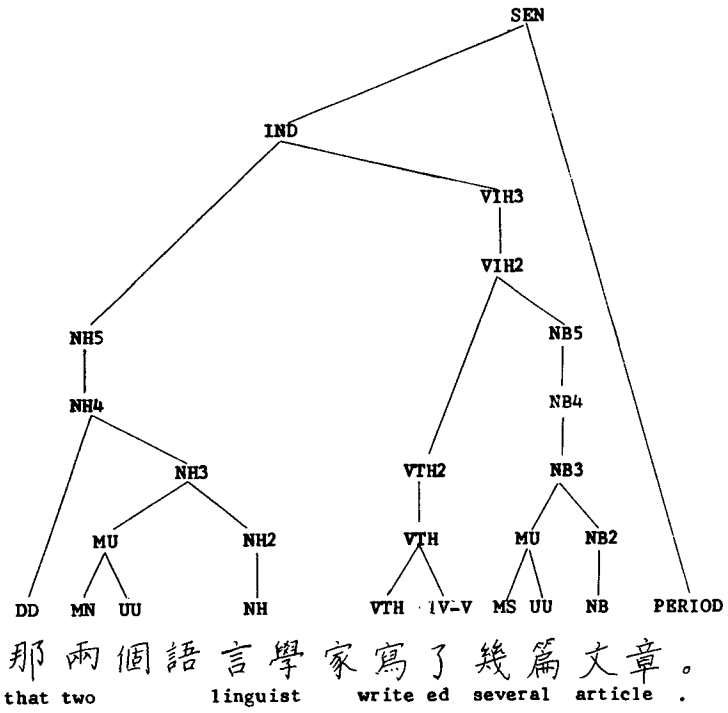


Diagram I

By applying the longest match to parsing, many possible ambiguities can be avoided, as indicated by the dotted lines in Diagram II.

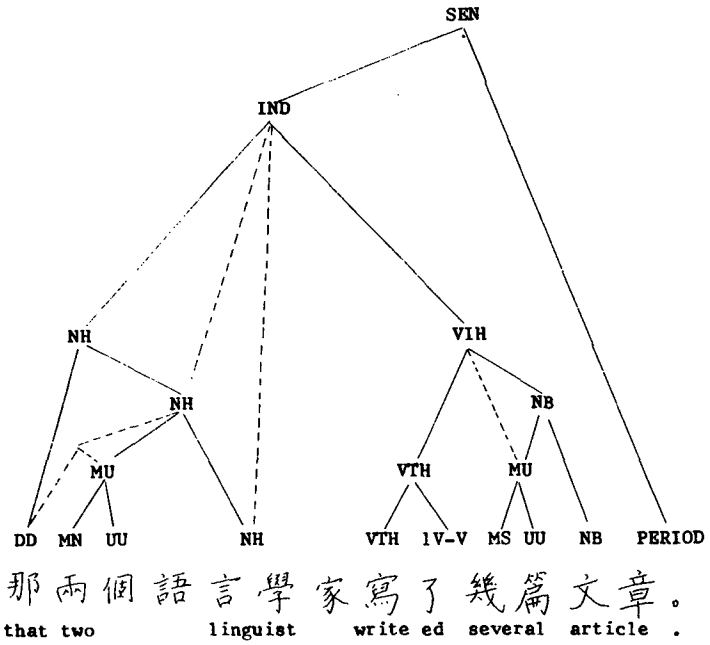


Diagram II

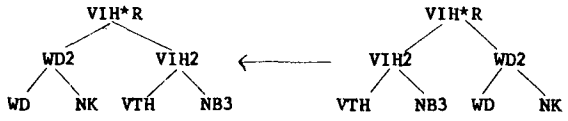
The third way of reducing the ambiguities is to adopt the principle of the longest match in the dictionary lookup. Just take the four characters yeu yan shyue jia 'linguist' as an example; all four of them can be used freely or individually. The first three characters are both nouns and verbs. As nouns the first two mean both 'language' and 'word', and the third means 'learning'. As verbs, the first two mean 'speak' and third, 'learn'. The fourth one means '-ist', 'family' and 'home'. The compound of the first two means 'language' and that of the last two means '-ist'. The compound of the first three characters means 'linguistics'. Sixty unwanted constitutes can result from these four characters alone. With the longest match with the dictionary entry, the four characters act as a single unit, and thus sixty ambiguities can be eliminated.

(3) The syntactic conversion rules from Chinese to English are formulated on the basis of comparative study of the structures of both Chinese and English sentences which represent the same meaning. It is assumed that if the structure of the English sentence or phrase is the same as that of the Chinese sentence or phrase, no syntactical conversion is necessary. Simple lexical substitution and the application of the English morphological rules will render the correct translation. Diagram I shows that the structure of the English is the same as that of the Chinese. By placing the English equivalents in the location of the grammar codes (The grammar codes are used in parsing.), the translation is already rendered in the citation forms. The last thing that needs to be done is to apply

the English morphological rules to the citation forms, so that 'that' is changed to 'those' and 'linguist' to 'linguists' and 'write' '-ed' to 'wrote'.

The structures of many English phrases are not the same as those of the Chinese phrases. In cases like these, some syntactic conversions, in the forms of permutation, addition or deletion are needed. The following examples will illustrate the logic on the basis of which the syntactic conversion rules are formulated.

The verb phrase which consists of a prepositional phrase and a main verb calls for a simple permutation in the conversion. The code for such a verb phrase is VI*R (or VIH*R if the human element is added) where * indicates that a conversion is needed and R indicates that the conversion is in the form of permutation of the two constituents.

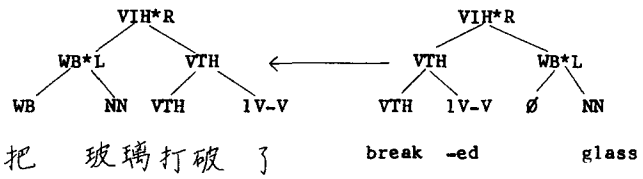


為 母親 作 一件事 do one thing for mother

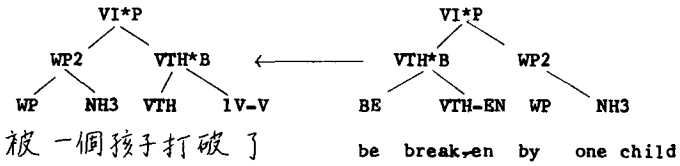
The arrow (used in the sense of programming language) indicates where the conversion occurs. The two items on the left side of the arrow are replaced by the two items on the right side. The constituents of these two constituteswill follow accordingly. This conversion can be easily manipulated by the machine.

Even though the Chinese pretransitive preposition, bae or jiang, is the same as other prepositions structurally, it is deleted in

the English structure. Therefore an additional conversion rule WB*L is needed. The L after the * indicates that the left constituent is to be deleted.



The conversion rule for the passive construction on the other hand requires addition. For example:



All the syntactic conversion rules are either unary or binary (some longer ones are forced to be binary) in order to save table space in the computer. The table of constitutes resulted from the automatic parser consists of three columns: the array of the constitutes, the array of the left constituents and the array of the right constituents. A search through the array of the constitutes, the computer will know where and what conversion rules are to be applied. For example:

	Before Conversion		After Conversion	
Column 1	Column 2	Column 3	Column 2	Column 3
VIH*R	WD2	VIH2	VIH2	WD2
VIH*R	WB*L	VTH	VTH	WB*L
WB*L	WB	NN	Ø	NN
VI*P	WP2	VTH*B	VTH*B	WP2
VTH*B	VTH	1V-V	BE	VTH-EN

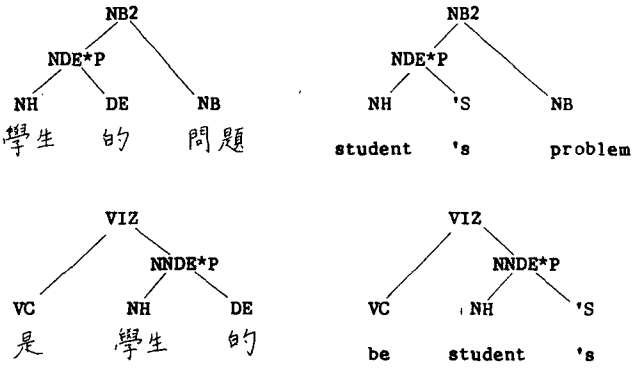
Another advantage of such a system is that further differentiation of meanings is made possible. It has been mentioned earlier that the meanings of a given lexeme can be differentiated by its context. When the context is as long as a clause or a phrase, it is best handled by syntactic conversion rules. For example: the character 的 de is translate to ''s', 'of', 'who (which, when or where)','the one who (which, when or where)' and 'Ø' under different conditions, a problem cannot be solved by dictionary lookup, but can be solved by the different constitutes it forms with other constituents.

The constitute NDE, whose left constituent is a noun or verb phrase of any level or a whole clause and whose right constituent is de, can be an adjective or a noun. It is an adjective when it is followed by a noun phrase and it is a noun when it is preceded or followed by a verb phrase. First of all its function should be determined by the automatic parser; the noun (NNDE) is distinguished from the adjective (NDE). The internal structures of both are the same in Chinese, but they represent different meanings which are represented by different English lexemes. Emphasis should be made here that this is not an ad hoc attempt to write a Chinese grammar on the basis of English translation, but an attempt to incorporate semantic

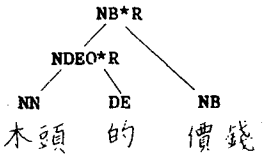
information into the system as demanded by the machine. Very frequently the different English translations point out the fact that a given Chinese lexeme represents different meanings in different contexts. A closer analysis of the contexts will reveal that the conditions which produce the same translation are usually consistently similar.

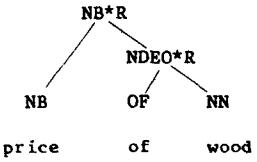
The two sub-classes of NDE and NNDE are not sufficient to differentiate the various meanings de represents. Further refinements have to be made within each subclass of NDE or NNDE. The following pairs (NDE and NNDE) of conversion rules will illustrate this point.

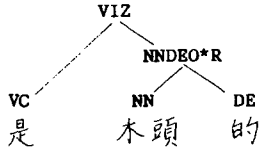
(a) Possession The meaning of possession is represented by de in Chinese and 's in English when they are preceded by animate nouns, especially human nouns,and pronouns. Both the adjective and the nominal forms can be converted the same way. For example:

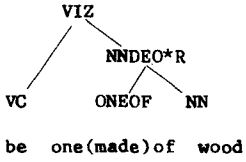


(b) Part of a whole When de is preceded by an inanimate noun or an abstract noun, it represents the meaning 'part of a whole' which is usually translated to 'of' in English. For example:



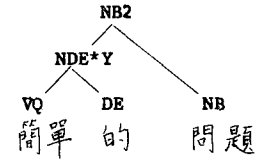


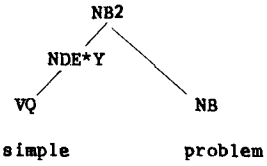




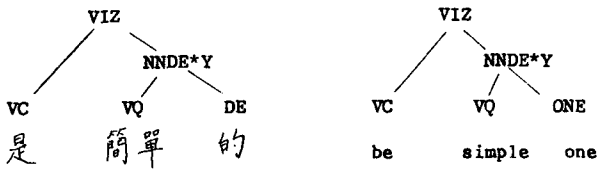
If the class of NN only includes material nouns, such as cloth, wood, metal and others, then the DE can be replaced by ONE MADE OF. Otherwise it can only be converted to ONE OF.

(c) A connective which represents no meaning When de is preceded by an adjective, it has no meaning. Therefore it is deleted in the conversion.



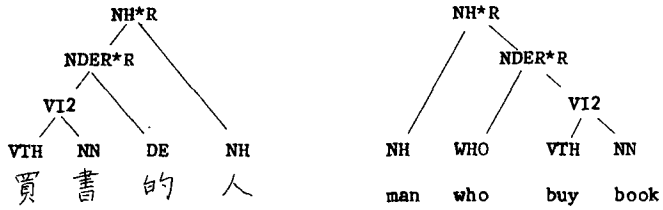


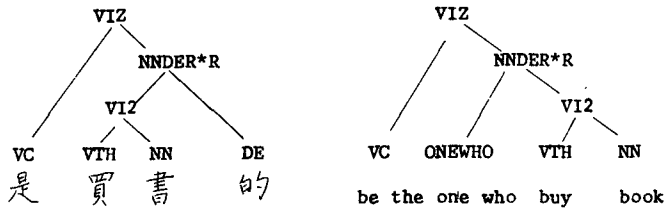
(e) The one that In a nominal phrase, the de represents the one which is modified by whatever is the left constituent, and it cannot be deleted.



(f) A translated syntactic connective The de connects a verb phrase to a noun represents no meaning, but it has an English equivalent in the form of a relative pronoun. The case of the pronoun depends on the constituent that precedes the de.

If the verb phrase consists of an intransitive verb phrase, the noun that the NDE modifies, or the NNDE implies is the subject of the verb phrase. For example:

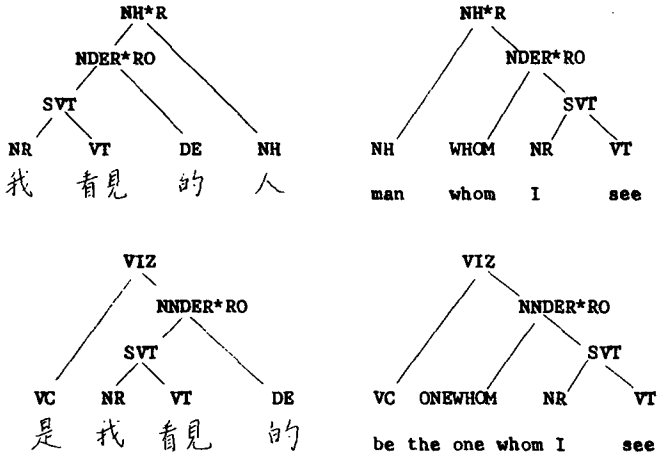




WHO and ONEWHO are used as codes for subjective relative pronouns. They may be where, which or when depending on their antecedents. In order to make the computer to choose between where, who,

which and when, more refinements have to be introduced to the codes of higher nodes.

If the verb phrase before the de consists of a subject (or agent) and a transitive verb, the noun that the NDE modifies and the NNDE implies is the object of the verb phrase. For example:



When the constitute is formed by a complete clause and de, it can only be an adjective which usually modifies the nouns such as time, place, method, instrument, manner and condition. Its conversion rule is the same as that for the intransitive verb phrase.

Of the millions of de phrases, ten conversion rules are sufficient to differentiate the many meanings that de represents, and render a rough but adequate translation. In fact not many conversion rules are needed. In addition to those mentioned above, there are those for comparative constructions, locative phrase and some minor ones.

(4) English morphological rules have been worked out by others. How they will be applied to machine translation from Chinese to English is a problem yet to be solved. Some of the solutions can be written into the dictionary, and some, incorporated into the syntactic conversion rules. A great deal of research have been done in this area; the main problem is to implement the information to the existing system.

(5) The Chinese text to be translated is encoded in the Chinese telegraphic codes. The problem of allographs (including abbreviations) are taken care of by the dictionary lookup. The allographs represent the same lexeme are referred to that lexeme by the system. Punctuations prove to be confusing. The period is used to terminate a sentence, any other usage of the period, such as marking the end of a sub-title, should be eliminated. Otherwise every sentence can be nominalized by the automatic parser.

With some knowledge of PL/1, I am tempted to say that machine translation from Chinese to English is not only possible, it is also easy to program. A program can be written to read the dictionary, the syntactic rules and the text, parse the sentences and store the constitutes. The program searches through the array of constitutes for the symbol of *, and then performs the syntactic conversions as indicated by the codes that follow the *. Once this is done, the program retrieves the constituents of each constitute until the constituents are the terminal lexemes. Then the translation is already rendered in the citation forms.

It is unfortunate that machine translation was considered

impossible a few years ago, and research efforts were curtailed. Correct machine translation of modern scientific writing is possible if enough organized linguistic data are stored in the computer. In the past twelve years the rigid demands imposed by the computer have accelerated the progress of linguistic research on many native languages. It is due to the demand of the automatic parser, the systematized syntactic rules were formulated. It is due the demand of choosing automatically the correct translation by the computer, the study of semantics was initiated. Now the meanings of a given lexeme can be differentiated in terms of its context. With more sophisticated linguistic data of both source and target languages, more efficient programming language, and bigger and faster computers, machine translation can be a reality in the near future.