Some Linguistic Problems Observed in English-Chinese Machine Translation

WU Xiaohong, ZHANG Yujie and ISAHARA Hitoshi
National Institute of Information and Communications Technology
{xiaohongwu, yujie, isahara}@nict.go.jp

1. Introduction

The work introduced here was based on a rule-based English-Chinese MT system by adopting controlled language (CL) technique as a supporting method. This system was focused on English-Chinese MT but was also oriented to a bi-directional MT environment by chosen two highly specialised medical sub-domains as the application field. Our selected texts (medical protocols (MP)), though lexically less ambiguous, are often composed by sentences with long and complicated structures, for instance, long and unclear pre- and/or post-modifications, unexpected ellipsis, complicatedly embedded clauses which are proved difficult to be correctly transferred between these two languages. These characteristics of the MPs made us feel the necessity of semantically and syntactically constraining them [1] [2]. To achieve this goal, we designed a CL with which we greatly reduced the linguistic complexity of these texts, especially those related to the language-specific phenomena.

As is the case, language-specific phenomena are often among the major causes of mistranslation or wrong matching because they are specific to each language and can not be transferred cross-linguistically, be the language concerned the source language or the target language. In the case from English to Chinese, as well as from other language(s) to Chinese or vice versa, examples include the wrong attachment of the Chinese structural particle DE (的), the failure to the construction of the Chinese BA-structure [3] and the failure to the arrangement of the prepositional phrases in Chinese sentences [4], and so on.

Our test using some free English-Chinese MT systems on Internet (Systran [5], Worldlingo [6], and Babelfish Translation [7]) showed a very encouraging result. Compared with the test to the uncontrolled texts, the controlled ones are much better translated by these systems. All the controlled sentences can be considered pretty understandable with some problems mainly caused by lexical choices and/or the language-specific phenomena we have already mentioned. The result also indicates that special attention must be paid to the language-specifically bound phenomena as such.

In the following sections, we will briefly introduce how we specified some of these phenomena in our work. Some suggestions will thus be proposed to tackle the similar problems in the related work.

2. The Chinese Structural Particle DE (的)

The use of the structural particle DE (的) in Chinese is flexible and thus becomes very problematic like many other particles. It is generally considered as the marker of the attributives of nouns or the marker of the adjectives and it is often attached after them. The particle DE (的) has another special usage: to form a language unit: the DE-structure (的字结构). In this case, it is also attached after a word or a phrase to substitute the central word or constituent and is no longer used as the marker of attributive (we will not discuss this in this paper). As the marker of adjectives, the best proof is that in all English-Chinese dictionaries the Chinese correspondences to these words have the particle DE (的) attached after them, ignoring the real grammatical property of certain words in the Chinese language, e.g., "parasitical" as "寄生虫的" (寄生虫 is always a noun in Chinese); "beautiful" as "美丽的/漂亮的". However, this kind of practice is useless in many cases and might look strange or make the sentence ungrammatical if we attach the "DE (的)" after each counterpart adjectives in the lexicon used for MT.

In Chinese grammar, we have special rules for the use of the particle "DE (的)". The employment of this particle after an attributive/adjective is not mandatory. In many cases, we can leave out the DE or we must leave it out unless the ellipsis might produce ambiguity, e.g., “父亲母亲 (father and mother)” versus “父亲的母亲 (father’s mother)”; “生物历史 (biology and history)” versus “生物的历史 (the history of the biology)” and “我们学生(we/us/our students; an ambiguous structure)” versus “我们的学生 (our students)” [8]. However, we found that most of the MT systems we used have not fully resolved this attachment problem. For example, we found the following wrongly translated sentences:

1) *If antigen detection is negative* …
   如果抗原侦查(检测)是消极的(呈阴性),
2) *Portable ultrasound equipment that has a 3.5 – 5 MHz probe*
   (带)有一根 3.5 – 5 兆赫探针的便携式的超声波设备

Note: the character(s) in the parenthesis are the corrections.

Leaving aside the lexical problems (lexical choices), we can see from the above examples (1 and 2) an obvious trace of the word “DE” integrated as the marker of the adjectives in the lexicon. Keeping the DE as the marker of adjectives makes the whole sentence look both redundant and ungrammatical. However, the missing of this word would also make the sentence ungrammatical, e.g.,

3) *Safety and reliability of PAIR depend on the training of the medical staff, relevant indications, the observance of technical rules and safety rules.*
   对(PAIR) 的安全(性)和可靠性取决于(对)医疗职员(人员)的训练(培训) ①, (对)相关的征兆(适应症)②, 技术规则和安全规则遵守③。

The first ① and the second ② parts are nearly correctly translated into Chinese but for the third one ③ we see immediately that between the two conjoined noun-phrase attributives “技术规则和安全规则 (technical rules and safety rules)” and the head word “遵守 (the observance)”, the DE is missing, resulting in a strange rendition: “技术规则和安全规则遵守”.

To solve this problem, we constructed a special database which was used to polish the translated sentences by adding the missing or deleting the redundant particle DE in the sentences. As for the equivalents of the English adjectives we left out the particle DE as an obligatory element in the lexicon. In other words, the information concerned with the use of this particle should be stored as separate datum in the grammar database.

3. Chinese Noun Phrases

Chinese noun phrases (NP) exhibit some special characteristics distinct from many European languages, which makes it difficult to apply directly the methods proposed for these languages to the analysis of Chinese NPs. For example, in English an NP must have a noun as its head. If a word of other grammatical category is going to function as the head noun in the phrase, it must get a legal status, e.g. by means of inflection or derivation or this word possesses the property of multiple parts-of-speech. Same rule is applied to other grammatical phrases too.

However, in Chinese a grammatical constituent can be taken by words of different grammatical categories without overt indications. This is due to the fact that Chinese words do not exhibit phenomenon like morphology seen in most indo-European languages. Chinese language is heavily dependent upon context. Without context, we have no sound reason to claim that the Chinese counterpart of an English noun is also a noun. As a result, many Chinese counterparts of the English NPs have the property of other phrases. Without context, it is extremely difficult to define the grammatical property of a Chinese phrase.

In Chinese grammar phrases are usually referred to as “structures”, namely, “subject-predicate structure”, “verb-object structure”, “verb-complement structure”, “modifier-modified structure” and “conjunctional structure”. A subject position can be occupied by almost all of the above structures. Moreover, as Chinese phrase and sentence constructions are generated using the same set of principles, it is difficult to distinguish explicitly most of the Chinese syntactic structures between a phrase and a sentence. Therefore, the grammatical function of these phrases is uncertain and heavily depends on how they are used in the sentence. For example,

4) A. 进行血清对照 。

Perform serological control.

B. 进行血清对照非常必要。

To perform the serological control is very necessary.

While Example A can be considered as a sentence corresponding exactly to an English imperative sentence, in Example B the underlined part (identical to Example A) can only be considered as a verb-object structure functioning as the subject and thus getting the role of a noun phrase.

Our analysis combines the methods applied in the analysis of English phrases and a method which deals with the characteristics of the Chinese phrase structures. In brief, we allow phrases headed by words of other grammatical categories correspond to the English NPs. That is, we specified the above mentioned Chinese structures also in the form of phrases (relatively fixed syntactic structures) which can correspond to the English NPs. For example, an English NP can be replaced by any of the above mentioned Chinese phrase structures. In so doing these phrases thus possess the property of a noun phrase and get a legal status when functioning as NPs. This way of treating Chinese phrases is favourable to the bi-directional MT application, with special rules guiding the process of such unparallelism.

Similar method is also applied to other phrases. For example, an English prepositional phrase (PP) can have a Chinese NP and/or VP as equivalent or after a Chinese preposition there might be a VP instead of a NP complement, etc.

4. The Chinese prepositions

Prepositions are by nature polysemous and it is almost impossible to find a constant equivalent in two languages. In the case from English to Chinese and/or vice versa, one English preposition might face all the following three situations: having zero equivalent (mainly related to time), several equivalents, and/or an equivalent of other grammatical category (mainly verbs). However, the PP attachment problem is less or sometimes not ambiguous in Chinese due to the different arrangements of these phrases in the sentences. Yet, we find other problems related to the Chinese prepositions which are caused by the complex nature of the Chinese preposition itself. Here we focus only on two of such problems.

First, most Chinese prepositions come from verbs and they exhibit the characteristics of verbs. In some cases it is hard to define the real grammatical status of a Chinese equivalent of a particular English preposition. This suggests that some of the Chinese prepositions can be used both as a verb or a preposition. This is not at all the case of English prepositions. The Chinese prepositions which can also function as verbs are sometimes called ‘coverbs’ which can stand alone as main verbs. For example,

5) *John saw the man with a telescope.*

A. 约翰用望远镜看见了一个人。

B. 约翰看见了一个拿着望远镜的人。

In the translations of this English ambiguous sentence, the two Chinese equivalents (underlined) to the English preposition “with” can be used directly as verbs. This is especially the case with the preposition “with” as it has to be translated into different verbs according to the NP that follows it. Our practice with such prepositions is that we treated the Chinese counterparts of such English PPs as verb-object structures (mainly) which can be transferred into either an English PP or an English relative clause under certain conditions, if it is involved in the case of Chinese-English MT. Therefore we specified different semantic rule sets to guide the processing and the correct matching between such Chinese phrases and the English PPs.

Second, generally speaking, one English preposition has one Chinese equivalent as has shown in the above example. However, sometimes the Chinese counterparts are consisted of two discontinuous elements. In other words, an English preposition has to be represented by two elements in Chinese which are separated by its NP complement. For example, "in the street": "在街上"; "from the classroom": "从教室里". In Chinese linguistics the second element is usually considered as a noun indicating direction, location, etc. Thus for an English PP structure such as PP → P + NP; its Chinese correspondence might become one as: PP → P + NP + N. This property makes it necessary to generalise special rules for the uses of these separated elements accordingly. In our work we did not follow the common practice, i.e. to treat the second element as a noun. We considered the second element as one discontinuous unit of the preposition and generalised a special rule to correspond to this kind of situation, e.g. (PP → P- + NP + -P). It means that these two elements are regarded as one lexical unit and are directly stored in the lexicon as the correspondence of one English preposition with semantic specifications to cope with the other situations.

Therefore, to tackle such problems we specified different rules for different usages of one preposition in order to deal with the potential problems for the purpose of obtaining a better translation quality.

However, though we have successfully solved some of the semantic and syntactic problems concerning the transferring between English PPs and Chinese counterpart structures, there are still other problems we have not fully resolved, e.g. how to exactly differentiate the semantic nuances of some prepositions. This is a tedious work which has to be accomplished with intensive efforts in the future.

5. Conclusion

In conclusions, when facing texts which are written in complicated ways, it is a wise choice to restrict them both in vocabulary and grammar so as to facilitate the MT application. While analysing a language, first, the language-specific problems have to be specially studied. It is better then to explore and construct a grammar which can cope with the characteristics of the language concerned and which can best resolve these problems. Secondly, we must bear in mind that there is no universal rules which can deal with all linguistic phenomena, especially in the application of MT. Finally while applying the commonly recognised linguistic theories we should creatively adapt them to our real practice by combining the most suitable methods in order to tackle the specific problems that jeopardise our work.

References:

1. CARDEY, S., GREENFIELD, P., WU X. H., 2004. "Desinging a Controlled Language for the Machine Translation of Medical Protocols: the Case of English to Chinese", In *Proceedings of the AMTA 2004, LNAI 3265*, Springer-Verlag, pp. 37-47
2. WU, X. H., 2005. "Controlled Language – A Useful Technique to Facilitate Machine Translation of Technical Documents", In *Lingvisticoe Investigationes* 28:1, 2005. John Benjamins Publishing Company, pp. 123-131
3. WU, X. H., CARDEY S., GREENFIELD P., 2006, "Realisation of the Chinese BA-construction in an English-Chinese Machine Translation System", *the 5th SIGHAN Workshop on Chinese Language Processing, Sydney, Australia.*
4. WU, X. H., CARDEY, S., Greenfield, P., 2006, "Some Problems of Prepositional Phrases in Machine Translation", *FinTal: -5th International Conference on Natural Language Processing, Turku, Finland.*
5. http://www.systransoft.com/Corporate/SWS.html
6. http://www.worldlingo.com/en/products_services/worldlingo_translator.html
7. http://world.altavista.com/tr.
8. HU, Y. S., 1987. "Modern Chinese" 《现代汉语》, Educational Publishing House, Shanghai. ISBN 7-5320-0547-X-G·466, p. 345