The Processing of English Compound and Complex Words in an English-Chinese Machine Translation System

Shu-Chuan Chen* **and Keh-Yih Su****

***BTC R&D Center**
2F, 28 R&D Road II
Science-based Industrial Park
Hsinchu, Taiwan, R.O.C.

****Department of Electrical Engineering**
National Tsing Hua university
Hsinchu, Taiwan, R.O.C.

ABSTRACT

In a machine translation system the information of the words of the source language should be available before any translation process can begin. The information of simple words can be obtained only by entering a word with all its relevant information into the lexicon. On the other hand, compound words and complex words, it seems, can be handled in a satisfactory way by lexical redundancy rules, and will thus help keep down the size of the lexicon. This paper argues that lexical redundancy rules are not as useful as they may seem to be for a machine translation system, and both their limitations and functions will be examined in depth. In addition, detailed discussions on the various problems that may arise during analyzing and translating of compound and complex words are presented.

1. Introduction

In a machine translation (MT) system the information of the words of the source language should be available before any translation process can begin. The information of simple words can be obtained by entering a word with all its relevant information into the lexicon. On the other hand, compound words and complex words at one extreme may be expected to be exhaustively listed in the lexicon [Zhan87] or at the other extreme be handled in a satisfactory way by lexical redundancy rules. However, it is obvious that since new compounds and complex words are created from day to day, they are impossible to be exhaustively listed. And as it would be made clear in the following discussions that the predictability, or regularity, of derivational words, inflectional words, or compounds is of limited use as far as translation is concerned, the lexical redundancy rules used to account for these words often fall short of their functions when applied in a MT system. Competent strategies are needed to successfully handle these two types of words to guarantee correct parsing and translation, and also help keep down the size of the lexicon.

In this paper, the various problems encountered in the morphological analysis module of the BTC English-Chinese MT system are discussed and possible solutions are proposed. The discussion will focus on the role of lexical redundancy rules in a MT system; and the issue as to whether English compound and complex words used in such a system should be derived from their stems solely through lexical redundancy rules. At last, we will look into the problems of processing multi-affix words and compounds of different formations.

2. English Compound Words, Complex Words, and Lexical Redundancy Rules

In English, new words may come into being through the process of derivation, inflection, or compounding. These processes, distinct from other less productive word formation devices, e.g. clipping, acronym, etc., create new words by adding new morphemes. Compounding creates words by adding one base to another and the forms created are called compounds. Derivation and inflection produce words by adding an affix to a base. Complex words, often used by linguists to mean exclusively for formations by the addition of derivational affixes to compounds, will be used in this paper to cover forms with either derivational or inflectional affixes for the reason that they are both created by affixation and thus require similar operations in a MT system. Words formed by these processes are large in number and bear a fixed phonological, syntactic[1], and semantic relation either to the stem of the complex word or to the grammatical head of the compound word.

For a MT system like ours whose input is written strings rather than spoken words, the syntactic and semantic predictability of compound words and complex words is of special interest. Due to the predictability, it appears that these words can be recognized and analyzed by lexical redundancy rules, and need not be listed as separate lexical items in order to reduce the memory space of lexicon. Lexical redundancy rules are intended to assign default form class, semantics, and other attributes to a group of words that share formal and functional resemblance. As an example, complex words ending in the suffix -ment, such as *arrangement*, *puzzlement*, etc., are all nouns and have a common meaning of "the result of " the action indicated by the verb base, and so on [Quir85]. And these words can be generated (in a

language generation system) or analyzed (in a language analysis system) by a redundancy rule of the following simplified form:

V + -ment —> N
Meaning : the result of V
Chinese translation : same as V

The arrow indicates that the word created by adding the suffix *–ment* to a verb is a noun with its Chinese translation identical to the base. An example involving compounds can readily be cited as well. However, lexical redundancy rules are not as useful as they appear in a MT system for the reasons to be discussed in the following section.

3. Limitations of Lexical Redundancy Rules in a MT System

Ideally, the use of lexical redundancy rules in a MT system will help restrict the lexicon to a reasonable size, and thus keep down the space allocated for storing lexical items. Nevertheless, the question as to whether a compound or complex word should be entered into lexicon, or whether they should be analyzed by rules, is not merely a matter of the size of lexicon, especially when the time spent on searching and analyzing a lexical item and the memory taken up by lexicon is trivial. (It is observed that in our system morphological analysis, including I/O, takes up less than 5% of the total processing time.)

The main concern of a MT system is to render a suitable translation from the source language to the target language. To this end, several conditions have to be satisfied, and they are the determining factors as to whether a compound or complex word should be built into lexicon or not.

1. Compositionality of translation. The translation of a multi-affix word is not necessary compositional, meaning the translation of such a word is not necessarily the composite of the respective translations of the affixes plus that of the stem [Zhan87]. For one thing, the suitable translation is subject to Chinese word-formation rules[2]; for the other, if there already exists in Chinese an established term for the same idea expressed by a given English word, the established word are most likely to be used as the corresponding translation. In the absence of compositionality, correct translation can not be obtained by general rules. In this case, the multi-affix word has to be entered into the lexicon. As an example, the derivational word *reconfigurability* is formed by attaching *re-* to the root *configure*; then *-able* to *reconfigure*; and finally *-ity* to *reconfigurable*. Provided that *re-* is given the default translation " 重新 ", *–able* " 可以 ", *-ity* " 性 " and *configure* " 配置 ", the Chinese translation of *reconfigurability* is not likely to be the composite of the translations of *re-*, *-able*,, and *-ity* plus that of *configure*, that is " 可以重新配置性 ". A potential candidate is " 重構性 "[3]. The same criterion goes for compounds. An example of it is *flesh-and-blood*. When translated compositionally, it would be " 肉和血 ". However, the corresponding institutionalized translation of the compound is " 血肉之軀 "
2. Adequate information for rendering correct translation. A variety of morphological, syntactic, and semantic information is needed for an English word to be correctly translated into Chinese. If any single piece of information of an English lexical item failed to be obtained through default assignment by lexical redundancy rules, this word has to be

entered into the lexicon. For example, it is necessary to consult the subcategorization restrictions of the Chinese words (Mandarin, to be more precise, since the BTC MT system is actually English to Mandarin) " 老 " and " 舊 " to determine which word the English word *old* should be translated. The rule is basically that if *old* modifies an animate noun, it should be translated as " 老 " and never " 舊 ". Suppose that the word *dancer* is generated by a redundancy rule that derives nouns by adding *-er* to verbs and stipulates that the derived noun must indiscriminately be of the attribute "inanimate" (which can be animate as well), *dancer* will be erroneously labeled as an inanimate object. This will result in the translation of *old dancer* into " 舊的舞者 " , and not the correct " 老的舞者 ". Therefore, words like *dancer* must be listed in the lexicon.

3. Ability to identify elements in a compound. Compounds may take the form of two separate words, such as *hard copy*, or a hyphenated word, such as *hard-copy*. The elements of a compound can also be combined together as a single word, such as *hardcopy*. The last form proves to be very difficult in identifying its composite morphemes. In this case, despite the regular syntactic and semantic ties between a compound and its elements, compounds of this makeup have to be built into the lexicon.
4. Productivity of affixes. The productivity of an affix also plays a role in the admissability of a word into the lexicon. Words derived by an affixation of limited productivity should be entered as separate lexical items, because they are very few in number. Otherwise, the addition of a non-productive morphological rule may increase the complexity of the system and the processing time as well. For example, the prefix *step*, denoting kinship, is no longer productive [Baue83], and we should enter all the words with this prefix[4] into the lexicon. However, a risk in deciding to leave out a marginally productive affixation rule is that as it is not actually extinct, occasional coinings are still possible. There is a tradeoff to be made in this regard.

4. Functions of Morphological Redundancy Rules in a MT System

The above criteria will eliminate most compound words and some complex words from the possibility of being analyzed by lexical redundancy rules. Although the use of redundancy rules is restricted by the concern for rendering a correct translation, they are still important in three areas. First, since inflectional morphemes preserve the category of their stems, and the corresponding Chinese translations of inflected forms are highly regular, the majority of them should be handled by rules.

Second, redundancy rules can be used to predict the possible category for a word not in the lexicon by examining just the affixes attached to it. For example, if the word *absentmindedness* is not in the lexicon, while *absentminded* is, a rule that identifies a word which is made up of an adjective plus the suffix *-ness* is a noun, the word *absentmindedness* will be given the correct category. This makes it possible to assign a correct category to a word, and which is one of the prerequisites in producing a correct parse tree for a given construction. Once the right structure is obtained, the whole construction will be correctly translated as a result.

Third, redundancy rules are of avail in giving a suitable translation to words not in the lexicon by considering the semantic relation they bear to the stem of a complex word or to the head of a compound word. In line with the rule that gives category to *absentmindedness*, the Chinese translation of the same word can be obtained by giving *-ness* a default translation.

Thus lexical redundancy rules are helpful in providing as much information as possible in both parsing and transfer phases. This is the general idea and technique behind the "fail-soft" in a MT system [Benn85]. The conclusion to be drawn regarding the use of lexical redundancy rules is that: Compound words and complex words should be built into lexicon if good translation is not available through default assignment by lexical redundancy rules. On the other hand, since new words are constantly created, lexical redundancy rules are indispensable.

In Sections 5 and 6, we will examine the internal structure of compound and complex words, and the effect it has on the processing of these words in a MT system.

5. Processing of Compound Words

Among the three types of compounds noted in section 3, only compounds with elements separated by a space or hyphen are of interest as far as processing is concerned. The type of compounds spelt as a single word will all be listed as lexical items in our system, because they are difficult to process.

As with compounds composed of elements separated by a space, quite a few are established compounds, and this type of compounds is rather productive in coining new ones, found particularly in the terminology used in a specific field of study. As for the kind of rules needed to analyze these compounds, it can be either a morphological rule or a syntactic rule. The former will recognize the compounds at the phase of morphological analysis, which is prior to syntactic analysis. But this can alternatively be done during syntactic analysis; that is, compounds of separate elements are treated like a phrase in order to eliminate the need of an extra operation during morphological analysis. Thus, the processing of a noun compound like *prototype development system* can be left until syntactic analysis phase to be parsed as a noun phrase. This can be done because phrases and compounds share quite a lot of common ground in their internal structure; in other words, word syntax is on a par with phrase structure [Tang88].

The most frequently encountered compounds are the hyphenated compounds, and it is this type of compounds for which lexical redundancy rules are of the greatest use. Hyphenated compounds used as adjectives are extremely productive and most of them are the instances of occasional coinage. Established ones are fewer in comparison to occasional creations. For instance, compounds made of cardinal plus noun, such as *40-word* in *40–word lexicon*, are extremely productive.

Formally, several individualities of hyphenated compounds are noteworthy. First of all, they may take an entire phrase as its elements, e.g. *higher-than-average (an adjective phrase)* in *higher-than-average wages*, and *do-it-yourself (a verb phrase)* in *do-it-yourself approach.* Second, suffixes may be attached to the last element of a compound which does not normally take such suffixes when used as an independent word. For example, the noun in a compound expressing physical attribute might take the past participle ending, e.g. *leg* in *three-legged table.* And in compounds that express fraction, ordinals might take the plural ending, e.g. *third* in *two-thirds*[5]. In addition, a number of grammatical relationships are possible between the components in a compound, and different types of meaning and translation will thus result. For example, in a noun-verb compound, the grammatical relation between noun and verb may be instrument-action, such as *petrol-lighter*, in which *petrol* is the instrument the lighter uses. Whereas, in *fire-lighter*, *fire* is the object of the action *light.*

The formal and semantic attributes of hyphenated compounds observed above have the following effects on processing and translating these words. First, for words like *three-legged* and *two-thirds*, special rules have to be constructed for handling the irregular inflection.

Second, detailed rules have to be worked out to pinpoint the grammatical relation between the elements of hyphenated compounds in order for them to be correctly translated into Chinese. For instance, the corresponding translation of the instrument-verb compound *voice-controlled* may employ " 由 " to express the instrumental case, such as " 由聲音控制 ". While the corresponding translation of the object-verb compound *letter-writing* is simply placing the object after the verb " 寫信 ".

In view of the fact that the grammatical relationship of the elements within a compound is difficult to define, and the translation is far from certain even if the precise relation can be identified, therefore, the vast majority of these words have to be built into the lexicon. Nevertheless, there are two cases in which correct translation is possible without resort to lexicon. For a group of compounds that have the same stem and the stem also has a fixed translation in Chinese, translation rules can be constructed specifically for this stem. For example, there are a lot of compounds involving the stem *oriented* in their formation, such as *screen-oriented*, *row-oriented*, *column-oriented*, to name just a few. A rule to the effect that *noun-oriented* will be translated to "noun- 導向 " will be sufficient. On the other hand, compounds like *three-legged* which contains the same items and word order as in a corresponding noun phrase *three legs* can be handled by the very set of transfer rules constructed for translating English phrases into Chinese. So, the phrase *three legs* when translated into Chinese needs a classifier " 隻 " before " 腳 " to give " 三隻腳 ". The same holds for a compound containing identical elements and functioning as a modifier of another noun, i.e. *three-legged* as in *three-legged table*, whose translation will thus be " 三隻腳的桌子 "

Third, English phrase compounds are phrase in nature and, when used as a modifier of nouns, correspond closely to the structure of Chinese noun phrases: when modifying a noun, phrasal modifier and clausal modifier, are pre-modifier rather than post-modifier of nouns in Chinese. For example, in English *three-year-old* can be a noun or a modifier of noun, as in *a three-year-old girl*, which is equivalent to *a girl who is three years old.* Both the phrase compound *three-year-old* or the noun phrase *three years old* will be translated identically as " 三歲 ". Hence, the translation of phrase compound can be taken care of by transfer rules as well.

6. Processing of Complex Words

An English complex word exhibits several characteristics that are pertain to the processing of complex words in a MT system. First, English inflectional affixes are all suffixes, while derivational affixes can be either prefixes or suffixes. Second, in terms of the number of derivational and inflectional affixes, a complex word may consist of more than one derivational affix, with an additional inflectional suffix outside these derivational affixes. For example, *configurabilities* is formed by adding derivational suffixes *–able* to *configure*, *-ity* to *configurable*, and the inflectional suffix *–s* to *configurability*. Third, in terms of the number of prefixes or suffixes, a complex word may have more than one prefix or suffix. For example, *unrerunability* has two prefixes *un-* and *re-* and two suffixes *–able* and *-ity*. Fourth, suffixation, but not prefixation, may cause changes in the orthography of the stem forms[6]

For example, the suffix *-able* when attached to a verb ending in *e*, will sometimes delete the final *e*, e.g. *consume* becomes *consumable*.

Based on the characteristics of complex words observed above, the processing of words with only prefixes, words with only suffixes, and word with both prefixes and suffixes each requires different operations. For words with prefixes alone, de-prefixation is followed by dictionary look-up to check if the stem can be found in the lexicon. If the word is found then no prefix should be further removed. If a stem can not be located in the lexicon, two things are possible. First, there is no such word in the lexicon and thus it should be assigned the category specified by the rule in order for the sentence in which it occurs to be successfully parsed. Second, if there is another prefix after the current one, further de-prefixation will unravel the stem. The operations de-prefixation requires are depicted in Figure 1:

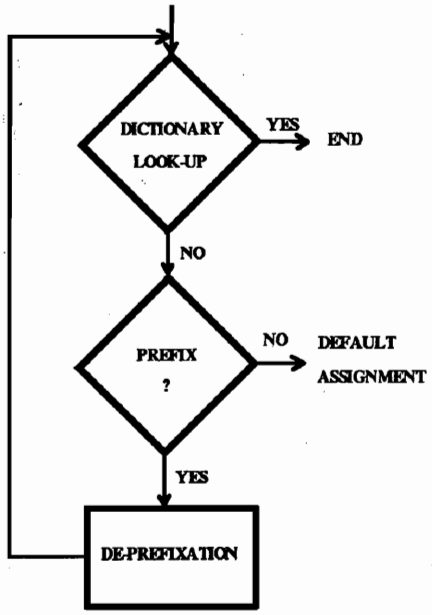

Fig. 1 THE FLOW OF DE-PREFIXATION

For words with suffixes alone, de-suffixation is likewise followed by dictionary look-up to check if the stem can be found in the lexicon . However, if a stem can not be found in the lexicon, three things are possible. First, it may due to the fact that there is no such word in the lexicon and a suitable category should be assigned. Second, it may be that there is another suffix before the current suffix. In this case, further de-suffixation is needed. Third, it is also possible that suffixation process has altered the orthography of the stem; and only after the original form has been restored, can dictionary loop-up be performed to see if another suffix should be removed. For example, after *-able* is removed from *consumable*, the form *consum* is not a word, and an *e* has to be restored. De-suffixation requires the following operations in Figure 2:

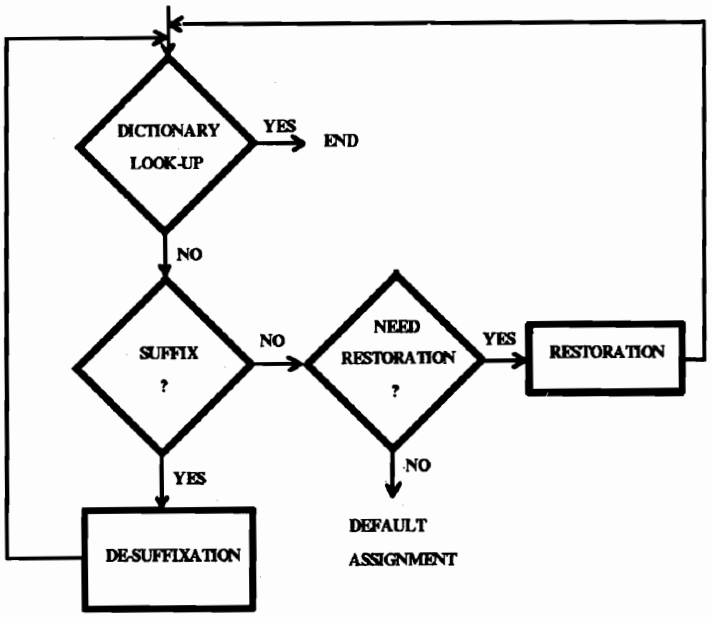

Fig. 2 THE FLOW OF DE-SUFFIXATION

To make the situation more complicated, words containing prefixes in addition to suffixes call for both de-prefixation and de-affixation. This involves a back-and-forth check on both ends of a word. The check can be initiated at either end. Once the category matches the specification in a prefixation rule or that in a suffixation rule, de-affixation has to be done. For example, to de-affixize *reconfigurable*, either de-prefixation or de-suffixation can be tried first. If we start with the prefix, de-prefixation will fail since *re-* is a prefix that must be attached to a verb, but *configurable* is an adjective. At this point, we have to restart with de-suffixation, *–able* will be removed to yield *reconfigure*. After *-able* is removed the remaining form is *reconfigure*, de-prefixation can now be executed to remove *re-* and leaves *configure*

De-suffixation, and de-affixation in general, includes a check in category. If the stem does not match the category specified by the rule concerned, de-suffixation should not be carried out. The importance of matching category in the process of de-suffixation lies in the fact that it helps determine if restoration is in order. For example, after taking off the ending *-ing*, *using* will become *us*, which is not a verb and cannot be the right stem. Therefore it is obvious that a final *e* must be missing.

The processing of complex word, especially multi-affix complex words, may pose a number of problems:

1. The major problem is that for words that are not in the lexicon, there is no way of telling if they contain affixes or not. For a newly-coined word *prerechit* , we are not sure whether it contains a prefix, *pre-*; or two prefixes, *pre-* and *re-*; or no prefix at all. In this case, the principles of assigning default category and default Chinese translation may result in wrong guesses.

2. As pointed out above, after a suffix has been removed, if the remaining part cannot be found in the lexicon, it is likely that there is another suffix before the current suffix. In this case, if there is no change in orthography, further de-suffixation will unravel the stem. On the other hand, if there is a change in spelling, it is hard to detect if the remaining part is a word not in the lexicon or a word with more suffixes. There are, however, two possible ways to solve this problem. For instance, after *-ity* is detached, the remaining part of the

word *executability* is *executabil*, because the suffix *-able* has been transformed. In this case, we can either stipulate that if there is a string *abil* occurs before *-ity*, then *abil* should be restored to *-able*. The other way is simply to treat *ability* as a single suffix. Thus no further analysis of the internal structure of *ability* is necessary. The latter is a better solution for the sake of simplicity in processing.

3. The restoration of base words can be time-consuming. For example, the rule that derives nouns by adding the suffix *-sion* to a verb can cause a base form to lose its final *e*, such as *confusion*; or *t*, such as *conversion*; or *de*, such as *explosion*, etc. Every possibility has to be tried to restore the verb. To remedy this problem, if a given operation applicable to only a handful of words, these words might as well be listed in the lexicon. If we choose to do so, however, new words can not be accounted for if they happen to need restoration of this sort. Here, we are faced with another tradeoff.

7. Conclusions

In this paper, we provide a comprehensive look at the functions and limitations of lexical redundancy rules used in analyzing compound and complex words in a MT system. The conclusion is that since redundancy rules most of the time cannot guarantee correct translation of compound and complex words, it is suggested that redundancy rules be reserved for analyzing words that are occasionally coined in order for the construction to be parsed successfully. In the paper, various problems concerning the processing of compound words and complex words are examined, and possible solutions are proposed. Nevertheless, the problems presented in this paper are by no means exhaustive, and there are other difficulties in processing compound and complex words that are worth noting, such as the treatment of words like *passers-by*, which has an inflected form as the first element, and so on. In addition, idioms or collocations can also be regarded as a special case of compounds and are needed to be studied further. These issues, however profound they may be, are out of the scope of the current paper.

8. Acknowledgement

We are grateful to Professor Ting-Chi Tang at Foreign Languages Department of National Tsing Hua University for his helpful discussions and critical comments on this paper. Special thanks are due to the colleagues in BTC R&D Center for their support and encouragement.

NOTES

1. What we mean by syntactic relationship is mainly about the relationship in the word class between the composite elements and the whole compound or complex word.

2. Discussion of Chinese morphological rules is beyond the scope of this paper. For detailed discussions of Chinese morphological rules, please refer to [Tang88].

3. The Chinese translation of *configure* given in the English-Chinese Dictionary of Computing Technique is " 配置 ", and one of the translation of *reconfiguration* is " 重新配置 ". However in the compound *reconfiguration system*, *reconfiguration* is translated as " 重構

". This is also true for five other compounds containing *reconfiguration*. Based on this, *reconfigurability* is given the translation of " 重構性 "

4. There are only eight of them: *stepbrother*, *stepchild*, *stepdaughter*, *stepfather*, *stepmother*, *stepparent*, *stepsister*, and *stepson*.

5. Strictly speaking, the suffixes are added to the compound as a whole when functioning as an adjective , not to an individual component.

6. The prefix *in* also causes changes in spelling to the initial consonant of the base through an assimilation in pronunciation, e.g. *in* becomes *il* before the lateral *l* as in *illegal*; *in* becomes *im* before a labial as in *impossible*, etc. Since the prefix is no longer in productive use due to the competing prefix *un*, it is safe to state that prefixation does not cause any changes in the spelling of the base.

REFERENCES

[Baue83] Bauer, L., *English Word-Formation*, Cambridge University Press, Cambridge, Great Britain, 1983.

[Benn85] Bennett, W.S., "The LRC Machine Translation System," *Computational Linguistics*, Vol. 11, NOs. 2-3, pp. 111-119, April-September 1985.

[Biss85] Bissantz, A.S. and K.A. Johnson ed., "The Minimal Units of Meaning: Morphemes", *Languages Files*, The Ohio State University Department of Linguistics, 3rd ed., Advocate Publishing Group, Ohio, U.S.A., 1985.

[Hutc86] Hutchins, W.J., *Machine Translation: Past, Present, Future*, Market Cross House, West Sussex, England, 1986.

[Quir85] Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman Group Limited, Essex, England, 1985.

[Tang88] Tang, T-C., *Studies on Chinese Morphology and Syntax*, Student Book Co., Ltd., Taipei, Taiwan, 1988.

[Vasc85] Vasconcellos, M. and M. Leon, "SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization" *Computational Linguistics*, Vol. 11, Numbers 2-3, pp. 122-136, April-September 1985.

[Zhan87] Zhang, Liangping, and Shengxin Chen, "Ambiguity Processing in English-Chinese Machine Translation", Conference on Translation Today, Hong Kong, 1987.

名山出版社 (左宜有, 左宜德主编) 電腦資訊科學辭典 (English-Chinese Dictionary of Computing Technique (Data & Information)), 臺北, 1983.