Chinese Sentence Generation in a Knowledge-Based Machine Translation System

Tangqiu Li, Eric H. Nyberg, Jaime G. Carbonell
Center for Machine Translation
Carnegie Mellon University

ABSTRACT

This paper presents a technique for generating Chinese sentences from the Interlingua expressions used in the KANT knowledge-based machine translation system. Chinese sentences are generated directly from the semantic representation using a unification-based generation formalization which takes advantage of certain linguistic features of Chinese. Direct generation from the semantic form eliminates the need for an intermediate syntactic structure, thus simplifying the generation procedure. The generation algorithm is top-down, data-driven and recursive. The descriptive nature of the pseudo-unification grammar formalism used in KANT allows the grammar developer to write very straightforward semantic grammar rules. We also discuss some of the crucial problems in Chinese language generation, and describe how they can be dealt with in our framework. This technique has been implemented in a prototype Chinese sentence generation system for KANT. Some implementation details and experimental results concerning the prototype are presented at the end of this paper.

1. Introduction

Natural language generation in general contains several processes: content planning; sentence planning; and surface realization (sentence generation), each of which also contains several subprocesses (Nirenburg, et al., 1989). Content planning, or strategic generation, involves determining what to say; sentence planning organizes the content to be conveyed into sentence-sized units, selecting proper tense, voice, mood, etc, as well as the proper words to be used. On the other hand, surface realization, or syntactic generation, determines the syntactic structure for a sentence, inflecting words properly when necessary, and placing words in the proper order. The latter process is sometimes referred to as tactical generation in some literature.

In knowledge-based machine translation systems such as KANT (Mitamura, et al., 1991), the content planning process is less crucial than it is in explanation generation for expert systems or database dialogue systems; the input to the generation module in KANT is usually a sentence-sized unit of meaning representation, which already contains the source information for sentence planning. Hence the scope of this paper is limited to the issue of word selection and sentence generation.

In this paper we present a technique for generating Chinese sentences from the semantic representation encoded in the semantic frame representation (Interlingua) used in KANT. Based on the features of the Chinese language, we have constructed a unification-based generation grammar which generates Chinese output directly from the semantic representation, which eliminates the need for an intermediate syntactic structure and thus simplifies the generation procedure. The generation algorithm used is top-down, data-driven and recursive (Nyberg, et. al, 1989).

Before giving a detailed explanation, we first give an outline description of our technique and the structure of the interlingua text (ILT) (Section 2); then we discuss some features of Chinese and explain why we choose the technique proposed in this paper (Section 3). Then we proceed with an explanation of the semantics-based generation technique itself, including lexical selection, the pseudo-unification framework for knowledge-base representation, and the generation control algorithm (Section 4). Then we discuss some of the crucial problems in Chinese language generation and how we deal with them in our framework (Section 5). Finally, we present an overview of the system's implementation and conclude with a discussion of some open issues related to our technique. All of

the examples in the paper are drawn from our system's current output.

2.Features of Chinese and Generation Strategy

Chinese is very different from English and other European languages. Besides the obvious difference in the number of elements of character set and their appearance, there are many differences in syntax, and in the relationship between syntax and semantics. First of all, Chinese does not manifest a high degree of morphological complexity in term of types of grammatical morphemes when compared to other languages. Hence it is characterized as an isolating language (Li and Thompson, 1982). For example, there is no inflections which function as number markers on nouns, such as the plural morpheme "+s" in English; there is no inflection of verbs to signal difference in number, person, tense or aspect, such as the English forms "give, gives, given, giving" for the verb "give". There is no number and gender agreement between subject and verb. To convey these different meanings, Chinese uses open class words themselves, with the help of word order cues and some closed-class functional words..

Secondly, Chinese is not an easy language to classify in term of relative word order between subject, object and verb, such SVO, SOV, or VSO. This means that sentences with verbs at the beginning, in the middle, and at the end of a sentence can be found in the language. The order in which basic words and phrases occur is governed to a large extent by considerations of meaning rather than of grammatical functions. For example, in general, a pre-verbal locative phrase signals the location of an action, while a post-verbal locative signals the location of a person or thing as the result of the action. The following examples show that these positions are not interchangeable in Chinese sentences, although in English their corresponding prepositional phrases are all at post-verbal position:

1.a 他在车房里修理汽车.(He is repairing a car in the garage)

1.b 他把汽车停在车房里.(He park the car in the garage)

Moreover, Chinese is a topic-prominent language. The topic always comes first in the sentence, and signals what the sentence will be about. The topic of a sentence can be the subject or object of the sentence, a time phrase, a locative phrase or even something which does not necessarily relate directly to the verb in the sentence. The role of topicalization and its effect on semantic content is such that the movement of constituents occurs much more frequently than in English. The following illustrates another very typical example:

2.a 我买了书了.

2.b 我书买了.

2.c 我把书买来了.

2.d 书我买了.

Where 我 means *I*, 买 means *buy*, 书 means *book*, and 了 is a particle to signal perfect aspect or a sentence-final particle to signal a new situation. Superficially, all the sentences listed above talk about the same thing, while their precise meanings are quite different. The first sentence is a neutral way to say "I have bought a book." The second one, an SOV structure, is a way to say "I have bought the book." But this sentence is typically used in a situation in which what is being conveyed is contrary to the expectation held by the listener. The third example is the 把 construction, emphasizing what has been done to the object or what is the result of the 'doing', where 来(come) is a verb originally, but here it combines with 买 and functions as a directive. The final example is the topicalized form of the sentence, signaling that the current topic of the conversation is the book. In the example, we can see that the same words can occur in many different orders, implying not just different ways of saying the same thing, but ways of achieving quite different communicative goals. Therefore, these patterns are not generally interchangeable in a given context.

These distinct features of Chinese give rise to a lot of challenges in both analysis and generation of Chinese sentences. During analysis of Chinese, the syntactic means does not provide much clue. In the example context given above, both 2.b and 2.d can occur; these are difficult to analyze without appeal to semantics and pragmatics. In sentence generation, the generator is forced to choose one of many possibilities to convey the original meaning properly, based on both the meaning to be conveyed and the context in which the meaning is to be expressed.

In short, syntactic clues are not adequate for Chinese processing, whether it is sentence analysis or sentence generation. Both processes are heavily dependent on word meaning, pragmatics and other sources of knowledge. Based on this observation, we adopt a sentential semantic representation based on concept frames as the basis for the sentence generation, instead of more typical syntactic structures. We map the semantic structures directly into linear structures (sentences) without first converting them into f-structures as in a typical KBMT application (Goodman and Nirenburg, 1991). This technique allows us to construct a semantic grammar which models the direct relationship between semantic predicate-argument relations and the linear realization of the sentence.

3. Generation From Semantic Frames

The input to the generation module varies greatly from system to system. The difference between input representations usually reflects the difference between choices of machine translation methodology. As mentioned before, our system utilizes a version of the interlingua representation used in KANT.

The KANT interlingua is based on the notion of *concept frames*. Each concept frame represents a a given unit of meaning along with its specific properties and/or its relationships with the other concepts reflected in the utterance. There are several kinds of concepts in the system's ontology (Leavitt, et al., 1994). The basic concept types are objects, events and properties, which are the basic elements of our interlingua text (ILT) and typically represent nouns, verbs and adjectives. The representations of a single source language input is a kind of semantic network which contains instances of concepts from the ontology. Like other artificial and formal languages, the interlingua has its own lexicon and syntax. A BNF-like specification of the structure of the concept frame is briefly shown as follows:

```
<concept frame> := (<concept name> <relationship with other
                    concepts>*)
<relationship with other concepts> := (<relation> <concept frame>)|
                                      (<property> <value>)
<relation> := agent|patient|theme|location|time|predicate|...
              |reason|purpose|condition|sequential-event|...
              |attribute|type|object|proposition|...
              |intensifier|...
<property> := mood|tense|aspect|negation|...
              number|reference|...
<value> := <symbol>|<string>
```

A concept frame contains a concept name and an arbitrary number of slot-value pairs representing the properties the concept has and/or the relationships of this concept with other concepts in the sentence. Both the concept names and other slot names of the encoding use English strings; however, they represent language-independent concepts in the sense that both English speakers and other language speakers can use them for meaning assignment to lexical items in their respective languages. Different concepts, such as objects, events and properties use the same framework of representation with different sets of relation and property labels. Sentences, for example, are typically represented by an event frame which contains information about the meaning of the proposition, an instantiation of a predicate and its arguments, as well as non-propositional information about modality, speech acts, focus, discourse relationship and so forth. See the example below:

NOTE: The truck must be parked on a level surface.

```
(*E-PARK
    (MOOD DEC)
    (PASSIVE +)
    (MODAL NECESSITY)
    (COMPULSION +)
    (LABEL (*O-NOTE))
    (THEME
       (*O-TRUCK
          (REFERENCE DEFINITE)))
    (LOCATION
       (*O-SURFACE
          (REFERENCE INDEFINITE)
          (ATTRIBUTE (*P-LEVEL)))))
```

The task of sentence generation, of course, is to map the ILTs to a target language sentence or phrase according to the concept type of the ILT. Traditionally, this is done in three phases in KANT: lexical selection, f-structure creation, and syntactic generation. In the lexical selection phase, the most appropriate lexical item or items are selected for each frame in the ILT. Then the ILT and the set of appropriate candidate lexemes are analyzed to determine and produce a syntactic functional structure (f-structure) for the target utterance. Finally, the syntactic generation phase produces a properly inflected and ordered output sentence according to the target language generation grammar.

The method adopted in this paper uses only two phases of generation, instead of three: lexical selection and semantics-based generation. The functions of the first phase stay almost the same. The major difference is that the target utterance is generated directly from the concept structure in one step during the semantic generation phase. We will describe both of these phases in more detail in the following sections, while keeping our focus mainly on semantic generation.

4. Semantics-Based Generation Using Pseudo-Unification

4.1 Lexical Selection

Propositional meanings are typically realized by open-class lexical items, with the assistance of some non-propositional information about modality, speech acts, focus, discourse relations, etc.
As mentioned above, the ILT contains information about the meanings of propositions as well as closed-class items. Non-propositional meaning can be realized with the help of lexical items, word order or selection of specific syntactic structures. The lexical selection process, especially with respect to open-class lexeme selection, is one of the most important tasks in machine translation (Goodman and Nirenburg, 1991). We give only a brief description of lexical selection here; for more details on lexical selection in KANT, see (Mitamura, et al., 1991; Nyberg, et al., 1989).

The basic means by which lexical selection is performed is via thematic-role sub-categorization. Each noun and verb in the generation lexicon is sub-categorized for its appropriate thematic roles. To find the most preferred lexical choice, the system must compare the thematic roles of various lexical candidates with the roles of a given input or proposition in the ILT to be realized. To provide the knowledge needed in this phase, the generation lexicon has to information about the desired subcategorization fillers for each lexical item where appropriate. The structure of our generation lexicon is as follows:

(<concept name>
 <sub-categorization information and its concept hierarchy in domain onto>
 <Chinese correspondent word>
 <transformation rules>)

The <concept name> is the same as the concept name used in the ILT, which serves as the index which matches corresponding Chinese lexicon items. The <sub-categorization information and its

concept hierarchy in the domain ontology> is the main source of information to be compared with the sub-categorization information in the ILT to be realized. When the system selects a Chinese word to realize a concept, it finds all the candidates through its <concept name> and then selects the one which best matches the ILT. The badness of the match between a canditate and the ILT is judged by the number of semantic constraint violations in its thematic roles. If two candidates have the same number of semantic contraint violations, their order in the lexicon will decide which one will be selected. Otherwise, the match with fewest violations is chosen. Once a candidate lexeme is selected, its <Chinese correspondent word> is selected to realize the concept. The <sub-categorization information and its concept hierarchy in the domain ontology> indicates how to transfer the thematic roles in the ILT into corresponding roles in the Chinese word concept frame. When the interlingua concept does not exactly match the concept represented by the Chinese word, the <transformation rules> will be used to transform the resulting sub-categorization to the proper form.

For example, some concepts in the ILT, such as *e-make, will have two different interpretations in Chinese depending on the context; one is '做', the other is '制造':

3.a If you make alterations, you must provide adequate lifting devices.

如果你做改变, 你必须提供足够的起重设备.

3.b The conditioner and the antifreeze is not made by this company.

调节剂和反冻结剂不是这公司制造的.

Our generation lexicon will contain all the possible interpretations of the concept 'make', including the following to help the system to make the right selection:

```
(*e-make
  ((is-a (value *mental-action))
   (agent (sem (*or *human *institution)))
   (theme (sem *abstract)))
  ((root "做")
   (cat (value v)) (subcat (value vt)))
  ())
(*e-make
  ((is-a (value *action))
   (agent (sem (*or *human *institution)))
   (theme (sem *merchandise)))
  ((root "制造")
   (cat (value v)) (subcat (value vt))
  ())
 ...
```

When the the system processes the ILT of 3.a, the first lexeme will be selected because 'alternation' is an abstract concept, while the second lexeme matches better with the ILT of 3.b, since the 'conditioner' and the 'antifreeze' are merchandise.

In the lexical entries shown above, no transformation rules are specified, which means that there is no transformation needed for this concept. In some cases, however, transformation is required. In that case, the transformation rule set will provide the necessary information about how to map the concept structure in the ILT into the proper form. The details of this lexical transformation are omitted here.

4.2 Generation Based on Semantic Function

Almost all language generation systems adopt some kind of grammar formalism to fill the gap between their internal meaning representation and the linear structure of their corresponding target language. This is usually done in two steps. First, they convert their internal semantic representation

to some sort of intermediate syntactic structure, like the f-structures typically used in KANT. Then, taking the intermediate syntactic structure as input, a syntactic generation procedure (perhaps based on a unification grammar) is used generate the sentence of a target language.

Essentially, the principle of this formalism of generation is to decide the strict positions of constituents in the surface string via their grammatical functions. But whether the grammar functions in the language can uniquely decide the position of the constituents in the surface string is not easy to answer, because the answer usually varies form language to language. Even though the answer may be yes, we still have the problem of how to find ways of mapping the internal semantic representation into the intermediate structure, say f-structure, with high accuracy.

The semantic-based generation scheme presented in this paper is different. We generate sentences form the semantic internal representation, the ILT, directly. The basic idea behind the technique is to combine the two steps mentioned above into one. The generation procedure takes the ILT as the input and generates a Chinese sentence in one step. The principle of this technique is to decide the positions of constituents in the surface string not by their syntactic functions, but by their semantic functions.

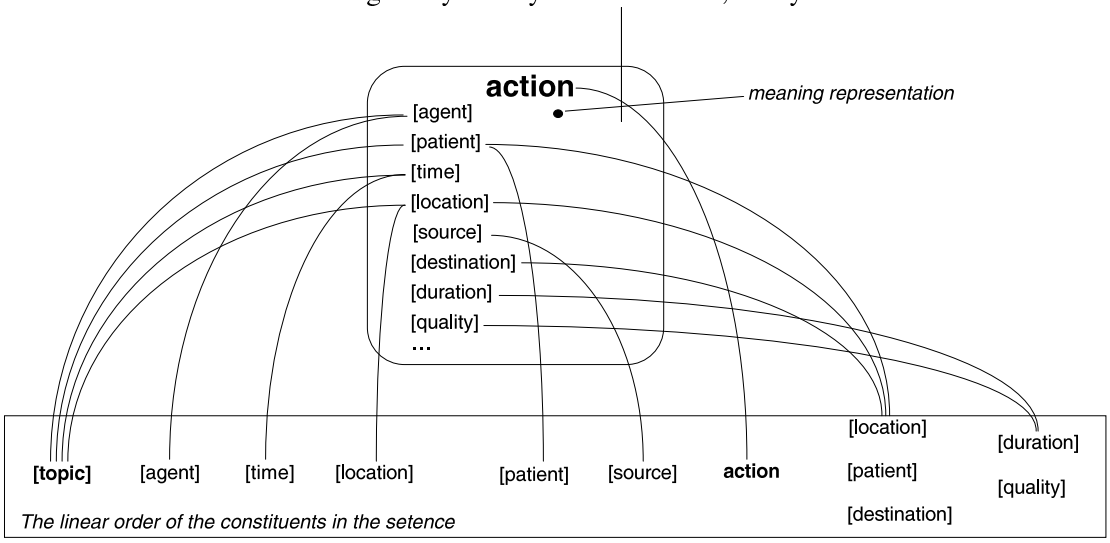

Figure 1. Semantics-based generation.

The implementation of the technique is based on a grammar formalism called Pseudo Unification Grammar, as implemented in GENKIT (Tomita and Nyberg, 1988), which is the same formalism used in the Generalized LR Parser (Tomita and Carbonell, 1987). The Pseudo Unification Grammar formalism resembles that of PATR-II. Although it is originally motivated by and designed for syntactic analysis and generation based on unification, its notation and the control structure are general enough and powerful enough to be adopted in semantics-based generation. The generation rules are designed to process a ILT by decomposing it into smaller ILT pieces, such that each substructure corresponds to a particular phrase-structure category in the target language.

Each rule of the formalism consists of a context-free phrase structure description and a cluster of pseudo equations. The non-terminals in the phrase structure part of the rule are referenced in the constraint equations as x0 ... xn, where x0 is the non-terminal in the left hand side and xn is the nth non-terminal in the right hand side. Originally, the pseudo equations are used to check certain attribute value, such as verb form , person and number agreement, and to construct or disassemble f-structure. Now it is also used to check semantic attribute values and to disassemble ILT instead of f-structure. Following is a sample rule of our generation system, which seems very similar to a rule in a syntactic-based generation system.

```
(<dec-act-s> --> (<np> <vp>)
   (((x0 mood) =c dec)
    ((x0 subcat) =c vt)
    ((x0 passive) = *UNDEFINED*)
```

```
((x0 agent) = *DEFINED*)
(x1 == (x0 agent))
(x2 = x0)))
```

When the rule is activated, the <dec-act-s>, or x0, will contain the ILT to be realized as a declarative active sentence. The constraint equations check the mood of the ILT, the category of the verb chosen, the voice of the ILT. Then they move the agent from the ILT to x1, to realize it as an <np>, and put the remaining part of the ILT into x2, to realize it as a <vp>. The order of the elements in the right-hand side of the context-free phrase structure description of the rule decides the order of the two constituents in the linear structure of the sentence to be generated.

All the generation rules are written in this unification-based style. The generation rules, then, serve as the knowledge sources for language generation. GENKIT will compile them into LISP code which can be further compiled into binary code for run-time use.

4.3 Control Structure

The basic control structure implemented in our generator is Top-down and Left-right. When GENKIT compiles the generation knowledge base, that is, the rule set, it compiles all the rules that have identical left-hand sides in their context free phrase structure rule part into a single Lisp function. Therefore, the generator consists of different functions named according to different left-hand sides of the context free rule part of the rules which appear in the rule set. The GENKIT compiles the rules in a particular way, such that when the function is invoked the evaluation of these rules in the function is ordered according to the order of the rules in the original rule set. When generating a sentence, the generator takes an ILT to be realized as its input and fire the appropriate rules to generate a sentence based on the order in the grammar. A higher-level rule usually disassembles the ILT and calls lower-level rules in the order in which they occur in the right-hand side of its context free phrase structure rule part, producing a string and returning that string to the calling rule as a substring to be assembled; otherwise a failure will be reported and the next rule will be tried. The generator continues generation until a rule with the highest non-terminal in its hrase structure left-hand side (e.g., <START>) succeeds. In that case, a sentence will be produced and the generator will stop.

Obviously, the goal of grammar design is to achieve the widest grammatical coverage possible. Keeping this goal and GENKIT evaluation order in mind, we must order ensure that the rules in the grammar are ordered properly. The principle governing rule ordering is to consume as much as possible of the input ILT. To accomplish this, there are several strategies whch can be used.

The first strategy is to consider the most complex case first. By 'complex case' we mean the case in which more constituent structures are involved. For example, a complex sentence is more complex than a simple sentence, because a complex sentence has more roles to be realized. Hence when we generate a sentence, we must check if it is a complex sentence or not and place such rules before the rules which deal with only simple sentences. Using recursive rule definition, it is very easy to do. The following sample rules demonstrate this principle:

```
(<s> --> (<reason-event> <s>)
    (((x0 reason) = *DEFINED*)
     (x1 == (x0 reason))
     (x2 = x0)))

(<s> –> (<when-event> <s>)
    (((x0 when-event) = *DEFINED*)
     (x1 == (x0 when-event))
     (x2 = x0)))
...
(<s> –> ... )  ;;; simple sentence.
```

The second strategy is to consider the most specific case first, and then a more general one. The

specific case is not a case which has more constituent structures, but a case which needs a special treatment. For example, in the category of declarative active sentences there is a group of sentences which contain the main verb "BE" and which need special treatment in generation of Chinese. Therefore, in the design of the rule set to deal with declarative active sentences, the rules which account for such cases must be placed before others.

Thirdly, differentiating different cases to divide them into different groups is a good way to promote processing efficiency. Recall that when GENKIT compiles the rule set, it compiles all the rules that have identical left-hand sides in their context free phrase structure rule part into single Lisp function. Introducing a new nonterminal symbol for a subset of rules which have similar features and need special treatment will result in abstracting a subset of such rules from a larger, more general set. This helps to narrow down the search space dramatically during rule search. It also improves the modularity of rule design and makes incremental modification easier. The following samples show the combination of the last two strategies. The first rule is more specific than its following rules, so it is placed before other general rules; it also introduces a new rule name called <predicate-s>, which results in abstracting these specific cases from the rule set name <dec-act-s> and grouping them into a new subset of rules named <predicate-s>:

```
(<dec-act-s> --> (<predicate-s>)
    ((*EOR* (((x0 predicate) = *DEFINED*))
            (((x0 predicated-of-theme) = *DEFINED*)))
     (x1 = x0)))

(<dec-act-s> -->...)  ;;; More general declarative sentence.
 ...
(<predicates-s> --> ...)
 ...
(<predicates-s> --> ...)
```

In the terms used in rule-based systems design, these strategies are called complexity ordering, specificity ordering and rule grouping respectively.

5. Dealing with Some of the Important Issues in Chinese Sentence Generation

In this section we will explain some of the crucial problems in Chinese sentence generation and show how they can be treated fairly well in the framework of semantics-based generation. These problems have great influence not only on the quality of the translation but also on the acceptability of the sentence generated. These problems are (1) the position of the locative and directional phrases; (2) the position of the direct object; (3) the treatment of passive sentences.

5.1 Position of Locative and Directional Phrases

The locative and directional phrases in Chinese correspond to a large subset of prepositional phrases in English. Sentential prepositional phrases in English are ordinarily located in clause-final position. The locative and directional phrase in Chinese, however, may occur in either the pre-verbal or post-verbal position. The position of these phrases is decided both by the semantic functions of these phrases and the meaning of the main verb in the sentence. The different positions usually convey different meanings. In order to generate a correct sentence, one of the important tasks in sentence generation is to position these phrases properly.

The structure of the locative and directional phrase are similar. Both consist of a particle, a noun phrase specifying the location concerned, and an optional locative particle specifying a spatial relation, such as 中(inside) , 上(on, above), 外(outside), 下(under, below) and so on.

Locative Phrase:	Directional Phrase:
在 <noun phrase> [<locative particle>]	从 <noun phrase> [<locative particle>]
. at	from

到 <noun phrase> [<locative particle>]
to

4a. 他们 在 房间 里 读 书。
they at room inside read book
They are reading in the room.

4b. 我 想 从 匹兹堡 开 车 到 旧金山。
I went from Pittsburgh drive vehicle to San Francisco.
I am going to drive from Pittsburgh to San Francisco.

Under the framework of semantic-based generation, the directional phrase can be dealt with rather straightforwardly. First, directional phrases specify the source or the destination of an action and very well match with the source thematic role and the goal or destination thematic role in the ILT accordingly. Secondly, the position of directive phrases is rather fixed, with the source directive phrase always positioned before the verb and the destination directive phrase after it. See example 4b and 5:

5. Fuel should be drawn from the tank that is closest to the engine.

燃料应当从最靠近发动机的箱抽取.

The position of locative phrases, however, is much more difficult to decide. In Chinese the locative phrase may occur in either the pre-verbal position or the post-verbal position. In the pre-verbal position, it has a general locational meaning which specifies the general location at which the event or state occurs. Since nearly any event or state can have a location, most verbs allow a pre-verbal locative phrase. Post-verbal locative phrases, on the other hand, address the locational consequences derived from the action or prolonged effects of the state specified by a verb, and are restricted to certain types of verbs. The single factor that determines whether the locative phrase occurs before or after the verb is the meaning of the verb.

In fact, there are few semantic classes of verbs which require post-verbal locative phrase: the so called verbs of posture, such as 站(stand), 停(park, stop), 住(have residence) and some others; a verb so called verbs of appearing, such as 发生(happen, occur), 产生(occur), 形成(form), etc.; the so called verbs of placement, such as 放(put), 写(write), 藏(hide), 停(park as a transitive verb), and so on.

The solution for this problem in our system is to appeal to semantic distinctions between verbs. In our concept lexicon, the action concepts are divided into several subclasses accordingly, so that these distinctions can be used in sentence generation. For example, in generation of a verb phrase, we can fire a rule like the following before the theme or the patient of the verb phrase is generated, to deal with this issue:.

```
(<vp> --> (<vp> <pp>)
    (((x0 location) = *defined*)
     (*EOR* (((x0 is-a) = *posture-event))
            (((x0 is-a) = *appearing-event))
            (((x0 is-a) = *placemant-event)))
     (x2 == (x0 location))
     (x1 = x0)
     ((x1 complement-generated) = +)))
```

As for the examples, please refer to 1.a and 1.b in section 2.

The post-verbal locative phrase has such a great semantic 'intimacy' with the verb that there must be no intervening element, not even the object noun in a verb object compound, between the verb and the post-verbal locative phrase. Therefore, if a post-verbal locative phrase has been generated, the object, if any, should be positioned before verb using the 把 construction (explained in following section). The last equation in the rule is used to signal this fact, and to guide further generation of the verb phrase accordingly.

5.2 Positioning the Direct Object

We have shown that there are several possible positions for the direct object of a sentence in Chinese. Its most natural position is the one immediately after the verb. Depending on the context, however, it can occur in other positions as well. It can be the topic and be positioned at the sentence-initial position, whether or not the subject is expressed; it can occur before the verb with or without the particle 把.

We will talk about some other uses of the topic structure in the following subsection. The SOV form is typically used in dialog to express a meaning which is contrary to the expectation held by the listener. In text generation, it is hardly used. The remaining problem, then, is how to choose between the remaining two forms correctly: the form in which the direct object occurs after verb and the form in which the direct object occurs before verb with particle 把. Since these two forms are used most often in Chinese and they are not interchangeable in some situations, correct treatment of this distinction is important for accurate translation.

In order to determine the correct structure, we have to make clear what characteristics these structures have and what communicative functions they serve. First, it is not too difficult to find in Chinese that the noun phrases after the verb usually are indefinite, and are used to introduce a new thing or a new concept which has not been talked about in the current context, while noun phrases following 把 and appearing before the verb are generally definite or generic, and are understood to refer to something about which the speaker believes the hearer knows. On the other hand, some Chinese grammarians provide an essential clue in characterizing the 把 construction as the 'disposal' form (Li, 1974; Wang, 1947). It is noted that sentences using the 把 construction have a meaning of disposal. Roughly, disposal sentences have to do with addressing how do deal with and what happens to the direct object. This characterization of the 把 construction nicely explains many phenomenona related to the 把 construction and provides further evidence that Chinese grammar is very much semantically driven.

But the question of how to decide whether a sentence has the disposal meaning must be answered. Even though we can group all the verbs into different classes according to whether they have the disposal meaning or not, the characterization of sentences is not so clear cut, because some sentences still have the disposal meaning even though they contain a main verb not associated with the disposal meaning. Consider the following example:

6.a 你必须停车. (you must stop a vehicle)

6.b*你必须把车停. (you must stop the vehicle)

6.c 你必须把车停住. (you must stop the vehicle completely)

6.d 你必须把车停在水平表面上. (you must stop the vehicle on a level surface)

Since 停 is a verb which lacks the disposal meaning, sentence 6.b is not grammatical, while sentence 6.a is. When the verb becomes 停住 in sentence 6.c, and when there is a post-verbal locative phrase as in sentence 6.d, the whole sentence conveys a disposal meaning, and the 把 construction should be used.

Obviously, the method of simply dividing verbs into different semantic classes is not feasible. However, there do exist some clues which tell us when the disposal meaning appears in the sentence and a 把 construction must be used in the sentence generation:

1). When a verb having absolute disposal meaning such as 最小化(minimize), 最优化(optimize) :

7. When you are using No. 2 diesel fuel in cold weather, the following additional devices can minimize starting problems and fuel problems:

当你在冷天气正在使用2#柴油燃料时, 下述附加设备能把起动的问题和燃料

问题最小化:

2). When there is a post-verbal directive or locative phrase in the sentence, as shown in previous subsection, and the patient is definite or generic:

8. In order to prevent engine damage, do not add coolant to an overheated engine.

为了防止发动机损坏, 不要把冷却剂添加到过热的发动机.

3). When there is a post-verbal resultive adverb or complex stative clause and the patient is definite or generic:

9. Non-thermostatically-controlled fuel heaters can heat the fuel above 149 *F (65 *c)

非温度自动控制的燃料加热器能把燃料加热到高于149华氏度(65摄氏度).

Like the situation we mentioned in the previous subsection, these kinds of heuristics can be easily implemented in our generaiton rules by checking certain semantic features of the main verb and its arguments, and/or the presence or absence of certain semantic constituents.

5.3 Treatment of Passive

The structures of the passive sentence in Chinese are as follows:

<noun phrase> 被 <noun phrase> <verb>

<noun phrase> 被 <verb>

Similar to English, the noun phrase in sentence-initial position refers to the patient of the action, while the noun phrase after coverb 把 refers to the agent.

The passive form is widely used in English and other Indo-European languages. On the contrary, the use of passive form in Chinese is very limited. There are several semantic reasons for this. First, the 被 passive is used essentially to express an adverse situation, in which something unfortunate has happened to the patient. In addition to adversity, the 被 construction also expresses disposal in the same manner as the 把 construction. In other words, the 被 construction describes an event in which a patient is dealt with, or manipulated in some way. Due to the influence of Indo-European languages, the number of 被 constructions that are not used to express adverse situations is increasing in modern Chinese. This loosens the standard of acceptability of the 被 construction to some extent, but overall the usage of the 被 construction is much rarer than that of the passive in English.

Another observation is that the English passive usually does not correspond to the 被 construction in Chinese. Rather it usually corresponds to the topic-comment structure in Chinese. In generation of ILT containing passive voice, we can simply put the patient role in the sentence-initial position as the topic of the sentence, and generate the other constituents as usual in most cases; e.g.:

10. Some objects should be removed with lifting fixtures.

一些物体应当用起吊固定装置移走.

11. This water and sediment should be drained at each oil change.

这水和沉淀物应当在每次换油时排出.

12. If a filter is mounted outside the frame rails or in any location that is exposed to wind, there will be persistent problems in cold weather.

如果过滤器安装在框架尾部外面或在暴露在风的任何位置, 在冷天气下将有持续的问题.

Some other situations in which a passive in the ILT should not be translated to the 被 construction are when the passive verb is really used as a predicate, or when the focus of the sentence is on the agent of the action verb. Some passive formations can be represented directly as predicative adjectives; e.g., 有限(limited), 推荐 (recommended), 要求(required), 适合(suited), etc. The second situation happens when the agent of the action verb is present and the sentence is very short. In these situations, the "是....的" constructionsshould be used. The following are two examples:

13. Quantities of No. 1 diesel fuel are limited.

1#柴油燃料的量是限制的.

14. The conditioner and the antifreeze in the cooling system were not made by this company.

调节剂和冷却系统的反冻结剂不是这个公司制造的.

Considering all the factors mentioned above, our system implements several heuristics to deal with ILTs with passive mood. They can be summarized as follows; we have found that they work well in the majority of instances of the passive, especially in technical text:

1). If the patient is an animal, or human being, then use a passive form. Since most actions are performed by animalsor humans, non-passive translation may cause confusion.

2) If the verb has an obvious adverse meaning, then use the appropriate passive form, such as 杀死 (kill), 破坏 (destroy) , etc.

3). If the verb can be used as a predicate, or the agent of the main verb is specified and the ILT is simple (only agent, patient are specified), then use the "是....的" construction.

4). Otherwise, use the topic-comment construction.

6. Implementation

The technique presented in this paper has been implemented in our prototype Chinese sentence generation system, whose generation knowledge base contains about two hundred rules for sentence generation and about one hundred and fifty rules for noun phrase generation. Since Chinese is an isolating language to some extent, the system does not need a separate rule set for word inflection. All the knowledge about particle or coverb selection is contained in the rule sets for noun phrase and sentence. Also, the particle and coverb selections are dealt with on a semantic basis. The primary goal of the system development was to investigate the feasibility of generating Chinese sentences using the interlingua text generated from English, already in existance and previously used for generation of other languages. Since Chinese has a lot of distinguishing features, we are also interested in developing some techniques especially for Chinese. We evaluated the system using a corpus of 215 sentences and phrases from heavy machinery technical documentation, some of which are complex sentences. The results were reviewed by an expert in the field, and all of the corpus examples got very good translations. All the example translations in this paper are taken from those sample sentences.

The concept lexicon and Chinese lexicon are written as frames, and generation rules are written in the pseudo-unification formalism used by GENKIT. All the knowledge sources are compiled into Lisp code which can be further compiled into binary code. The system can run on any machines which support Common Lisp. Our most recent system is running on a SUN Sparc II and takes 1 - 2 seconds

for each sentence.

7. Discussion and Further Work

As mentioned above, there are two main features of the technique presented in the paper: generating sentences or phrases of Chinese from a semantic representation encoded in a concept frame structure; and mapping the semantic structures to sentences or phrases mainly in a single step, based on the relationships between concept elements and the corresponding variations in surface order.

Most people believe that the conceptual structures of different language speakers are similar, although their language grammars differ greatly. There different preferences, though, concerning what kind of representation scheme is best for semantic representation. Some researchers are in favor of non-hierarchical semantic representations, such as conceptual graphs (Nicolas, et al., 1995) more so than hierarchical ones, such as tree-like representations. Their arguments are that dominance relations between nodes in semantics often stem from language-specific considerations, and are not always preserved across languages, and if the semantic input comes from other applications, the dominance relations may not be explicitly available. It is probably true that dominance relations may not be explicitly available in some cases, but if we agree that concept structure always reflects the relationships between concepts and inherently contains dominance relations, then these relationships can be inferred during generation.

According to our observation, the concept dominance structures inferred from the source language are very useful in target language generation. In some cases, however, there do exist some differences in the way in which different language speakers view things or express their ideas, and those differences may be reflected in the interlingua and cannot be easily preserved across languages. Passive voice is a good example of a construction which is not generally preserved across languages unless they are very simlar in structure. Negation of objects in English is another example, which never happens in Chinese and whose semantic content is expressed differently. However, as we have shown, it is possible to write semantic grammar rules which create the appropriate target language structures from Interlingua frames representing English source text with a very different grammatical structure.

Mapping the semantic structures to sentences or phrases directly enables us concentrate on the relationships between concept elements and their order of realization in the utterance. This eliminates the need to transform the concept structure into a syntactic f-structure as an intermediate step. This method is especially suitable for generation of languages like Chinese, whose basic word orders are mainly determined by semantic rather than grammatical considerations.

Our continuing research direction is to enlarge the rule set to deal with more semantic phenomena according to the KANT interlingua specification, so that the Chinese generation system is ready to be used in a practical machine translation domain. Another research direction is to incorporate more semantic inference mechanisms for implicit knowledge, such as the 'disposability' and 'adversity' mentioned above, to further improve the quality of generation.

8. Conclusion

We have presented a technique for top-down case-frame based generation using a semantic unification grammar, in which target language is generated directly from the interlingua without converting it to an intermediate grammar structure. Under this framework, we can consider natural language generation in terms of mappings between meaning and and its linear realization, and many crucial problems in Chinese generation can be treated naturally and conveniently. Although the method is motivated by Chinese generation, it is general enough to be used in knowledge-based machine translation systems for generating any language in which semantic factors play a major role in surface realization.

References

Goodman, K. and S. Nirenburg (eds) (1991). A Case Study in Knowledge-based Machine Translation, San Mateo, CA: Morgan Kaufmann. pp. 231-261

Leavitt, J., D. Lonsdale and A. Franz (1994). "A Reasoned Interlingua for Knowledge-Based Machine Translation," Proceedings of CSCSI-94.

Li, C. and S. Thompson (1982). Mandarin Chinese: A functional reference grammar, University of California Press, Los Angeles, CA

Li, Yingche (1974). "What Does 'Disposal' Mean? Features of the Verb and Noun in Chinese," JCL 22:200-218

Mitamura, T., E. Nyberg and J. Carbonell (1991). "An Efficient Interlingua Translation System for Multilingual Document Production," Proceedings of the Third Machine Translation Summit, Washington, D.C.

Nicolas, N., C. Mellish, G. Ritchie (1995). Sentence Generation from Conceptual Graphs, in Conceptual Structures: Applications, Implementation and Theory, pp. 74-88

Nirenburg, S., V. Lesser and E. Nyberg (1989). "Controlling a Language Generation Planner," Proceedings of the 1989 International Joint Conference on Artificial Intelligence, Detroit, MI, August

Nyberg, E., R. McCardell, D. Gates and S. Nirenburg (1989). "Generation," Computers and Translation: 4

Shapiro, S. (1982). Generalized Augmented Transition Network Grammars For Generation From Semantic Networks, American Journal of Computational Linguistics, Volume 8, Number 1, pp. 12-25

Tomita, M., and E. Nyberg (1988). Generation Kit and Transformation Kit, Version 3.2 User's Manual, CMU-CMT-88-MEMO

Tomita, M. and J. Carbonell (1987). The Universal Parser Architecture for Knowledge-Based Machine Translation, in Proceedings of 10th International Joint Conference on Artificial Intelligence, Milano, 1987

Wang, Li (1947). 中国现代语法 (Modern Chinese grammar), Shanghai, Zhonghua shuju.