Chinese Generation System

---System design and implementation

Student: Wei Fang

Last semester I analysed the CGS system which was implemented by Dr. Li.
This semester, I continued my research in the Chinese Generation System (from interlingua to Chinese), and tried to develop it in a more systematic way.
As Eric suggested, first I did the Chinese generation for NP phrases. I selected about 60 representative NP phrases from the Catalyst domain, and implemented the Chinese generation for them. I should mention here, these interlingua is very different from what Dr. Li did two years ago. The result is good, and this approach helps me a lot to understand the interlingua better and more systematically. Then I started doing Chinese generation for the CNBC domain. Although the interlingua in this domain is different from the one I used to do NP generation, it is much easier for me to do the whole sentence generation after I get familiar with the NP generation.

In this report, I will explain what I did in the NP generation in Catalyst domain, and then the whole sentence generation in CNBC domain.

Section 1. Chinese Generation for NP phrases

The architecture I used here is very similar to Dr. Li's.

```
generator.lisp  ---  load everything we need :  load kant
                                                compile the grammar file
                                                load newgen-sys.lisp file

newgen-sys.lisp ---  build hashtable for lexicon (newlexicon.chinese)
                                    and interlingua (newworking.ir)
                     get Fstructure from the interlingua
                     use generator function to do generation

newworking.ir   ---  the interlingua file

newlexicon.chinese ---  mapping from interlingua lexicon to Chinese
                        for countable noun, we specify its unit
                        for adjective, we specify a feature no-de
```

In English, we say "two children", "three tables"... But it is very different in Chinese. We still need a unit for the countable nouns. For example: we say 两(two) 个(unit) 孩子(children) in stead of :
两(two) 孩子(children). The units for different countable

nouns are different. The most common unit is 个.

For adjective, we also treat them in different ways. Usually when adjective modify a noun, we put a special character 的 between them, so we get: adjective 的(de) noun. But some adjectives don't have to use this special character, we can just say: adjective noun. That is why we need some special features in our lexicon.

The system works as follows:

Step 1. Build a *sentence-table* for the interlingua NP phrases, and build a *lexicon* for the mapping from interlingua lexicon to Chinese.

Step 2. Use the sentence number index to get the corresponding interlingua, use function "lexicon-lookup" to translate the interlingua to the Fstructure which replaces the lexicon in interlingua with Chinese lexicon.

Step 3. Use Genkit compiled grammar for generation.

For a noun phrase, a lot of modifiers (function slots or features) can be attached to it, such as:

1. QUANTITY

English : two (QUANTITY) dipsticks (NP)
Chinese : 2 (two) 个(unit) 油面杆(dipsticks)

As I mentioned before, we need a unit word to put the quantity in front of the noun phrase.

2. DISTANCE

English : this (DISTANCE NEAR) dipstick (NP)
Chinese : 这(this) 油面杆(dipstick)

3. NUMBER

English : these (DISTANCE NEAR) (NUMBER PLURAL) dipsticks (NP)
Chinese : 这些(these) 油面杆(dipsticks)

We must consider the NUMBER feature sometimes to get the right translation.

4. ATTRIBUTE

English : new (ATTRIBUTE) dipstick (NP)
Chinese : 新(new) 油面杆(dipstick)

English : the rear (ATTRIBUTE) sump (NP)
Chinese : 后面(rear) 的(de) 油箱(sump)

Here we noticed: sometimes "de" is used to attach to an adjective in order to modify a noun phrase, but sometimes it is not required.

And also we should consider the DEGREE feature of the adjective.

English : the taller (ATTRIBUTE)(DEGREE COMPARATIVE)(COMPARISON MORE) truck (NP)
Chinese : 较(DEGREE COMPARATIVE)(COMPARISON MORE) 高(tall) 的(de) 卡车(truck)

English : the tallest (ATTRIBUTE)(DEGREE SUPERLATIVE)(COMPARISON MORE) truck (NP)
Chinese : 最(DEGREE SUPERLATIVE)(COMPARISON MORE) 高(tall) 的(de) 卡车(truck)

5. Q-MODIFIER (PP)

English : the number (NP) of turns (Q-MODIFIER)
Chinese : 旋转(turn) 的(de) 数目(number)

Here we don't need to translate the prep (such as "of"), but under some other circumstances we must do the translation for the prep, otherwise the translation is not complete, and even not understandable.

English : dipsticks (NP) with (prep) Readings (NP)
Chinese : 带(with) 读数(reading) 的(de) 油面杆(dipstick)

6. SUCCESSOR

English : another (SUCCESSOR +) engine (NP)
Chinese : 另一(another) 发动机(engine)

7. POSSESSOR

English : your (POSSESSOR) engine (NP)
Chinese : 你(you) 的(de) 发动机(engine)

We need a word 的(de) to put the POSSESSOR in front of the noun phrase.

8. CONJUNCTS

English : the rear sump (NP) and (conj) the front sump (NP)
Chinese : 前油箱(the front sump) 和(conj) 后面的油箱(the rear sump)

9. ORDINAL-ATTRIBUTE

English : the first (ORDINAL-ATTRIBUTE) oil change (NP)
Chinese : 第 1 (first) 次(unit) 换油(oil change)

10. NUMBERED-NUMBER-IDENTIFIER

English : No. 1 (NUMBERED-NUMBER-IDENTIFIER) diesel fuel (NP)
Chinese : 1#(No. 1) 柴油燃料(diesel fuel)

11. SUBSTANCE (*G-PARTITIVE)

English : one (QUANTITY) of (*G-PARTITIVE) the engines (SUBSTANCE)
Chinese : 发动机(engine) 中 的(de) 1 个(one)

12. QUANTIFIER

English : all (QUANTIFIER) engines (NP)
Chinese : 全部(all) 发动机(engine)

We have a lot of QUANTIFIER, such as:

none (QUANTIFIER) of the engine (NP)
a few (QUANTIFIER) hundred (QUANTITY) trucks (NP)
several (QUANTIFIER) trucks (NP)
a little (QUANTIFIER) oil (NP)
any (QUANTIFIER) water (NP)
both (QUANTIFIER) engines (NP)
each (QUANTIFIER) engine (NP)
either (QUANTIFIER) engine (NP)
every (QUANTIFIER) engine (NP)
less (QUANTIFIER) engines (NP)
most (QUANTIFIER) engines (NP)
neither (QUANTIFIER) engine (NP)

... and so on.

13. QUOTE-TAG

English : "FULL" (QUOTE-TAG) mark (NP)
Chinese : "FULL" ("full") 标志(mark)

Apart from dealing with these function slots and features, we should also consider the order of translation when several of them appear together. For instance, when ATTRIBUTE and POSSESSOR appear together, how to put them in some order in the Chinese generation is a big concern. Sometimes a predefined order doesn't work, so we must deal with it in the grammar very carefully.

In general, a good start for NP generation will give us a base for whole sentence generation. In the NP generation, I didn't deal with relative clause. But as whole sentence is generated, relative clause is easy to handle. So my next step is doing Chinese generation for the whole sentences.

Section 2. Chinese Genartion in CNBC domain

The system architecture I used here is the same as in Section 1.

```
generator.lisp  ---  load everything we need :  load kant
                                                compile the grammar file
                                                load newgen-sys.lisp file

cnbc-sys.lisp   ---  build hashtable for lexicon (cnbclexicon.chinese)
                                    and interlingua (cnbcworking.ir)
                     get Fstructure from the interlingua
                     use generator function to do generation

cnbcworking.ir  ---  the interlingua file

cnbclexicon.chinese  ---  mapping from interlingua lexicon to Chinese
                          for countable noun, we specify its unit
                          for adjective, we specify a feature NO-DE
                          use feature SUBCAT to classify the lexicons which are
                          under the same category according to their
                          characteristics
```

In the whole sentence generation, we must consider a lot of details in order to get a good translation. But most of all is the sentence structure we should handle. Since the sentence head is the verb, and we have a feature ARGUMENT-CLASS for the verb, we can use Mapper in the Genkit to map the semantic slots to the syntactic slots. I haven't done this yet, what I did is doing the mapping in the grammar file. In this summer, I plan to use the Mapper in Genkit to get the Fstructure from the Interlingua.

Here is what I did for CNBC domain so far:

1. I translate proper noun as the same in English.

English : TERRY
Chinese : terry

2. DISCOURSE

Put the discourse part in front of the sentence.

English : AND (DISCOURSE) I AM SUSIE GHARIB

Chinese : 还有(and) 我(I) 是(am) susie harib

3. Q-MODIFIER

This Q-MODIFIER is not for NP, but for the whole sentence, or the verb, so I treat it in a different way.

English : ON WALL STREET(Q-MODIFIER)THE DOW INDUSTRIALS MISSED "RECORD NUMBER NINE".

Chinese : 在华尔街上(on wall street), ...

4. MANNER

Put it right in front of the verb.

English : LET'S RUN DOWN THE FINAL NUMBERS NOW(MANNER) FROM WALL STREET.
Chinese : ... 现在(now) 过一遍(run down) ...

5. (FORM IMPERATIVE)

English : LET(FORM IMPERATIVE)'S RUN DOWN THE FINAL NUMBERS NOW(MANNER) FROM WALL STREET.
Chinese : 让(let) ...

6. (TENSE FUTURE)

In Chinese, we have a special character for tense. For example, we use 将 for future tense, use 了 for past tense, and use 已经 for perfect tense ...

For future tense, we should put 将 right before the verb.

English : TOSCO WILL(TENSE FUTURE) BECOME THE NATION'S LARGEST INDEPENDENT REFINERY.

Chinese : TOSCO 将(will) 成为(become) ...

7. For those features and function slots which modify NP, I did it the same way as I mentioned in Section 1.

But as we deal with the whole sentence structures, we must consider the co-effects of different words, this is the most difficult part in whole sentence generation. In this summer, I will improve the current version, and systematically analyse the interlingua phenomena in the CNBC domain. The way it works is: we must consider some specific features such as: REFERENCE, NUMBER ... to find some regularity which can achieve a good performance for the whole sentence generation.

So my plan for this summer is:

1. Use Mapper in Genkit to get the Fstructure from the Interlingua.

2. Do the Chinese Generation for whole sentences in CNBC domain.