The Whole Sentence Chinese Generation in CNBC domain

Student: Wei Fang

After I did above 10 sentences Chinese Generation in the CNBC domain last semester, this summer I continued my research in this field, finished about 50 sentences' Chinese generation, and have made several progresses. In this report, I am going to illustrate the approachs I used to do the whole sentence Chinese generation, and I also would like to discuss some remaining or hard to resolve issues.

Section 1. System Architecture

The system architecture I used here is as follows:

```
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
                          1. for countable noun, we specify its unit
                          2. for adjective, we specify a feature NO-DE
                          3. use feature SUBCAT to classify the lexicons
                             which are under the same category according
                             to their characteristics
                           ** see as follows
```

Section 2. System implementation

一. PP attachment

First of all I am going to discuss the lexicon I improved in this summer. For PP attachment, there are a lot of issues we should concern:

1. The order

"XP PP" (XP could be NP, VP) is always right in English, but it is not the case in Chinese. There might be more situations we should handle, but in these 50 sentences I found several different circumstances we should consider.

a. VP PP (English) --> PP VP (Chinese)

THEY CLOSED (VP) AT 59 7/8 (PP). -->
他们 在59 7／8(PP) 结束了 (VP).

b. VP PP (English) --> VP PP (Chinese)

LOCTITE SAYS IT IS LOOKING AROUND (VP) FOR OTHER BUYERS (PP). -->
LOCTITE说它[这]正在 四处找寻(VP) 另外买家(PP).

In this cases, "look around for sth." is a verb phrase, it doesn't make sense if we seperate the "for sth" apart from the verb.

UNOCAL SAYS IT WILL USE SOME OF THE PROCEEDS (VP) TO PARE DOWN DEBT (PP).
UNOCAL说它[这]将 使用一些的收入(VP) 来(TO) 缩减债务(NP).

Here PP is used to express the goal of the VP, so it should follow VP in Chinese.

c. NP PP (English) --> PP de NP (Chinese)

AND BILLIONAIRE MARVIN DAVIS HAS SWEETENED HIS TAKEOVER OFFER (VP) FOR CARTER-WALLACE (PP).
而且亿万富翁的marvin davis已经更加优惠 给予CARTER-WALLACE(PP) 的(de) 他的接管提供(NP).

d. NP PP (English) --> NP de PP (Chinese)

TEXT "IT WAS A SCHIZOPHRENIC KIND (NP) OF DAY (PP) OF TRADING ALL DAY LONG:
它[这]一整天是SCHIZOPHRENIC类型(NP) 的 (de) 交易的日子(PP):

2. The mutual impacts

When PP is attached to NP and VP, they have some impacts on each other, and sometimes, these impacts are significant. In order to get the understandable and accurate translation, we need to consider the mutual effects.

First, for a specific HEAD (NP or VP), when different preposition is attached to it, the translation of the prep is different. We do have the translation for a specific preposition, sometimes it is determined by its HEAD, sometimes it should consider the whole sentence. I deal with this issue in the lexicon.

Here are some examples:

A. The impact of HEAD on Preposition

a. (*A-CLOSE (CAT V) (ROOT "结束") (WITH ((ROOT "以"))))

SHARES OF CARTER-WALLACE CLOSING WITH A GAIN OF 1 1/4 AT 16 DOLLARS A SHARE.
CARTER-WALLACE的股份 以(with) 1 1／4的赢利在每股份16美元结束 .

If the current PP head matches one of the PP definitions in the HEAD, we use the defined translation, otherwise we use some default translation for the preposition.

b. (*A-BUY (CAT V) (ROOT "买下") (FOR ((ROOT "以"))))

MATTEL BUYING TYCO TOYS FOR $755 MILLION.
MATTEL 以(for) 755million美元买下TYCO TOYS .

Usually "for" is not translated as "以" in Chinese, but if it is attached to "buy", it has a special meaning.

c. (*A-TIE (CAT V) (ROOT "系紧") (ba +) (WITH ((root "与") (ba +))))

THE NATION'S NUMBER-ONE TOYMAKER, KNOWN FOR ITS BARBIE DOLLS, WILL TIE THE MERGER KNOT WITH THE COMPANY FAMOUS FOR ITS MATCHBOX CARS.
以它[这]的巴比木偶闻名的这国家的第一位的玩具制造商 把(ba) 与(with) 以它[这]的火柴盒轿车著称的公司的合并扣将 系紧(tie) .

In this case it is more complicated. "tie A with B" should be translated in Chinese in this way:

把 A 与 B 系紧

"With" has a lot of translation in Chinese, so its meaning is determined by the Verb or Noun it is attached to.

d. (*O-TAKEOVER-OFFER (CAT N) (ROOT "接管提供") (FOR ((ROOT "给予"))))

AND BILLIONAIRE MARVIN DAVIS HAS SWEETENED HIS TAKEOVER OFFER FOR CARTER-WALLACE.
而且亿万富翁的marvin davis已经更加优惠 给予(for) CARTER-WALLACE的他的 接管提供(takeover offer) .

When preposition is attached to Noun phrase, its translation is changed accordingly.

e. (*O-KIND (CAT N) (ROOT "类型") (OF ((phrase +) (root "*GAP*"))))

IT WAS A SCHIZOPHRENIC KIND OF DAY OF TRADING ALL DAY LONG:
它[这]一整天是SCHIZOPHRENIC 类型(kind) 的 交易的日子(day of trading) :

When the HEAD is "kind", and "of" is attached to it, it always means this is a phrase in Chinese. Its translation of "of" and the order of the translation can be determined in some way. But I am considering this sentence:

I am a kind of sleepy today.

I don't know what the interlingua might be, but the common translation "类型" for "kind" can never appear in this case.

B. The impact of Preposition on HEAD

a. (*O-DAY (CAT N) (ROOT "天") (OF ((root "*GAP*") (headroot "日子"))))

DAY OF TRADING
交易的日子

Usually the translation of "day" is "天", but in some situation this translation is not suitable. In this case, "of" is attached to "day", the presence of the preposition indicate the translation for "day" should be "日子". I believe this kind of situation occurs often in Chinese, but in the current data we only found this example.

二. Other issues

Since I already discussed the NP related generation issues, here I only discuss the whole sentence generation issue.

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 ...)

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) ...

For past tense, we put 了(le) right after the verb.

English : ON WALL STREET THE DOW INDUSTRIALS MISSED "RECORD NUMBER NINE".
Chinese : 道尔工业指数在华尔街上错过 了(le) ″九号记录″ .

For perfect tense, we put 已经(yijin) right before the verb.

English : THE DEFENSE DEPARTMENT HAS CHOSEN BOEING AND LOCKHEED MARTIN TO DEVELOP PROTOTYPES OF A NEW COMBAT AIRCRAFT .
Chinese : 国防部 已经(yijin) 选择BOEING和LOCKHEED-MARTIN来发展新的战斗机的原形 .

7. REL-QUAL: relative clause

English : UNOCAL(NP) ALSO KNOWN AS "76 PRODUCTS" COMPANY(S)

Chinese : 而且作为76个产品公司闻名(S) 的(de) UNOCAL(NP)

We express relative clause in Chinese as the adjective, and it functions as an adjective, so we translate it this way: relative-clause de NP.

8. COMPLEMENT

English : UNOCAL SAYS IT WILL USE SOME OF THE PROCEEDS TO PARE DOWN DEBT.
Chinese : UNOCAL说 它[这]将使用一些的收入 来(to) 缩减债务 .

In this sentence, "it will use ..." is the complement for the whole sentence, and "to pare down debt" is the complement for the subsentence "it will use ...".

9. SUB-QUAL

English : AS CNBC'S TIM TINDALL REPORTS(sub-qual), THE DECISION IS A STUNNING DEFEAT FOR MCDONNELL DOUGLAS AND A SURPRISE WIN FOR BOEING.
Chinese : 据CNBC的tim tindall报道(sub-qual), 这决定是对于MCDONNELL-DOUGLAS的震惊的失败和对于BOEING的惊喜的成功 .

10. TIME-MODIFIER

English : HE UPPED THE ANTE TO 20 DOLLARS A SHARE {JUST TWO WEEKS} (time-modifier) AFTER THE CONSUMER PRODUCTS MAKER TURNED DOWN HIS INITIAL 18 DOLLARS A SHARE BID.
Chinese : 在 仅2星期(time-modifier) 这消费产品制造商拒绝了他的最初的每股份18美元的竞价之后 , 他提价了到每股份20美元 .

If TIME-MODIFIER is to modifier verb or the whole sentence, put it before the verb.

三. The remaining issues:

I use semantic roles to generate the sentence. The feature ARGUMENT-CLASS give us a lot of hints to get the semantic meaning from the interlingua. So I use this feature in the grammar to write the specific rules. But the problem is what's the coverage of these rules? How do I know if these rules can cover the situation for those unseen data? This problem is not only for ARGUMENT-CLASS, since our rules is generalized according to the data, but is it really general enough to predict the unseen data? Some rules are very general rules, but some of them is written to deal with some specific cases in our interlingua. These issues might also be the common issue we discuss in Knowledge based system. In CNBC domain, what's the restricted writing space? I think we definitely can't deal with unrestricted writing using knowledge based approach.