[*MT Summit III*, Washington, DC, USA, July 1991]

Translation Accuracy and Translation Efficiency

Wanying Jin
Computing Research Laboratory
New Mexico State University
P. O. Box 30001/3CRL
Las Cruces, NM 88003
wanying@nmsu.edu

April 22, 1991

Abstract

ULTRA (Universal Language Translator) is a multi-lingua] bidirectional translation system between English, Spanish, German, Japanese and Chinese. It employs an interlingua] structure to translate among these five languages. An interlingual representation is used as a deep structure through which any pair of these languages can be translated in either direction. This paper describes some techniques used in the Chinese system to solve problems in word ordering, language equivalency, Chinese verb constituent and prepositional phrase attachment. By means of these techniques translation quality has been significantly improved. Heuristic search, which results in translation efficiency, is also discussed,

1 Introduction of the ULTRA system:

ULTRA (Universal Language Translator) is a multi-lingual bidirectional machine translation system developed at Computing Research Laboratory, New Mexico State University. It includes English, Spanish, German, Japanese and Chinese. The system employs an interlingual structure, i.e. an intermediate representation (IR) is used as a pivot through which any pair of these languages can be translated in either direction. Each language component works independently to convert its own language string from/into the intermediate representation regardless of the linguistic expressions in any other languages. Symmetric grammar rules are employed in all languages to perform bidirectional translation. The vocabulary consists of 10,000 word senses for each language. In addition, common linguistic information is encoded into the interlingual lexicon specifications shared by all languages.

The intermediate representation (IR) of ULTRA developed by Farwell [Farwell & Wilks, 1990] is a data structure produced dynamically during the parsing of an input string. It contains syntactic, semantic and pragmatic information to represent what is said; how it is said; and why it is said during communication. The intermediate representation is hierarchically organized. IR tokens which refer to definitions in *Longman Dictionary of Contemporary English* [Procter et al, 1978] are encoded for word sense disambiguation. An intermediate representation for a particular sentence:

The printer is placed to the left side of the display.

will be represented as an IR in the following form:

```
[prdctn,[type,indpnt] ,
        [class,dcl],
```

```
        [form,fin],
 [prop, [type , indpnt],
        [class,dcl],
  [pred,[tense,prs],
        [aspect,simp],
        [mood,indic] ,
        [voice,pasv],
        [pol,pos],
  [rel,[type,dyn],
      [s_case,agnt] ,
      [o_case,pat] ,
      [io_case,none],
      [s_class,human],
      [o_class,p_obj] ,
      [io_class,none],
  [r_desc,place2_l]]],
[arg,[g_rel,subj],
      [k_rel,agnt],
      [t_rel,top],
  [ent,[type,nrm] ,
     [class,p_obj] ,
     [agree,ts] ,
     [det,spin],
     [quant,unq],
   [e_desc,printer1_3]]],
  [p_mod,[g_rel,oo],
        [k_rel,loc],
        [t_rel,none] ,
   [ent,[type,nrm],
        [class,p_prp],
        [agree,tu] ,
        [det,spin],
        [quant,unq],
     [e_desc,left_side_x] ,
      [e_mod,[type,oo],
             [class,posr] ,
       [ent,[type,nrm],
            [class,p_obj],
            [agree,ts],
            [det,spin],
            [quant,unq],
        [e_desc,display_x]]]]]]].
```

where the predication (prdctn) at the top level, basically, represents the sentential information and the syntactic structure, the proposition (prop) represents the sentential information and the syntactic structure at the clause level, the predicate (pred) represents all functional information of the verb, the argument (arg) represents the information relating to a subject, direct object or indirect object, accordingly, the p_mod is a proposition modifier and the e_mod is an entity modifier. For detailed explanation see [Farwell & Wilks, 1990 or Jin, 1991].

This paper focuses on the techniques employed in the Chinese system of ULTRA. The system consists of a Chinese grammar with about 200 rules and a Chinese lexicon with 10,000 word senses. The information about semantic preferences, which is common to all languages, is encoded separately and shared by the Chinese component as well as the other language components.

The Chinese grammar covers a range of syntactic patterns. It also incorporates semantic and pragmatic information to cope with sense disambiguation, elided phrases and idiomatic expressions. The Chinese system has the form of a Semantic Definite Clause Grammar [Pereira & Warren 1980], [Huang 1985, 1988] in order to represent its grammar in concise manner. It is symmetric in the sense that both parser and generator share the same set of grammar rules in order to perform bidirectional translation [Jin & Simmons 1986], [Wilks, 1990]. Top-down heuristic search algorithm is employed to achieve translation efficiency,

The Chinese lexicon is specially designed to correlate Chinese word senses with IR tokens. Some information, which appears in the Chinese string but not in the IR structure, (such as noun classifier, 个, 张, 种,... etc.), or appears in the IR structure but is not used in the Chinese string, (such as information about determiners *the, a*), is specially encoded into the Chinese lexicon and used implicitly for the purposes of translation accuracy. Two versions of the Chinese lexicon, Pinyin and Chinese characters, work identically. Thus, Chinese translation in both a Romanized version and with Chinese characters is available.

The focus of this paper is on the discussion of techniques successfully used in the Chinese system to achieve high quality and efficiency in the translation.

2 Translation accuracy:

Language ambiguity occurs in various aspects in natural language processing, i.e. **lexical ambiguity, case ambiguity** and **referential ambiguity**. The following example illustrates these ambiguities:

Please let him know when his book can be published.

Two interpretations are possible:

请告诉他一下什麼时候他的书可以出版.

(Please let him know **at what time** *his book can be, published*).

当他的书可以出版了请告诉他一下.

(**At the time** *that his book can be published, please let him know).*

The word when has different senses. Also, without providing a context the pronouns *him* and *his* are ambiguous as to whether they refer to the same person or different person.

During the design of the Chinese system, several issues were considered in the attempt to arrive at translation accuracy: i.e. word ordering, language equivalency, verb form constituent, and prepositional phrase attachment.

2.1 Word ordering:

- **Clause order in a compound sentence with subordinate clause**: the IR for a compound sentence with subordinate clause has the same representation shown below regardless of the placement of the dependent clause in any languages:

```
[prdctn...
[prop...[conj...], [prop_dpnt...]],
[prop. . .]]
```

In Chinese the position of dependent clauses in a sentence depends on the semantic nature of the conjunction used. Two orders are possible:

```
prdctn::= prop [conj prop_dpnt] |
          [conj  prop_dpnt]  prop
```

In some cases the dependent clause must be placed first, such as:

如果打印机被放在显示器的左边,干扰可能在显示器上发生.

(**If** *the printer is placed to the left side of the display, interference may occur on the display* .)

In other cases the dependent clause must be placed after main clause, such as:

打印机被放在显示器的左边,因此干扰在显示器上发生.

(*The printer is placed to the left side of the display,* **thus***, interference occur on the display.)*

An ordering flag is encoded in the Chinese conjunction lexicon to control the clause order. The lexical entry for a Chinese conjunction has the following form:

```
conj (IR_token,Chinese,Items,Flag)

conj(if_1,如果,prop,first).
```

- **Ordering in a sentence with proposition modifier (p_mod):** The position of proposition modifier in a Chinese sentence also depends on semantics. Three possible orderings for two propositional modifiers which corresponds to the same IR may occur, i.e.

```
[prop,...[pred...],[arg...],
         [p_mod...], [p_mod...]]

prop ::= arg p_mod pred p_mod |
         arg p_mod p_mod pred |
         arg pred p_mod p_mod
```

调试系统现在运行正常.

(*The debugging system* **now** *executes* **normally**.)

图像处理系统在日本飞速发展.

(Graphics processing systems **in Japan rapidly** *develop.)*

这辆赛车跑得又快又稳.

(The racing car runs **fast** *and* **smoothly**.)

These three sentences apply three different rules. The semantic case relation in the p_mod is used to restrict the rules applied. Generally, in Chinese p_mod is placed preceding a predicate if its semantic class is *location, temporal, direction, method, iteration* or *source.* Otherwise, it is placed after the predicate, if its semantic class is *extent, time-extent* or *destination.*

- **Reordering in a sentence with a ditransitive verb**: the IR for a sentence with a ditransitive verb has a fixed order as follows:

```
[prop...[pred...],
        [arg(subj)...],
        [arg(dobj)...],
        [arg(idobj)...]]
```

Two Chinese readings can be produced from the above IR. i.e.

你给我这本书.

(you give me this book.)

你把这本书给我

(you give this book to me.)

In the first reading the word order in Chinese is the same as that in English, in which emphasis is placed on **you**. In the second reading the direct object is placed preceding **give** and a particle 把 is inserted in front of the direct object to make a proper Chinese sentence. In this case emphasis is shifted to the **book**. The two readings are legal sentences. In the Chinese system verb transitivity information is encoded into the lexicon to control the word order.

2.2 Language equivalency:

Language equivalency problems occurs at both the structural level and the lexical level. Below is a detailed discussion of the techniques used in the Chinese system.

- **Problems of structural equivalency**: Some IR structures, which closely match the structure of English or other languages, may not match the structure of Chinese. Therefore, structural transformation is required before generating/parsing a Chinese sentence from/into these IR structures, An example of this is:

English: *Thank you for your reply.*

Chinese: 谢谢你的回信. *(thank your reply.)*

Information encoded in IR: *I thank you for your reply.*

The following transformation is made implicitly to produce appropriate translation:

```
[prop...[pred...],[arg...],
        [arg...],[p_mod]]
 <==>   [prop...[pred...],
        [arg...]]
```

- **Idiomatic expressions vary from language to language**. A typical examples include date expressions. The information received from the IR structure:

```
[ent,[type,prop],
     [class,date],
     [agree,ts],
     [det,spin],
     [quant,unq],
 [e_desc,august_x,1,1986]]
```

will produce

in English: *August 1, 1986*

in Chinese: 1986年8月1日
(1986 year 8 month 1 day).

Special rules are needed to insert the proper words, 年(year), 月(month) and, 日(day) in the appropriate place.

- **Problems of lexical equivalency occur in three different ways**: one-to-many correspondence, many-to-one correspondence or no correspondence.

  - One-to-many correspondence: The appropriate interpretation is determined by the semantic class. For example IR token more_x corresponds to several Chinese interpretations as shown below:

```
espec(more_x,tu,更大的,force).
espec(more_x,tu,更加,a_prp).
espec(more_x,tu,很多,a_obj).

a_prp  = abstract property
a_obj  = abstract object
```

  These constraints (e.g. force, a_prp, a_obj), which indicate the semantic class of the entity modified, allow the system to produce the appropriate translations as follows:

```
more lift   <==>更大的升力
more detail <==>更加详细
more paper  <==>很多稿件
```

  - Many-to-one correspondence: The occurrence of many IR tokens corresponding to a single Chinese lexical item is usually a case of synonymy, If the IR specifications of those tokens are compatible, different readings which have the same meaning will be produced. This case is acceptable in most machine translation systems. For example, if

```
verb(begin_1,T,开始).
verb(start_3,T,开始).
```

  are in the Chinese lexicon, two English readings:

  They begin to separate from the wings.

  They start to separate from the wings.

  will be produced from Chinese input string,

  它们开始与机翼脱离.

  No constraint is necessary as long as the IR specifications for *begin_1* and *start_3* are compatible.

  — No appropriate correspondence at lexicon level: In this case a paraphrase is provided to represent the equivalent meaning. Two examples are:

  如果你能详细地把它解释一下我们非常感激.

  We would greatly appreciate it if you can explain it in detail.

  我们很希望邀请你担任英语分组的主席.

  We would like to ask you to act as the chairman of an English session.

  The individual word senses for *would, like,* and *appreciate* do not apply for the collocation *would like* or *would appreciate.* Therefore, language equivalency is applied to convert *would like* into 很希望 or *would appreciate* into 非常感激 to assure the equivalency.

2.3 The verb constituent:

Unlike English, the Chinese verb does not have morphology to reflect tense, aspect, mood, voice or polarity. Instead, auxiliary verbs and special pre-particles or post-particles are used and a separate verbal constituent is made to represent this information as shown in the list below: (special particles are showed in Chinese character).

TENSE		
present	past	future
verb	verb+过 verb+了	将+verb
e.g. *eat* 吃	*ate* 吃过 吃了	*will eat* 将吃

ASPECT		
simple	perfect	progressive
verb	已经+verb+了 已经+verb+过 已经+verb+过了	正在+verb verb+着 正在+verb+着
e.g. *eat* 吃	*have eaten* 已经吃了 已经吃过了	*eating* 正在吃 正在吃着

VOICE	
active	passive
verb	被+verb+了
e.g. *check* 检查	*is checked* 被检查了

The Chinese system makes use of the information received from the IR to produce its verb constituent based on the rules listed above.

2.4 Prepositional phrase attachment:

There are no IR tokens corresponding to prepositions in the IR structure. Instead, the semantic class constraints, which play a role in restricting prepositional phrase attachment, are placed in the IR. Thus, Chinese propositional phrases are constructed and attached to the modified phrase based on the information available in the IR and IR lexicon matching against the information encoded in the Chinese lexicon. For the sample sentence:

The printer is placed to the left side.

the information about the proposition modifier p_mod provided in the IR is:

```
[p_mod,[g_rel,oo] ,
       [k_rel,loc],
       [t_rel,none],
  [ent,[type,nrm],
       [class,p_prp],
       [agree,tu],
       [det,spin],
       [quant,unq],
       [e_desc,left_side_x]]]
```

It indicates that the phrase *to the left side* has semantic class **p_prp** (physical property). It attaches to the predicate *place2_1* by case relation **loc** (location). The semantic class of the predicate *place2_1* is **action** which is encoded in the IR specification. The Chinese preposition lexicon has the following information:

```
prep(to_p,在,nil,action,p_prp,loc).
```

Once the information provided by the IR structure and the IR specification matches against the preference encoded in the Chinese prepositional lexicon, the right prepositional phrase attachment is carried out.

In summary, issues in word order, language equivalency , verb form constituency and prepositional phrase attachment all affect translation accuracy.

3 Translation efficiency:

As the coverage of the system is extended, the translation efficiency becomes increasingly more critical. Obviously, blind search results in high cost and inefficiency during computation. Linguistic knowledge (such as: a compound sentence must contain at least one conjunction; a complex sentence may contain more than one verb; a discourse modifier usually appears at the beginning of a sentence; and, an imperative sentence begins with a verb, ... etc.) can be used as a heuristic in order to reduce the

cost and increase efficiency. The Chinese system employs a top-down, depth-first searching algorithm- Therefore, the choices at the highest level of the grammar hierarchy have the greatest effect on efficiency. The system uses the heuristic in the following ways:

- Organizing grammar rules in a hierarchy depending on the syntactic consideration. Search is controlled by the following heuristic:

  - Search for a conjunction in the input string at the top level to avoid applying compound proposition rules to a simple sentence and simple proposition rules to compound sentences. For example:

    飞行员可以增加飞机的速度或者他可以增加冲角.

    (*The pilot can increase the speed of the airplane* **or** *he can increase the angle of attack.)*

    The discovery of a conjunction 或者 (**or**) between two clauses directs the search to the compound proposition rules.

  - Count the number of verbs in the sentence to direct the application of proposition rules with either simple or complex arguments as is appropriate given the count. Following sentences give the examples in this case:

    请寄一份在会议室参加讨论的人员的名单.

    (Please **send** *a list of the people who* **participated** *in the discussion in the conference room.)*

    (number of verb = 2, a case of an argument with relative clause.)

    我们很希望邀请你担任英语分组的主席.

    (*We* **would like** *to* **ask** *you to* **act** *as the chairman of an English session.)*

    (number of verb = 3, a case of an infinitive propositional argument.)

    请寄给我一份名单.

    (Please **send** me *a list.)*

    (number of verb = 1, a case of simple argument.)

  — Check the first word in the input string at the top level to identify sentences which begin with a discourse modifier. e.g.

    然而我需要更多些信息.

    (**However**, *I need some more information.)*

    The same strategy is also applied at propositional level to check the first word in an input string to identify sentences beginning with a verb. e.g.

    谢谢你的1986年8月1日的回信.

    (Thank *you for your reply of August 1, 1986.)*

- Block an inadequate search branch at an early stage by checking semantic match constraints for each constituent pair to avoid unnecessary search.

- Find the clause boundary for compound sentences in advance by searching for a conjunction or by checking transitivity features for verbs to avoid misapplying inappropriate rules. There are two types of compound sentences: In the first case, in which the conjunction is embedded between two clauses, the clause boundary can easily be discovered by searching for a conjunction. In the second case, in which a conjunction appears as the first word in the string, the clause boundary is determined by checking the transitivity feature of the verb. This constraint prevents the direct object of the first clause from being mistakenly treated as the subject of the second clause because of the unknown clause boundary.

In summary, the use of these heuristics results in translation efficiency. The detailed description is in [Jin, 1991].

4 Conclusion:

The Chinese system works either independently, parsing an Chinese input string into an IR structure and generating an Chinese string from an IR, or dependently as a subsystem of ULTRA, translating a Chinese string from/into any of the other four languages. Intermediate representations (IRs) make it possible for Chinese to translate from/into additional languages without further modification. Symmetric grammar rules, initially developed on a limited corpus [Jin, 1986], have been scaled up to handle various types of texts including expository texts, business letters, e_mail messages as well as menu type documentation. As the coverage is extended, high quality and efficiency are assured by using the techniques described above. Further research is needed on the issues involved in keeping translation accuracy while relaxing constraints, and in maintaining translation efficiency while reducing redundancy for the purpose of robustness. Furthermore, by integrating other NLP techniques into the system, such as techniques for the modeling the pragmatics of belief ascription, natural language semantics, as well as techniques for the large-scale extraction of meaning from dictionaries and other texts, a fully automatic advanced MT system looks promising in the future [Wilks & Farwell 1990].

5 References:

[Farwell & Wilks, 1990] D. Farwell and Y. Wilks. Ultra: a Multi-lingual Machine Translator, MCCS-90-202, Computing Research Laboratory, New Mexico State University.

[Procter et al, 1978] P. Procter et al. *Longman Dictionary of Contemporary English,* Longman Group Limited, Harlow, Essex, England.

[Pereira & Warren 1980] F. Pereira and D. Warren. Definite Clause Grammar for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition network, *Artificial Intelligence,* 13, pp. 231-278.

[Huang, 1985] X. Huang. Machine Translation in the SDCG (Semantic Definite Clause Grammars) Formalism. *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* Colgate University, New York, pp.l35-144.

[Huang, 1988] X. Huang. Semantic Analysis in XTRA, an English-Chinese Machine Translation System. *Computers and Translation,* 3, pp. 101-120.

[Jin & Simmons 1986] W. Jin and R. F. Simmons. Symmetric Rules for Translation English and Chinese, *Computers and Translation,* Vol 1, No. 3, pp.153-167.

[Wilks, 1990] Y. Wilks. Where am I Coming From: The Reversibility of Analysis and Generation in Natural Language Processing, MCCS-90-195, Computing Research Laboratory, New Mexico State University.

[Jin, 1991] W. Jin. Translation Techniques in the issue of Accuracy and Efficiency, MCCS-91-208, Computing Research Laboratory, New Mexico State University.

[Jin, 1986 1 W. Jin. Machine Translation between Chinese and English, *Proceedings of the sixth Canadian Conference on Artificial Intelligence,* Montreal, Canada, pp.129-133.

[Wilks & Farwell, 1990] Y. Wilks and D. Farwell. A White Paper on Research in Pragmatics-based Machine Translation, MCCS-90-188, Computing Research Laboratory, New Mexico State University.