计算机科学 1998 Vol. 25 No. 5

一个基于中间语言的汉英机器翻译系统*)

A Interlingua-based Chinese-English Machine Translation System

周会平 王 挺 史晓东 陈火旺 齐 璇

(国防科技大学计算机学院 长沙410073)

摘 要 To translate a natural language accurately, a mount of knowldege of grammar, semantic and context is needed. In this paper, we present ICENT, a Interlingua-based Chinese-English Natural-language Translation system. ICENT uses different methods in connection with the different features of Chinese to get good results. We also discuss the goal we should attain and the implement approach of the system.

关键词 Machine translation, Interlingua

一、引言

随着生活的信息化、计算机的普及以及Internet网的迅速发展,人们每天要接触和处理的信息量越来越大。这些信息很大部分都以各种自然语言(如汉语、英语、德语等)为载体,如果能让计算机把其它语种的资料自动翻译成自己的母语,这将给人们带来极大的方便。因此机器翻译(MT)一直是信息处理领域中一个极有意义且极富挑战性的课题。

我们对机器翻译的研究始于1992年史晓东开发的Matrix英汉翻译系统,1996年6月,我们又承担了国家863课题"多语种机器翻译系统的研究和开发"的汉英机译系统部分。本文介绍了ICENT的研究的目标和实现机制,包括如何针对汉语的特点进行准确有效的分析、如何有效地组织和管理词典、设计精确有效的中间语言表示以及如何生成准确的译文等。

二、ICENT的研究目标

ICENT是基于中间语言的,其流程如图1。汉语首先经过分词和标注,然后进行句子结构的分析,分析的结果存入一种中间表示结构,即中间语言。英语生成部分从中间语言中获取信息,生成译文。

我们的研究计划是实现一个基于知识和中间语言的汉英翻译系统。它不限制汉语的语法,能接受我们当前所使用语料的所有句型。之后,我们将用其它

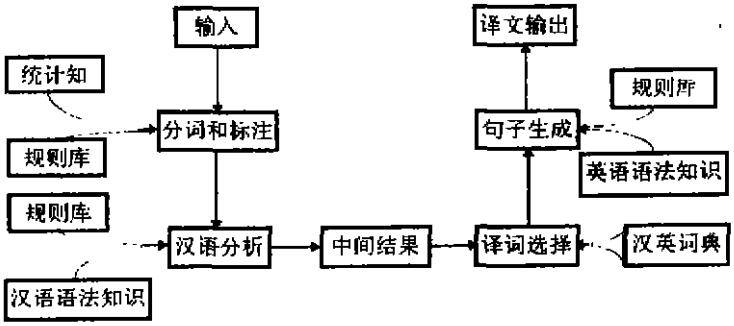


图1 汉英机译系统流程

*) 本课题受863高科技计划和国家自然科学基金资助。

语料对它进行扩充，让它能处理所有的汉语句型，针对这种考虑，我们制定了以下开发目标。

·不需要人工预编辑。系统能处理一般的汉语句子，不限制源语言的语法，设计中间语言时尽量考虑了汉语的所有可能句型。

·自动分析和消歧。对于输入的汉语句子，系统能自动对它进行分析，自动处理分析过程产生的歧义，不需要人工的干预。

·高速的在线翻译。为了实现快速翻译，系统必须要有好的算法分析句子的结构，良定义的中间语言表示以及从中间语言到英语句法结构的快速转换。

·良好的容错性。汉语分析时尽可能地抽取重要的信息，这样即使遗漏了某些信息，只要它们不影响句子的主要结构，系统仍然能生成可读的译文。

·系统的可扩充性。系统的词典可扩充，可以增加新的词汇和义项；中间语言结构可扩充，容易被扩充以表示被遗漏的句法结构；规则可扩充，为了适应新句法结构的要求，可以对规则库进行扩充和修改。

三、ICENT 的设计

和传统的基于中间语言的机译系统一样，ICENT 的翻译过程分成四个部分：汉语的分词和标注、汉语的分析、汉语到中间语言结构的转换和英语的生成。

3.1 汉语的分词和标注

汉语的切分和标注是汉英机器翻译的第一步。目前，国内已有不少单位在这方面做得很好，我们利用了北京大学的研究成果，并对它进行了适当的修改。首先，我们增加了一个切分标注的后处理模块，它将排除一些固有的错误。如对于“一天下…”的分词就存在歧义，有“一/m　天下/n”和“一天/t　下/v”两种切分方法。在后面接名词时（如“一天下雨”）系统将它错误切分为“一/m　天下/n　雨/n”，这种错误可以在后处理模块中排除。另外，对于某些标注时的歧义，为了减少消歧时的错误，我们将它保留下来，留到句法分析时再去处理。如“别过河”和“说过话”，系统把在动词后面的“过”都作助词处理，则“别/v　过/u　河/n”就是错误的。对于这些可能导致错误的消歧，我们不让它去做，而将其结果“别/v　过/vu　河/n”保留到句法分析时，借助于其它知识来帮助消歧。

3.2 汉语的分析

汉语分析是系统的关键，其结果直接影响到中间语言表示和英语生成的质量。在进行分析之前，我们首先要了解汉语的特性。汉语是孤立语。朱德熙先生曾指出[1]：“汉语的名词、动词、形容词都是‘多功能’的，不象印欧语那样一种词类只跟一种句法结构对应，”掌握汉语和英语在语法上的区别对于如何去表示和分析汉语都是非常重要的。

·汉语的名词除了作主、宾语外，还可以作谓语和定语，

1. 今天星期天。

2. 语法结构和词类有关系。

句子1中没有动词，其中的两个名词，“今天”是主语，“星期天”是谓语。句子2中的名词“语法”作定语修饰“结构”。

·形容词除了作定语外，经常直接作谓语和补语，有时还能作主语和宾语。

3. 爱漂亮是一种普遍的现象。

4. 她很漂亮。

形容词“漂亮”在句子3中是动词“爱”的宾语，在句子4中则作句子的谓语。

·动词除了作谓语外，作主语和宾语也都是常见的。

汉语中这些词语在充当不同的句子成分时没有时态或体的变化。因此一个汉语句子经常含有多个动词，而且各动词之间缺少明显的语法约束，也没有时、体等信息来说明它们之间的主次关系。这些现象都导致了汉语分析的复杂度非常高。

·具有相同词性顺序的两个句子可以有完全不同的句子结构。

5. 他坐车去学校。

6. 他叫大家去学校。

句子5和6具有相同的词性顺序，都是“代词＋动词＋名词＋动词＋名词”的结构。但句子5中的两个动词“坐”和“去”是并列关系，都是描述主语“他”的。而在句子6中，动词“叫”是主动词，“去”是兼语结构，描述宾语“大家”。

·组成句子的各语法成分之间的语法约束比较弱，语义约束往往比较强。

7. 杯子打破了。

8. 我打破了杯子。

“杯子”在7中是“打破”的主语，在8中则是打破的宾语。在语法上“杯子”在两个句子中担任不同的语法成分，但从语义上看，它都是动词“打破”实施的对象，是受事宾语。现在，很多语言学家都认为语义研究对于汉语的分析理解有着非常重要的作用，甚至

有人建议抛弃现有的语法体系直接进行汉语的语义研究。但不管从现在的理论水平还是从实践水平上看，用计算机直接去处理汉语的语义都还很不成熟。根据汉语的特点，我们采用语法和语义相结合的方法，实现汉语的分析。

·汉语的省略现象非常多。从一段文章中单独抽出一个句子，不考虑它的上下文，则它一般都是不完全的。这是因为汉语里没有名词的指称信息，主语和宾语的承前省略也很多，所以对汉语句子的理解离不开对上下文的分析。

·汉语以词组为本，句子中比较稳定的是各种词组的结构与搭配。在句子一级上，汉语词组的作用非常灵活，但词组的内部结构一般都比较稳定。如体词性偏正词组都由形容词、名词或"的"字结构加名词构成。

针对汉语的特点，我们采用多种策略相结合的方法分阶段地对汉语进行分析。汉语的词组结构比较稳定，这部分是汉语分析的第一步，系统用模式匹配和规则的方法快速抽取其中的词组信息。

9. 依存/n 文法/n 直接/a 描述/v 单个/b 词/n 之间/f 的/u 关系/n 。/w

名词"依存"和"文法"组成体词性词组"依存文法"，动词"描述"和修饰词"直接"组成谓词词组"直接描述"，"单个"和名词"词"组成体词性词组，然后和方位词"之间"组成一个表示方位的体词性词组，通过助词"的"修饰名词"关系"。句子9的分析结果是：

10. 依存文法/ti 直接描述/wei 单个词之间的关系/ti 。/w

英语句子中有时态和语态的信息，汉语句子中没有明确的表示，它借助于一些特殊的词语。第二步，系统对汉语句子中的虚词、助词和时间词等进行适当的语义分析，同时引入一些上下文信息，共同确定句子的时态和语态。如动词后面的助词"了"表示动作的完成，副词"正"、"正在"表示动作的正在进行，时间词有"昨天"、"明天"、"星期天"等，其中"昨天"和"明天"是相对时间词，能直接确定句子的时态，"星期天"是绝对时间词，需要借助于和当前时间的比较来确定句子的时态。第三步是在词组的基础上进行句子结构的分析，生成句子的中间表示结构。系统首先判断输入是否属于某一特殊句型，是则用模式匹配的方法直接转换到中间语言结构，否则采用自动分析的方法生成可能的中间结构，然后借助于单个词汇的语法特性（这些信息可以在系统词典中获取），排除大部分错误的中间结构，最一对剩下的结构进行概率估值，求出最有可能的一个或多个语法树，作为汉语句子分析的结果。

3.3 汉语分析的文法

汉语以词组为本，词组的结构和搭配都比较稳定。词组在句子中的语法功能非常灵活，主谓词组和句子具有相同的结构，而且词组中的动词和句子的主动词之间没有任何形态上的差别。依存文法（Dependency Grammar）适合于描述词和词之间的关系，依存关系树和语义网之间存在比较简单的对应关系。现在，许多语言学家都认为语义研究对于汉语的分析理解有着非常重要的作用。我们采用扩充后的依存文法作为汉语的表示和分析的文法体系。

依存文法是由 Tesniere 于1959年提出的一种语法理论[2]，1965年 Gaifman 对它进行了形式化，并正式提到了语言学界[3]。1970年，Robinson 提出了在转换生成框架中使用依存文法的可能性，同时给出了四条公理[4]：

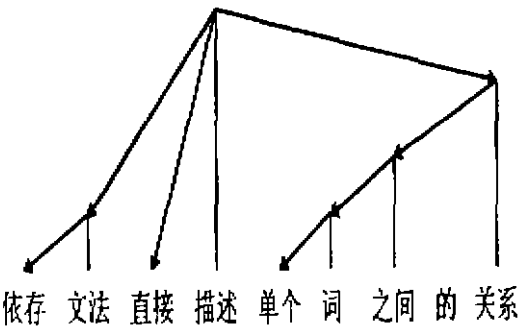
1. 有且只有一个元素不依存于其它元素；

2. 其它元素都直接依存于一个元素；

3. 任何元素都不直接依存于两个或两个以上元素；

4. 如果元素 A 直接依存于元素 B，元素 C 在它们之间（指句子的词序），则 C 直接依存于元素 A 或 B 或其它在 A、B 之间的元素。

依存文法直接描述单个词之间的关系，依存关系就是指中心词与修饰词、动词与施受对象之间的关系。如果用箭头描述依存关系，句子9就有下面的结构：



由上面的图是可以看出，"描述"是句子的中心词，直接依存于它的词有三个，其中："文法"是它的主语，"关系"是宾语，"直接"是修饰词，直接依存于"文法"的词只有"依存"，是它的修饰词。

这种依存文法只描述了词之间的依存关系，没有包含它们之间的语法关系。直接用它来表示和分

析汉语句子，其结果显得太粗，忽略了太多的信息。我们对它进行了改进，在其中加入一些语法功能函数，来描述词之间的语法关系。我们将上面的公理形式化为：

X(A(f(A)),B(f(B)),…)

其中A、B等都是直接依存于X的元素，f(A)、f(B)分别描述了A、B与X之间的语法关系。f可以是SUBJECT、OBJECT、QUANTITY等值。这样对句子9就可以描述如下：

(1)描述(文法(SUBJECT)，直接(MANNER)，关系(OBJECT))

(2)文法(依存 DISCRIPTIVE))

(3)关系(之间(RESTRICTED))

(4)之间(词(DISCRIPTIVE))

(5)词(单个(QUANTITY))

改进后的依存文法清楚地描述了句子中各词语之间的关系以及它们在句子中的语法功能，使句子的框架也一目了然。加入的语法功能函数和系统的中间语言结构对词汇的语法功能的描述是一一对应的，便于从依存文法表示到中间语言结构的直接转换。

3.4 词典的组织

系统的汉语句子分析和汉英词语转换部分使用同一部词典。词典包括汉语词汇的语法信息和英语译词信息，采用多层次的组织方法。首先，词典根据汉语词汇的词性对词语进行分类，如名词、动词、形容词、时间词等，对某些重要词类再细分为多个子类。如动词可分为带宾语的动词和不带宾语的动词，带宾语的动词又分为带谓词宾语的动词、带体词宾语的动词、带双宾语的动词等。然后根据语法功能和语义对词语进行义项上的分类。义项是词典中的最小的语法单位，一个词语可能包含多个义项。例如副词"都"包括"全部"、"甚至"等多个义项。词汇的英文译词是基于义项的，如果一个义项存在多个译文选项，则词典给它们一个优先级上的排序。如"变化"有change和vary等译词，词典把change放在vary前面，当一个义项存在多个用法不同的译词时，词典在每个选项前加上了简单的判断条件，系统可以根据条件更好地选择适当的译文。

动词是汉语句子分析的关键，弄清楚句子中各动词的关系，也就知道了句子的结构。不同的动词带不同的参数时一般都有不同的含义，因此词典中要有关于动词语法功能的描述。如动词"给"有三个参数，subject、object1和object2。它可以只带一个宾语(如"他给了我。")、也可以带两个宾语("他给了我书。")。如果只带一个宾语，该宾语是直接宾语。带两个时，则第一个是直接宾语，第二个是间接宾语。词典中必须提供这些信息。

3.5 中间结构的设计

中间语言是系统的核心部分。源语言分析的所有结果都体现在中间语言表示结构上，目标语言生成所需要的所有信息也都从中间语言结构中获取。在设计中间语言时我们采取循序渐进的方法，汉语分析、中间语言的补充和修改、英语生成这三个部分同时进行，中间语言可以不断从其它两个部分获得反馈信息，不断地得到完善。

中间语言是连接汉语分析和英语生成的桥梁，其设计应当充分体现汉语的语法特征和英语生成的需要。ICENT的中间语言表示结构的表示采用框架的形式，这有利于在开发过程中不断地补充和完善，也有利于今后对系统进行进一步扩充。英语在语法上和汉语有很大的区别，只有充分地了解了两种语言的语法，才知道如何去满足汉语分析和英语生成两个过程的需要。汉语有分句，英语只有从句。介词短语是英语中比较复杂的成分，汉语中很多句法结构都对应英语中的介词短语。介词短语可以作状语、定语，也可以作表语。作不同的语法成分时修饰句子的不同部分。如：

11. Thank you for reminding me of it.

谢谢你提醒我。

介词短语"for reminding me of it"作状语，修饰thank。而在汉语中"提醒我"则表现为兼语形式。

12. The products on display are new ones.

展示的产品是新的。

"on display"在这作定语，修饰products。在汉语中"展示"也是作定语修饰"产品"，但它在形式上和其它定语没有什么区别。系统要确定什么样的结构对应英语中的介词结构。

汉语的修饰成分有很多种不同的作用，如状语可以表示时间点、时间段、地点、程度、目的、方式、频率等等。表示不同修饰作用的状语一般都对应英语中的不同词法结构。如时间点状语对应英语中的at或on介词结构或时间词，地点状语对应于英语中的at、in等介词结构。也有的状语(如程度状语)对应于英语中的一般的程度副词。因此，系统必须明确各修饰成分的作用。另外时态、语态、冠词等信息也都是英语生成时所不可缺少的。系统的中间结构包含下面这些信息：(1)分句间的关系；(2)句子的主、谓、宾结构以及它们的修饰成分；(3)各修饰成分的作用；

(4)谓语的时和体的信息;(5)名词的数以及特指信息;(6)多个动词之间的关系等。

下面以一个例子来说明中间语言的形式:

13. 他们/r 俩/m 找/v 老牛/n 评理/v 。/w

```
((PATTERN SIMPLE)
(MOOD DECLARATIVE)
(TENSE PRESENT)
(PREDICATE ((HEAD ((CENTER ((CAT VERB)
                  (SENSE 找)))))
            (SUBJECT ((CENTER ((CAT PRON)
            (SENSE 他们)))))
            (OBJECT ((CENTER ((CAT
            NOUN)(SENSE 老牛)))))
(EVENT ((HEAD ((CENTER((CAT VERB)
                  (SENSE 评理)))))(OBJECT
                  NIL))))))
```

中间结构表示使用框架的形式,以主动词为中心词,用槽 PREDICATE 表示。PATTERN 是对句型(复句或单句)的描述。MOOD 是对主动词的语态描述,其值有陈述句(DECLARATIVE)、祈使句(IMPERATIVE)、感叹句(EXCLAMATORY)和疑问句(INTERROGATIVE)。TENSE 表示句子的时态,其值有过去时、现在时和将来时,CAT 描述词性,ROOT 描述词根等等。

3.6 英语的生成

在好的中间语言表示结构的前提下,要生成高质量的目标语言,系统还必须满足以下要求:

·要很好地掌握目标语言的词法和句法,知道如何由短语构成句子和必要时由多个句子结构构成复杂的句子。

·精确的转换规则。中间语言的每个槽或框架都对应英语中的一个或多个句法结构,转换规则帮助系统选择正确的词汇和句法结构,充分获取中间语言中的信息。

系统的中间结构是完全针对英语需要的,当一个句子分析完成后,所获得的信息基本上能满足生成一个英语句子所需要的所有要求。系统使用规则的方法,对中间结构各属性的值一一进行分析,逐步生成英语中相应的词句结构。系统的规则优先获取重要的信息,如主语、谓语、宾语等,保证即使在遗漏某些信息时,只要不破坏句子的结构,系统仍然能输出可读的译文。

结论 我们的目的是开发一个基于中间语言的汉英机器翻译系统 ICENT。我们对现有语料进行分析,建立了系统的词典和句型库。中间语言的设计包含了句型库中的所有句型,并尽可能包含不在句型库中的句型。ICENT 能自动完成对汉语句子的分词、标注、句型分析、到中间语言的转换和最后的英语译文生成。它不需要人工译前和译后的编辑,达到快速地较精确地翻译。我们首先完成的是一个限定语料的原型系统。之后,我们将用其它语料对它进行扩充,使它可以处理所有一般的汉语句子。

参 考 文 献

[1] 朱德熙,语法答问,商务印书馆,1985

[2] L. Tesniere, Element de Syntaxe Structurals, Paris, Klincksieck, 1959

[3] H. Gaifman, Dependency Systems and Phrase-Structure Systems, Information and Control 8, 1965

[4] J. J. Robinson, Dependency Structures and Transformation Rules, Language 40, 1970

[5] 冯志伟,自然语言机器翻译新论,语文出版社,北京,1994

(上接第50页)

[7] A. Lavie, GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language, PhD's thesis, CMU, 1996 5

[8] J. W. Amtrup, Layered Charts for Speech Translation, TMI'97, Sante Fe, NM, 1997. 7

[9] M. Rayner and D. Carter, Hybrid Language Processing in the Spoken Language Translator, Proc. of ICASSP'97, Munich, Germany, 1997. 4

[10] Yoshinori Sagisaka, Recent Advances in Speech Translation Research at ATR-ITL, Proc. of C-STAR I 96: ATR Intl. Workshop on Speech Translation, Kyoto, Japan, 1996. 9

[11] Finn Dag Buo, FEASPAR-A feature Structure Parser Learning to Parse Spontaneous Speech, PhD's thesis, University of Karlsruhe, 1996, 9

[12] S. Wermter, V. Weber, SCREEN: Learning a Flat Syntactic and Semantic Spontaneous Language Analysis using Artificial Neural Networks, J. of Artificial Intelligence Research, Vol. 6, 1997