

中一日机器翻译系统 J-北京

周 明

清华大学计算机科学与技术系

摘 要: J-北京¹是作者在访问日本高电社期间主持设计的中一日机器翻译系统。该系统 1996 年底开始研制,1998 年 4 月完成第一版,1998 年 12 月完成第二版。现有词汇量 22 万条,中文句法分析规则、中一日转换和日本生成规则 2000 余条。运行在 Windows98/95 日文版或者 Windows NT4.0 日文版。在 Pentium 166 计算机上翻译速度达到 50 字/秒。本文将简要介绍该系统的设计思想和实现技术。

关键词: 中一日机器翻译

J-Beijing Chinese-Japanese Machine Translation System

Ming Zhou

Dept. of Computer Science, Tsinghua University

Abstract J-Beijing is a commercial Chinese-Japanese machine translation system designed by the author during his visit in Kodensha Ltd. of Japan. The project was started in the end of 1996, and the first version was finished and put into market in April, 1998, with 160,000 entries in its lexicon. The second version was finished and put into market in December 1998 with 220,000 entries in its lexicon. Over 2,000 rules including Chinese parsing rules, Chinese-Japanese transfer rules and Japanese generation rules. The system runs on Japanese Windows95/98 or Windows NT 4.0. The translation speed reaches as high as 50 Chinese characters per second. This paper will focus on the design and the implementation of this system.

Keywords Chinese-Japanese machine translation

1 引言

近年来,随着中、日两国政治、经济、科技、文化的交流日趋密切,有关中一日机器翻译和日一中机器翻译的研究也逐渐引起重视。1996 年底到 1999 年 3 月,我应日本高电社公司的邀请,在该公司主持了中一日机器翻译的研究项目。1998 年 4 月完成了第一版,词汇量为 16 万。1998 年 12 月完成了第二版,词汇量达到了 22 万,主要是增加了许多香港和台湾的用词以及常用的科学技术词汇。具备自动识别字体(GB,BIG5 和 Chinese Writer 字体)的功能。

¹J-北京, Chinese Writer 是株式会社高电社(<http://www.mesh.ne.jp/KODENSHA/>)的注册商标。

除了具备通常的正文翻译功能之外,还具备实时地翻译中文的 Homepage 和网上中文报刊的功能。在日本市场反应良好。

本系统在设计上的主要特点是①采用了转换式的设计,对输入的原文经过中文分析,中一日转换和日文生成三个阶段生成日语;②在中文分析上采取了基于块的依存文法,该文法综合了传统的短语结构分析和依存分析的优点,具有鲁棒性好,表达汉语的特定句型十分方便等特点。本文将简要介绍该系统的设计思想和实现技术。

2 中一日机器翻译系统的总体设计

中一日机器翻译的难点主要体现在:

- ①中文的理解或者中文的分析难度极大。中文在自动分词、词性标注、短语结构分析、依存分析以致语义分析等几个阶段都存在着不可逾越的困难。
- ②中文和日文分别属于不同的语系。中文属于独立语系,日文属于粘着语系。中文缺乏屈折变化,日文屈折变化十分丰富。中文没有或者极少有明确表示句法功能的助词,而日语的助词十分丰富。对句法功能的提示十分明显。中文的语序基本上是主一谓一宾,而日语的语序基本上是主一宾一谓语。两种语言的语序差别很大,转换的难度也很大。

但是在设计中一日机器翻译系统的时候,除了看到上述的难点之外,还要分析两种语言的相似之处,以有助于减少不必要的工作。下表列出了中、日两种语言的相似点。

		例
字符集	共同汉字有2000以上	中、国、日、本
词汇	常用名词、动词、形容词中有50%形同义同或者日本人可知其意[3] 成语容易翻译	中国、日本、安全第一 大器晚成一>大器晚成(漢式成語) 心懷叵測一>腹に一物(和式成語)
多个定语的排列顺序	基本一致	我的三个红色的书包一> 私の三つの赤いかばん
多个状语的排列顺序	基本一致	昨天在报纸上显著地一> 昨日新聞に著しく
日期表示	基本一致	1998年4月28日一>1998年4月28日
数字表示	基本一致	一千九百八十一 一>一千九百八十一
专有名词	可以用汉字直接对应	山西、香港特别行政区、董建华
模糊对模糊	例如“的”和“の”的对应程度比较强	香港特别行政区的董建华的办公室一> 香港特别行政区の董建华の事務所

根据中、日两种语言的对比分析,我们认为直接式的翻译难以达到较好的译文质量;同时也认为由于中日两种语言的相似之处,也没有必要使用所谓的中间语言方式。所以采用了转换方式。另一方面,我们认为由于中文的难点,根本做不到深层的语义级分析;而且句法级分析基本可以满足中日翻译的需求,所以分析的深度到句法分析为止。中文分析的方法主要有基于短语结构的分析方法和基于依存的分析方法。本系统分析了两者的特点,提出了综合两者优势的新的分析体系—「基于块的依存分析法」。

所谓基于块的依存分析实际上是低层延用了短语结构分析,高层上延用了依存分析。由

于句子的分析结果是几个具有依存关系的大块。使得句子的表达比短语结构和依存结构都显得清楚。在具体实现上, 去掉原来短语结构分析的根结点(S), 只保留了短语级的非终结结点, 采用了部分分析。这样做, 继承了短语结构分析效率较高的优势, 可以利用现成的分析算法, 同时由于部分分析, 不必分析到根结点, 分析的鲁棒性大大提高。如果低层也采用依存表示, 从道理上也可以, 但是在短语结构内部的依存表示过于零碎, 不利于向日语转换。从中一日转换来考虑, 块是一个相对稳定的信息单位。由于中日语言的历史背景等因素, 在中文属于一个块的成分, 转换成日文也大多在一个块内。所以以块为基本单位, 再加上块之间的依存信息, 特别适合于中文的分析和中一日的分析和转换。

根据以上的分析, 可以确定分析和转换的界面为一个基于块的依存分析树, 它的描述如下:

①描述每个块的子树的集合 T

{T 1, T 2, …… T n }

②子树之间的依存关系集合 D

{<gov1,dep1,rela1>,<gov2,dep2,rela2>,... <govm,depm,relam>

翻译处理流程是:

- 1 对输入的句子, 进行自动分词;
- 2 前缀和后缀的处理
- 3 人名等专有名词的识别;
- 4 查中文分析词典
- 5 确定词性
- 6 块的分析(即部分分析)
- 7 依存分析
- 8 块的转换
- 9 依存转换(句型转换, 特殊词转换, 一般依存关系转换)
- 10 日文形态素的生成
- 11 输出译文

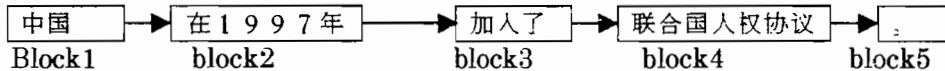
下面用一个例子说明翻译的流程。

输入句子: 「中国在1997年加入了联合国人权协议。」

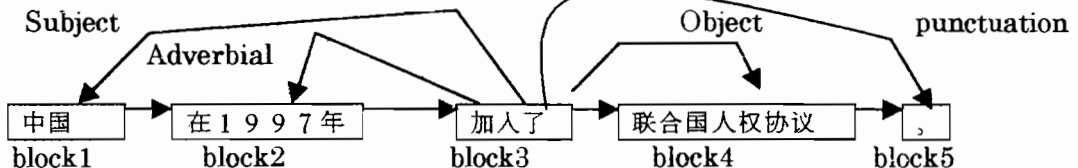
中文分词: 中国/在/1997/年/加入/了/联合国/人权/协议/。/

确定词性: 中国 N/在 I/U/9U/9U/7U/年 S/加入 V/了 E/联合国 N/人权 N/协议 N/。 P/

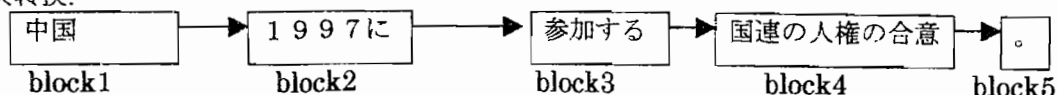
块分析:



依存分析:



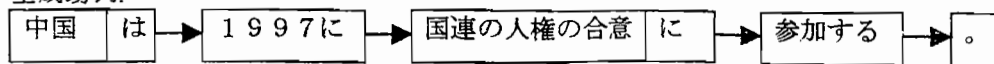
块转换:



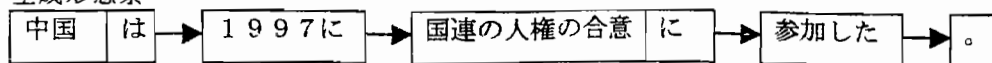
依存转换:



生成助词:



生成形态素



输出译文:

⇒「中国は1997年に国連の人権の合意に参加した。」

3 基于块的中文依存分析法

3.1 搭配结构的辨识

将汉语的搭配结构(也称为框式结构)识别出来,并合并为一个结点。例如:①{在,不在,正在,于}+{时,时候,同时,期间,时间,之际},②{从,自,在,不在,正在,于}+{方位词(前,后,中)},③从,自}+{来看,来说,看来},④{对}+{说来,来说},⑤{象,跟,和,与,同}+{一样,似的},⑥「」,⑦“”,⑧‘ ’,⑨[],⑩(),⑪{ },⑫{ },⑬< >等等。

3.2 块分析

块分析主要包括:①名词块(NP),②数字块(UG),③数量块(UP),④时间块(ITL),⑤时间长度块(ITP),⑥形容词块(AP),⑦副词块(FP),⑧动词块(VP)

分析方法:

- 「1」利用上下文有关的模式描写各种类型的块
- 「2」利用 ATN 文法识别各种类型的块

3.3 介词、方位词块的分析

- [1] 对介词要设法找到介词短语块的右边界

[2] 对方位词要设法找到方位词短语块的左边界

3.4 搭配结构 顿号结构 方位结构 介词结构 内部的分析 递归地重复上述的块分析过程

4 依存分析

我们定义了如下 17 类依存关系。

①	SUB	主語
②	OBJ1	第 1 宾语(间接宾语)
③	OBJ2	第 2 宾语(直接宾语)
④	COMP	补语
⑤	NUM	数量关系
⑥	TOP	主题
⑦	ADV N	近状语(副词、助动词、连词)
⑧	ADV F	远状语(介词、方位词构造、搭配结构状语)
⑨	QT	动词之前的闲杂成分
⑩	HT	动词之后的闲杂成分
⑪	PUNC	标点
⑫	PIVT	兼语
⑬	SOC	兼语补语
⑭	VAA	后连动关系
⑮	VAB	前连动关系
⑯	G	助动词关系
⑰	LOG	句子之间逻辑关系(假设, 因果等)

按照宾语的种类, 可以把中文的动词化分成如下 6 种类型:

- | | |
|---------------|--------------------|
| ①表示判断的动词 | 例:「是」「为」「不是」「是不是」 |
| ②可以带小句做宾语的动词 | 例:「认为」「希望」「相信」「批准」 |
| ③可以带兼语的动词 | 例:「请」「有」「使」「让」「催促」 |
| ④带双宾语的动词 | 例:「给」「递」「赠送」「告诉」 |
| ⑤只能带一个体词宾语的动词 | 例:「打击」「写」「听」「看见」 |
| ⑥不能带宾语的动词 | 例:「睡觉」「写字」「听话」「打球」 |

中文可以做谓语的成分:

- | | |
|--------------------|----------------------|
| ① 动词 | //例:「他睡觉」「他以为这样去不太好」 |
| ② 形容词 | //例:「花红了」 |
| ③ 「在...」「于...」介词短语 | //例:「他在北京。」 |
| ④ 一部分象声词 | //例:「风呼啸」 |
| ⑤ 数量词、数字短语 | //例:「大米三十公斤」 |
| ⑥ 时间名词 | //例:「今天星期三」 |

确定谓语的方法:

STEP1 对句子中具备谓语潜能的块, 取消那些虽然具备谓语潜能但是在当前句子中不可能做谓语的考察资格。

STEP2 在候选的具备谓语资格的块中, 选取最有可能具备谓语资格的块为谓语。

依存分析算法:

先考虑全句中只有一个具备谓语资格的块的情况。

块(1)	块(2)	块(k)	块(m)	块(n)
------	------	-------	------	-------	------	-------	------

[1] 在谓语前面寻找并且建立主语类(SUB, TOP)、状语类(ADV, ADVF, LOG, G)的依存关系。

[2] 在谓语后面寻找并且建立补语类(COMP)、宾语类(OBJ1, OBJ2)的依存关系。

如果一句话中有两个具备谓语资格的块, 块(k)和块(m), 其中假设块(k)被确定为谓语。

块(1)	块(2)	块(k)	块(m)	块(n)
------	------	-------	------	-------	------	-------	------

[1] 在块(k)之前寻找并且建立块(k)的主语类、状语类的依存关系

[2] 在块(m)之后寻找并且建立块(m)补语类、宾语类的依存关系

[3] 根据块(k)和块(m)的动词类型对块(k)和块(m)之间的区域内进行细致的分析:

[4] 在确定块(k)和块(m)的依存关系。

5 日语的转换和形态的生成

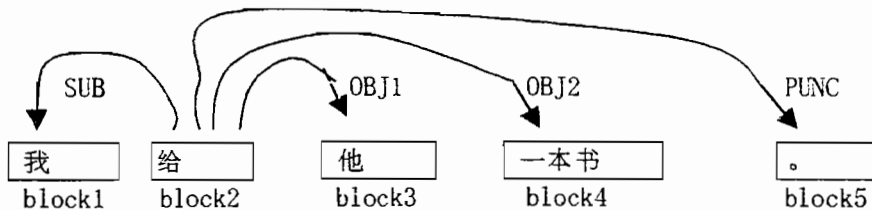
5.1 块的转换

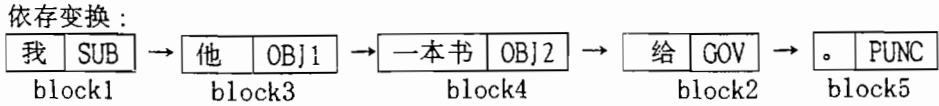
每个块的转换通过一组转换规则来进行。涉及调序, 加词和减词等操作。

5.2 依存变换

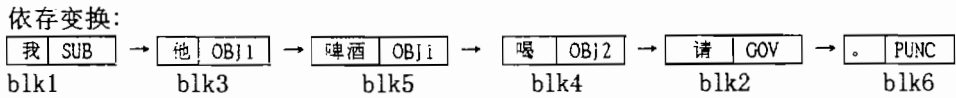
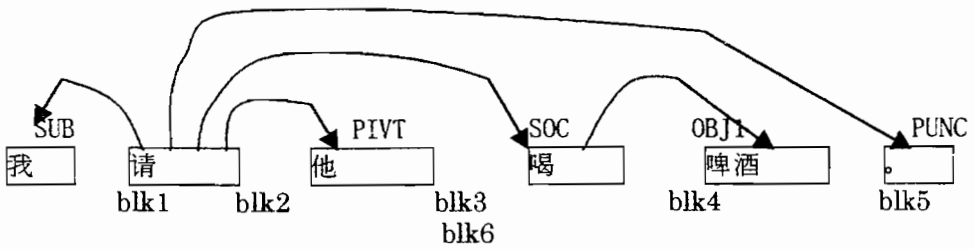
首先处理句型的转换, 包括: 把字句、被字句、兼语句、请字句、有字句、单宾句、双宾句、无宾句、比较句、连动句、宾语从句、主语从句。然后按照依存关系进行缺省的转换。

<例①> 原文「我给他一本书」





<例②> 原文「我请他喝啤酒。」



5.3 确定助词

以主谓关系 SUB 为例, 说明助词的确定方法

「我吃」→「私は」

「吃饭是」→「ご飯を食べるのは」(主語が動詞の場合、名詞化する。)

「我也吃」→「私も食べる」(「也」がある場合、訳語の助詞に「も」を使用する。)

助词确定规则

1. 主语后面有「也」「也要」「亦」「亦要」「也是」「亦是」等場合⇒助詞「も」
2. 主語表示疑问(例:「誰」「哪里」「什么」)⇒助詞「が」
3. 从句的主语の場合⇒助詞「が」
4. 其他場合⇒助詞「は」

5.4 特殊词驱动的转换

针对一些特殊的词, 要专门描写它的转换规则, 例如:

①「莫不如」(動詞): 译语「…ほうがよい」 依存关系:(莫不如, *, OBJ1)

原文:「莫不如去。」⇒译文「行くほうがよい。」

②「虽然」(接續詞): 译语「…が」 依存关系:(*, 虽然, LOG)

原文:「虽然他马上回去了。」⇒译文「彼はすぐに帰ったが、」

③「可以」(助動詞): 译语「…ことができる」 依存关系:(*, 可以, G)

原文:「你可以去。」⇒译文「あなたは行くことができる。」

5.5 日语的形态生成

在分析时,获得了时态,体,否定等信息。在上述的转换过程中,又可以进一步确定形态生成的必要信息。包括内部形态信息:①NEG(否定),②ASP(進行形),③VOI(受動態),④TEN(過去形),⑤SHIYI(使役体)。以及外部表现形式:①FORM=d(中頓形),②FORM=t(連体形),③FORM=z(終止形),④FORM=y(連用形),⑤FORM=N(体言形),⑥FORM=J(假定形),⑦FORM=M(命令形),⑧FORM=l(推量形)。形态生成的步骤是:①使役形生成,②受动态生成,③可能态生成,④否定型生成,⑤时态/体生成,⑥根据FORM进行语尾变化

6 下一步的发展

(1) 分词系统的组合歧义,是造成许多分析错误的原因。

组合歧义不仅包括常用词的歧义。例如“将来”,“马上”,这一类常用词组合歧义的特点个数有限,首先可以在词典中标记出它是组合歧义。给出各种可能的切分方法。在一定程度上,利用前后文的约束分别予以处理。但是在一个大规模的词典里,由于存在大量的复合词,如果再考虑到用户词典和专业词典里可以加入很长的词组,组合歧义的解决要困难得多。比如下列复合词如果加入到词典中,将造成组合歧义。

研究工作	我们正在研究工作/这个研究工作十分重要
研究过程	我们正在研究过程参数是否设的对/在这个研究过程中,
研究计划	他们正在研究计划明天去旅游
研究结果	我们正在研究结果对不对/这个研究结果

实际上在j-北京22万词条的词典里,几乎每一个长一点的词都有可能产生组合歧义。这个问题的处理必须要有系统化的解决方案。决非可以用简单标记的方法通过逐个设立规则来解决。

(2) 含有动词的复合名词的动态判定。

机器翻译研究
教研组改革问题
教研组改革计划安排

今后希望能针对这些难点能做一点工作,以便进一步提高系统的水平。

参考文献

- [1]『日语语法』,王曰和,商务印书馆,1987年北京
- [2]『汉译日基础教程』,迟军,北京大学出版社,1986年
- [3]『汉日翻译教程』,苏琦,大新書局,台北1993年
- [4]『汉语计算语言学』,吴蔚天,罗建林,电子工业出版社,1993年。
- [5]『现代汉语句法分析』,吴竞存,侯学超,北京大学出版社,1982年。