日汉机器翻译系统中的断段分析法研究与实现

曹睿妮 陈群秀

清华大学计算机科学与技术系

摘 要：本文介绍了用于日汉机器翻译的断段分析翻译方法，着重讨论了以下几个问题：(1)日语子句的概念及其切分；(2)基于日语句节的日语段的概念及其属性；(3)日语段的切分及其属性识别；(4)独立日语段的译文生成；(5)日语段译文的次序调整；(6)日语附属词的多义问题等。用此方法我们设计并实现一个分析器。

关键词：日汉机器翻译 断段分析法 子句 段

Research and Implementation of the Partitioning Method in JCMTS

Cao Ruini Chen Qunxiu

Department of Computer Science & Technology, Tsinghua University, Beijing

Abstract This paper introduced the partitioning method used in Japanese-to-Chinese machine translation and discussed the following five problems: (1)Determination and cutting of the Japanese clauses; (2)Determination of the Japanese chunks and their attributes; (3)Cutting and analysis of the chunks; (4)Target generation of the chunks; (5)Rearrangement of the sequence of the target chunks; (6) About the ambiguities of the Japanese auxiliary words, etc. Then we designed a parser using this method and implemented it.

Keywords JAPANESE-TO-CHINESE MACHINE TRANSLATION, PARTITIONING METHOD, CLAUSE, CHUNK

一、引言

近年来，国际计算语言学界越来越重视对大规模真实文本的处理[1]，而提高分析器的鲁棒性和开放性是实现这一目标的关键。为了设计出实用的日汉机器翻译系统，我们采取了多重知识与多种方法相结合的分析策略，以期得到较高的翻译效率和较好的翻译结果。我们的系统首先选择了基于动词配价模式、辅以格框架和名词语义分类的语言模型[4]。为了提高了系统的开放性和适应性，又对基于实例的翻译方法进行了研究[5]。虽然系统中已经开发了以上两个分析器，但在对实际文本的测试中也暴露出一些不足之处。

动词配价模式非常清晰地表达了句子的结构，减少了分析和生成过程中的许多歧义，有利于提高译文的质量和翻译速度。但因为其规则的概括能力稍差、匹配机制灵活性较小，

要解决系统的开放性和鲁棒性还有问题。比如，动词的配价模式中主要书写的是句子的骨架成分，但是，日语句子中经常出现由副词构成的状语成分，这些成分在一个句子中可能出现多次，并且可能在句子的任何地方出现，配价模式不可能概括所有这些现象，否则模式的数量将出现爆炸现象，且失去了其格式清晰的优点。再比如，对于出现在谓语动词后面的表示句子时态、语态、体态、语气、肯否定等信息的助动词和补助动词，在动词的配价模式中也不能穷举，这主要是因为这些词可以任意组合，形成各种各样的涵义。而这些信息对于理解句子是非常重要的，忽视了这些信息将无法得到完整的译文。另外，对于日语句子中大量出现的由系助词表示的提示成分，配价模式也无法全部概括。这些系助词还有副助词几乎可以与句子中任何其他成分重叠使用，也可能与句末某些词呼应，改变成分在句子中的句法功能及句子的涵义。

基于实例的翻译方法在很大程度上提高了系统的开放性和适应性，能容纳输入句子的较大变化，但也存在着信息丢失等缺点。它只使用实例库中最相似的实例进行翻译，其他实例中有可能包含对句子残余部分的翻译有效的信息，却丢失了。这样，基于实例的方法也有可能得不到完整的译文。并且，这种以语料库和统计方法为基础的分析方法在某些情况下也难以避免时间和空间的组合爆炸。

研究断段分析翻译方法主要是为了弥补系统中以上两个分析器的不足，进一步提高系统的鲁棒性。主要任务如下：(1)处理句子中可能出现的枝叶成分，如状语等；(2)处理助词主要是系助词与其他成分重叠使用的情况；(3)分析句末的时态、语态、体态、语气等信息；(4)解决日语附属词的多义现象；(5)处理由助词、助动词表现的独立式惯用型；(6)处理助词、助动词同现所构成的分立式惯用型；(7)处理复杂句等。

“断段分析翻译方法”最初是由孙国钦在《速成科技日语》一书中作为一种学习方法提出来的，其核心思想是利用日语中助词、助动词的语法功能，把日语句子切分成小段，然后进行分段分析和生成，即化难为易地进行翻译[6]。受其思想启发，我们进行了日汉机器翻译的断段分析法研究并研制了一个基于断段分析法的分析器。断段分析翻译方法的研究主要有以下几个难点：(1)寻找子句的断点；(2)寻找段的断点；(3)如何生成各段的译文；(4)依什么规则调整各段次序；(5)各段译文如何合并；(6)各子句译文如何合并；等等。下面将对这几个问题进行具体讨论。

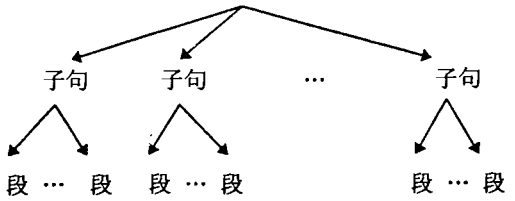
二、断段分析法

断段分析翻译方法先将句子切分为子句，然后再将各子句切分为一个个段。本系统中“句子”、“子句”和“段”的概念是上下级关系，即认为句子由数个子句构成，子句由数个段构成（如图一）。

下面将对断段分析翻译方法中的几个问题进行详细讨论。

A、日语子句的概念及其切分

句子



<句子>:={<子句>}+

<子句>:={<段>}+

图一 本系统中日语句子、子句、段的关系

本系统中“子句”的概念与传统语法中“从句”或“分句”的概念类似，有一点不同的是，传统意义上的“定语从句”在本系统中并不属于一个子句。也就是说，在将句子切分为子句时并不把定语从句切出来。子句的断点结构如下：

<用言活用形>[接续助词|连接词]{[标点符号]}

设计接续助词词典的时候，包含了“左接词活用形”一项，这样，在切分子句的过程中，若句子中出现上述断点结构的单词串，且与接续助词词典中某项匹配，则认为是一个断点。在子句的译文生成过程中，要在子句的最前面添加接续助词对应的汉语词（在词典中为“左加词”项），在子句的最后面也要添加相应汉语词（在词典中为“右加词”项）。

比如下面一个例子： 日本では,この日,子供たちがねがいごとをかみにかいて,竹につるす.

分词结果如下：

日本 N で XN は XN,T この L 日 N,T 子供 N たち N が XN ねがいごと N を XN かみ N に XN かく VT7 て XC2,T 竹 N に XN つるす VT2.T

共有两个子句，其断点分别为： かく VT7 て XC2,T 和 つるす VT2.T

译文如下： 在日本，这天，孩子们写愿望在纸上，挂在竹子上。

B、基于日语句节的日语段的概念及其属性

“段”是断段分析法中的一个基本概念。本系统中“段”的定义是建立在日语语法中关于句节的定义的基础上的。

日语句节的定义如下： <句节>:=<独立词>{<附属词>}*

日语段的定义如下： ①段由一个或多个句节组成；②表示并列和限定的句节与其后面的句节不断开，其余句节与其后面的句节断开。

之所以这样定义，一方面因为表示并列和限定的助词对应汉语没有左加词，另一方面也因为只有这样定义才能保证段的独立性，参见D。

段有如下属性：

①有种类之分：每个段有一个类型标识，表明该段是什么句法成分。不同类型的段在生成时存在着细微差别。

②有隶属关系：其他类型的段都隶属于某个谓语段。段的次序调整是在该段与其所隶属的谓语段之间进行的。

③有层次性：定语从句中各段的层次比其所限定的段的层次低一级。在生成过程中非0层次的段译文要调整至其所限定的单词的紧前面。在最终的译文结构中，低层次段浓缩为一个节点，被高一层屏蔽。

④有序性：其他类型的段与谓语段有前后位置关系。每段有一个是否调位的标识，取值为真则进行次序调整，否则不调整。

C、日语段的切分及其属性识别

段的切分从后往前进行，这是由日语的特点决定的。根据以上对于段的定义，段的切分大体指的是句节的切分。由于日语句节的组成具有鲜明的特点，即总是由一个独立词或一个独立词加数个附属词组成，所以，可以根据日语词的词性辅以语义分类（参见[11][12]）来进行断段。系统中断段的依据是附属词词典，在设计词典的过程中，包含了切分段、分析段属性、段次序调整等信息（参见[7]）。下面是助（体）词词典的格式及样例：

/————————条 件 部 分————————\ /————结论部分————\

词形	词性	前接体言语义属类	用言类型	时态	体态	语态	语气	肯定/否定	语法功能	左加词	右加词	是否调位	副影响

“前接体言语义属类”指该助词所附属独立词的语义属类。

“用言类型”指该助词所在子句的谓语用言的语义属类。

“时态”、“体态”、“语态”指该助词所在子句的谓语的态（由谓语用言后面的助动词、补助动词来表示）。

“语气”指该助词所在子句的语气（由句末终助词来表示）。

“肯定/否定”指该助词所在子句是肯定句还是否定句。

“语法功能”指该助词表示什么句子成分。

“左加词”、“右加词”指生成汉语时在独立词的前面和后面分别所需添加的相应的汉语译词。

“是否调位”指该助词所属段在生成汉语时是否需要调到谓语段的后面。

“副影响”指该助词对谓语的态、子句的语气及肯/否定信息是否造成反面影响。

格助词	义项号	前接体言语义属类	谓语用言类型	时态	体态	语态	肯/否定	语法功能	左加词	右加词	是否调位	副影响
…												
で	1	场所	/	/	/	/	/	状语	在	/	否	无
で	2	交通工具1	/	/	/	/	/	状语	乘	/	否	无
で	3	交通工具2	/	/	/	/	/	状语	骑	/	否	无
…												
に	1	场所|地点	存在|静止	/	/	/	/	状语	在	/	是	无
に	2	场所|地点	移动	/	/	/	/	状语	/	/	是	无
…												
に	11	人|动物	/	/	/	被动	/	状语	被	/	否	有
に	12	人|动物	/	/	授受	使役	/	状语	请允许	/	否	有
に	13	人|动物	/	/	/	使役	/	状语	让	/	否	有
…												

需要说明的一点是，碰到表示并列和限定的助词时并不断段；其他助词与其重叠使用时同样也不断段。比如：　　私 は / 日本 と アメリカ へ / 行った /.

↑

此处不切分

再比如：　　これ は / 日本 へ の ひこうき / だ /.

↑

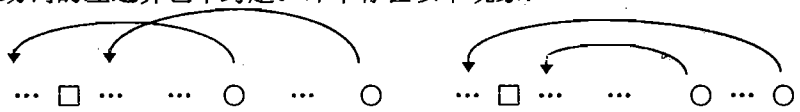
此处不切分

段的隶属关系的确定是一个核心问题，它会影响到段的次序调整。本系统中搜索各段所隶属谓语段的过程简述如下（可参考[8]对这个过程做进一步完善）：

(1)碰到主语段，需跳过一个谓语段；
(2)碰到宾语段，跳过所有由自动词表示的谓语段；
这样找到的第一个谓语段就是该段所隶属的谓语段。

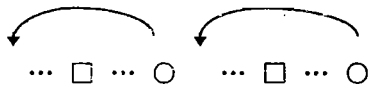
D、独立日语段的译文生成

段的生成中主要难点是判断日语助词的左边界。日语助词的左边界不交叉。如果把表示并列和限定的助词除外，并且不考虑其他助词与表示并列和限定的助词重叠使用的情况，日语助词的左边界也不跨越。即不存在以下现象：



其中，□：独立词　○：助词

这样，在除去表示并列和限定的助词及其他助词与这两种助词重叠使用的情况之后，助词的左边界只可能是依次顺序排列，如下所示：



从后往前搜索助词左边界时，碰到表示限定的助词肯定跨越。当碰到表示并列的助词时要进行相似度的计算，再根据计算结果确定是否跨越。相似度的计算请参考[5][9]。

基于以上分析，段的生成可从前往后进行。先找到第一个非表示并列或限定的助词，作一标记，寻其左边界，在此添加其左加词，顺序生成中间所有词的译词，再添加其右加词；从标记处继续往后，重复以上操作，直至段尾。

谓语段的生成较特殊，即：首先添加终助词的左加词，接着按照从后往前的次序依次添加各助动词的左加词，再生成用言的译词，然后又按照从前往后的次序依次添加各助动词的右加词。如下所示：

原文结构：　□　○$_1$　○$_2$ … ○$_N$　/　⊙

其中，□：用言　　⊙：终助词

○$_I$：第I个助动词或补助动词　　/：段分隔符

译文结构：〈⊙　〈○$_N$ … 〈○$_2$　〈○$_1$　□　○$_1$〉　○$_2$〉 … ○$_N$〉　/　⊙〉

其中，□：用言译词

〈○$_I$：第I个助动词或补助动词的左加词　　〈⊙：终助词的左加词

○$_I$〉：第I个助动词或补助动词的右加词　　⊙〉：终助词的右加词

E、日语段译文的次序调整

段的次序调整，是在某个非谓语段与其所隶属的谓语段之间进行的。首先从层次最低一级开始调整，调整完成后，进入次低一级调整，如此下去，直至最高层。比如：

原　　文： 彼は日本をほとずれる外国の旅行者のために便宜をはかっている。
分词结果： 彼 P は XN 日本 N を XN ほとずれる VT3 外国 N の XN 旅行者 N のために XN 便宜 N を XN はかる VT7 ている XV2 。 T
分段层次： 彼は / 日本を / ほとずれる / 外国の旅行者 のために / 便宜を / はかる ている / 。
各段译文： 他 日本 访问 为外国(的)旅客 方便 提供着 。
次序调整1：（注意：仅在低一级进行） 他 访问 日本 为外国(的)旅客 方便 提供着 。
次序调整2：（注意：在高一级进行） 他 访问 日本 为外国(的)旅客 提供着 方便 。
定语从句放在所修饰词的前面： 他 为外国(的)旅客 提供着 方便 。
↑ 访问 日本 (的)
译　　文： 他 为 访问 日本 (的) 外国(的)旅客 提供着 方便 。

F、日语附属词的多义问题

日语中大量存在的具有语法功能的一类词即附属词，确实给日语的分析带来很大方便，但由于该类词中某些词也存在着多义现象，必须予以考虑。为了解决这个问题，我们在设计词典格式时考虑了上下文信息（参见[7]）。

另外，在书写词典时引入了模糊思想，即：在选择日语附属词所对应的汉语译词时，尽量选择那些概括性强的词；若有些字可能有也可能无，则可用括号括起来，放在译文中用户自己可做出判断（比如，日语中的格助词“の”在生成汉语时，根据汉语的表达习惯可能加“的”字，也可能不加[10]，译词就可书写为“（的）”）；在有些很难对付的情况下，括号中甚至可以出现候选译词；等。

三、分析器设计

断段分析翻译器的结构图如图二所示。分析器的运行机制如下：

⑴基于接续词词典，从左向右扫描一遍句子，寻找出子句的断点，将各子句的起始位置及结束位置记录入Section结构中的相应槽；

⑵对各子句进行⑶～⑬处理；

⑶从右往左扫描当前子句，分析句末终助词，将语气信息记录入Section结构中的语气槽；

⑷分析各助动词（若出现），将时态、体态、语态、肯/否定等信息记录入谓语段的相应槽中；

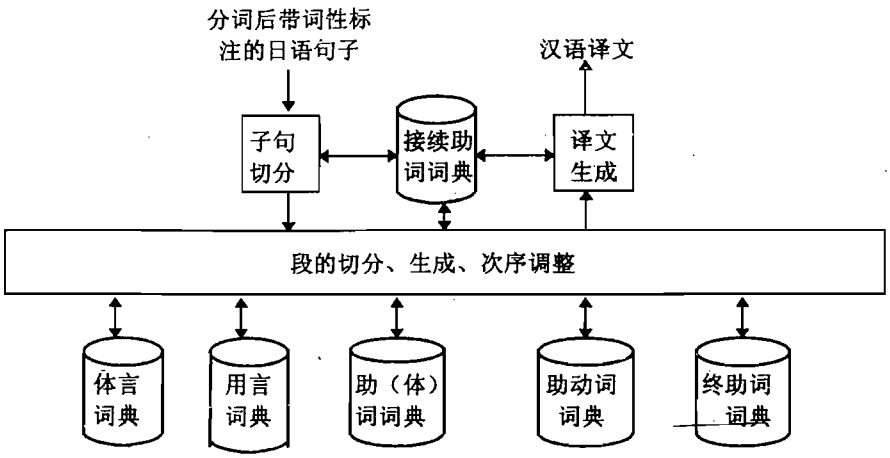
⑸若出现谓语动词，记录谓语段的起始、终止位置等信息；

⑹分析格助词，将句子中格助词的左邻接词与词典中该格助词各义项的相应条件项进行匹配；

⑺往右寻找该段所隶属谓语段；

⑻将该谓语段所记录的时态、体态、语态、肯/否定等信息与格助词词典中的相应条件项进行匹配；

⑼若匹配成功，记录该格助词被选中的义项号，将该义项中的成分类型、调位信息、副影响等记录入Chunk结构中；



图二 断段分析器的结构图

⑽重复从⑹执行，直至子句最左端；

⑾从左向右扫描当前子句各段，根据各段的成分类型生成其译文，格助词的左、右加词也在此时添加，并且根据副影响信息作相应修正；

⑿根据各段调位信息，从最低级至最高级进行次序调整；

⒀将同层次各段译文合并，层次非0的段放在其所修饰词的紧前面；

⒁添加接续助词及连接词的左、右加词，将各子句译文合并；

⒂输出译文。

四、实现结果及翻译实例

系统基本词典的录入工作正在进行中，其他各部词典的大小如下：接续助词词典共有158个词条，助（体）词词典约有260个词条，助动词词典有240个词条，终助词词典有30个词条。我们对一些包含分立式惯用型的句子进行了测试，并从《标准日本语》（中级Ⅰ）中选择了一篇文章“七夕”（共47句）进行了翻译，从翻译结果看，系统对日语原文的分析大体正确，并且译文通顺、机器味少。以下是几个翻译实例：

このL 日N は XN 女性 N にかぎって XN 休む VI3 ことができる XV2 。T
===>这（个）天只限女性能够休息。
日N が XN たつ VI3 につれて XC2, T なやみ N が XN 増える VI7 てくる XV7 た XV2 。T
===>随着天推移，烦恼增加起来（了）。
天女N の XN 服N が XN おく VT7 てある XV2 から XC2， T 赤い AJ3 服 N を XN 持つ VT7 て XC2， T 隠れる VI1 なさい XV2 。T
===>仙女（的）衣服放着，拿红色（的）衣服，请隐藏。
そのL 服N の XN 持ち主N は XN あなた P の XN お嫁さん N に XN なる VI3 人N だ XV2 。T
===>那（个）衣服（的）主人是变成（的）你（的）新娘子（的）人。
次の日 D 若者 N が XN 湖 N へ XN 行く VI7 てみる XV3 と XC2， T 牛N が XN 言う VT7 た

XV3 とおり XC2，T 天女 N たち N が XN 水浴び N を XN する VT7 ている XV7 た XV2 。T ===>第二天青年往湖去一看，正如牛说（了），仙女们正在做着（了）洗澡。

五、结束语

断段分析翻译方法面向最一般的句子、长难句与新句式，将其切分为小的段进行分析和翻译，做到了化难为易、化繁为简，为提高机器翻译系统的开放性和鲁棒性探索出了一条新路子。不过，断段分析翻译方法需要上一级日语分词的支持，分词的质量（包括词性标注）直接影响到分析的正确率，应设法增强抵抗分词干扰的能力。

参考文献

[1]. 黄昌宁，“关于处理大规模真实文本的谈话”，《语言文字应用》，1993年第2期。

[2]. 曹睿妮，《实用型日汉机器翻译支援系统的设计与实现》，清华大学计算机系毕业设计论文，1994年6月。

[3]. 陈利人，“实用型日语自动分词系统的研究与初步实现”，清华大学计算机系毕业设计论文，1993年6月。

[4]. 孙勇、陈群秀，“基于动词配价模式日汉机器翻译系统的设计与实现”，全国第三届计算语言学联合学术会议论文集《计算语言学进展与应用》，1995年，上海。

[5]. 陈利人、陈群秀，“基于实例的日汉机器翻译方法中的句子相似度计算研究”，全国第三届计算语言学联合学术会议论文集《计算语言学进展与应用》，1995年，上海。

[6]. 孙国钦，《速成科技日语》，天津科学技术出版社，1980年。

[7]. 曹睿妮、陈群秀，“日汉机器翻译系统用日语附属词词典的设计与实现”，第七届日本学中日学术交流会论文，1996年10月，北京。

[8]. Yeun-Bae Kim and Terumasa Ehara, “A Method for Partitioning of Long Japanese Sentences with Subject Resolution in J/E Machine Translation”, Proceedings of the 1994 International Conference on Computer Processing of Oriental Languages, May 10-13, Taejon, Korea

[9]. Sadao Kurohashi and Makoto Nagao, “A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures”, Computational Linguistics, Jun 1992

[10]. 陈群秀、李咏玖，“日汉机译系统中有关汉语生成的几个问题及处理方法”，全国首届计算语言学联合学术会议论文，1991年，杭州。

[11]. 陈群秀，“有关语义分类体系研究的几个问题”，92年全国机器翻译学术会议论文集《机器翻译研究进展》，1992年。

[12]. 陈群秀、张普，“信息处理用现代汉语语义分类体系（之一）：属性分类”，全国第二届计算语言学联合学术会议论文集《计算语言学研究与应用》，1993年，厦门。