面向机器翻译的日语形态素解析方法

隋福民 黄德根
大连理工大学计算机系 大连 116024
E-mail: fuminsui@sina.com

摘 要：日语形态素解析是日文信息处理特有的研究课题，也是语言信息处理领域中最成熟的技术之一。针对机器翻译的特点，提出了一种面向机器翻译的日语形态素解析方法。该方法采用最长次长匹配法建立切分路径有向图，运用用言活用匹配及品词接续等语法规则进行歧义的解消。

关键词：日语形态素解析，最长次长匹配法，机器翻译

MT-Oriented Japanese Morphological Analysis

Sui Fuming Huang Degen
Department of Computer Science and Technology, Dalian University of Technology, Dalian 116024
E-mail: fuminsui@sina.com

Abstract: Morphological analysis is a particular research subject in Japanese information processing. It is the foundation of machine translation, information retrieval and document classification. According to the feature of machine translation, the paper presents a implementation of Japanese morphological analysis based on maximum matching and second-maximum matching.

Keyword: Japanese morphological analysis, Maximum matching and second-maximum matching method, Machine translation.

1 引言

形态素(morpheme)[1]是日语中具有最小意义的语言单位，日语中的单词由一个或多个形态素组成。日语形态素解析是指利用形态素词典及形态素构词规则，把日语句子划分成单词序列，并判定各个单词在句子中作用的过程。日语形态素解析主要有三个问题：切分单词、确定品词（类似汉语中的词性）、未登录词的识别处理。

至今为止，研究者已经提出了许多形态素解析方法，如最小代价法[2]、最小分割法、基于马尔可夫模型的方法[3][4]等，并且发布了 chasen[5]这样的日语形态素解析系统。对于句子「昨日は休日ではありませんでした」，chasen 的解析结果如图 1 所示。

昨日	昨日	名詞-副詞可能		
は	は	助詞-係助詞		
休日	休日	名詞-副詞可能		
で	だ	助動詞	特殊ダ	連用形
は	は	助詞-係助詞		
あり	ある	助動詞	五段ラ行アル	連用形
ませ	ます	助動詞	特殊マス	未然形
ん	ん	助動詞	不変化型	基本形
でし	です	助動詞	特殊デス	連用形
た	た	助動詞	特殊タ	基本形

图 1 chasen 解析结果

昨日	昨日	时间名词
は	は	系助词
休日	休日	一般名词
ではありませんでした	ではありませんでした	助动词 终止形

图 2 机器翻译系统期望的解析结果

单纯从解析精度来看，以 chasen 为代表的日语形态素解析系统已经达到很高的精度，并在信息检索、文本分类等领域得到广泛应用。通过分析比较，发现现有的日文形态素解析系统的输出结果颗粒度很细，它满足了信息检索、文本分类等应用领域的要求。但对于机器翻译系统而言，颗粒度过细，会导致部分单词失去原有的意义，增加了其后句法分析的复杂度。例如，上述例句中的「ではありませんでした」，chasen 将其解析为 7 个形态素单元，给句法分析、中文译语的生成增加了难度。事实上，对机器翻译系统而言，「ではありませんでした」等助动词作为一个形态素单元来处理更为理想。

2 形态素词典的设计

2.1 日语单词的特点

与汉语不同，日语单词有词尾变化。我们把单词中不变的部分称为“词根”，把有变化的部分称之为“词尾”。例如：单词「起きる」，词根是「起き」，词尾是「る」；而单词「言葉」，只有词根「言葉」，没有词尾。根据单词是否存在词尾变化，我们把单词分为两大类：有活用和无活用单词。有活用单词包括形容词、动词、形容动词、助动词四类，其词尾活用形可以是未然形、连用形、终止形、连体形、假定形、命令形和推量形；除此之外的单词为无活用单词。

按照品词的不同，我们把单词分成如下大类：

<单词>::=<独立词>|<附属词>

<独立词>::=<动词>|<形容词>|<形容动词>|<名词>|<代词>|<数词>|<量词>|<数量词>|<副词>|<连体词>|<接续词>|<感叹词>|<接头词>|<接尾词>|<独立语>

<附属词>::=<助词>|<助动词>

对于日语中的常用短语、惯用语等，从传统的语法角度来说，虽然它们不属于日语品词体系中某一类品词，但是，根据机器翻译处理的需要，也把它们作为一个单词登录到词典中。例如，起名词作用的「いよいよというとき」、「もしものこと」等作为一个名词登录；起动词作用的「いよいよとなる」等作为动词登录；起助词作用的「という目的で」等作为助词登录；在句子中起谓词作用的「ではありません」、「ではいけません」等作为助动词登录。

在本文提出的形态素解析中，单词和单词之间存在的接续关系是以品词来描述的，品词分类越细，单词接续规则所达到效果也越好。为此，我们对动词、名词、代名词、助词、

接头词、接尾词、连体词、形容动词、副词等品词进行了更进一步的分类，并称之为子品词。

2.2 形态素解析词典

考虑到上述特点，在词典中，我们包括了下面的信息：

单词词根，词尾，品词，子品词，概念分类，特征，中文译语。

- 单词词根：为单词中不变的部分，如单词「食べる」的词根为「食べ」。
- 词尾：单词中可发生活用变化的部分，如单词「食べる」的词尾为「る」。
- 品词：品词相当于汉语中的词性。一个单词可以具有多个不同的品词，不同品词的次序是根据各个品词出现的频率由大到小排序的。主要分为：〈动词〉、〈名词〉、〈形容动词〉、〈形容词〉、〈代词〉、〈数词〉、〈量词〉、〈数量词〉、〈副词〉、〈连体词〉、〈接续词〉、〈感叹词〉、〈接头词〉、〈接尾词〉、〈独立语〉
- 子品词：为了进行严格的接续检查，对部分品词进行了更进一步的细分类。比如名词可细分为：一般名词、形式名词、时间名词、方位名词、自动词サ变名词、他动词サ变名词、自他动词サ变名词、固有名词、一般疑问名词、时间疑问名词、方位疑问名词
- 概念分类：主要根据「日本語の意味分類体系」[6]对单词的概念进行分类。
- 特征：是一个 32bit 的二进制数，用于表示某一类品词与其他品词的前后接续情况。
- 中文译语：日语单词所对应的中文译语。

3 活用匹配和接续检查

形态素解析过程中除了需要使用词典外，还需要形态素相关的语法知识，主要指活用知识和接续知识。无论采用统计还是规则的方法，一般都需要对单词进行活用匹配和接续检查，活用匹配和接续检查直接影响解析的精确度。

3.1 形态素解析过程

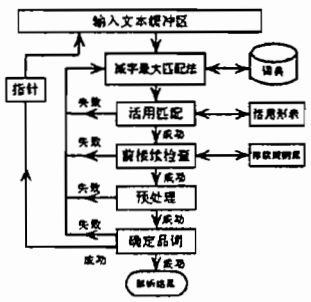

图 3 形态素解析过程

最长匹配法是一种得到广泛应用的分词方法。考虑到日语句子中的单词一般和前后单词都有活用、接续的关系，在形态素解析中，引入了预处理过程，其解析过程见图 3。

预处理是指在当前单词的活用匹配和前接续检查均成功的情况下，对其右边的单词进行预处理，得到当前单词右边的单词。若当前单词和预处理得到的右单词存在接续关系（称之为后接续），则认为当前单词可能是正确的；若不存在接续关系，则判定当前词是不正确的，返回处理错误标志，对当前词重新进行减字最大匹配。

3.2 活用匹配

活用匹配是针对有活用单词进行的，对匹配成功的品词确定活用类型，对匹配失败的，删除该品词。如果单词中所有品词都被删除，则认为当前单词活用匹配失败。进行匹配的时候，首先通过单词信息中的词尾找到词尾原形表中的位置，然后将单词的实际词尾与该活用变化行中的每一项进行对照，若匹配成功，则将设置相应的用言活用类型。例如，对句子「行きますか」进行活用匹配时，首先通过查词典，得到词根为「行」，词尾为「く」的单词(即动词「行・く」)，由单词信息知道「行・く」为五段か行动词，查找下述五段か行动词的活用表，即可得出「き」为动词「行・く」的连用形活用词尾。

{“か”，“N”，“き”，“い”，“く”，“く”，“け”，“け”，“N”，“こ”}

3.3 接续检查

接续检查主要有两个功能：一是可以排除大部分不正确的切分，包括删除含有多品词单词中不正确的品词；二是为文节划分和后续的文节系受关系分析提供依据。

根据品词接续的强弱，将相邻品词之间的接续关系分为“确定的接续”和“不确定的接续”。如果当前品词的接续是“确定的接续”关系，但是和后边相邻单词的品词不接续，那么，就认为接续失败，即当前品词错误。例如：接头词的接续关系是“确定的接续”关系，它必须有后接续；副词的品词接续是“不确定的接续”，在句子中它不一定与后边相邻的品次有修饰关系。

接续检查，一部分是通过品词接续规则来实现的，如：

“形容词连体形＋名词”

是可接续的规则，这些规则适合具有某一类品词的所有单词。但是，有些接续并不适合某一类品词的所有单词。这种情况无法采用上述通用的接续规则，而必须针对具体的单词来建立，为此，在单词信息中建立 32 位的特征字，利用特征字来控制该单词是否可以与其它品词接续。目前，本文主要对助词、助动词、接尾词等数目较少，且接续不规范的单词设置了特征位。以助词「て」为例，其特征字的值为 0x00002180，从助词特征位的定义可知，助词「て」允许前接形容词连用形、动词音变连用形、动词连用形。助词特征字的定义如下：

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16
							w	c	u	y	d	p			g

g 为形容词语干

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
a	a	a	a	v	v	v	v	v	v	b	m	q	r	艹	n

假定 音连 连用 原型 推量 假定 原型 音连 连用 未然

3.4 确定品词

在活用匹配、接续检查及预处理之后，还可能保留下来多个“合法”的品词，在确定文节之前需要选择其中一个正确的品词。确定品词实际上就是根据品词活用及前后接续等信息对同根异形品词进行选择。它的难点在于未登录词的干扰以及具有“不确定的接续”关系品词的干扰（主要是助词、副词、时间名词、形容动词等品词）。

4 最长次长匹配的形态素解析模型

4.1 有向图路径的选择

任意一个句子，经最长次长匹配进行切分后，可得到一个有向图[7]。例如，对于句子「この問題はある出来事に関連している」，其解析后的切分路径有向图如图 4 所示，该切分路径有向图中存在 6 条可能的切分路径，但其中只有 1 条是正确的，因此需要遍历该有向图，选择正确的切分路径。

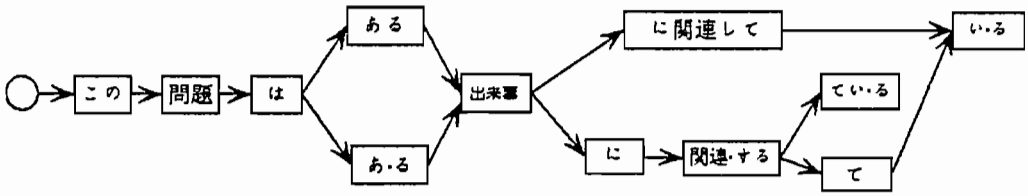

图 4 基于最长次长匹配的形态素解析

为叙述方便，取有向图的一个子图进行描述。设 N 为有向图中的某一结点（即切分序列中的某一可能的切分字段）所代表的单词，从单词 N 的结束位置进行最长、次长切分，分别得到最长单词 FN 和次长单词 SN；又分别从 FN 及 SN 出发，进行最长、次长切分，分别得到 FN 的最长单词 FFN 及次长单词 FSN，SN 的最长单词 SFN 和次长单词 SSN，同样可得 SFN 的最长单词 SFFN 和次长单词 SFSN。

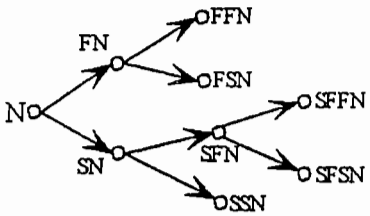

图 5　切分路径有向图的子图

设函数 Len（N）为节点 N 所指单词的长度，有向图路径选择算法如下（以图 5 有向图的子图为例）：

（1）　若 N 已经是终止结点，则处理结束。

（2）　若 SN 为空（只有最长单词），则选择指向 FN 结点的路径，转（7）；若 SN 不为空，

则执行下一步

（3） 若 SFN 不为空，且 (Len(FN) == Len(SN)+Len(SFN))，则可能存在组合歧义，由组合型歧义处理模块检查是否有满足条件的规则，若没有满足条件的规则，则认为不是组合歧义，选择指向最长单词的结点 FN；若存在满足条件的规则，则按规则选择路径。组合歧义处理结束后转(7)。

（4） 若 Len(FN)+Len(FFN)< Len(SN)+Len(SFN))，根据路径最短法原则，选择指向 SN 和 SFN 的路径，转（7)。

（5） 若 Len(FN)+Len(FFN)=Len(SN)+Len(SFN))，表示存在交叉歧义型歧义，依据交叉型歧义处理规则选择路径，交叉型歧义处理结束后转（7)。

（6） Len(FN)+Len(FFN)= Len(SN)+Len(SFN) +Len(SFFN)，且 Len(FFN)≠Len(SFFN)，则可能存在非对称的交叉歧义，需要进行交叉型歧义处理。

（7） 结点 N 指向新确定的结点，转（1）重新进行路径选择。

4.2 交叉型歧义的处理

图 7 是一种典型的交叉型歧义切分，从单词 N 经最长次长匹配后，得到最长单词 FN 和次长单词 SN，单词 FN、SN 经最长匹配后切分出来的单词 FFN 和 SFN 有相同的终点，出现了 FN+FFN=SN+SFN 的交叉型歧义。

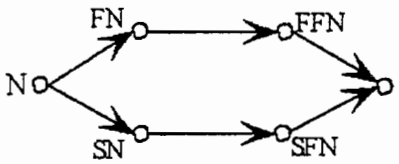

图 7 交叉型歧义

在日语中，不同品词间具有不同的接续强度，通常动词词干和词尾的接续强、名词和副词的接续弱[8]。交叉型歧义处理就是在图 7 中选择出正确的路径。路径的选择主要根据单词的形态素信息和品词接续、用言活用等制定的规则进行处理。根据品词接续强度的优先级，制订了路径选择规则，举例如下：

1) 若歧义路径的接续分别是“助词を+名词+助词に”、“助词を+连用形+助词に”类型的，则选择接续为“助词を+连用形+助词に”的路径。

2) 若歧义路径的接续分别是“连体词+助动词そうだ”、“动词+助动词そうだ”类型的，则选择接续为“动词+助动词そうだ”的路径。

3) 若歧义路径的接续分别是“接头词+动词连用形+ になる”、“接头词+名词+になる”类型的，则选择接续为“接头词+动词连用形+になる”的路径。

4.3 组合型歧义的处理

组合型歧义的处理需要利用语法规则和上下文信息，目前主要是针对歧义出现的上下文单词及品词信息建立的规则进行处理。例如："名词性短语 + でもいい"，选择"名词性短语 + でも + いい"，即次长路径。如果没有相匹配的规则，则取最长路径。

5 实验结果与展望

本文提出的日语形态素解析是面向中日机器翻译应用的解析方法。考虑到颗粒度过细的解析结果会增加句法分析汉语生成的复杂度，提出一种粗粒度的日语形态素解析方法。

系统所建立的规则主要依据《标准日本语》中所出现的语法，经闭式测试，精确率达到 99%（无未登录单词）。为了比较，我们从《朝日新闻》上随机抽取了 3000 多字的新闻语料作为开式测试集，如果不考虑未登录词对切分结果的影响，开式测试的精确率达到 97% 以上，若计算未登录词所造成的切分错误，开式测试的精确率为 93.10%。

参考文献

[1] "形态素解析"，神奈川工科大学情報工学科石井研究室，1999。

[2] 小松英二、安原宏："コスト最小法形態素解析のコストルールの作成方法"，《自然言語処理》，No.085-001.2001。

[3] 浅原正幸、松本裕治："統計的日本形態素解析に対する拡張 HMM モデル"，《自然言語処理》，No.137 - 006.2001。

[4] 北内啓、宇津呂武仁、松本裕治："誤り駆動型の素性選択による日本語形態素解析の確率モデル学習"，Vol.40 No.05 - 041.2001。

[5] 松本裕治、山下远雄等："日本语形态素解析システム「茶筌」"，version 2.0 使用说明书第二版，1999 年 2 月。

[6] 荻野孝野："日本語の意味分類体系"，《計量国語学》，Vol16.No.3.1987。

[7] 黄德根、朱和合、王昆仑："基于最长次长匹配的汉语自动分词〔J〕"，《大连理工大学学报》，Vol.39 No.6 1999。

[8] "コスト最小法による曖昧性の絞り込み"，日本語形態素解析システム Maja。