基于依存关系分析的日语句法分析器*

简幼良　唱红涛　王秀坤　黄德根
（大连理工大学计算机技术研究所　116024）

摘　要：本文介绍的日中翻译句法分析器，提出以依存关系分析为基础，利用多种知识来消除歧义的方法．在必要时可选择优先解，并保留必要的候补解．该分析器已应用于一个商品化的日中机器翻译系统"CJ-TRANS"系统中，提高了系统的分析能力和鲁棒性．

关键词：机器翻译　分析器　依存关系　歧义

A Dependency-relation-based Parser to Japanese-Chinese Machine Translation

Jian Youliang　Chang Hongtao　Wang Xiukun　Huang Degen
(Research Institite of Computer Technology, DUT)

Abstract: This paper introduces a parser of Japanese-Chinese machine translation. A method of elminating ambiguity that used knowledges and based on dependency relationis is presented. The prior solution can be selected when necessary, and alternate solution is also reverved. This parser is applied to the pratical Japanese-Chinese translation system(CJTRANS), and raised efficiency of analysis.

Keyword: machine translation　parser　dependency relation　ambiguity

一、引言

日语和中国语虽然都使用汉字，但是从语言学上来说，它们所属的语系不同．汉语是孤立语，没有词形变化．词与词间的关系，主要通过词序和部分虚词来表达．而日语则属于粘着型语言，有词尾变化，附属词很丰富，词序比较自由，词和词间的关系主要通过附属词来表达．

在日语的机器翻译中基本上采用桥本文法[1]．该文法认为，句子的直接构成要素是句节，句节的构成要素是单词．句节的构造可表示如下：

〈句节〉::＝＜独立部＞|〈句节〉(〈附属部〉)　〈附属部〉::＝附属词＊

〈独立部〉::＝(〈接头词〉)〈独立词〉(〈接尾词〉)　()表示可以省略，＊表示可以有多个．

在句节中的独立词主要提供概念信息．而附属词主要提供句节间依存关系，或说话者的判断、叙述方式等构造意义上的信息．一个句节是以"对"的形式，提供句法结构和语义这两种

＊ 本文受国家自然科学基金资助

信息的单位．因此，在日语的句法分析中，采用依存关系分析是比较适合的．本文主要讨论如何根据日语句节的性质，来计算出它们之间的依存关系，以及消除依存关系上的歧义．

本文采取的方法是，先计算出句中各句节间在满足依存关系公理条件下的一切可能的依存关系的组合，然后，利用多种知识对依存关系上出现的歧义逐个进行解消．一个歧义的消除可能同时消除多个歧义．到最后，如果有的歧义实在无能为力，可以选择最优解，保留其它的关系作为候补．其优点是保证分析不致于失败，而且还可以提供系统回朔的可能．

二、依存关系分析

2.1 依存公量

依存文法体系由四大公理支持，结合日语特点，日语句子中的依存关系具有如下性质：

(1) 只有句末句节是独立的，它不依存于任何其它句节．

(2) 句中各句节(句末句节除外)能且仅能依存于其后方的某一句节，但可以承接它前方的多个句节．

(3) 一个句子中各依存关系彼此不能相互交叉．即如果将各句节间的依存关系用从系侧向受侧画一方向弧来表示，那么这些方向弧不能交叉．

(4) 一个句节不能承接多个具有同一格的句节．

合理的日语句子都应满足上述的4条依存性质，它们构成了句子文法结构的制约条件．

2.2 依存关系分类及依存矩阵

依存关系分连用性和连体性关系两大类，从机器翻译句法分析的立场出发，有必要将这些关系根据构成句节的句法地位进行细分类．依存关系的分类如表1所示．

表1 依存关系分类

依存关系			例文
连用关系	连用修饰	连用格关系	花が咲いた；私が言う
		连用样相关系	深刻に悩む；美しく咲く；
		连用副词修饰关系	きつと伺う；そとて笑った；
	连用接续关系		見て驚く；起こると困る；
	连用并列关系		明るく正しい；
连体关系	修饰	连体格关系	热いお茶；走る车；
		体言一体言关系	私の本；
	连体并列关系		AかBか；AとB；

对于一个句节的依存关系的性质，在日语的句节切分阶段，就尽可能地利用独立部和附属部的语法关系加以区分．

对于独立部从体言性和述语性来考虑，分成以下两类．

n：担负名词性概念的表现(如“花”，“仕事；，“談み．こと”，“難し．さ”等)

p:担负述语性概念的表现(如“来る”,“勉強.する”,“学校.である”,“きれい.だ”等)

N:以 n 为独立部的句节(如“人/に”,“する.の/さえ”,“車/関して”,等)

P:以 p 为独立部的句节(如“美しかっ/た”,“飲む/のか”,“本である/か”等)

对于担负结构信息的附属部,包括助词,助动词,补助用言以及相当于助词,助动词的惯用复合表现,根据节间依存关系,和述语内的接续和活用的性质加以分类.

(1) 关系的表现

关系表现主要出现在 N 或 P 句节的附属部,指示本句节依存于句中的后继的哪种类型的句节. 具体分类为:

Rnp1 指示 N 依存于 P 的格助词或副助词表现(如“ご飯/を 食べる”,“私/も 言う”)

Rnp2 指示 N 依存于 P,并把前面的 n 暗化为 p 的表现(如“美人/なので”)

Rpp1 指示 P 依存 P 的接续助词表现(如“見/て 驚く”,“する/ために”)

Rpp2 指示 P 依存于 P 的引用关系的表现(如“良い/と 思う”)

Rpp3 指示 P 和 P 的并列关系的表现(如“読む/し 書く”)

Rnn1 指示 N 和 N 的连体性的修饰表现(如“私/の 本”,“外国/たおける”)

Rnn2 指示 N 和 N 的并列关系的表现(如“山/と 川”)

Rpn 指示 P 和 N 的连体性的修饰表现(如“上る/といった 現象”)

(2) 助述表现

助述表现包括助动词(补助用言)和终助词,它们从句末侧接续到述语上. 它们出现在述语的独立词之后,除表达说话者的判断和叙述方式外,它对句节性质有影响,也指示句节间的关系.

App1 可以连接到 P 上的助动词的活用或非活用表现(如“聞い/たことがある”)

App2 可以连接到 p 上的终助词的非活用表现(如“走る/か”,“食べ/なさい”)

Anp1 可连接到 n 上,暗化 n 为 p 的助动词的活用表现(如“人/かもしれない”)

Anp2 可连接到 n 上,暗化 n 为 p 的终助词的非活用表现(如“システム/か”)

在句节切分结束时,由它本身的性质(如 Rnp 等中的第一个字母),可以确定它所承接的前方的句节的类型. 由它的依存性质(如 Rnp 等中的第二个字母)可以确定它向后依存的句节类型(有少数句节既可以依存于体词性句节,也可以依存于述语性句节).

2.3 满足依存公理的依存结构分析

我们要求得的句子的依存关系,是每个句节只存在一个向后的依存关系. 现在的问题是,在依存歧义的情况下,以依存公理中头三条为制约条件,如何计算出所有可能存在的依存树.

〈算法分析与描述〉

设欲分析的句子为 S,从句首起句中各句节的句节号依次为 i=1,2,…,n. 根据各句节的分类,可以确定 i 句节与它后面的句节 j(j=i+1,…,n)的依存关系 a_{ij}。如存在依存关系,则 $a_{ij}=1$,否则 $a_{ij}=0$. 这样得到一个依存矩阵 $M(S)=(a_{ij})$. 从日语依存文法特点——依存的后方性可知,依存矩阵是对角线以下元素全为 0 的上三角矩阵.

[句子 S 的部分句]

从句头开始数,第 i 个句节到第 j 个句节为止(i<=j)的部分列,称为 S 的部分句. 记为

S(i,j). 则 S=S(1,n)(n=S 中句节的数目)

[依存结构]

部分句 S(i,j)的一棵依存树记为 D(i,j)，它是满足依存文法的非交叉条件限制的句节 i 到句节 j 之间的依存关系的集合，由 M(S)的要素构成，且满足下面两个条件：

$D(i,j)=\{a_{l,kl}\}$

(a) $a_{l,kl}=1$，$l<kl<=j, l=i,i+1,\cdots,j-1$

(b) D(i,j)的所有要素都是非交差的.

S(i,j)的所有依存树 D(i,j)的集合记为 T(i,j)，T(i,j)的浓度|T(i,j)|称为根据非交差条件分析的部分句 S(i,j)的依存结构歧义数. 这里所谓浓度，即是 T(i,j)集合中元素的个数，即依存树的棵数.

如果 M(S)中的元素 $a_{i,k}=1$(存在 i 系于 k 的依存关系)，而且存在部分句 S(i+1,k)的无交叉依存树的集合 T(i+1,k)和部分句 S(k,j)的无交叉依存树的集合 T(k,j)，则可知 T(i+1,k)中的任何依存树 D(i+1,k)中的元素均与 T(k,j)集合中的任何依存树 D(k,j)中的元素满足不交叉条件，且与 $a_{i,k}$ 不相交叉，由此构成的集合 D(i,j)满足条件(a)(b)，是句节 i 到句节 j 的一个棵依存树.

$D(i,j)=D(i+1,k)\cup D(k,j)\cup\{a_{i,k}\}$

可见，句子的依存树可以由部分句的依存树生成.

求解 T(i,j)的递推公式如下：

在 T(i,j)的要素中，含有 $a_{k,l}$ 的要素的集合，用 $T(a_{k,l})(i,j)$ 表示.

$T(i,j)=\bigcup T(a_{i,k})(i,j)\quad(k=i+1,\cdots,j)$

$$T(a_{i,k})(i,j)\begin{cases}\{D(i+1,k)\cup D(k,j)\cup\{a_{i,k}\}\mid D(i+1,k)\in T(i+1,k), D(k,j)\in T(k,j)\} \\ \qquad a_{i,k}=1\ \&\ |T(i+1,k)|>0\ \&\ |T(k,j)|>0 \\ \Phi \qquad \text{其它}\end{cases}$$

$|T(i,j)|=\Sigma|T(a_{i,k})(i,j)|\ (k=i+1,j)$

$|T(a_{i,k})(i,j)|=a_{i,k}*|T(i+1,k)|*|T(k,j)|$

定义 T(i,j)的浓度为 N(i,j)，

$N(i,j)=|T(i,j)|=\Sigma a_{i,k}*N(i+1,k)*N(k,j)\ (k=i+1,j)$

$N(i,i)=1$

$T(i,i)=\{\Phi\}$

根据以上公式可以依次求得长度为 2,3,…的部分句的可能依存树，最终求得全句 S 的依存树的集合 T(1,n)，并可求得依存结树的数目，即歧义数 N(1,n).

花子/は　橋/で　泳ぐ/0　人/を　見る/.

1　　2　　3　　4　　5

M(S)=

	1	2	3	4	5
1	0	0	1	0	1
2		0	1	0	1
3			0	1	0
4				0	1
5					1

计算得 N(1,5)=3 棵依存树，T(1,5)的结果为：

1 组：1—3,2—3,3—4,4—5

2 组：1—5,2—3,3—4,4—5

3 组：1—5,2—5,3—4,4—5

三、消歧策略

上述方法得到的依存关系一般都存在歧义，即存在多棵依存树，还必须运用表层结构以及文法知识、语义知识来消除歧义．

3.1 利用依存公理制约的传递作用消歧

根据依存关系的非交差性公理可知，若某两个依存关系（如 a_{ij} 与 a_{kl}）不满足非交叉性，则此两种依存关系不能同时存在．若 a_{ij} 是确定的，则判定 a_{kl} 必不能成立，于是消除一个歧义．同时这一个歧义的解消，还可能连带地导致其它歧义性的解消，即这种制约具有传递的效应．

如某依存网络中，1 句节可能依存于{3,4,6}，2 句节可能依存于{3,4}，3 句节可能依存于{4,6}，存在歧义．根据非交叉性制约，依存关系(1,3)与(2,4)，(1,4)与(3,6)，(2,4)与(3,6)不能共存．如果依据某种方法判断出依存关系(2,4)成立，则根据(2,4)与(3,6)交叉，推断 3 句节不可能依存于 6 句节，3 句节只能依存于 4 句节．从而 3 句节的歧义也解消了．同理可消除(1,3)这一依存歧义解．这就是文法制约的传递作用，后文的消歧机制中将利用这一传递作用，在此将给出其传递算法的形式描述．

[相交叉的依存关系]

对于依存矩阵 $M(S)=[a_{ij}](i=1,2,\cdots,n-1;\ j=i+1,\cdots,n)$（n＝句节数）

如果　$a_{ij}=1$，$a_{kl}=1$ 且 $k<i\&i<l<j$　||　$i<k<j\&j<l$

则称　a_{ij} 与 a_{kl} 相交叉，它们不能共存．

与 a_{ij} 交差的储存存关系的集合记为 $contrad(a_{ij})$.

[传递算法]

已知：$a_{ij}=1$ & $N(i)=1$

N(i)为句节 i 的向后依存关系数．

```
SpreadContract(aij) beging
    for each apq ∈cotract(aij) in M(S)
        if apq=1{
            apq←0,N(p)--;
            if N(p)=1{
                p 的向后储存存关系是唯一的，检查 M(S)，求得为 apl;
                SpreadContract(apl):
            }
        }
```

3.2 综合利用各种可能的知识来消除依存关系上的歧义

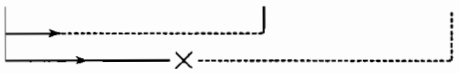
句子分析的结果是要得到一个唯一的一棵依存树．上面的计算，使我们在进行句法分析时了解到每实行一步歧义解消后，还存在几棵依法树，以及在那些成分上存在歧义，可以从何

处着手消歧．实践证明，对于消歧，一般说来，不可能利用一种方法彻底解决．而是利用多种知识，渐近地逼近．这里所说的多种方法，包括句子表层的知识，语法和语义方面的知识，以至从语类库中获得的知识[7]．下面介绍我们系统中已采用的几种方法．

(1) 呼应表现

日语中有许多惯用接续句型．将前后呼应的接续助词等信息，作为确定句的标识，子句中句节的依存关系不能超越子句范围．

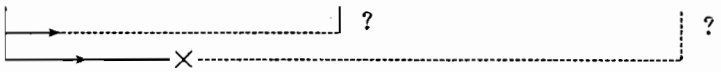
例1：　　もし　雨が　降れば　私は　行きません



(2) 利用标点等信息

在日语句中可以发现这样一种倾向：没有逗点的句节，它越过满足特定条件的逗点和特定的语句而系于后面的句节的可能性很小．本机制中利用这种倾向，找出具有限定其它依存范围作用的标识(加逗点)，并根据其特征，消除超过此范围的依存歧义解．

例2：　　キーボードから　指示を　与えると，解析結果が　表示される．

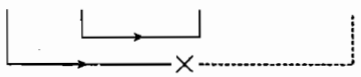
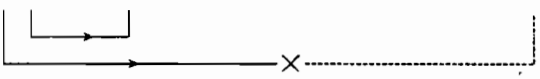
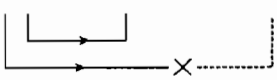


应该指出的是作为限定范围的标识，会因文章的性质不同而有些差异，因此应该根据翻译文章的类别而制定此类消歧规则．

(c) 利用强依存关系消歧

在某些句子中存在强依关系，往往可以成为很好的消歧手段，举例如下．

表2　强依存规则

No.	分类	例　文
1	邻接的，连体修饰要素形式名词(和后面的词构成并列的除外)	LEDを　駆動する　ことによる　消耗 
2	邻接的连体修饰要素和“關/内/前/后”等近似形式名词的词(和后面词构成并列的除外)	動作す　間に　割りこみを　受けない　装置 
3	邻接的连体词和名词(与后面的词构成并列的除外)	この　本は　子供のために　書いた 

(c) 并列成分的依存关系分析

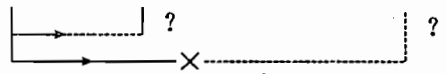
当句子中出现多个并列成分时，确定它们的依存关系分析往往是一个比较困难的问题．

我们曾对此问题进行过专题的探讨[6]. 利用句子中并列成分间在结构和语义上的类似度来选择并列成分的依存关系和并列成分的范围，可以作为依存歧义解消的一个有力的根据.

(d) 利用动词的配价模式进行消歧

配价语法认为，句子中动词是核心，其它成分都受动词的支配，这种关系可以看成类似化学中的价的关系. 可以把动词的句型，转换成一种配价模式. 在日语中由于助词在句法结构中的特殊地位，使得从表层上判断动词的配价比较容易. 如果把表层的配价关系和深层的格语法结合起来，在消除句子的依存歧义上是一个很有效的方法[8]. 例子如下.

例：　　私/は　山/に　登っ/て　景色/を　見る.



例文中，"山に"系于"登って"还是"見る"存在歧义.

∵ 动词"登る"的配价模式中有一种是：

pred：〈VERB〉〈OBJ〉

OBJ：格标＝"に"

　　语义分类＝LOC

VORB：登

句节"山に"满足以上关系，故"山に"系于"登ぬて"是确定解，消除系于"見る"的歧义解.

(e) 对于修饰成分，在一般情况下可以选择优先解

这里主要是指体言的连体性修饰成分和用言的副词性修饰成分. 修饰成分后面存在多个体言时，往往都出现歧义. 当然如果对修饰成分有可靠的制约根据时，可以消除歧义. 但很多情况下[5]可选择它的最近的依存关系作为它的优先解. 同时其它解可以保留作为候补.

四、分析实例

下面用一个实例来说明利用多种知识来消歧的结果. 句子为：

表層格や前置詞から予測される深層格と、バスが含まれる深層格は適切な対応がある.

	1	2	3	4	5
S＝	表層格/や	前置詞/から	予測/される	深層格/と	バス/が
	6	7	8	9	10
	含ま/れる	深層格/は	適切/な	対応/か	ある/。

〈歧义解消之前〉：有 6 个句节依存关系存在歧义，组合的无交叉依存树为 318 棵。

〈歧义解消过程〉：

(1) 依存范围限定：句节 4 判定为制限标识，消除超过 4 的连体性依存关系。由此确定 3—4 的依存关系。句节 7 的"は"，限制句节 6 不能越过 7，从而确定 6—7 的依存关系。

(2) 并列算法判定句节 1、4 并列，4、7 并列，因此选取最优依存 1—4，4—7。

(3) 短语捆绑：8—9。

(4) 配价模式匹配：确定 2—3，5—6，7—10，9—10.

(5) 消除与最优先解交叉的歧义解.

(6) 计算无交叉依存候补.

〈歧义解消后〉:1—4,2—3,3—4,4—7,5—6,6—7,7—10,8—9,9—10

按功能结构线性展开,得到译文:

表层格与从前置词预测的深层格,与路径包含的深层格存在适当的对应。

参考文献

[1]. 橋本正吉,国文法体系論,岩波書店,1941

[2]. 正員吉田将,二文節間の系り受けを基礎とした日本語文の構文分析,情報処理学会論文志,Vol. 30, No. 8, 1991

[3]. 長尾確,制約と選好としての知識を動的に統合して行う構造的多義性の解消,情報処理学会論文志,Vol. 32, No. 10, 1991

[4]. 平井章博,機械翻訳向け前編集のための日本語系り受け構造の曖昧性検出方式,Vol. 31, No. 10, 1990

[5]. 任福継,范莉馨,日本機械翻訳にあける系り受け構造の可保留曖昧関系たついて,情報処理学会論文志,Vol. 34, No. 8, 1993

[6]. 简幼良,高健,王秀坤,基于语境类似度的并列成份的分析,中文信息学报,1997. 1., Vol. 36, No. 1, pp51—58

[7]. 黄昌宁,关于处理大规模真实文本的谈话,语言文字应用,Vol. 6, No. 2, 1993

[8]. 孙勇,陈群秀,基于动词配价模式日汉机械翻译系统的设计与实验,机械翻译研究与发展,1995, pp. 232—238