ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.3, March 2007, pp.538−546
DOI: 10.1360/jos180538

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于句子对齐的汉语句法结构推导的计算模型*

王厚峰+, 王 波

(北京大学 信息科学技术学院 计算语言学研究所,北京 100871)

A Computational Model for Chinese Syntactic Structure Induction Based on Sentence Alignment

WANG Hou-Feng+, WANG Bo

(Institute of Computational Linguistics, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62753081 ext 106, E-mail: wanghf@pku.edu.cn

Wang HF, Wang B. A computational model for Chinese syntactic structure induction based on sentence alignment. *Journal of Software*, 2007,18(3):538−546. http://www.jos.org.cn/1000-9825/18/538.htm

Abstract: This paper introduces an unsupervised learning framework of Chinese syntactic structure based sentences similarity. First, all sentence pairs in the Chinese sentence corpus are aligned, and each pair is partitioned into similarity segmentations and different ones which alternately occur, Then, aligned similarity segmentations or different ones are selected as potential constituent candidates based on the strategy of similarity priority or of difference priority respectively. As the boundary friction may be introduced in the later step, its disambiguation is further carried out. Finally, by inducing sentence constituents, the syntactic structures are learned. In order to reduce word sparseness in the process, some words are replaced by classes in advance. Three forms of the sentence units, such as the sequence of words, the sequence of POS (part of speech)-tags and the sequence of words with POS-tag, are examined and the learned syntactic structures are evaluated respectively. The results show that different priority strategy achieves a better performance than the similarity one, and the Fs are above 46% for all three forms, with the best one being 49.52%, which is better than those having been reported.

Key words: sentence alignment; unsupervised learning; boundary friction; similarity priority; difference priority; Chinese syntactic structure induction

摘 要: 基于句子的相似性,提出了无指导的汉语句法结构推导方法.基本思想是:首先,在汉语句子库的基础上,通过句对之间的对齐,得到交替的相同片断和相异片断.然后,根据相同片断优先或相异片断优先策略,选取相应的对齐片断作为句子成分候选,并对可能因片断交叉而导致边界摩擦的候选进行歧义消解.最后,通过逐步归约句子成分,推导出汉语句法结构树.为了避免对齐过程中词的稀疏问题,还对部分具有明显规律的词事先作了归类处理.分别以词、词性以及词联合词性作为句子基本构成单元,评测了推导的句法结果.测试结果表明:对于 3 种构成单元,相异片断优先归约得到的结果的 *F* 值都超过了 46%,均优于相同片断优先归约所得到的结果,最好的达到了 49.52%,好于已报道的结果.

* Supported by the National Natural Science Foundation of China under Grant Nos.60473138, 60675035 (国家自然科学基金)
Received 2006-01-26; Accepted 2006-04-12

关键词: 句子对齐;无指导学习;边界摩擦;相同优先;相异优先;汉语句法结构推导
中图法分类号: TP18 文献标识码: A

具有句法结构的语料库称为树库.树库建设是自然语言处理中非常重要的一项基础性工作.树库不仅可以用于自动评测句法分析的结果,也是获取句法结构知识的重要基础.然而,完全通过人工来构造树库是一件十分繁琐的工作,不仅耗时费力,开发人员也需要掌握足够多的语言知识,而且还可能产生难以预料的人为错误,特别是不一致问题.近年来,随着语料库语言学的不断发展,研究人员也开始尝试如何从大规模语料库中自动获取句法知识,并构建树库.

从给定的树库中推导出句法结构知识(或文法)的方法,称为有指导的学习方法.比较典型的有 Brill 的基于变换的错误驱动方法[1]和 Pereira 的 Inside-Outside 方法[2].此外,Nakamura 采用增长式的方式,先构造一个正例集和一个反例集,在已有的一个"小"的文法基础上,通过用 CYK 算法分析正例,补充新的规则;通过分析反例,自动删减"不合理"的规则[3].

无指导的学习方法直接基于原始或者初级加工的句子,不使用人工加工后的结构信息或结构规则.这种方法大体上分成两类:

(1) 基于压缩的方法.压缩方法实际上是提取"公因子",将多次出现的多词词串代之以"成分(或称为非终结符)".比较典型的有 Grunwald 的最小描述长度(MDL)方法[4]和 Wolff 的最小长度编码(MLE)方法[5].但已有的研究表明,单纯的压缩方法在文法推导中并不能达到很好的效果.一个直接的原因是,貌似"公因子"的词串,实际上并不一定能够抽象为成分.

(2) 基于分布(distrbution)的方法.按照 Harris 等语言学家的基本思想,当两个不同的词串所在的上下文具有一致的分布特点时,它们很可能就具有了可替换的特点;此时,可以将两个不同的词串用一个非终结符表示.分布方法可以分为局部分布和全局分布两种:局部分布只考虑某个词序列前后相邻的词的特征.如 Stanford 大学 Klein 和 Manning 的工作[6,7],他们以句子的词性标注序列作为输入,通过对词性(序列)的上下文(主要是相邻的词)信息来判断两个词是否有相似.他们研究了依存结构和成分结构树的推导,分别对英语、德语和汉语进行了测试,也是我们唯一看到的关于汉语句子结构推导的研究.英国 Sussex 大学的 Clark 用到了与此类似的思想[8],在带有词性标注的语料基础上,根据词性的上下文分布将其聚类为非终结符,推导文法规则.Clark 在处理过程中结合了 MDL 方法.他们的方法对英语测试也取得了较好的结果.局部分布的最大特点是只考虑前后相邻的信息,在语料库不是非常庞大时比较适用;但在一个较小的窗口内,所得到的信息毕竟不够充分.例如,在英文中,"IN(介词)+DT(冠词)+NN(名词)"的模式,很可能将 IN+DT 归约一个结构(互信息值可能更大),而实际情况应该是由 DT+NN 先结合.扩大词的左右窗口范围,在一定程度上可以避免这一问题,在极端情况下,可以将范围扩展到整个句子.荷兰 Amsterdam 大学的 Adriaans 设计的 EMILE 系统[9]和英国 Leeds 大学 Zaanen 的基于对齐的学习都是以整个句子作为考察对象的[10,11].EMILE 的基本思想是将一个句子看成 3 部分:cl+e+cr,cl 在 e 的左部,cr 在 e 的右部,称为 e 的上下文.对于一个句子,e 可以取其中的任何词串,剩下的部分就形成其上下文.在文法推导时,从句子库中抽取所有可能的模式,然后再进行聚类.而 Zaanen 的思想与 Bilkent 大学的 Cicekli 等人在翻译模板提取中的思想有很大的相似性[12],都通过多个相同片段和不同片断交错对齐的基本方法,只是 Zaanen 进一步推导出了句子的层次结构.Zaanen 研究了英语句子结构的推导,在结构推导中,不对英语句子作任何其他预处理(如词性标注).这种思想虽然易于实现,但如果词的词性兼类现象比较严重,而训练语料又不足够大,即使是找到了对齐,也不一定能保证是正确的对齐.如果事先对句子作适当的预加工(如词性标注和简单的语义归类),并加入一定的对齐约束(如词性约束),则是可以减少明显不合理推导现象发生的.

无指导句法推导还有很多其他方法,如 Smith 就研究了模拟退火的方法[13].

本文受上述方法的影响,特别是受 Zaanen 与 Bilkent 方法的启发,提出了无指导的汉语句库构建和汉语句法结构知识推导的方法.测试表明,该方法达到了较好的实验效果.

1 句法结构的学习框架

1.1 基于句子对齐的相似度计算

在自然语言中,相同的上下文可以用不同的词替换出现.当两个不同的词(或词组)出现在相同的上下文中时,我们可以假设这两个词(词组)具有句法上的可替换性.要确定两个不同的词(或词组)是否出现在相同的上下文中,可以简单地借助于句子之间的对齐来判断.下面,我们先引入几个基本概念和表示.

设 $S=a_1a_2\ldots a_n$ 是长度为 n 的句子,其中,$a_i(1\le i\le n)$是基本单元.如果以词为单位,则 a_i 代表句子中的词;如果以词性为单位,则 a_i 代表句子中的词性.此外,也可以用带词性的词作为基本单元.

例 1: 句法结构推导是一项困难的研究课题.

将上面句子表示词序列,则为"S=句法 结构 推导 是 一 项 困难 的 研究 课题";若表示为词性序列[14],则为"S=n n vn v m q a u vn n".以词+词性表示,则为"S=句法/n 结构/n 推导/vn 是/v 一/m 项/q 困难/a 的/u 研究/vn 课题/n".

定义 **1**. 如果句子 $S=a_1a_2\ldots a_n$ 可以按基本单元顺序分段为 $S=P_1P_2\ldots P_k$ 形式,其中,

$$P_1=a_1a_2\ldots a_{m_1},P_2=a_{m_1+1}a_{m_1+2}\ldots a_{m_2},\ldots,P_{i+1}=a_{m_i+1}a_{m_i+2}\ldots a_{m_{i+1}},\ldots,P_k=a_{m_{k-1}+1}a_{m_{k-1}+2}\ldots a_n,$$

则称 $P_1P_2\ldots P_k$ 是对 $a_1a_2\ldots a_n$ 的片段划分,每个 $P_i(1\le i\le k)$称为一个片断.

定义 **2**. 设句子 S 和句子 T 分别可以划分为片断序列 $S=P_1Q_1^SP_2Q_2^S\ldots P_kQ_k^SP_{k+1}$ 和 $S=P_1Q_1^TP_2Q_2^T\ldots P_kQ_k^TP_{k+1}$,其中: $Q_i^S\ne Q_i^T\ (1\le i\le k)$ 且 Q_i^S 和 Q_i^T 不同时为空;除 P_1 和 P_{k+1} 可能为空外,其余的 $P_i(1\le i\le k)$不为空,则对 S 和 T 的这种片断划分称为一种对齐;$P_i(1\le i\le k+1)$称为 S 和 T 关于这种对齐的相同片断, Q_i^S 和 Q_i^T $(1\le i\le k)$则称为该种对齐的一对相异片断.

对于给定的句子对,可能会存在多种可能的对齐模式.表 1 给出了两个句子的 3 种对齐结果.

Table 1 Different alignment results between two Chinese sentences

表 **1** 汉语句对之间的不同对齐

Sentence 1	从 首都 北京 到 上海
Sentence 2	从 上海 到 首都 北京
Mode 1	[从] [⇑] [首都 北京] [到 上海] [从] [上海 到] [首都 北京] [⇓]
Mode 2	[从] [首都 北京 到] [上海] [⇑] [从] [⇓] [上海] [到 首都 北京]
Mode 3	[从] [首都 北京] [到] [上海] [从] [上海] [到] [首都 北京]

在表 1 中:符号"⇑"表示需要增加句子 2(sentence2)对应的片断;"⇓"表示需要删除句子 1(sentence 1)对应的片断.

当一对句子有多种可能的对齐时,需要选择"最合理"的对齐模式.为了度量合理性,我们采用了编辑距离最小的策略.所谓编辑距离最小,是指将一个句子 S 转化为另一个句子 T 时所需编辑操作的代价最小.编辑的基本单元随具体应用的不同而不同.在句法结构推导中,编辑的单元选取句子的基本构成单元,本文中可以是词、词性、或者词+词性.

将一个句子变化为另一个句子,有 4 种基本的编辑操作,即插入、删除、替换和保持;当保持不变时,不需要花费编辑代价.若以 $cost(.)$函数表示编辑代价,则插入、删除和替换所需的代价表示为

- $cost(X\to Y)$:X 替换为 Y 的编辑代价;
- $cost(X\to\varepsilon)$:删除 X 的代价;
- $cost(\varepsilon\to Y)$:插入 Y 的代价.

特别地,我们用 $cost(X\to X)$表示单元直接复制,即保持不变.

若以 $D(i,j)$表示为句子 S 的第 1~i 个基本单元到句子 T 的第 1~j 个基本单元的编辑代价,则代价的计算公

式为

$$D(i,j)=\begin{cases}0, & i=j=0\\ i*cost(S[i]\to\varepsilon), & j=0\\ j*cost(\varepsilon\to T[j]), & i=0\\ \min\begin{pmatrix}D(i-1,j)+cost(S[i]\to\varepsilon)\\ D(i,j-1)+cost(\varepsilon\to T[j])\\ D(i-1,j-1)+cost(S[i]\to T[j])\end{pmatrix}, & \text{otherwise}\end{cases} \tag{1}$$

其中:$S[i]$表示句子 S 的第 i 个单元;$T[j]$表示句子 T 的第 j 个单元.

我们使用的编辑代价取值为

$$\begin{cases}cost(X\to X)=0\\ cost(X\to\varepsilon)=1\\ cost(\varepsilon\to Y)=1\\ cost(X\to Y)=cost(X\to\varepsilon)+cost(\varepsilon\to Y)=2\end{cases}.$$

当 $D(i,j)$中的 i,j 分别为 S 和 T 的长度(如词的个数)时,其值表示将 S 变化为 T 的最小编辑代价.

在大多数情况下,利用公式(1)计算得到的对齐都是比较合理的.但是,对一些特殊情况也不能保证完全合理.例如,在表 1 所示的 3 种对齐模式中,“模式 1(mode 1)”具有的公共子串长度和最大,所需的编辑代价为 4,其余的两种对齐模式(mode 2 和 mode 3),其编辑代价都为 6.虽然“模式 1”的代价最小,但是,这种方案显然不甚合理.从形式看,两个句子中的“首都北京”是通过较大的位移才对齐的.因此,如果将相对位移较大的两个相同的词串赋以较高的代价,可以在一定程度上避免这样的对齐[10].这可以通过在编辑距离计算中引入位移代价因子来实现.于是,可以简单地将编辑代价值修改为如下形式:

$$\begin{cases}cost(X\to X)=\dfrac{1}{2}\left|\dfrac{i_X^S}{|S|}-\dfrac{i_X^T}{|T|}\right|\times(|S|+|T|)\\ cost(X\to\varepsilon)=1\\ cost(\varepsilon\to Y)=1\\ cost(X\to Y)=2\end{cases} \tag{2}$$

其中:i_u^v 表示基本单元 u 在句子 v 中的位置索引;$|S|$,$|T|$分别表示句子 S 和 T 所含基本单元数目.

1.2 句子成分的假设和结构规约

在句子对齐后,其相同部分与相异部分一定交替出现.此时,对应的相同部分与对应的相异部分就形成关联.如果以相互关联的不同部分作为句子成分进行归约,这种策略便称为相异片断优先(difference priority).相异片断优先的基本假设可以从 Harris 的可替换思想中发现.与此类似,也可以考虑将相同部分作为句子成分进行归约,即相同片断优先(similarity priority).相同片断优先的基本假设则是:多次出现的词串很可能就代表一种相对固定的结构,此时,可以将其独立出来,作为一种成分看待.图 1 分别对两种规约策略作了实例说明.

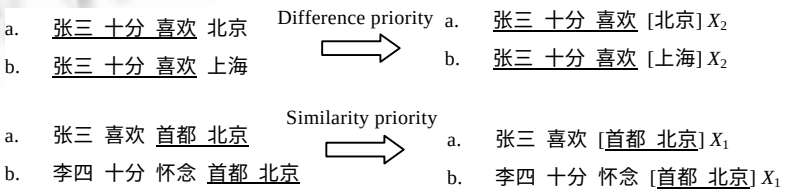


Fig.1 Two kinds of sentence constituent selections

图 1 两种句子成分选择

在图 1 中,中括号表示的内容被抽象成为成分,分别用变量 X_1 和 X_2 表示.变量实际上表示的是非终结符.随着归约的不断进行,变量的数目也将不断增加,句子的层次结构也随之产生.

为了避免变量的重复引入,我们使用如下两种策略:

其一,有结构的句子对无结构句子的泛化处理,即把存在的非终结符类型作为无结构句子中的片断类型.如在下面的例子中,“美国纽约”泛化为变量 X_2:

a. [张三 喜欢 [北京]X_2]X_1
b. [张三 喜欢 美国 纽约]X_1

⇒

a. [张三 喜欢 [北京]X_2]X_1
b. [张三 喜欢 [美国 纽约]X_2]X_1

其二,两个有结构句子的变量的归并处理,即把可替换部分的类型用相同的非终结符号标识,同时更新系统中相应的变量表示.如下面例子中,变量名为 X_3 的“美国纽约”与变量名为 X_2 的“北京”归并为统一的 X_2:

a. [张三 喜欢 [北京]X_2]X_1
b. [张三 喜欢 [美国 纽约]X_3]X_1

⇒

a. [张三 喜欢 [北京]X_2]X_1
b. [张三 喜欢 [美国 纽约]X_2]X_1

经过对齐和结构归约,就可以层层引入句子成分,其基本框架见算法 1.

算法 **1**. 汉语树库的学习算法.

对汉语句库中的每个汉语句子 s_1

对该汉语语句库中与 s_1 不同的每一个句子 s_2

对齐 s_1 和 s_2

计算它们之间的相同片断和不同片断

根据相同片断优先法则或相异片断优先法则引入非终结符号

1.3 汉语文法的获取

当句子的层次结构生成之后,就可以在此基础上推导出概率上下文无关语法.在文法表示上,我们把双亲节点作为产生式左部(LHS),把该节点对应的所有子女按从左到右的顺序连接,作为产生式的右部(RHS).产生式(s)的概率计算方式为:该产生式出现的次数除以与之有相同左部的所有产生式数目.

$$P(s)=\frac{|s'\in G: LHS(s')=LHS(s)\wedge RHS(s')=RHS(s)|}{|s'\in G: LHS(s')=RHS(s)|} \tag{3}$$

2 边界摩擦消歧

句子对齐中一个困难的问题是边界摩擦.

定义 **3**. 设有 3 个句子 S_1,S_2 和 S_3,P 和 P'是 S_1 分别与 S_2 和 S_3 对齐时形成的两个片段,如果 $P\neq P'$,但 S_1 中的同一基本单元 a_i 同时出现在 P 和 P'中,则称基本单元 a_i 在两种对齐中发生了边界摩擦.

直观上看,边界摩擦是指对齐部分存在部分重叠的成分.为了作进一步的说明,我们给出了符号化的 3 个句子,其中的大写字母 A,B,C,D,E 表示片断,相同字母表示相同片断.

句子(1): A D

句子(2): B C D

句子(3): B E

图 2(a)给出了句子(1)和句子(2)的对齐,其对应关系为 A↔[B C],D↔D;图 2(b)给出了句子(2)和句子(3)的对齐,其对应关系为 B↔B,[C D]↔E.

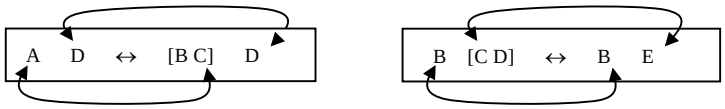
当上述两种对齐都出现时,句子“B C D”中的 C 就发生了摩擦:此时的 C 是与左边的 B 结合成[B C]还是与右边的 D 结合成[C D]?当一个句子和多个句子对齐时,很容易出现边界摩擦的问题,如下是出现辩解摩擦的 3 个具体例句:

他 喜欢 吃 红 苹果

李四 爱 吃 红 葡萄

张三 爱 吃 苦瓜

“吃”可以和左边的“爱”联合形成片断,也可以和右边的“红”联合形成片断.



(a) (b)

Fig.2 Two kinds of alignments

图 2 两种对齐

为了消解边界摩擦歧义,我们根据片断出现的概率大小进行抉择.设 P 是某个片断(如上面的[B C]或者[C D]),U 是所有可能的对齐片断集合,$freq(P)$表示片断 P 在实例中出现的次数,$freq(U)$表示 U 中的所有元素出现的次数总和,于是

$$Prob(P)=\frac{freq(P\in U)}{freq(U)} \tag{4}$$

公式(4)的基本假设是:片断出现的次数越多,其独立性也越强,表明对该片断的划分也就越合理.

3 实 验

3.1 测试语料

测试的语料来自我们收集的汉英双语句对库,主要包括科普读物、新闻报道、新概念英语等.从约 5 万句对中抽取出汉语句子,其长度绝大多数在 10 个词左右,少量句子达到了 15 个或 15 个以上的词.这样的限制,主要是为了避免结构过于复杂.此外,去掉了不完整的、口语化的句子,剩下的基本都是完整的单句.最终的测试集共有句子 5 506 个,其中 5 000 个没有结构信息,506 个句子通过手工加工构造出了句法结构.带结构的句子主要用作评测比较的参考答案.但在学习时,我们去掉了 506 个句子的结构信息,与 5 000 个句子一起作为学习的样本.

在句法知识学习之前,我们使用北京大学计算语言学研究所的分词和词性标注[14]工具对所选的语料作了预处理.我们的测试分别以词、词性和词+词性作为句子的基本构成单元进行对齐.

由于句子中单元在表层上可能已经表现出了明显的规律性,因此我们对这些单元事先作了归类处理.例如:“人名”在不同的例子中,形式上并不一定相同,但是作为语言单元,“人名”彼此具有可替换的特点,我们将这样的词(或多个词)直接归类.在测试中,我们对如下词作了归类:

- 表示人的词,包括人名、表示人的普通名词以及人称代词归类为“人名词/np”;
- 数词或数量词归类为“数字词/m”;
- 时间词(序列)归类为“时间词/t”;
- 地名(序列)归类为“地名词/ns”;
- 机构名归类为“机构词/nt”;
- 其他专有名词归类为“专名词/nz”.

3.2 结果与分析

评测主要针对 506 个句子,判断推导的句子结构与人工标注的结构有多少是一致的.人工构造的句法树共有 2 562 个内部节点.为了便于计算,我们引入如下符号:

$S=a_1a_2\ldots a_i\ldots a_j\ldots a_n$ 是含有 n 个基本单元的句子($1\le i\le j\le n$);

$A_IN(S)$是句子 S 经过自动推导所得到的句法树的内部节点集合;

$M_IN(S)$是句子 S 经过手工标注的句法树得到的内部节点集合;

$Coverage(node)$表示句法树的一个内部节点 $node$ 所覆盖的基本单元序列.

定义 **4**. 设 $A\_node \in A\_IN(S)$,$M\_node \in M\_IN(S)$.如果 $Coverage(A\_node)=a_i \ldots a_j$ 且 $Coverage(M\_node)=a_i \ldots a_j$,则称节点 *A_node* 和节点 *M_node* 是一致的.

我们定义的节点一致性,不考虑节点的非终结节点名是否一样,即忽略掉了节点标记.这种评估方法与 D.Klein 的方法是一样的,也是目前无指导句法推导评测中采用的方法.当然,在推导句法结构时,每个内部节点是有标记的,我们以数字编码来表示.

定义 **5**. 设 *C*(corpus)表示评测的句子集合,*P*(precision)表示自动推导的句法结构的准确率,*R*(recall)表示自动推导的句法结构的召回率,则

$$P = \frac{\sum_{S \in C}(A\_IN(S) \cap M\_IN(S))}{\sum_{S \in C} A\_IN(S)} \tag{5}$$

$$R = \frac{\sum_{S \in C}(A\_IN(S) \cap M\_IN(S))}{\sum_{S \in C} M\_IN(S)} \tag{6}$$

在 *P* 和 *R* 的基础上,我们也引入了调和平均值 *F*,即

$$F = \frac{2PR}{P+R} \tag{7}$$

首先,我们针对对齐后的相同片断归约方法进行了评测,结果见表 2.

Table 2 Results of structure induction based on the strategy of similarity priority

表 **2** 相同片断优先归约的测试结果

Sentence unit	*P* (%)	*R* (%)	*F* (%)
Word	53.57	36.34	43.30
POS (part of speech)	53.15	34.93	42.16
Word with POS	53.57	36.34	43.30

从表 2 中我们可以看到:以“词为单元”和以“词+词性”为单元,其测试结果完全一样.其主要原因在于测试的语料出现兼类的词非常少,因此,增加词性并没有增加区分性.此外,按照相同串规约优先的原则,有词性区分的情况正好不被归约.

我们也针对对齐后的相异片断归约的结果作了评测,结果见表 3.

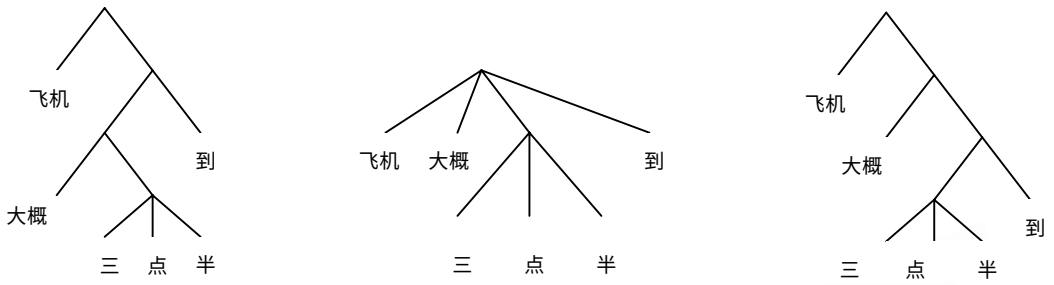
Table 3 Results of structure induction based on the strategy of difference priority

表 **3** 相异片断优先归约的测试结果

Sentence unit	*P* (%)	*R* (%)	*F* (%)
Word	49.62	43.87	46.57
POS (part of speech)	49.63	49.41	49.55
Word with POS	49.71	43.64	46.48

在表 3 中,以“词”为句子的单元和以“词+词性”为句子单元的结果非常近似,这同样是因为词的兼类现象较少.但是,“词+词性”的 *P* 值最高.可见:联合词性之后,在提高准确率上起到了较好的作用.另外,我们还看到:以“词性”为基本单元得到的 *F* 值最好,似乎与我们想象的以“词+词性”应取得更好的结果相矛盾.其主要原因是语料的规模不足够大,许多“词+词性”都只是出现一次,有较大的稀疏性,因而不容易体现出“共性”.只用词性,则避免了这一问题.

两组结果进一步表明:基于相异片断优先的策略,其 *F* 值明显要高于相同片断优先的测试结果.虽然后面一组结果(见表 3)的 *P* 值相对低一些,但 *R* 值高很多.这也从一定程度上说明:在相同上下文中,不同片断具有可替换的特点.相反地,由于语料规模有限,多词相同的片段并不很多.为了更直观地加以比较,我们从测试语料中选择了一个句子“飞机/n 大概/d 三 m 点/q 半/m 到/v”,以树结构的形式表示(如图 3 所示),3 棵树分别表示手工构造的结果、相同片段优先和相异片断优先推导的结果.两种推导策略都是以“词+词性”为基本单元.



(a) Reference answer (b) Result by using similarity priority (c) Result by using difference priority
(a) 手工结果 (b) 相同优先策略 (c) 相异优先策略

Fig.3 Result comparison among manual tree, tree based on similarity priority and difference priority

图 3 手工树与基于相同优先策略树、相异优先策略树比较

图 3 表明:片断相异优先导出的树结构与手工构造的结果非常相似.特别是两种结构究竟哪个更好,可能也有不同看法.

汉语句法结构自动推导的研究并不太多.我们看到的仅仅是 Stanford 大学 Klein[7]的研究.他从滨州树库(Upenn)中选择了长度不超过 10 个词的 2 473 个汉语句子,以“词性”为句子的基本单元,对多种策略和策略的组合进行了测试,得到的最好测试结果是,P=35.9%,R=66.7%,F=46.7%.从数值来看,我们的 F 值更好,而且我们所选取的句子并没有严格地将词数限定在 10 个以内.

D.Klein 曾在英语语料 ATIS 上针对 Zaanen 的 ABL 方法进行过测试,得到的结果是 P=43.6%,R=35.6%,F=39.2%[7].ABL 方法与我们的方法中“以词为单位”的“相异片段优先归约”类似.从表 3 看出:这种方法并不是最好的;但相比英语语料 ATIS 的测试结果,我们的结果值更好.我们认为有两方面原因:一,英语语料 AITS 可能没有我们测试的句子规范,我们测试的句子基本上是完整句子;二,我们对部分词进行了归类处理,这在很大程度上降低了噪音干扰.当然,语料不同,特别是语言不同,并不能作简单数值的等同比较.

4 结论与今后的工作

本文提出了基于对齐的汉语句法结构自动推导的无指导学习框架,是汉语句法结构自动推导的一种尝试.我们分析、比较并实现了相同片段归约优先和相异片段归约优先的两种规约方法.通过我们的语料测试,相异片段优先归约得到了更好的实验效果.在相异片断优先策略中,以“类”(词性)作为基本单元,又得到了明显好的 R 值.由此可以看到:在有限的语料中,归类的作用是显而易见的.因此,选择更加合适的类,将是我们要进一步探讨的问题.

今后的工作包括 3 个方面的内容:(1) 进一步改进现有的模型,特别是要加强归类处理;(2) 构造更大规模的树库,以便更准确地测试系统性能;(3) 基于对齐的单语言结构知识获取,也是双语翻译模板获取的基础.我们将结合汉英双语,并充分利用双语关系,研究双语翻译模板自动获取的方法.

致谢 感谢吴云芳博士和彭爽博士为句法树的构造所提供的帮助!感谢评审专家提出的宝贵意见!

References:

[1] Brill E. Automatic grammar induction and parsing free text: A transformation-based approach. In: Proc. of the 31st Annual Meeting of the Association for Computational Linguistics. 1993. 259−265. http://acl.ldc.upenn.edu/P/P93/

[2] Pereira F, Schabes Y. Inside-Outside reestimation from partially bracketed corpora. In: Pros. of the 30th Annual Meeting of the Association for Computational Linguistics. 1992. 128−135. http://acl.ldc.upenn.edu/P/P92/

[3] Nakamura K, Matsumoto M. Incremental learning of context free grammar. In: Adriaans P, *et al*., eds. Proc. of the Grammatical Inference: Algorithms and applications (ICGI-2002). LNAI 2484, Springer-Verlag, 2002. 174−184.

[4] Grunwall P. A minimum description length approach to grammar inference. In: Wermter S, Riloff E, Scheler G, eds. Proc. of the Symbolic, Connectionist and Statistical Approaches to Learning for Natural Language Processing. LNCS 1040, Springer-Verlag, 1996. 203−216.

[5] Wolff GJ. Unsupervised grammar induction in a framework of information compression by multiple alignment, unification and search. In: de la Higuera C, Adriaans P, van Zaanen M, Oncina J, eds. Proc. of the Workshorp at ECML/PKDD2003: Learning Context-Free Grammars. 2003. 114−124. http://ilk.uvt.nl/~mvzaanen/ECMLPKDD/talks.html

[6] Klein D, Manning CD. A generative constituent-context model for improved grammar induction. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. 2002. 128−135. http://acl.ldc.upenn.edu/P/P02/

[7] Klein D. The unsupervised learning of natural language structure [Ph.D. Thesis]. Stanford University, 2005.

[8] Clark A. Unsupervised induction of stochastic context-free grammars using distributional clustering. In: Daelemans W, Zajac R, eds. Proc. of the CoNLL 2001. Morgan Kaufmann. 2001. 105−112.

[9] Adriaans P, Trautwein M, Vervoort M. Towards high speed grammar induction on large text corpora. In: Hlavac V, Feffrey G, Wiedermann J, eds. Proc. of the SOFSEM-2000, Theory and Practice of Informatics. LNCS 1963, Springer-Verlag, 2000. 173−186.

[10] van Zaanen M. Bootstrapping syntax and recursion using alignment-based learning. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann. 2000. 1063−1070.

[11] van Zaanen M, Adriaans P. Alignment-Based learning versus EMILE: A comparison. In: Krose B, de Rijke M, Schreiber G, van Someren M, eds. Proc. of the Belgian-Dutch Conf. on Artificial Intelligence (BNAIC). 2001. 315−322. http://www.ics.mq.edu.au/~menno/research/publications

[12] Cicekli I, Guvenir HA. Learning translation templates from bilingual translation examples. In: Applied Intelligence, vol.15. 2001. 57−76.

[13] Smith NA, Eisner J. Annealing techniques for unsupervised statistical language learning. In: Proc. of the 42nd Annual Meeting the Association for Computational Linguistics. 2004. 487−94. http://acl.ldc.upenn.edu/P/P04/

[14] Yu SW, *et al*. The Grammatical Knowledge-Base of Contemporary Chinese—A Complete Specification. 2nd ed., Beijing: Tsinghua University Press, 2003 (in Chinese).

附中文参考文献:

[14] 俞士汶,等.现代汉语语法信息词典详解.第 2 版,北京:清华大学出版社,2003.

王厚峰(1965－),男,湖北天门人,博士,教授,CCF 高级会员,主要研究领域为自然语言处理.

王波(1982－),男,硕士生,主要研究领域为自然语言处理.