汉语依存树库的构建

赵怿怡，关润池

（中国传媒大学　北京　100024）

摘　要：语料库是计算语言学进入新时代的基础。构建依存树库是目前国际计算语言学研究的流行趋势。本文总结国内外树库建设经验，通过建设汉语依存树库的实践，对树库建设中的基本问题进行阐述，并尝试运用统计的方法来分析一些语言现象。

关键词：　树库；标注；依存关系；依存理论；并列结构

Building a Chinese Dependency Treebank

Zhao Yiyi , Guan Runchi

(cuc， Beijing, 100024)

Abstract: The corpus is a foundation of computational linguistics in the new era. Building dependency treebank is a trend in international computational linguistics researching today. This paper summaries experiences in building treebank, and presents some basic points via the practice of building a Chinese dependency treebank. In addition, the authors try to use a statistical method to analyze problems in the building process.

keywords: treebank,; annotation; dependency relation; dependency theory; coordination

1. 导言

短语结构语法被证明是一种描述固定词序语言的有效方法，同时我们也在寻找一种更具普遍适用性的语法，实现跨语言的研究。为了探索汉语的依存关系，我们尝试建设了小规模汉语依存树库，并力图在树库建设中发现并研究更多有价值的语言现象，为汉语其他领域的研究提供参考。

2. 树库建设

2.1 构建目的

目前，构建树库的目的有研究或改进语言资料（编纂词典、语法等）、训练或评估标注工具、服务于语言教学等领域等，构建目的更为多样化，也更为实际，

2004 年，我们尝试建设一个小规模的汉语依存树库，希望在以下领域有所收获：

1 研究汉语的依存关系。依存语法是一种结构语法，该语法的创始人特思尼耶尔认为结构句法的目的在于句子的研究，句子是一个“有组织的整体”，其组织性体现于构成句子的词和词与相邻词之间的“联系”，所有这些联系构成了句子的框架[1]。目前针对汉语依存关系的研究还不成熟，构建汉语依存关系的树库，无疑为此项研究提供了资源。

2 训练和评价剖析器。利用依存语法对语言结构进行分析，最终实现自然语言的计算机处理是依存语法的

作者简介：赵怿怡（1982-），女，天津，硕士研究生在读，zhaoyiyi@cuc.edu.cn.

关润池（1981-），女，沈阳，硕士研究生在读，rrcc0303@yahoo.com.cn

重要作用之一。

3 为汉语教学提供辅助工具。汉语教学的热潮经久不衰，同时也使汉语教学实践者面临着严峻的挑战。利用依存语法的研究成果，为汉语教学提供帮助是计算语言学成果走向实践的方向之一。

2.2 语料的选择

语料的选取是树库构建的重要步骤，我们要分析的语言现象不只是那些“咬死猎人的狗”之类的生僻句，因而，我们选择中央电视台“新闻联播”的部分语料，即2004年4、5、6、12月的新闻联播语料，约2万词，内容范围广泛，涉及政治、经济、文化、体育等各方面。此类语料，长度大，有助于对长、短句的分析研究；较为规范，是经过初步加工的带有词性标注的熟语料——我学院传媒语料库中的部分语料，这有利于树库的构建与研究，体现了研究的连贯性并提高了资源共享度。

2.3 标注的选择

2.3.1 标注规则

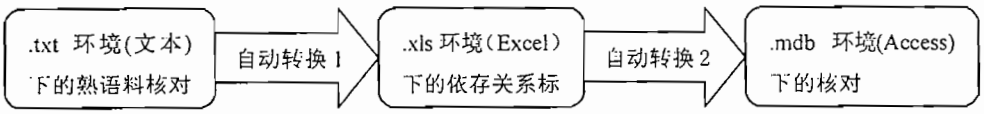
标注过程中遇到的问题要依据标注规则来判断、处理，但实际标注中仍有许多新的问题，这就需要标注者在标注过程中不断分析解决难题，积累经验、添加新规则。此外，我们也不能机械运用预设的规则，要对其灵活运用。不但要对其他树库的现有的标注规则有所承袭，而且要有新的见解。

由于国内外没有通用的依存语法理论标注体系，所以我们按照刘海涛老师的“汉语依存关系暂行方案”进行标注，内容包括词类划分、依存关系等。词类划分为12个大类，情况复杂的大类下还要再分小类，例如动词，以便将来提取更为精确、细致的特征。依存关系分为补足语关系（20类关系符号），说明语关系（14类关系符号）两种。

2.3.2 标注环境

标注在MS office环境下流转进行，充分运用了office组建通用性强的特点。通过工作环境的转换达到了人机结合、标注核对的最优化。

工作环境流程：



构建树库，劳动强度较大，标注过程中容易导致不必要的错误的出现，因而工具的使用对树库的构建很重要。在现有条件下，我们研发出初级标注工具，如在自动转换1中，胡凤国老师开发的“语料分割程序”为文本转换节省了人力，体现了工具研发在树库建设中的重要性；自动转换2中，使用的是office组件中集成数据库导入功能，充分利用了已有工具。

2.4 构建方法

树库的构建当前经常用人工与工具共同构建的方法，先进行自动标注，而后人工校正。有一些树库是同时对各标注层面进行校正的，如Negra树库[2]，另外还有一些树库是分别进行标记和剖析的,如布拉格树库[3]。

我们吸取现有树库的建设经验，采取人机结合的方式，标注的一致性是我们要解决的关键问题。

1 POS阶段：自动分词+人工核对

我们使用的熟语料，处理程度比较高，所以在该阶段的工作主要是对相关的符号替换和对个别现象的再处理。这一阶段，还要检查原来分词阶段的切分错误的现象。

2 句法标注阶段：人工标注+人工核对

人工标注： 20标注者在excel环境下，每人标注15-20个句子的文本。

人工核对阶段，第一轮：20人自由组合，互相核对；第二轮：由一名语言学研究生进行一致性的核对；第三轮：由指导教授最后进行校对。遇到标注不一致时，通过小组讨论共同商讨出一个比较一致的结果，这样极大地提高了标注的准确性，力求保持标注者之间的一致性。

3. 实践与问题

标注过程中，标注者们对主要涉及确定词性、依存结构（支配词的确定）、依存关系等方面的问题意见不同。需要针对性的深入研究。下面举两种结构的例子作为说明。

3.1 并列结构

并列连词（或标点）引入的 c-X 关系，X 指该并列结构和外部结构的依存关系。c-结构是较难把握的，虽然标注手册中对 c-的规则描述得很清楚，但是在实际语料中（尤其是长句较多的新闻中）并列成分纷繁复杂。

依存关系标注中， c-关系 的下位关系 c-atr、c-adva、c-baobj、c-comp、c-fc 等多达 11 类，虽然小类关系越多，句子成分之间的关系描述越精确，就越有助于词与词之间的深层次的研究，但是小类关系越多，正确率也越低，再加上这么多的小类大多出现在长句中，关系的确定非常容易混乱，加大错误率。如图 1 这类句长较大的句子，在新闻体语料中比比皆是。

	A	B	C	D	E	F	G	H
1	句子编号	句中词序	词	词性	支配词序号	支配词	支配词性	依存关系
2	s1	1	参加	v	3	的	usde	dec
3	s1	2	座谈会	n	1	参加	v	obj
4	s1	3	的	usde	5	、	bnd	atr
5	s1	4	何鲁丽	np	5	、	bnd	c-atr
6	s1	5	、	bnd	25	发言	v	subj
7	s1	6	丁石孙	np	7	、	bnd	c-atr
8	s1	7	、	bnd	5	、	bnd	c-atr
9	s1	8	成思危	np	9	、	bnd	c-atr
10	s1	9	、	bnd	7	、	bnd	c-atr
11	s1	10	许嘉璐	np	11	、	bnd	c-atr
12	s1	11	、	bnd	9	、	bnd	c-atr
13	s1	12	蒋正华	np	13	、	bnd	c-atr
14	s1	13	、	bnd	11	、	bnd	c-atr
15	s1	14	杜宜瑾	np	15	、	bnd	c-atr
16	s1	15	、	bnd	13	、	bnd	c-atr
17	s1	16	韩启德	np	17	、	bnd	c-atr
18	s1	17	、	bnd	15	、	bnd	c-atr
19	s1	18	林文漪	np	19	、	bnd	c-atr
20	s1	19	、	bnd	17	、	bnd	c-atr
21	s1	20	黄孟复	np	21	、	bnd	c-atr
22	s1	21	、	bnd	19	、	bnd	c-atr
23	s1	22	林毅夫	np	21	、	bnd	c-atr
24	s1	23	等	ur	22	林毅夫	np	auxr
25	s1	24	先后	d	25	发言	v	adva
26	s1	25	发言	v	26	。	bjd	s
27	s1	26	。	bjd				

图 1

红色部分是易出错的地方。由此可见，句内组块衔接的地方是出现错误概率最高的地方，

3.2 Coor 和 Va 关系：

图 1 中已经见到了 coor 关系，较之于 c-，coor 指几个相同性质的成分直接并列时的关系。如：

表 1

	支配词(governor)	从属词(Dependent)	例句
COOR	V	V	他整天吃 喝 玩 乐
	N	N	为工 农 兵 服务
	A	A	勤劳 勇敢的人民

注: 引自标注手册《现代汉语依存关系》刘海涛

名词和形容词的并列是比较好处理的，而在多个动词并列出现的时候，它们之间的关系就不好确定了，容易与之混淆的是 va（连动）关系。 如“李长春考察四川强调深入学习贯彻三个代表重要思想。”这个句子，不同的同学可能会标注出两种依存关系：coor、va。

H1 ▾ 依存关系

	A	B	C	D	E	F	G	H
1	句子编号	句中词序	词	词性	支配词序号	支配词	支配词性	依存关系
2	s1		1 学习	v				
3	s1		2 贯彻	v	1	学习	v	coor
4	s2		1 学习	v				
5	s2		2 贯彻	v	1	学习	v	va
6								

图 2

如图 2 所示，对依存关系认识的不同给标注的一致性带来很大的问题。为此我们对树库的 21385 个依存关系中的 va，coor 关系进行提取和统计，得出连动（Va）的比率（1.72%）要远高于并列关系（Coor）的比率（0.36%）的结果。两种关系比例较小，对整个树库的准确性不会产生太大的影响。

依存距离指的是支配词和从属词之间的线性距离，即一个句子中存在依存关系的两个词之间的词位置之差，这里用 k 表示。我们按照从属词和支配词的距离进一步比较两动词连用时两种关系的数量关系：

表 2

	k=1	k>1	k 的平均值
Coor 关系数	53	24	1.56
Va 关系数	156	211	3.20

注：k=从属词序号 — 支配词序号

如表 2 所示，当依存距离大于 1 的时候，va 关系占到 89.79%；当 k=1 的时候比例虽有下降，仍高达 74.64%。而且在对 coor 关系分析时，对于类似表 3 动词间的关系，虽然标注为 coor，但仍有人认为也可以做 va 解释。

表 3

ID	句子编号	句中序号	词	词性	支配词序号	支配词	支配词词性	依存关系
11789	s393	39	控制	v	38	监测	v	Coor
12680	s424	21	违规	v	20	失职	v	Coor
9977	s325	13	研究	v	12	分析	v	Coor

根据以上统计分析，我们认为遇到两动词连用问题时，应该注意以下原则：

第一，当依存距离大于 1 的时候，如有混淆，优先考虑 va 关系，这个是实证的结果。

第二，当依存距离等于 1 的时候，除非常明显的语义承接关系标注 va，优先考虑 coor。

在标注的实践中不断解决问题，丰富标注规则是树库建设过程中很重要的步骤，这些薄弱点的发现是语言理论完善的关键，也是树库建设者值得共享的经验。

4. 结语

依存句法树库是近年来树库构建的重要趋势。本文从实践出发，梳理了树库建设的理论问题，对汉语依存树库建设实验的基本情况进行了介绍。通过对汉语动词连用的依存关系分析，说明了标注规则的不断完善在树库建设中的重要性。

参考文献：

[1] 刘海涛.依存语法和机器翻译[J]. 语言文字应用，1997，3：89-93.

[2] Brants T, Skut W, Uszkoreit H. Syntactic Annotation of A German Newspaper Corpus[A]. Abeillé A. Treebans. Building and Using Parsed Corpora[C]. Netherlands :Kluwer Academic Publishers,2003. 73-87

[3] BÖhmová A, Hajič J, Hajičová E, et al. The Prague Dependency Treebank.A Three-Level Annotation Scenario[A]. Abeillé A. Treebanks. Building and Using Parsed Corpora[C]. Netherlands :Kluwer Academic Publishers, 2003. 103-127