基于依存语法的汉语并列结构自动分析研究*

赵怿怡 高松 刘海涛
中国传媒大学应用语言学研究所 北京 100024
E-mail:zoyiyi@163.com

摘 要：并列是一种普遍存在的语言现象，也是一种极难处理的语言结构。本文在国内外众多语言学理论研究并列结构的基础上，理出了依存语法理论处理并列结构的三种方案，并对含并列结构句（共计1000句33049词次）进行了依存句法标注。通过依存句法分析器的自动学习，证明不同的分析方法对句法分析器精度的影响有所不同，最优方案分析精度比最差的方案高出3.7%。这说明语言描述方式和分析方法的改进可以提高自动分析的精度。两个相应的实验实现了最优方案整体精度1.1%和1.9%的提高。

关键词：汉语并列结构，依存句法，文本标注，自动分析

Automatic Parsing of Coordination Structure in Contemporary Chinese

Zhao Yiyi Gao Song Liu Haitao
Institute of Applied Linguistics, Communication University of China, Beijing 100024
E-mail:zoyiyi@163.com

Abstract: Coordination is an important structure with high frequence in human languages. Based on treebank, Data-Driven Parser is an efficient method to analyze and test the idea how to process language structures. This paper proposes three annotation methods to analyze Chinese coordinating strcuture. We annotate a coordination corpus using these three schemes and use MaltParser to parse sentences including Chinese coordinating structure. The result shows that the different analysis of coordination influence accuracy of a parser. Two experiments prove that it is feasible and efficient to improve accuracy of a parser through linguistic means.

Keyword: coordinating structure, dependency grammar, annotation, parser

1 引言

并列结构是一种复杂的语言现象，也是句法处理的难点。通常一个典型的并列结构有三个成分：两个并列体/一个并列标记，如“你和我”由三个词构成，其中“你、我”是并列体，“和”是并列标记。但实际语料中出现的并列结构往往较为复杂，如“……对飞船**舱载医学系统、环控生保系统以及结构和结构系统**进行充分的考核……”（选自《人民日报》2000）。并列结构“舱载医学系统、环控生保系统以及结构和结构系统”由12词构成，它受“飞船”修饰充当介词“对”的宾语成分。并列结构内部的三个并列体“……系统”由多词短语构成，且第三个并列体的定语部分是有明显标记的并列结构，这说明并列连词“和”与其他并列标记处于不同层次，并列结构内部需要多层分析。这样复杂的结构，人工处理起来尚难分清，机器处理就更困难了。所以大部分句法理论都有针对并列结构的特殊加工和处理。在诸多句法理论中，用依存语法(Dependency Grammar)处理并列结构尤为困难，这是因为依存语法是一种通过词间的从属关系来实现结构分

* 本文为中国传媒大学‘211工程’三期重点学科建设项目阶段成果。

析的语法，词间关系是二元不对称的。基于依存分析的理论，如词语法(Word Grammar)、功能生成描述(Functional Generative Description)、动态依存语法(Dynamic Dependency Grammar)都对并列结构有特别的处理。那么，这种表示从属关系的依存语法能否描述并列体间语义上平等、对称的关系呢？如果能，并列结构内部该由谁来充当支配成分呢？不同的处理方式对基于机器学习的依存句法分析器的精度有什么影响呢？

为了回答这些问题，本文参照各个理论处理并列结构的实践制定了并列结构的三个标注方案。目的是从语言分析的角度考察不同处理方式哪个更有利于句法分析器的训练和学习，探索改善自动句法分析效果的可行性。论文第二部分从依存句法的基本理论出发拟定相应的标注方案对并列结构进行分析并构建相关树库；第三部分通过数据驱动的依存句法分析器自动学习测试实现对不同语言分析方法的评价；第四部分小结了数据比较和三方案评价的结果。

2 并列结构的处理和相关树库的构造

Nivre (2006)指出并列结构的依存分析第一类方法就是把并列连词视为并列结构的中心词。这受语义驱动的分析是布拉格功能生成描述派所采用的方法。我们认为连词核心肯定了功能词在句法中的地位，并列连词与并列结构外部发生从属关系。它兼顾了连词地位和并列结构第二层分析的思想。我们分别用C1、C2、C3表示各个并列体，cc表示并列连词。制定方案一（如图1），在含两个并列体的并列结构中并列标记分别支配其他并列体。在含多个并列体的并列结构中，并列标记顺次支配后一个并列标记和其前面的并列体。

图1 方案一并列结构的处理

第二类分析方法为Mel'čuk (1988)采用，他认为整个并列体的中心词或头(head)是第一个并列体，并列连词只是它后面并列体的支配词。Maxwell(1995)继承Mel'čuk的分析方法。在他的处理中并列连词只是形式上区别并列结构的标志，他让并列连词依照顺序从属于第一个并列体，并支配第二个并列体。数据库语义学(Database Semantics)是针对语义分析的重要理论，其创始人Hausser(2007)把第一个并列体放在核心位置，第一并列体以外的并列结构被忽略，不直接进入动词的论元结构。芬兰学者提供了一个依存句法分析的在线资源[1]同样选择一个并列体为核心，不同在之处在于他们按照并列结构与支配词（或从属词）的线性顺序选择离支配词（或从属词）较近的并列体作并列结构的中心。哈尔滨工业大学信息检索研究室(2006)在中文依存句法分析[2]方面也作了有益的探索。在并列结构处理上采取第一并列体为核心的思想，并列连词作为附着成分从属于它后面的并列体。据此我们制定方案二（图2）：第一个并列体为并列结构核心。在并列结构内部，如果只有两个并列体，第一个并列体支配第二个并列体，并列连词从属于第二个并列体，这主要是考虑方案二应突现具有语义信息的实词地位，考虑到并列连词通常是后面并列体出现的标记，应从属于其后的并列体。若出现多个并列体的结构，我们采取并列体顺次支配、并列连词从属处理方法来减小依存距离。

[1] Machinese Syntax, connexor natural knowledge http://www.connexor.eu/technology/machinese/demo/syntax/ (2009-3-15)

[2] 哈工大信息检索研究室. 中文依存句法分析. http://ir.hit.edu.cn/phpwebsite/index.php (2009-3-15)

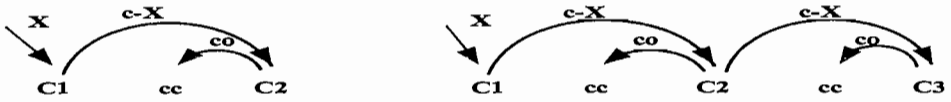

图2 方案二中三并列体的两种处理方法

第三类方法源于Tesnière(1959)关于并列结构的最初设想，他认为并列体分别从属于上位的动词作宾语，并列连词只是表示两个并列体的连接关系。两个并列体分别与外界发生从属关系，这样就不存在并列结构第二层分析的思想，是对每一个并列体并列句法功能的充分肯定。词语法(Hudson 2003)融合了短语结构的思想来处理并列结构。Hudson把复杂的带并列的架构看成以并列体为中心的短语，连词不是形态句法的核心。这样就把并列结构的分析划归为第二层次的分析。可以说上面两位语言学家注重的是对语义信息或概念的处理。在方案三（如图3）中，为了遵守依存语法的基本规则：从属词词不能有两个支配词。当并列体处于支配地位，我们规定由离从属词最近的并列体支配从属词，以遵守依存语法基本原则并防止交叉弧出现。

图3 方案三并列结构处理

我们提取了《人民日报》2000年部分语料中含并列结构的句子，采用刘海涛提出的《现代汉语依存关系语法》和三种并列结构处理方案进行标注，最终形成并列结构依存树库三个，分别含句子1000个，词33049个，平均句长33，不含循环句、非投影句和非联通句。

3 结果和讨论

本研究采用瑞典韦克舍(Växjö)大学Nivre提出的归纳依存句法分析方法(Nivre 2006)和在此基础上实现的依存句法分析器MaltParser (Nivre et al. 2007)为工具。Maltparser是与具体语言无关的数据驱动的依存分析系统，目前已经应用于多种语言的处理。

我们采用基于记忆的模式（MBL）和M4策略（该策略采用5个词类特征、4个关系特征和2个词性特征）进行学习。M4策略已被刘海涛（2007）证明为适用于小规模树库的自动学习的策略。我们把并列结构依存树库中1000句中900句作为训练集剩余的100句作为测试集，得到的试验结果如表1所示。

表1 三方案自动分析的总体分值[3]

	方案一	方案二	方案三
UAS	0.728	0.736	0.743
LAS	0.661	0.698	0.697
UnSent	31	29	24
bnd	0.601	0.681	0.681
cc	0.477	0.748	0.761
v	0.499	0.499	0.501
n	0.652	0.661	0.652

[3] UAS无标记依存关系识别精度；LAS有标记依存关系识别精度；Unsent非联通句。

识别无标记依存关系UAS的分值依次递增分别是0.728、0.736、0.743，UAS反映机器寻找两个具有依存关系的词对的能力。识别有标记依存关系LAS的分值，后两种方案可以说不分上下，比方案一高了近4%，该指标反映机器通过学习寻找具有依存关系的词对并判断它们间关系的能力。通过不同方案标注的语料的学习和分析，我们发现方案二整体性能略好于其他，这种方法也是某些短语结构语法和依存语法研究者普遍采用的；而方案一符合并列结构语义处理的需要，但因其内部关系层次复杂，显现出不容易被机器学习的劣势；方案三是并列结构和外部关系规定最复杂的一种，强调实词之间的关系，忽略并列结构的整体性，在句子联通性上表现最优。

从具体词类依存关系的学习情况看，三个方案和并列结构有关的虚词和主要实词（动词、名词）的精度低于总体分值。三方案进行比较发现，方案一并列连词cc的精度远落后于后两个方案，而方案一的标点bnd精度较后两个方案也有较大的差距。考虑到这种差异很可能由标点标记中混杂并列标记"、"造成，我们进行了"严格句法功能与词类标记对应的实验"。替换原树库中顿号的标点标记bnd为并列连词标记cc，900个句子的训练集中有580顿号标记被替换，100个测试句中有39个顿号，重新学习测试得到了三种标注方法训练分析器识别精度变化情况（表2）。

表2 三方案修改顿号标记bnd为cc后部分词类LAS分值变化情况

H&D PERTAG	方案一		方案二		方案三	
	精度	变化	精度	变化	精度	变化
LAS（总体）	0.663	+0.002	0.709	+0.011	0.705	+0.008
bnd	0.69	+0.089	0.718	+0.037	0.747	+0.066
cc	0.392	-0.085	0.747	-0.001	0.747	-0.014
v	0.482	-0.017	0.521	+0.022	0.523	+0.022
n	0.664	+0.012	0.668	+0.007	0.642	-0.010

通过以上的调整，方案二的整体精度LAS提高幅度最为明显，达到1.1%，方案一也小幅提高。这个小改进证明严格词类与句法功能对应对提高分析精度产生了积极影响。但是我们注意到同时改动三个树库的等量标记，三方案整体精度变化幅度存在差异，体现了机器对三方案的学习难度存在差异。在具体词类表现上，顿号由标记bnd修改为cc后，三种方案中的bnd精度都得到提高，这显然是剔出并列标记对句内标点影响的结果，而并列标记cc的精度都有降低，方案一降幅最大达到8.5%，方案二降幅最小0.1%。这说明方案一中并列连词cc较后两个方案有更重要的句法地位。在承载语义信息的主要词类——名词和动词精度变化上，修改后方案一名词大类n的精度提高了1.2%，而动词大类v精度降低了1.7%；方案二名词大类精度提高了0.7%，动词大类v提高了2.2%；方案三名词大类精度降低了1%，动词大类精度提高了2.2%。综合这两大词类的变化情况，方案二中动、名词精度的同步提高是可喜的，这证明了我们可以通过语言学分析方法的改进实现主要数据的提高。

结论一：严格句法功能与词类标记对应的实验证明，从功能角度重新考察词类和词类标记是必要和有效的。这里我们把以往认为不重要的并列标点"、"从功能上给与了重新考量，将其标点标记替换为并列连词标记"cc"。使得三个方案总体分析精度得到提高，所修改词类的分析精度已经超过了总体精度。其中方案二的主要词类动词、名词的分析精度同时得到提高，这对以往依存对词类精度此消彼长的分析经验是一个惊喜的突破，这意味着我们可以通过严格功能与词类标记的对应来实现主要实词精度的提高，为深入的语义分析做出更好的铺垫。

在依存关系类型精度比较中涉及两个指标：精确率，它是系统实际正确标记关系数和系统实

际标记的关系的比值；招回率，它是系统实际正确标记关系数和标准树库中某类关系数的比值，反映系统漏标、查全的能力。从并列结构相关依存关系类型的分析情况看，方案一共有45种依存关系类型，其中20种并列结构内部关系，出现5种c-c-X关系，没有co关系。方案一并列结构内部依存关系的分析效果普遍较差，在训练集中出现100次以下的并列关系类型（即c-X标记）的分析精度几乎全部为0。这种表现，一方面是因为并列结构内部标注本身的复杂性，牵扯到二层甚至多层分析，且方案一中核心并列连词cc具有代替多词类发生多类依存关系的能力，存在不可控性。另一方面的原因是：虽然我们尽量筛选包含并列结构的语料，但1578个并列结构仍然不能满足数据驱动的自动分析中因并列结构句法功能的丰富导致的数据稀疏问题，这证明了刘海涛、赵怿怡（2009）指出的：机器不能很好的处理各种并列结构的一种原因在于可用于训练的语料太少以致无法识别这些关系；方案二共有41种依存关系类型，并列结构内部关系16种，其中包含c-c-X的并列的关系类型4种。和方案一相比，方案二的变化使训练集中c-X关系的量大大减少，并且三层的并列依存类型有所减少；方案三共有38种依存关系类型，并列结构内部关系13种，没有c-c-X关系类型。这使得三方案标注的简单显现出来。对相同语料的分析呈现出不同数量和种类的关系类型体现了三个方案存在逐步简化分析的趋势。

出于解决数据稀疏和缩小三方案分析难度，更好地比较三方案并列结构关系类型的想法，我们对方案一、二中的复杂的依存关系类型进行了两步“简化关系标记”的实验。首先对方案一、二分析中出现的c-c-X关系进行简化，把c-c-X关系纳入到二层分析中，再对方案一、二进行完全简化，以co替代并列结构内部复杂的c-X关系。由于方案三是强调各并列体的同等地位，原本就不存在并列结构内部的关系类型，不予进行简化操作。表3反映了简化后的情况。

表3 简化c-c-X->c-X->co后的整体分值变化

	方案一			方案二		
	原始	简化c-c-X为c-X	简化c-X为co	原始	简化c-c-X为c-X	简化c-X为co
LAS	0.661	0.663	0.7	0.698	0.7	0.717
	准确率/招回率			准确率/招回率		
co	--	--	0.768/0.754	0735/0.728	0.746/0.727	0.742/0.706

方案一相继简化84个c-c-X关系类型为c-X和co关系后得到新的数据，LAS分值由0.661提高到0.7。方案二标注中出现的c-c-X并列关系数量上少于方案二，简化15个c-c-X关系，得到新的LAS分值0.717。简化关系标记后，两方案测试精度的提高说明：简化关系标注的层级和类型在一定程度上缓解了数据稀疏的问题，降低了分析器学习的难度。而方案一、二简化后的整体精度都超出了方案三的整体精度0.697。简化后并列结构内部关系完全由依存关系co表示使得整体依存关系数量大量减少，对于LAS分值的提高有很大的帮助，同时，简化后并列结构内部关系co的学习精度也得到大幅提升。方案一中原本没有co的关系类型标记，简化后co统一了所有的c-X关系，co分值0.768超越了方案一训练学习的总体分值。而方案二中co原本表示并列标记从属于并列体的关系，改进后co的准确率提高了0.7%，这说明简化使得分析器识别co关系的能力增强，但是招回率下降2.2%，是因为co关系基数的增大扩大了系统漏标的弱点。

结论二：在树库规模有限的情况下可以通过适当简化依存关系类型来提高分析的精度。但是这种简化必须保证对语言结构进行足够分析，如若将方案三中并列结构相关关系同样简化为co则无法判断并列结构的句法功能，所以不能盲目简化分析。并列结构标注方案一、二的简化实验说明了在句法标注阶段可以不对并列结构内部依存关系进行细致的描述，从而获得分析效率上的

提高。这也证实了动态依存语法(Dynamic dependency grammar)学者主张的并列结构需要额外的第二层的分析的结论。

4 结论

实验结果表明，不同的结构分析方法对分析器自动分析的效果会产生不同程度影响，方案二表现出最佳的机器学习效，精度达到0.698。方案一因结构层次分析复杂、依存关系类型繁多显现出较差的分析效果，精度为0.661。

通过对词类关系比较分析，我们发现方案一是一种注重并列结构内部词类和结构相似性的分析方法。词类标注和句法功能严格对应对于正确分析有重要的意义。方案二、三强调实词的作用，消除了并列连词充当多种句法成分的混乱，提高了部分词类的分析精度。在现有条件下，我们通过"严格句法功能与词类标记对应的实验"，证明了词类标注和句法功能严格对应能够提高所有方案自动句法分析的效率，这说明对并列结构其他成分进行严格标记与句法功能的分析应该是进一步研究的方向。通过对依存关系类型的比较，我们发现不同方案标注并列结构依存关系的复杂程度存在不平等性，方案三的高精度来源于其关系类型的最简分析，方案二的简化使得其精度提高1.9%，这说明在清晰描述语言结构的基础上采取句法阶段的简化处理有利于提高分析器自动分析的能力，为更好进入语义阶段的处理打好基础。

实践证明，将语料库和统计学习的方法运用于现代汉语特殊结构的处理是可行和有效的。计算的方法不仅可以筛选对语言结构认识上的分歧，更能为语言分析工作提供有益的指导。

参 考 文 献

[1]. Hausser, R. (2006) *A Computational Model of Natural Language Communication Interpretation, Inference, and Production in Database Semantics*. Berlin, New York: Springer.

[2]. Hausser, R. (2007) Handling Valency and Coordination in Database Semantics. *Valency: Theoretical, Descriptive and Cognitive Issues*. Berlin: Mouton de Gruyter. pp. 321-337.

[3]. Hudson, R.A. (2003) Word Grammar. *Dependency and Valency. An International Handbook of Contemporary Research*. Berlin: Walter de Gruyter. pp. 508-526.

[4]. Liu, H. and W. Huang (2006) A Chinese Dependency Syntax for Treebanking. *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation*. Beijing: Tsinghua University Press. pp. 126-133.

[5]. Maxwell, D. (1995) *Unification Dependency Grammar* (Draft).

[6]. Mel'čuk, I. A.(1988) *Dependency syntax: theory and practice*. Albany: State University Press of New York.

[7]. Nivre, J. (2006) *Inductive Dependency Parsing*. Berlin: Springer.

[8]. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E. (2007) MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2): 95-135.

[9]. Tesnière, L.(1959) *Eléments de la syntaxe structurale*. Paris: Klincksieck.

[10]. 刘海涛. (2007) 影响依存句法分析的因素探讨[C].第九届全国计算语言学论文集.北京:清华大学出版社. pp. 147-152.

[11]. 刘海涛, 赵怿怡.(2009) 基于树库的汉语依存句法分析[J].模式识别与人工智能, 22(1): 17-21.