并列结构在依存句法分析中的应用

石翠，王杨

（辽宁行政学院 信息技术系，沈阳 110161）

摘要：本文利用中文专利语料库，对中文专利文献中的并列结构进行了分析，主要分析了中文专利文献中并列结构的依存特征。根据中文专利文献中并列结构的依存特征，总结出并列结构依存处理规则，并根据并列结构依存处理规则对中文专利文献的依存分析结果进行了规则后处理，规则处理后提高了识别的准确率。

关键词：专利文献；并列结构；依存句法分析

中图分类号：TP301.2 文献标识码：A DOI：10.3969/j.issn.1003-6970.2014.04.017

本文著录格式：[1] 石翠，王杨 . 并列结构在依存句法分析中的应用 [J]. 软件，2014,35(4)：68-70

Application of Coordinate Structure in Dependency Parsing

SHI Cui, WANG Yang

(Department of information technology, Liaoning School of Administration, Shenyang Liaoning 110161*, China)*

【Abstract】This paper use of Chinese patent corpus analyzes the coordinate structures in Chinese patent corpus. It mainly analyzes dependency features of coordinate structures in Chinese patent corpus. According to dependency features of coordinate structures in Chinese patent corpus, this paper summed coordinate structure dependency processing rules, and according to coordinate structure dependency processing rules conducts post- processes to the dependency analysis results of Chinese patent literature in this paper. By the rule processing, the analysis results will be raised.

【Key words】Patent Literature; Coordinate Structure; Dependency Parsing

0 引言

专利文献是一种非常重要的技术资料，专利文献的文本格式比较固定，用语较为规范，除含有一些高频词和未登录词之外，还存在着大量的并列结构。并列结构是一种特殊的语言形式，它由两个或更多的并列成分组成，并列结构有时也称为联合短语 [1]。与非专利文献相比，专利文献中的并列结构主要有下面几点差异：包含嵌套并列结构多；不规则并列结构分布广泛；并列结构跨度大，甚至占据整个句子 [2]。

在专利文献中，由于并列结构一般表现为长距离关联或依存，目前广泛使用的统计句法分析器很难处理，分析效果较差。（下图 1（a）和图 1（b）分别为用专利语料训练的句法分析器分析结果和正确的分析结果，图中箭头指向的词为核心词。）

1 相关研究

1.1 并列结构相关研究

国内外的许多学者，对并列结构进行了研究。Hanamoto et al. (2012)[3] 结合 HPSG 句法分析器和并列结构的局部对齐特性对宾州树库中的并列结构进行了识别。吴云芳 [4,5] 利用现有的语言资源，从句法、语义两个层面详尽地考察了并列成分之间的约束关系，并对这些约束关系进行了形式化的描述，而后基于知识描述进行了并列结构的自动识别，基于并列词语进行了相似词语的自动聚类。李文杰，穗志方 [6] 在概念实例和属性的提取研究中，针对基于模式的方法召回率比较低的特点，提出了一种基于并列结构的概念实例和属性的同步提取方法。还有其他学者也针对并列结构做了相关的研究，本文不一一列举。

1.1 依存句法分析

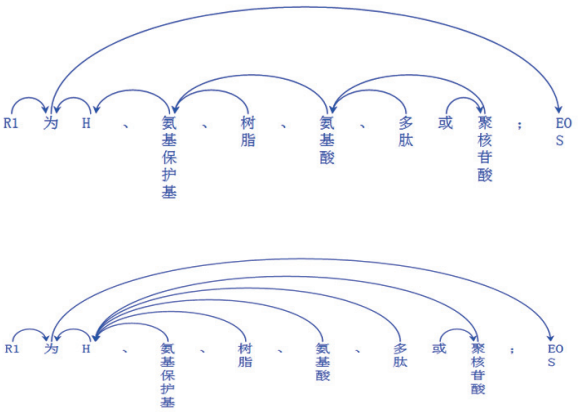

图 1 依存句法分析结果图

Fig.1 Results figure of dependency parsing

任何一种句法分析都是依赖于某种语法理论的。依存语法是通过词与词之间的依存关系来描述其语言结构。计算语言学家 J. Robinson 总结了依存语法的 4 条公理 [7]：（1）一个句子中只有一个独立成分不依存于其他任何成分；（2）句子的其他成

作者简介：石翠（1981-），女，讲师，主要研究方向：自然语言处理；王杨 (1981-)，男，讲师，主要研究方向：知识管理与知识工程。

分都必须依存于某一成分；（3）任何一个成分都不能依存于两个或两个以上的其他成分；（4）如果成分 A 直接依存于成分 B，而成分 C 位于 A 和 B 之间，则 C 依存于 A 或者 B，或者依存于 A 和 B 之间的某一成分。根据句法模型将一个句子中各个成分之间的关系表示为某种句法结构图的形式可直观的描述句子的形式模型，便于人对句子的理解以及机器的自动学习。图 2 示出了“她是优秀的学生”的三种依存结构：（a）依存树（b）有向图（c）依存投影树。

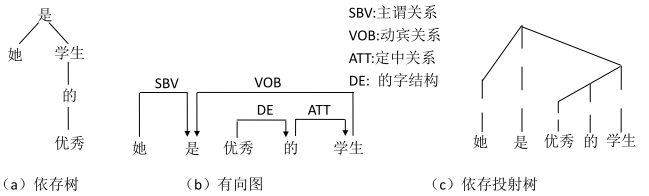

图 2　常用的三种依存结构图

Fig.2 Three commonly figures of dependency structure

2　并列结构在依存句法分析中的应用

为了能在并列结构的基础上，使得依存句法分析的准确率有所提高，下面先对并列结构的依存结构特征进行分析，在通过其结构特征改进依存句法的分析结果。

2.1　并列结构的依存结构特征

中文专利文献中并列结构的依存分析结构图如图 3 所示（方框所框部分为并列结构）。从图中我们可以看出，并列结构依存结构的特点是：并列结构可以作为一个整体进行依存分析。也就是说，并列结构是非核心词，则并列结构只有一条弧依存于并列结构之外的词；并列结构是核心词，则并列结构之外的词只能依存到并列结构中的某一个词，而不是多个词。并列结构内部的并列标记，如果是标点符号则不依存于任何词，如果是并列连词则依存于该并列连词引导的后并列成分的核心词。

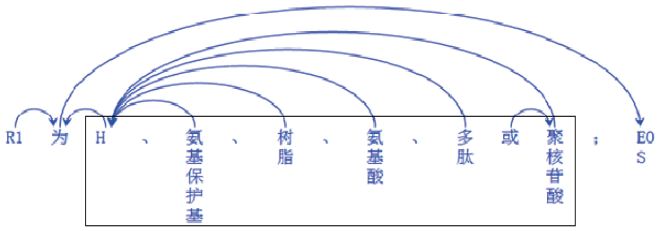

图 3　并列结构的依存分析结构图

Fig.3 Coordinate structure of dependency structure

2.2　基于并列结构的依存句法分析

本文应用文献 [8] 所使用的依存句法分析方法，考察并列结构对依存句法分析的影响。

我们进行了下面的对比实验，首先，在中文专利文献的基础上，运用文献 [8] 的依存句法分析方法，得出专利文献的依存弧准确率；其次，根据并列结构依存分析的特点，基于错误驱动的方法，对前面确定的依存结构进行规则后处理，得出处理后的依存弧准确率。该实验所用的语料为沈阳航空航天大学知识工程研究中心标注的，经自动分词、词性标注并人工校对的语料。语料中关于依存的标注是经过自动标注并人工校对的。

依存弧准确率（Unlabeled Attachment Score, UAS）的评价标准如公式 1 所示。

$$\mathrm{UAS}=\frac{\text{弧正确识别的词数}}{\text{所有词数}}\times 100\% \qquad (1)$$

基于并列结构的依存分析后处理规则：

1、找到并列结构内部的核心词（也就是第一个并列成分的核心词）。

2、并列结构内部的依存调整。各个并列成分都依存到并列结构的核心词。（如图 4、图 5 及图 6 所示，其中，方框所框部分为并列结构，打叉的弧为依存分析错误的弧。）

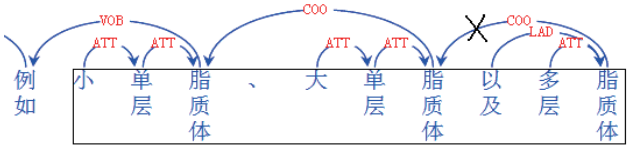

图 4　并列结构内部依存错误实例 1

Fig.4 Error example of coordinate structure internal dependancy 1

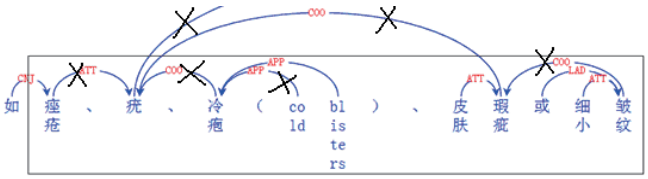

图 5　并列结构内部依存错误实例 2

Fig.5 Error example of coordinate structure internal dependancy 2

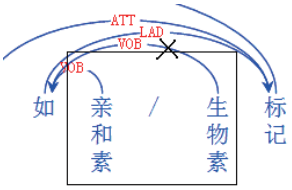

图 6　并列结构内部依存错误实例 3

Fig.6 Error example of coordinate structure internal dependancy 3

通过对错误实例的分析，我们将内部依存调整的规则制定如下：

判断并列结构的长度（并列结构中包含词的个数），并列结构长度为 3，直接调整，将非标点的并列标记依存到后面的并列成分，将后面的并列成分依存到前面的并列成分。

并列结构的长度大于 3，则进行下面的调整：

找核心词，即依存到并列结构外部的词。

找到一个核心词，判断该词前面是否有并列标记。①有并列标记，将并列标记前依存到该词的词设为核心词；且把该词后面的并列标记后的所有依存到该词的词都依存到核心词。②如果没有并列标记，该词就是核心词，找核心词后并列标记后

依存到核心词的词语，标记为词 W，寻找词 W 后并列标记后依存到词 W 的词 W1，将词 W1 的依存词改为核心词。

如果找到多个，暂不处理（并列结构包含多个核心词且有明显错误的情况多出现在并列结构长度为 3 的并列结构里，所以对于长度大于 3 的依存错误暂不处理）。

3、并列结构外部的依存调整。依存到并列结构内部的词都要依存到并列结构的核心词。（如图 7 所示，依存到并列结构的词应该依存到并列结构的核心词。）

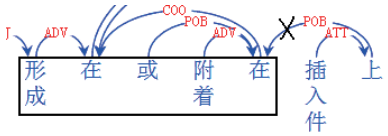

图 7　并列结构外部依存错误实例

Fig.7 Error example of coordinate structure external dependancy

基于并列结构的依存句法分析的对比实验结果如表 1 所示。

表 1　实验结果对比

Tab.1 Experimental results contrast

方法	UAS
文献 [8] 方法	93.72%
并列结构规则后处理	94.03%

2.3　实验结果分析

实验所用测试语料共包含 1000 个句子，1000 个句子中不是所有的句子都包含并列结构。在 1000 句测试语料中提出了 72 句根据并列结构识别结果进行依存句法调整的句子，72 句调整的句子的依存弧正确率的实验结果对比表如表 2 所示。

表 2　依存弧实验结果对比

Tab.2 Experimental results contras of dependency arc

方法	UAS
处理前	86.98%
处理后	90.71%

3　结语

本文在中文专利文献的基础上，运用文献 [8] 所用方法，进行的专利文献的依存分析实验。在上述实验结果的基础上，根据并列结构的依存分析特点，对依存分析结果进行了规则后处理，实验结果表明，基于并列结构的规则后处理提供的依存分析的准确率有所提高。

参考文献

[1] 冯文贺，姬东鸿 . 并列结构的依存分析与连词的控制语地位 [J]. 语言科学，2011，10(2)：168-181.

[2] 石翠，周俏丽，张桂平 . 面向中文专利文献的有标记并列结构的统计分析 [J]. 中文信息学报，2013，27（5）：43-50.

[3] Atsushi Hanamoto, Takuya Matsuzaki, Jun’ichi Tsujii.Coordination Structure Analysis using Dual Decomposition[C].IN: Association for Computational Linguistics. France: Avignon,2012:23-27.

[4] 吴云芳 . 并列结构的外部句法特征 [C] 机器翻译研究进展 -2002 年全国机器翻译研讨会论文集 .2002：110-116.

[5] 吴云芳 . 面向语言信息处理的现代汉语并列结构研究 [D]. 北京 : 北京大学，2009.

[6] 李文杰，穗志方 . 基于并列结构的概念实例和属性的同步提取方法 [J]. 中文信息学报，26(2):82-87.

[7] Robinson, J. Dependency structures and transformational rules[J]. Language , 1970, 46(2): 259-285.

[8] 朗文静 . 规则与统计相结合的汉语依存句法分析技术研究及其应用 [D]. 沈阳：沈阳航空航天大学 ,2012.

（上接第 67 页）

由于传统的网络设备（交换机、路由器）的固件是由设备制造商锁定和控制，而 SDN 将网络控制与物理网络拓扑分离。这必然引起设备制造商、服务提供商的某些不满和不快，他们必须做相应的改变。所以，平稳过渡是个大问题。如此新型的网络体系结构不可能在一天之内推倒重来，必须在现有的因特网的基础上逐步过渡。如果 SDN 将来能站得住，要全世界推广，它必须能够用局部的 SDN 连接目前全局的因特网。像谷歌的内部网络，用 OpenFlow 交换机把数据中心连到 SDN 一样。然后，才能扩大到全局。

上述粗略的介绍对于关心 SDN 的朋友也许会有一个概略的了解。但是，由于本人也是初学，而且，SDN 是一个新的网络体系结构，本人的理解错误难免，一定不全面。本人衷心希望内行和外行的朋友多多提出问题。这种交流也许本身就是一种研究。

参考文献

[1] Sixto Ortiz Jr.,”Software-Defined Networking: On the Verge of a Breakthrough?,” IEEE Computer, JULY 2013, pp.10-12

[2] Keith Kirkpatrick, “Software-Defined Networking,” Communications of the ACM, Vol. 56 , No. 9, September 2013, pp.16-19.

[3] Martin Casado, Nick McKeown, “ The Virtual Network System,” SIGCSE’05, ,St. Louis, Missouri, USA., February 23–27, 2005.

[4] Masayoshi Kobayashi, Srini Seetharaman, Guru Parulkar, Guido Appenzellerq, Joseph Little, Johan van Reijendam, Paul Weissmann, Nick McKeown, “Maturing of OpenFlow and Software Defined Networking through Deployments,” Preprint submitted to Elsevier, August 14, 2012.

[5] Thomas A. Limonceli, “OpenFlow: A Radical New Idea in Networking,” Communications of the ACM, Vol. 55, No. 8, August 2012, pp.42-47.

[6] Saurav Das, Guru Parulkar, Nick McKeown, “ Why OpenFlow/SDN Can Succeed Where GMPLS Failed,” ECOC Technical Digest © 2012 OSA.