基于字依存树的中文词法-句法一体化分析[1]

赵海　　揭春雨　　宋彦

香港城市大学 中文、翻译及语言学系，中国香港 九龙 达之路

E-mail: haizhao@cityu.edu.hk, ctckit@cityu.edu.hk

摘要：针对中文切分规范定义上的一些困难以及多层次处理的性能下降问题，本文提出了一种直接从字开始的依存关系表示用于中文的基本结构表示和分析。我们的分析表明，这一表示框架可以方便地用于建立一种词法-句法一体化的完整句子结构表示。通过标注词法依存，组合到已有的句法依存树库，我们获得了一个初步可学习的字依存树库，进而，我们通过实验比较说明，在相当的标注语料的学习性能上，字依存分析能够通过标准的依存句法分析有效地学习，在直接的数值比较上，其性能上优于传统的管线方式的联合学习策略，此外，字依存分析事实上涵盖了词法-句法分析，因而其输出具有更为丰富的语言学信息。

关键词：字依存，依存分析，词法-句法一体化分析

Character Dependency Tree Based Lexical and Syntactic All-in-One Parsing for Chinese

Hai Zhao　　Chunyu Kit　　Yan Song

Department of Chinese, Translation and Linguistics, City University of Hong Kong,

Tat Chee Avenue, Kowloon, Hong Kong SAR, China

E-mail: haizhao@cityu.edu.hk, ctckit@cityu.edu.hk

Abstract: Aiming at alleviate the difficulties in word definition and performance loss of multiple layer processing for Chinese, we propose character dependency tree for Chinese infrastructure representation and parsing. Our analysis shows that this type of representation is easy to help build an all-in-one infrastructure of Chinese for lexical or syntactic parsing purpose. Binding annotated lexical dependencies to an existing Treebank, a character dependency Treebank is roughly obtained. A series of experiments are performed on these data sets. The experimental results show that character dependency tree representation improves the performance through multiple-layer processing, and outperforms the typical cascaded joint learning strategy. In addition, character dependency parsing covers both lexical and syntactic information using one-off output, this brings more fruitful linguistic implications.

Key words: Character dependency, dependency parsing, lexical and syntactic all-in-one parsing

1 字依存树的引入

字依存树提出的动机在于缓解困扰当前中文信息处理的两个瓶颈问题所带来的困扰：一个是中文词定义的广泛争议性；一个是对于句子的词法-句法结构的深层处理的错误率不

[1]本文研究部分地由香港城市大学 SRG 项目 7002037 和香港特别行政区资局（UGC）的 CERG 研究项目 9040861(CityU 1318/03H)资助。

断递增导致整体性能极其低下。下面我们通过语言学实例说明第一点，对于第二点我们将通过实验比较来说明。

众所周知，中文词的严格定义无论在语言学上还是在计算上都是具有争议的。一个根本原因是中文天然是用字的串来书写的，而不是按照空格隔开的方式，也就是理想状态下的词串的方式来书写的。从语言学的角度来说，现代汉语的字-词的边界存在模糊的地方，词-短语的边界也存在模糊的地方。这种语言特性给需要严格的形式化定义的计算语言学带来了不小的障碍。最近几年开展的 SIGHAN-Bakeoff 评测的解决方案是简单直接的承认词的多重切分标准的存在[2]。然而，这并没有实质性的解决问题。

基于词的切分是面向应用的认识，我们观察几个语言学实例来说明我们所面临的困境：

a. 一/张/“/北京市京剧ＯＫ联谊会/会友/入场/证/”（Bakeoff:MSRA2005切分语料）

微软切分规范按照某种最长原则对输入串进行尽可能完整的切分，如同上面例句中的“北京市京剧ＯＫ联谊会”，这种切分规范的好处是能够有效的把握句子中的主要成分，然而，显而易见，这种切分规则强制地忽略了复杂的组织名的内部结构。

b. 中国/ 驻/ 南非/ 大使馆（Bakeoff:PKU2005切分语料）

和微软的切分原则相反，北京大学的切分标准将复杂的组织名切分为多个单元，这样切分的好处是能够正确地处理各个有意义的切分单元，然而，这样的处理对于句子的整体结构的把握不利。例如，如果以某种线性方式读取切分串来重组这些被切开的结构，我们可能获得“大使馆”属于“南非”这样的错误的理解模式。

下面是一个较为极端的例子（第九届计算语言学联合学术会议上董振东老师举的例子）：

c. 星期一三五开会

我们会发现，用多种方式试图切分这样一个串“星期一三五”都会导致困难。第一种切分：“星期一/三/五”。这种切分忽略了数字“三”和“五”需要在逻辑上跟从“星期”从而构成期望中的词“星期三”和“星期五”。第二种切分：“/星期一三五/”，这样做完全忽略了该串复杂的内部结构。

下面的例子是现代汉语的一种常见结构：

d. 洗了一个澡

尽管普遍意识到，“洗澡”是一个有效的词，但是在这样的常见结构中，我们失去了对这个“词”的认知和搭配关系的有效利用，原因还是很简单，常规的词定义限于相邻的字搭配，对于这种非线性的关系无能为力。

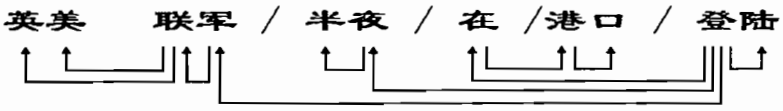

图 1 字依存树的例子

注：斜线代表常规的词切分

[2] http://www.sighan.org

我们的解决方案就是直接在句子中字符的基础上定义依存关系。图 1 是这样的一个例子。注意到这颗依存树是能够相对完备地反映所给句子中几乎所有结构关系，而不只是仅能取代词切分的表达形式。实际上这个字依存树已经涵盖了句法内容。这里我们所给出的是无标记的字依存树，如果辅助以依存标记，我们还能够表达更加复杂的句子内部结构。

值得注意的是，尽管字依存和词依存仅仅是一字之差，字依存确属词依存表示到字符级别的扩展，但是它们代表了完全不同的中文分析思路。词依存树只是传统的词/词性/句法分析链条上较高的一个环节，仅限于用于依存形式的句法分析，而且依赖于最开始的词切分结果，也不包含词内部的构词信息。字依存树是独立的完整的并且一致的基础结构表示方法，它本身是完备的，不依赖于其他的处理结果，它可以同步地表达传统意义上的词法/句法信息。

基于字依存树的表示，我们能够通过简单的方式解决上述的词切分标记所带来的一些困扰。例如，对于“星期一三五“这样的包含有非线性结构的串，可以简单地通过分别连接“一“、“三“、”五”到“星期”标记依存关系即可。对于上面的例子 c，‘洗’将自动地作为‘澡’的上位词，‘澡’作为‘一个’的上位词，我们可以用一种很自然的方式确定‘洗澡’这一个简单地搭配。推而广之，以这种模式，我们能够很容易捕捉被额外的修饰成分所分割的固定搭配，从而可以获得语言学意义上更为可靠的认识和统计信息。

如果必要，字依存树的确可以通过一棵树结构表达从词法到句法的全部信息，从而给带来一种一致的结构表达方式。这样，由于学习层次减少——仅需学习一个字串上面的依存关系图——将能有效地遏制学习的传播误差增长。我们将探索多种的字依存树表达方式，考察它们对语言学理解和学习算法的影响。

基于字依存树，我们已经通过实验验证，可以直接借用已有的成熟的依存句法分析工具和方法来对中文进行高效地处理。能够有效地通过依存方式表达并进行学习很大程度上得益于最近几年来在依存句法分析方面取得的进步。我们的初步工作表明，已有的这些技术可以无障碍地应用于字依存树的分析工作[9]。

2. 标注词法依存关系

字依存树的起点是字，因此，我们从词法的级别开始启动这项工作。为了降低标注的难度，验证字依存树的有效性，我们使用了一个简易方案：标注一个句法树库中的所有的词的内部依存关系（某种意义上代表了词法，因此我们也称之为词法依存），然后，将这些标注的内部依存加入到原有的句法树库，可以快速的建立一个字依存树库，原因是通常的词的数量在几万这一个数量级。在[9]中，我们验证了这种字依存树库的有效性，证明联合学习内部字依存和外部字依存，甚至可以帮助改进传统的分词性能[7,8,9]。

在上述的方案中,我们使用一种简单的策略来完成已有词表中词的内部依存的标注工作。具体来说，就是不断的标出中心成分，直至整个词中的所有的字都被标注完毕。对于译名和连绵词，使用简单的线性依存序列来解决这个问题，即后一个字连续地作为前一个字的中心字，以词的最后一个字作为词中心字（经过试验比较，如果使用最前一个字作为中心字确立这样一组线性依存关系，其学习性能将显著下降。）。注意到，对于无标记的

字依存标注来说，这是一个相对词依存标注更为简单的工作，仅需考虑两类关系：修饰关系和并列关系。原则上所有我们所要考虑的关系都能归结为这两类，对于修饰关系，根据修饰的方向确定依存弧的方向，对于并列关系，直接的引出或者引入两条平行的弧，但是这样将不可避免地引入高度复杂的非投影型关系。因此，在目前的研究中，我们继续使用线性化的简化策略：即后一个字连续地作为前一个字的中心字，以词的最后一个字作为词中心字。

3. 学习模型

为了说明字依存学习的效果，我们将比较字依存分析和传统的分词/依存句法分析的效果。我们采用 Nivre 方式的移进规约框架作为基本的字或者词依存句法学习框架[7][8][3]。基于转换的句法分析方法实际上是一种词对的分类方法，但仅限于处理投影型的输入句子。这一依存关系学习的优势是句子的分析效率具有优势。我们所考虑的中文依存句法树库(严格来说，是经过通常的成分-依存转换过程获得的[6]）是高度投影性的，因此较为适用于这一学习框架。在 Nivre 框架的句法分析中，分析器按照一定方向扫描输入的句子，同时保存已经分析过的部分句子的状态。具体来说，使用一个栈来维护已经得到分析的部分句子。在每个状态，分析器检查两个词（或者字），一个词（字）位于栈顶（通常用 TOP 表示），一个词（字）位于尚未处理的句子的首部（通常用 NEXT 表示）。根据分类器的输出，来决定是否在这个词（字）对之间建立一定的依存关系。如果我们用弧来表示依存关系，有两种弧来表达 TOP 和 NEXT 之间的关系，左弧代表后者是中心词（字）(上位词)，右弧代表前者是中心词（字）。分析器还需要移进和规约两个操作来完成扫描句子的操作，因此，在一个无标记的依存分析中，需要四类操作：（1）左弧（2）右弧（3）规约（4）移进。这里的四类操作意味着需要一个 4 类的分类任务需要分析器的分类器来完成。

我们继续使用经过轻微变形处理的伪投影化技术[2,4]来处理少数的非投影型句子。标准的伪投影化技术将非投影型依存的中心词转移。传统的基于转换的依存分析所使用的分类器通常是 SVM 或者其他基于边界或者基于记忆的方法。但是，这些分类器在依存学习中表现为训练时间长和解码低效，甚至比基于图模式的分析器要慢很多。我们在[4,5,9]中成功使用了最大熵作为分类器，并证明了：最大熵作为基于转换的依存分析的分类器，能够提供可比较甚至更优的性能。因此，在本文中，我们继续使用这一分类工具。

表 1　特征表示的基本记号

标记	含义
s	栈顶元素（字/词）
s'	栈顶下方第一个元素（字/词）
$s_{-1}, s_1, \ldots$	栈顶元素的串左（右）方第一个元素（字或词）
$i, i_{+1}, i_{+2}, \ldots$	未处理部分第一（二、三）个元素
h	中心字或者中心词
lm	最左依存（下位元素）
.	的。例如，s.lm 代表栈顶词的最左下位元素
+	串加法，合并两个串为一个串作为特征

依存句法分析所依赖的特征涉及多重因素，为了方便表示，我们定义了一组基本记号，如表 1 所示。依据表 1 中的记号，我们用于字依存关系分析特征集及其必要解释如表 2 所示，需要说明的是，该特征集已经根据我们在[5]中提出的特征选择过程进行过适度的优化选择。由于我们涉及的是一个初等的处理任务（尽管是句法分析！），只有字特征可以使用，因此该特征集实际上是各种位置的字形特征的组合。

表 2　用于字依存分析的特征集

标记	含义
i_n	未处理字串的 unigram, n=1,2,3
$i_n + i_{n+1}$	未处理字串的 bigram，n=-2,-1,0,1
$i_n + i_{n+1} + i_{n+2}$	未处理字串的 trigram，n=1
s_n	栈顶字相关的 unigram, n=0,1
$s_n + s_{n+1}$	栈顶字相关的 bigram，n=-2,-1,0
$s + s.h$	栈顶字和已识别的栈顶字的中心字
$s + s.lm$	栈顶字和已识别的栈顶字的最左依存
$s'_n + s'_{n+1}$	次栈顶字相关的 bigram，n=-2, 0
$s.curRoot + i$	curRoot 代表从栈顶字开始的那棵部分已分析树的根
$s + i_n$	n=-1,0,1
$s'_{-1} + i_1$	/
$s' + i$	/
s:i|pathChar	pathChar 搜集所有从 s 到 i 的串路径上的字并合并为特征串
s:i|pathCharBag	pathCharBag 搜集所有从 s 到 i 的串路径上的字去除重复后排序最后合并为特征串

4. 评估结果比较

我们比较两个基本任务的学习性能。一个是单一的字依存树分析学习，这一学习过程某种意义上可以认为是词法-句法合一的。另一个是分词/词性标注/依存句法分析，这一过程我们使用一个简单的层次管线，即首先做自动分词以及自动词性标注[3]，然后在其输出结果上作依存分析。词级的依存句法使用我们在[10]中的基线特征集的一个进一步的优化版本来完成训练以及测试。

对于词法依存关系，我们现已初步完成了宾州中文树库（CTB）的所有约 4 万词的标注工作，并在此基础上，集成到句法树中构造了一个粗略的汉语字依存树库。对于数据划分，我们根据传统，依据[6]中确定的标准的训练/开发/测试集的划分方式以及词一级的依存句法转化方式（基于版本 4.0）。经过词内部依存关系展开，用到的测试集包含约 1.4 万字。考虑到依存句法，无论是基于字的还是基于词的，对于标注的质量要求很高，因此该

[3] 使用我们在[16][17]中的基准分词系统 baseSeg 的 n-gram 特征作条件随机场的训练并测试，所用到的中文词性标注工具依据我们发布的词性标注工具 basePoS 的同样的特征集作最大熵模型的训练并测试。以上工具的二进制及源代码发布均可从 http://bcmi.sjtu.edu.cn/~zhaohai 下载。

标注还在持续的改进检验中，本文所依据的内部标注版本号为 0.10。

我们使用标准的度量来评估从分词/词性标准/句法分析三个模块的性能。但是由于从自动分词到依存句法管线学习是一个复杂的联合学习任务，某种意义上是一种新的学习任务。因此目前为止没有一个已报道的度量方式。考虑到该任务实际上涉及双重因素：词的正确率和依存关系判定的正确率。通常的依存分析精度事实上不能直接用于评估这个任务，原因是词的切分不一致可能出现在标准答案和系统输出之间。为此，基于传统的依存句法分析的 UAS（无标记分数，UAS, unlabeled attachment score），我们引入一种确切无标记分数（EUAS，exact unlabeled attachment score）来度量自动分词-依存句法管线学习的性能：在对应的两个输入句子中，搜索所有的词依存对，如果标准答案和输出的词对中对应的两个词完全相同并且依存关系判定一致，则正确的计数器加一。在此基础上，我们分别定义 EUAS 精度和 EUAS 召回率（因为系统输出和标准答案的词数可能不一致），最终的 EUAS 分数就是 EUAS 精度和召回率的 F1 值。

表 3 性能比较 (%)

		分词	词性标注	依存句法（+词性特征）	依存句法（-词性特征）
标准	精度 (F1/Acc./UAS)	97.2	88.0	84.9	73.4
	整句正确率	72.7		40.8[4]	28.7
自动	精度 (F1/Acc./EUAS)	97.2	86.0	69.2	65.5
	整句正确率	72.7		30.2	26.4

基于以上的实验设置，我们最后得到的各种管线处理结果如表 3 所示。表中的“自动”代表词性使用自动切分的结果，以及句法分析依据自动分词和自动词性标注的结果，“标准”代表所有的任务基于前一个任务的标准答案输入。根据以上的经验结果，可见尽管当前的分词技术已经获得极大进展，在基于词的性能上超过 95%，但是分词所导致的整句正确率的损失依然高达 1/4。从分词到词性标注的连续的错误传递导致最终的句法分析性能降低了近 15-18 个百分点，而整句正确率则降低了 7%-1/3。

下面我们报道字依存分析的具体性能，但是由于词依存分析涉及到词的因素，而前者不涉及，因此具体的性能数据事实上两者是不能直接比较的，因此在此仅作为参考。使用表二的特征，在测试集上的获得的 UAS 为 80.9%，整句正确率为 24.2%。注意对于字依存分析，这一精度数据的获得是在没有附加的句法辅助信息下获得的。在词级的句法分析中，我们可以依赖于词性这一附加语言资源信息，从表 3 可以看出，词性特征能够带来近 5 个百分点的分析精度增长，但是在字一级的分析中，我们目前尚没有“字性”这一辅助资源的协助。完全依赖于字形特征，我们所获得的超过 80%的词法-句法合一的性能数据，说明基于字的合一处理在计算上是完全可行的。仅从数值上的比较，这一结果远远优于分词/依存句法管线学习的 EUAS 分数。而且还应注意到，字依存分析同步地揭示了词法依存，因此就输出来说带来了更为丰富完整的句子分析结果，就训练的运算过程来说，其训练难度并不高于单一的分词/依存句法分析管线学习。这一尽管不是很严格的数据比较，依然说明了字依存词法-句法分析的

[4] 句法的整句正确率仅考虑分词和句法本身的错误，没有考虑词性错误。

优势所在。同时我们注意到，在该数据集上的整句正确率不高[5]，甚至轻微差于无词性特征辅助的词依存管线处理，这反映我们的词内部标注在该部分数据上存在较多的不一致，尚需进一步的改进规范和校验。

5. 结论

本文提出了一种直接从字开始的依存关系表示用于中文的基本结构分析。这一表示有助于解决中文的词定义的某些困难。同时，它能用于建立一种词法-句法一体化的完整句子结构表示。通过标注词法依存，组合到已有的句法依存树库，我们获得了一个初步可学习的字依存树库，进而，我们通过实验比较说明，在一致的标注语料的学习性能上，字依存分析有助于降低多层次处理的性能损失传递，在性能上优于目前的管线方式的联合学习策略，此外，字依存分析事实上涵盖了词法-句法分析，因而其输出具有更为丰富的语言学信息。

参考文献

[1] Joakim Nivre. 2003. An Efficient Algorithm for Projective Dependency Parsing. IWPT'2003: 149-160.

[2] Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In Proceedings of ACL-2005, pages 99–106, Ann Arbor, Michigan, USA, June 25-30.

[3] Joakim Nivre , Hall J., Kubler S., et al. 2007. The CoNLL-2007 Shared Task on Dependency Parsing. EMNLP-CoNLL'2007 (CoNLL shared task): 915-932.

[4] Hai Zhao and Chunyu Kit. Parsing syntactic and semantic dependencies with two single-stage maximum entropy models[C]. Proceedings of CoNLL-2008. 2008: 203 - 207.

[5] Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Exploiting Rich Features for Tagging Syntactic and Semantic Dependencies in Multiple Languages. In Proceedings of CoNLL-2009[C], June 4-5, Boulder, Colorado, USA.

[6] Qin Iris Wang, Dekang Lin, and Dale Schuurmans. 2007. Simple training of dependency parsers via structured boosting. In IJCAI 2007, Proceedings of IJCAI-2007, pages 1756–1762

[7] Hai Zhao, Chang-Ning Huang, Mu Li and Bao-Liang Lu. Effective tag set selection in Chinese word segmentation via conditional random field modeling [A]. In *PACLIC-20* [C], pp.87-94, Wuhan, China: November 1-3, 2006

[8] Hai Zhao, Chang-Ning Huang, Mu Li. An Improved Chinese Word Segmentation System with Conditional Random Field [A]. In *SIGHAN-2006* [C], pp.162-165, Sydney, Australia, 2006

[9] Hai Zhao, Character-Level Dependencies in Chinese: Usefulness and Learning, In *EACL-09* [C], pages 879-887, Athens, Greece, March 30 - April 3, 2009

[10] Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. Cross Language Dependency Parsing using a Bilingual Lexicon , ACL-IJCNLP 2009[C], Singapore, August 2-7, 2009

[5] 出于可比较的数据的原因，我们最终选用了 CTB4 的语料，该语料的划分沿用基于成分的句法分析的一些约定。但是注意该语料的规模较小，我们在较大规模的非标准划分的数据上的结果证明字依存的整句正确率同样优于词依存的管线学习。