中文字级依存结构标注规范

版本 0.98

赵海

上海交通大学 计算机科学与工程系, 中国上海

zhaohai@cs.sjtu.edu.cn

内山将夫 隅田英一郎

国立通信技术研究机构，日本京都

mutiyama@nict.go.jp, eiichiro.sumita@nict.go.jp

2012.04.23

本数据标注在国家自然科学基金 **60903119** 以及日本国立通信技术研究机构联合资助下完成。

摘要

本规范提供中文词内部的字之间的依存关系的基本标注原则。从一个修饰-被修饰的角度， 所有的中文词可以用一个迭代分析过程获得一个基于字的依存树结构。这样一个树结构的标注分析过程基于给定词的绝对词性以及词内部各个字的“字词性”。 而这两个词性也是由标注者所给出。依存类别标记可以直接从这些词性以及相关的依存关系导出。

1. 标注目标

本规范说明如何正确标注由论文(Zhao, 2009)提出的字级依存（建议标注者阅读该论文）。

我们假定标注者具有基本的汉语的读写知识。如果具有一定的古汉语知识则更佳。

大部分词语抽取自宾州中文树库版本 7.0。标注中，标注者可能需要参考树库原文理解该词语的确切含义，特别是那些专有名词。

标注目标是完成各个词内部依存关系，依存类别以及各个字的“字词性”。

2. 标注格式

假定有三字词,

天安门

各个字按照顺序给出位置标记 1 到 3，所以我们有

1 天

2 安

3 门

通过依次给出各个层次的中心字（head）可以完成全部标注任务。假定“ 天”的中心字是“安”,“ 安”的中心字是“门”，而“门”是整个词的根(root)。我们可以得到如下的整个词的内部依存结构输出如下

1 天 2

2 安 3

3 门 0

这里，指标 0 代表整个词的根(root)。

对于依存类别需要标注的情形，依存类别可以直接加在各个中心字标记数字的后方，如下

1 天 2 *n-n-v*

2 安 3 *nn*

3 门 0 *root*

以上涉及三个依存类别标记 *n-n-v, nn* 以及 *root*.

下面为简化书写，我们暂时用如下的格式表达上述数据,

天安门 2 3 0

或

天安门 2 3 0 *n v n n-n-v nn root*

其中 *n* 和 v 为相应的字词性。

以上的词和数字字母之间都需要有空格。

3. 词的类型

我们区分如下几类词用语加速和简化标注过程。

(1) 单字词

完全无需标记

(2) 完全由数字和字母组成的词，标注举例：

一百万 2 3 0 *g g root*

100,000 2 3 4 5 6 7 0 *g g g g g g root*

NATO 2 3 4 0 *l l l root*

各个字的标注按照增量进行：前一个字的中心字总是它右邻的字，最后一个字是根。依存类别标记：数字是 *g*, 字母是 *l* 。

注意，类似“一百多”这样的词不是纯数字词，需要进行更详细的标记。

(3) 命名实体/专有名词

人名、地名、组织名或其他的译名。

胡锦涛 2 3 0 *r r root*

上海 2 3 0 r r root.

纽约 2 0 *r root*

奥巴马 2 3 *0 r r root*

中心字标注规则和纯数字/字母词一样。*R* 代表此类依存关系。

注：对于(2) 和 (3)两类词，标注可以大体上省略，标注者只需简单对于整个词给出标记*_number, _letter,* 或 *_person, _loc, 以及 _org* 即可。基于给出的标记，我们将有程序自动完成完整的标注。

(4) 混合名

混合名典型地包括一个命名实体部分和一部分其他普通中文字部分。如下的标注示例：

纽约城 2 3 0 *r nn root*

上海市 2 3 0 r nn root

注：对于 (3) 和 (4)之中涉及到的部分命名实体词，特别是来自中国、日本、朝鲜/韩国以及越南的地名和组织名，如果其中部分字构成了明显具有意义的组合，则其内部结构需要详细标注，而不能简单地标为*_loc*、*_org* 等。这样的典型例子是“ 天安门、安定县、东京”。

4. 一般标注规则

以下说明如何根据字之间已有的句法和语义联系来标注基本的依存关系。

我们建议标注者采用一种**迭代标注策略** 。举例来说，我们有四字词,

abcd

假定 *abc* 首先被识别为一个形容词成分，继而 *d* 被视为名词成分，我们可以将前者标为修饰成分， *d* is 作为它的中心字。在 *abc* 之中，*bc* 进一步被识别为动词成分，而 *a* 视为副词成分，这样， *bc* 会被设为 *abc* 的中心成分。现在转向最后一个多字部分， *bc*，假定 *c* 可以取为其中心字。最后，我们有标注结果如下,

abcd 2 3 4 0

这里并无绝对的规则能确定某一个字和另外一个字的搭配中，其中必然某字必定是中心字。下面我们根据依存句法的基本思想，提出若干基本准则来指导标注。

为此，首先，我们需要定义词性标记。这一词性标记不仅能适应于给定的词，而且涉及词内部的字，我们称之为‘字词性’。

注意这里的词性是独立于词的上下文的，因此我们给出一组简化的词性标记集。

对于这里涉及的词性，我们称之为**绝对词性**，类别如下。

1. *p*:代词 这个(this one) 他们(they) …
2. *n*:名词 大门(main door) 汽车(automobile) …
3. *v*:动词 书写(write) 玩耍(play) 举行(held) …
4. *a*:形容词 热闹(alive) 邪恶(evil) 平坦(plain)…
5. *d*:副词 最近(recently) 很难(difficultly) 大概(about) …

注意中文是一个词性兼类非常严重的语言，各个词随着上下文的不同可能具有非常不同的实际词性。

因此，这里定义的绝对词性是从一个词的所有可能的词性列表中抽取出的最可能的那个词性。对于本标注体系所需要的绝对词性，以下是基本的消岐规则。

(1) 中文中最普遍的词性兼类是动词-名词兼类。在这种情形，该词的绝对词性是动词。

(2) 对于名词和形容词兼类，绝对词性是形容词。对于某些依然具有歧义的词，我们设定一个语义判据：名词是对于真实世界中的实体给出名称，形容词是涉及属性或者特性的描述。

(3) 对于形容词和副词兼类，绝对词性是副词。

字级词性标记

类似于词在句子中的词性，字级的词性标注将是动态的，或者是依赖于所在词的上下文环境的。注意这和上面定义的词的绝对词性非常不同。

我们定义如下的用语字词性的字级词性标注集:

(1) *p*: 代词 这(this) 那(that) 我(I) 一(one)…

(2) *n*: 名词 门(door) 体(body) 名(name)

(3) *i*: 数字、外文字母，其他非中文字符

(4) *v*: 动词 写(write) 跑(run) …

(5) *a*: 形容词 红(red) 苦(difficult) …

(6) *d*: 副词 很(very) 最(most)…

(7) *f*: 功能字 的(of) 们(-es) 在(at) …

注意，字一级的词性标注是一个相关词性的消岐过程。考虑下面的实际例子：

‘书’通常被认为是一个名词性的字，例如在 ‘读书(read book)’这个词中。但是对于‘ 书写(write, writing)’这个词， ‘书’实际上是动词性的。在古代汉语中，‘书’更多的是用于动词，典型的成语有‘奋笔疾书’。

中心字规则

假定已知词的绝对词性以及各个字的字词性，通过遵循下列中心字规则可以完成标注。

以下规则同样遵循传统的依存语法的基本准则

(1) 对于给定绝对词性为名词的词，其中的具有名词性字以及它的修饰成分，前者必须为中心字。在标注出来的整个依存结构树上，修饰-中心字关系总是具有最高的优先级。

红花 2 0 *ad root*

蓝天 2 0 *ad root*

一本书 2 3 0 *ad ad root*

(2) 对于给定的词，核心的动词性字应作为整个词的根或者至少是相关成分内的根字。

吃掉 0 1 *root cd*

快跑 2 0 *vv root*

(3) 对于并列结构，右手边的部分总是作为中心字(注意此时需要特定的依存类型标记，参见后文相关部分)。

玩耍 2 0 *vc root*

神仙 2 0 *nc root*

(4) 如果一个平级成分内涉及字词性为功能字的字，则该字必须为中心字。

好的 2 0 *af root*

我们 2 0 *pf root*

依存类别标记

依存关系的类别由三个词性的组合来自动确定。这三个词性是中心字、依存字以及它们组合后的成分。设有依存关系 *a* 是依存字 *b* 为中心字，相应的字串为 *a…b*。假定存在一些其它的字串在这两个字中间。 如果 *p1, p2* 以及 *p* 分别是 *a, b* 和 *a…b* 的词性，则依存关系 *a* <- *b* 的类别标记是 *p-p1-p2*。如果在某些情形 a 或者 b 属于某些不可分割的成分（如命名实体），或者，包含 a 或者 b 的最小成分的词性和 a 或者 b 的自身词性冲突，则需要以 a 或者 b 为根的那个成分的词性作为这里的替代词性。例如，“葡萄”中的“萄”，离开这个词就没有意义，即以“葡萄”这整个成分的词性作为萄字的替代词性。

(1) 简化起见，以下定义了一些依存类别标记的缩写。

ad n-a-n

af a-a-f

pf p-p-f

vd v-d-v

cd v-v-d

nn n-n-n

vv v-v-v

如果标注过程中所得到的类别组合不在以上缩写中，则使用原始的类别标记形式如 d-d-f 。

(2) 对于并列结构，分别针对名词、动词、形容词和副词，我们定义四类表达并列结构的标记 *nc* ， *vc* ， ac 和 dc。以下是两个例子：

凤凰 2 0 *nc root*

书写 2 0 *vc root*

因此, *nc* 和 *vc* 可以等价的看作是

nc n-n-n

vc v-v-v

注意这里的并列结构标记实际上不代表一种依赖关系，例如，nn 和 nc 的含义是不一样的。

(3) 对于不同的词类型，包括，数字、字母和不可分割的命名实体，我们用三类标记区分。

g 数字

l 字母

r 命名实体

在混合名中， *g*/*l*/*r* 可以视为 n 的一个特例用于词性的推导。如下面的例子，

ad n-a-g

5. 完整标注的例子

下面给出一组完整标注的字依存的例子。

绿色的各列代表输入，红色各列代表需要标注的输出。

注意，我们要求标注者将所给词的绝对词性作为一个后缀加在 root 后面。

Word 鲟鱼(sturgeon)

index	char	Pos-of-char	Head-index	Dep-label
1	鲟	*n*	2	*nn*
2	鱼	*n*	0	*root-n*

Word 天安门(Tian’anmen)

index	char	Pos-of-char	Head-index	Dep-label

1	天	*n*	2	*n-n-v*
2	安	*v*	3	*nn*
3	门	*n*	0	*root-n*

Word 普天同庆(the whole world is celebrating)

index	char	Pos-of-char	Head-index	Dep-label
1	普	*a*	2	*ad*
2	天	*n*	4	*n-n-v*
3	同	*d*	4	*vd*
4	庆	*v*	0	*root-n*

Word ５０余(more than fifty)

index	Char	Pos-of-char	Head-index	Dep-label
1	５	*i*	2	*g*
2	０	*i*	3	*a-g-a*
3	余	*a*	0	*root-a*

Word 拉美裔(Hispanic)

index	Char	Pos-of-char	Head-index	Dep -label
1	拉	*n*	2	*r*
2	美	*n*	3	*nn*
3	裔	*n*	0	*root-n*

Word 空对地(air to ground)

index	Char	Pos-of-char	Head-index	Dep-label
1	空	*n*	3	*nn*
2	对	*v*	0	*root-n*
3	地	*n*	2	*v-v-n*

Word 好人好事(good man and good deed)

index	char	Pos-of-char	Head-index	Dep-label
1	好	*a*	2	*ad*
2	人	*n*	4	*nc*
3	好	*a*	4	*ad*
4	事	*n*	0	*root-n*

参考文献

Hai Zhao, Character-Level Dependencies in Chinese: Usefulness and Learning, EACL-2009, Athens, Greece, April, 2009, *http://www.aclweb.org/anthology-new/E/E09/E09-1100.pdf*