中文词性标注规范

崔秀莲

2021 年 7 月 9 日

目录

1 引言 **1**

1.1 标注准则 1
1.2 词性标签集合 1
1.3 词性标签集合的大小 2
1.4 处理疑难案例 2
1.5 标记说明 3

2 词性标签详细介绍 **4**

2.1 动词 VA，VC，VE，VV 4
2.1.1 表语形容词 VA 4
2.1.2 系动词 VC 4
2.1.3 “有”作为主要动词 VE 5
2.1.4 其他动词 VV 5
2.2 名词 NR，NT，NN 5
2.2.1 专有名词 NR 5
2.2.2 时间名词 NT 6
2.2.3 其他名词 NN 6
2.3 方位词 LC 6
2.4 代词 PN 6
2.5 限定词和数词 DT，CD，OD 7
2.5.1 限定词 DT 7
2.5.2 基数词 CD 7
2.5.3 序列词 OD 7
2.6 量词 M 7
2.7 副词 AD 7
2.8 介词 P 7
2.9 连词 CC，CS 7

2.9.1 并列连词 CC . . . 8
2.9.2 从属连词 CS . . . 8
2.10 助词 DEC，DEG，DER，DEV，AS，SP，ETC，MSP . . . 8
2.10.1 “的”作为补语化成分或者名词化标记 DEC . . . 8
2.10.2 “的”作为所有格标记和关联标记 DEG . . . 8
2.10.3 表示结果的“得”DER . . . 8
2.10.4 方式“地”DEV . . . 8
2.10.5 动态助词 AS . . . 9
2.10.6 句末助词 SP . . . 9
2.10.7 ETC . . . 9
2.10.8 其他助词 MSP . . . 9
2.11 其他词性 IJ，ON，LB，SB，BA，JJ，FW，PU . . . 9
2.11.1 感叹词 IJ . . . 9
2.11.2 拟声词 ON . . . 9
2.11.3 长“被”结构 LB . . . 10
2.11.4 短“被”结构 SB . . . 10
2.11.5 把字结构 BA . . . 10
2.11.6 其他名词修饰器 JJ . . . 10
2.11.7 外文单词 FW . . . 10
2.11.8 标点符号 PU . . . 10

3 词性标签难点讲解 **12**
3.1 易混淆的词性组合 . . . 12
3.1.1 AD 和 AS . . . 12
3.1.2 AD 和 CC . . . 12
3.1.3 AD 和 CS . . . 12
3.1.4 AD 和 JJ . . . 12
3.1.5 AD 和 NN . . . 13
3.1.6 AD 和 NT . . . 13
3.1.7 AD 和 VA . . . 13
3.1.8 AD 和 VV . . . 13
3.1.9 AS 和 VV . . . 13
3.1.10 CC 和 P . . . 14
3.1.11 CS 和 P . . . 14
3.1.12 CD 和 DT . . . 15
3.1.13 CD 和 JJ . . . 15
3.1.14 CD 和 NT . . . 15
3.1.15 DT 和 JJ . . . 15

3.1.16 DT 和 OD 15
3.1.17 DT 和 PN 15
3.1.18 JJ 和 NN 16
3.1.19 JJ 和 P 16
3.1.20 LC 和 NN 16
3.1.21 LC 和 MSP 16
3.1.22 M 和 NN 16
3.1.23 NN 和 NR 17
3.1.24 NN 和 NT 17
3.1.25 NN 和 VA 17
3.1.26 NN 和 VV 17
3.1.27 NN 和 PN 18
3.1.28 P 和 VV 18
3.1.29 VA 和 VV 18
3.1.30 VE 和 VV 19
3.2 特殊的词语和搭配 19
3.2.1 的 19
3.2.2 等，等等 19
3.2.3 来 20
3.2.4 连 20
3.2.5 有 20
3.2.6 这样 20

1 引言

中文几乎没有屈折形态学（一门研究词形变化的学科）。也就是说，词语不会因为时态、人、数量或其他特殊情况而发生形态的变化。像英语，会有第三人称单数 +s，动词过去式，动词过去分词等形态变化。因此，很难根据词语形态来确定一段中文文本中某个词语应该被怎样标注。

这篇规范是针对中文宾州树库而设计的。中文宾州树库的目标是构造一个进行过句法分析的，包含 10 万个词的普通话中文语料库。其中标注分为两个阶段，第一阶段是分词和词性标注，第二阶段是句法分析。每一阶段都包括至少两轮，也就是说，数据会先由一个标注者进行标注，最终结果会由另一位标注者进行校对。

1.1 标注准则

词性标注应该基于词语的释义还是基于句法分布一直是一个焦点话题。这个话题从 1950s 开始不断地被讨论，但两种观点依然并存至今。举个例子，中文的“毁灭”一词翻译成英文可以是 destroy/destroys/destroyed/destroying/destruction，并且用法和英文中也是大致对应的。第一个观点认为词性标注只依赖释义。这些词虽然用法不同，但释义大致是一样的，应该都被标注为动词。第二个观点认为词性标注由语法分布决定。当“毁灭”是一个名词短语的开头时，它应该被标注为名词；当“毁灭”是一个动词短语的开头时，它应该被标注为动词。

我们的词性标注使用句法分布作为主要原则，因为它和当代语言学理论使用的原则是相匹配的。

很多人反对以句法分布为标注原则是因为，汉语中的很多动词可以出现在名词位置上，这就需要两个词性标签，会增加词典大小。我们认为这个观点不能令人信服，主要有两个原因。第一，两个词性标签让我们可以将动词区分开来，一类可以出现在名词的位置上，另一类不能。例如单音节的动词，AABB 和 A-不-A 等叠词形式的动词，都不能出现在名词的位置上。如果可以概括哪些动词可以出现在名词位置上哪些动词不能的话，就可以把这些概括表示成形态学的规则，那词典就可以自动扩展。如果不能概括的话，那这个名词化的过程就是特殊的词法现象，这些动词应该有两种词性。第二点，一些动词可以出现在名词位置上并不是汉语独有的，其他语言的标准处理中也给予了它们两种词性。

1.2 词性标签集合

我们的词性标签集合一共有 33 种词性。其中，由于“被”和“把”应该标注成介词还是动词仍然存在争议，所以我们会把它们分别标注成长被结构 LB，短被结构 SB，把 BA。

动词，形容词 VA, VC, VE, VV.

名词 NR, NT, NN.

定位词 LC.

代词 PN.

限定词和数词 DT, CD, OD.

量词 M.

副词 AD.

介词 P.

连词 CC, CS.

助词 DEC, DEG, DER, DEV, SP, AS, ETC, MSP.

其他 IJ, ON, PU, JJ, FW, LB, SB, BA.

1.3 词性标签集合的大小

假设我们现在只有一个广泛认同的很小的词性集合，包括名词，动词，副词，介词这种大类。那么我们该如何把某一个标签 T 替换成更具体的标签集合 $\{T_1, T_2, \cdots, T_i, \cdots, T_n\}$，这需要考虑以下几个因素：

1. 通常情况下，对应同一个词性标签的词语，它们的句法分布要完全一致，这就意味着我们需要一个比较大的词性标签集合。但是，词性标签集合越大，标注就会越困难。因为标注者不得不记住更多的标签，更多的用例，然后把它们很好地融会贯通。因此，我们在设计词性标签集合的时候需要做一下取舍。

2.$set(T)$ 是可以标注成词性 T 的词语的集合。如果不能把 $set(T)$ 中的每个词语都分配给某一个 $set(T_i)$，或者T_i 并不能够比 T 提供更多有用信息时，我们就不会把标签 T 分割为T_i。

3. 如果 $set(T)$ 是一个封闭集合，并且对于任意 i, j 而言，$set(T_i)$ 和 $set(T_j)$ 都无交集，也就是说，$set(T)$ 中的词语个数是可数的，并且每一个词语都只在唯一一个 set(T_i) 中出现，那么通过一个简单的转换程序就可以自动将某个标注成 T 的词改为标注成 T_i。因此，如果我们想要用T_i 中某个具体的标签来标注词语，这可以由程序自动完成，而无需由标注者手动完成。

1.4 处理疑难案例

有的时候，一段上下文中，某个词语是标注成 X 还是 Y 难以分辨。如果我们确切地知道这个词不在 $set(X)$ 和 $set(Y)$ 的交集中，那我们可以先简单地选择一个标签，比如 X。举个例子，“许多 学生”中的“许多”有可能是 JJ，也可能是 DT 或者 CD，但是它只能是其中某一个标签。我们可以简单地选择一个合适的标签，如果后面有需要的话可以自动替换。

假设在某一段上下文中，某个词语已经被标注为了 X，在另一段上下文中，这个词语应该被标注成 X 和 Y 还无法分辨。那么我们应该先确定，这个词语在这两个上下文中的句法分布是否一致。如果我们很确定它的句法分布是一致的，那么我们应该先把它标注成 X，否则应该把它标注成 Y。举个例子，(1)“又 高兴 又 难过”中的“又”可以是 AD，也可以是 CC。(2)“他 又 来 了”中的“又”很显然是 AD。因为我们不确定“又”在这两句中的表现是否一样，所以我们把（1）中的“又”标注成 CC。如果我们之后想把它改成 AD，只需要把“又/CC”换成“又/AD”即可。但是如果我们现在把它像（2）一样标注成 AD，并且之后我们想改成 CC 的话，那我们就需要区分一下这两个上下文，并且确认是否只有（1）中的“又/AD”需要替换成“又/CC”，而（2）中的不需要。

1.5 标记说明

本规范中使用的一些标记：

词性 N 表示所有的名词词性 NT/NN/NR。词性 V 表示所有的动词词性 VA/VV/VC/VE。“Det+M”是“DT+(OD|CD)+M”的简写，但 DT（限定词），OD（序数词），CD（基数词）可能出现也可能不出现。例如，“这三个”，“那一辆”都是 Det+M 结构。

“一个词语可以被否定”意味着“一个肯定意义的词语可以被否定”。类似地，“一个词语可以出现在 A-不-A 结构中”意味着“一个词语可以出现在疑问句结构 A-不-A 中”。

XP，YP，ZP 均指任意短语。

2 词性标签详细介绍

2.1 动词 VA，VC，VE，VV

通常，一个动词满足以下条件：

1. 除了助动词以外，动词可以充当一个句子的谓语，这个句子可以是主句也可以是从句。

2. 动词可以用“不”或者“没”来否定。

3. 大部分（非全部）动词后可以接体标记。体标记包括“着”、“了”、“过”等。

4. 大部分动词可以出现在 A-不-A 结构中。

如果一个可以被标注成 V 的词语是一个名词短语的开头，那它应该被标注成 N，而不是 V。如果一个可以被标注成 V 的词语是一个名词修饰语（除非这个动词是一个定语从句的开头），那它应该被标注成 N 或者 JJ，而不是 V。

2.1.1 表语形容词 VA

VA 可以大致地指向英语中的形容词和汉语语法文献中的状态动词。状态动词的定义大家莫衷一是。在我们的系统中，心理动词算作状态动词，比如“喜欢”，“了解”，“怨恨”等都会被标注成 VV。我们不会深入讨论汉语中形容词究竟是不是动词的子类。

我们的 VA 包括两类：

1. 没有宾语，并且可以用“很”来修饰的谓语。

2. 一些从第一类中引申出来的谓语，要么是叠词，比如“红彤彤”，要么是用 N+A 模式，意思是像 N 一样 A，比如“雪白”。这种类型的 VA 没有宾语，但是有一些不能被“很“修饰，因为它们本身就隐含着程度深的意思。

注：当一个可以被标注成 VA 的词语在没有“的/DEC”的情况下修饰一个名词，它应该被标注成 JJ 或者一个名词。当一个可以被标注成 VA 的词语有宾语时，它应该被标注成 VV。比如，这 项/M 活动 丰富/VV 了/AS 他的/DEG 生活。

2.1.2 系动词 VC

“是”和“为”应该被标注成 VC。当“非”的意思是“不是”，并且句子中没有其他动词的时候，“非”也应该被标注成 VC。

“是”有几种用法，所有这些情况中它都被标注成 VC。

1. 连接两个名词短语或主语，例如，他 是/VC 学生。

2. 在分裂句中，例如，他 是/VC 昨天 来 的/SP。分裂句指的是由单一的句子分裂成两个部分构成的句子。在汉语中，可以用“是”和“的”把一个句子分为两段，强调“是”后面的部分。如“张先生十五年前在上海读大学”就可以产生分裂句“张先生十五年前是在上海读大学的”。

3. 为了强调，例如，他 是/VC 喜欢 看 书。

第 2 点中的“是”后面可以接时间、地点等，参考英语中“it is …that …”形式的强调句。第 3 点中的“是”后面通常接动词，参考英语中的 do/does。

2.1.3 “有”作为主要动词 VE

只有当“有”，“没”，“无”作为主要动词时才被标注成 VE。包括表示所有、拥有的“有”，和表示存在的“有”，等等。

2.1.4 其他动词 VV

这一类包含了其他的动词，比如情态动词（如“能够”），提升动词（如“可能”），控制动词（如“要”，“想”），行为动词（如“走”），心理动词（如“喜欢”，“了解”，“怨恨”），等等。

2.2 名词 NR，NT，NN

一个名词可以是谓语或者介词的参数。通常来说，

1. 名词不能被表示程度和否定的副词修饰，比如“很”，“不”。

2. 很多名词可以被 Det+M 结构修饰。

3. 名词可以直接修饰名词，不用加“的”。

如果一个词语是一个名词短语的开头，它应该被标注成名词。有的时候很难说一个短语是否是一个名词短语，可以参考以下规则：

1. 如果一个短语被 Det+M 结构修饰，并且这个 Det+M 短语在其他上下文中只用来修饰名词短语，那么这个短语很有可能是一个名词短语。

2. 如果一个短语是一个动词或者介词的参数，并且在其他上下文中这些动词/介词的参数只能是名词短语，那么这个短语很有可能是一个名词短语。

3. 如果一个短语被“ZP（任意短语）的/DEG 或 DEC”修饰，那这个短语很有可能是一个名词短语。

2.2.1 专有名词 NR

专有名词是名词的子类。NR 指特定的人名，政治上或者地理上定义的地点（比如城市，国家，河流，山脉等），或者组织名（公司，政府组织或者其他组织）。一个专有名词通常是独一无二的，并且不能被 Det+M 结构修饰。

以下这些名字是 NR：

宗教，国家，郡县，城市；山脉，河流；报纸，杂志；组织，公司；学校，协会，基金会；人，家族。

以下这些名字不是 NR：

国籍（如“中国人”），人种（如“白人”），职称（如“教授”），疾病，职业，器官（如“肺”），乐器（如“钢琴”），游戏（如“足球”），花（如“玫瑰”），等等。

注：虽然 NR 和 NN 确实有一些区别，但我们作这样的区分主要的目的是为了实用。其一，它们在信息抓取这样的工作中很有用。其二，语料库中大部分未知的词语都是专有名词。要区分这两类有的时候有些困难，但其实把 NN 错标成 NR 或者把 NR 错标成 NN 对语法结构来说没有影响。

2.2.2 时间名词 NT

时间名词可以作为介词的对象，如“在”，“从”，“到”，“等到”。它们可以用“这个时候”来指向，也可以用“什么时候”来提问。它们可以直接修饰动词短语或主语。和其它名词一样，NT 可以是一些动词的宾语。

时间名词也包括“1990 年”，“一月”，“汉朝”这种，或者“PN 代词 +LC 方位词”（如“此后”），“N 名词 +LC 方位词”（如“战后”），“DT 限定词 +N 名词”这种形式。

例: 一月，汉朝，当今，何时，今后。

2.2.3 其他名词 NN

NN 包括其他名词，除了地点名词以外，一般不能够修饰动词短语，无论这个动词短语是否包含“地/DEV”。

2.3 方位词 LC

很多名词不能单独作为介词（如“在”，“到”）的参数，或者直接修饰动词短语/主语。方位词的一个功能是连在名词短语/主语后面，这样整个短语可以作为介词的参数或者直接修饰动词短语/主语。

方位词有两类：

1. 方位词：这一类的 LC 指的是方向，地点之类的词。它们是从名词中分出来的。一些可以单独作为介词/动词的参数。一些可以被“最”修饰。它们不能用 Det+M 修饰。

单音节的方位词，如，前，后，里，外，内，北，东，边，侧，底，间，末，旁。

双音节的方位词是在单音节方位词的基础上加上“以”，“之”等。如，之间，以北。

双音节的方位词也有可能由两个单音节方位词组成，如，前后，左右，上下，东北。

2. 其他：

为止：到 目前 为止/LC

开始：从 四月 开始/LC

来：5 年 来/LC

以来：1998 年 以来/LC

起：一九九三年 起/LC

在内：包括 他 在内/LC

2.4 代词 PN

代词是名词短语的替代，可以指代事先说明的或者从上下文中可以理解的人或物。它们一般不能被 Det-M 或者形容词修饰。

PN 包括人称代词（如“我”，“你”）；单独用作名词短语的指示词（如“这”，“此”）；所有格代词（如“其”）；反身代词（如“我自己”，“自己”）。

2.5 限定词和数词 DT，CD，OD

2.5.1 限定词 DT

这包括指示代词（如“这”，“那”，“该”），和一些像“每”，“各”，“前”（前/DT 7 个/M 月），“后”的词。DT 不包括基数和序数。

2.5.2 基数词 CD

CD 包含了基数词，也就是数字。基数词的后面可以接像“来”，“多”，“好几”这样的概数词。CD 也包括一些如“好些”，“若干”，“半”，“许多”，“很多”这样的词语。

例：1245，一百。

2.5.3 序列词 OD

序列词会被标注成 OD。我们把“第 +CD”看作一个词，标注成 OD。例如，第一百/OD。

有时候，虽然没有“第”，但隐含了顺序，如“一/OD 季度/NN”，在上下文中特指一年中的第一季度，所以“一”被标注为 OD。

2.6 量词 M

量词跟在限定词/数字后面来构成 Det+M 结构，来修饰名词/动词。它们包括量词（如“个”），群体量词（如“群”），和像“公里”、“升”一类的计量单位。

一些量词可以被有限的几个形容词修饰，如“一/CD 小/JJ 瓶/M 水/NN”。临时量词可以被名词或者形容词修饰，如“一/CD 铁/NN 箱子/M 书/NN”。

2.7 副词 AD

副词是一个很大的类。它包含了方式副词，频率副词，程度副词，连接副词等等。副词的表现形式各有不同。大部分副词的主要功能是修饰动词短语或者主语。

例：仍然，很，最，大大，又，约。

2.8 介词 P

一个介词可以以名词短语或者子句作为参数。

注：“把/BA”和“被/LB 或 SB”不被标注为 P，详见 2.11。

例：从，对（如“对/P 我 来 说”）。

2.9 连词 CC，CS

注：汉语语法中传统的连词根据它们的句法分布不同会被标注为 CC、CS 或 AD。CC 连接两个等价的，功能相同的部分（名词短语，子句或者其他）。CS 是在子句前面出现。连接副词通常在主句中出现，和从属连词 CS 成对出现（如“如果/CS …就/AD”）。

2.9.1 并列连词 CC

并列连词连接两个部分，每一部分的功能差不多，连起来作为一个整体。CC 的主要模式是 “XP , CC XP”，其中 XP 是任意短语。

例：与，和，或，或者，还是。

2.9.2 从属连词 CS

若一个子句从属于另一个句子，那么连接这两个子句的词语应该被标注为从属连词 CS。CS 有以下两种模式：

CS S1, S2.

S2 CS S1.

其中 S1 是从句，S2 是主句。第一种模式中，CS 也可以出现在 S1 的主语后面，如 “你 要是/CS 不 去，我 就 去 了/AS”。

例：如果/CS …就/AD …

2.10 助词 DEC，DEG，DER，DEV，AS，SP，ETC，MSP

2.10.1 “的”作为补语化成分或者名词化标记 DEC

这一类只包括作为补语化成分或者名词化标记的 “的” 和 “之”。例如，吃 的/DEC。模式是主语/动词短语 DEC {名词短语}。

注：“的” 还有其他标注：

DEG：他 的/DEG 车。

SP：他 是/VC 一定 要 来 的/SP。

AS：他 是/VC 在 这里 下 的/AS 车。

2.10.2 “的”作为所有格标记和关联标记 DEG

这一类只包括作为所有格标记和关联标记的 “的” 和 “之”。模式是名词短语/介词短语/JJ/DT DEG {名词短语}。JJ 是其他名词修饰器，详见 2.11.6。DT 是限定词，详见 2.5.1。

注：“的” 也可以标注为其他词性，如 DEC，SP 和 AS。

2.10.3 表示结果的“得”DER

在 “V-得-R” 和 “V-得”（如 “他 跑 得/DER 很 快”）结构中，“得” 应该被标注为 DER。V 是动词，R 是结果。

注：一些以 “得” 结尾的 “V-得” 结构的词语不是 DER，而是动词，如 “记得”，“获得”。

2.10.4 方式“地”DEV

这一类只包括出现在 “XP 地 VP” 结构中的 “地”，并且短语 XP 是修饰动词短语 VP 的。在一些古典文学作品中，“的” 也有这种用法，这时候 “的” 也需要标注成 DEV。

例：高兴 地/DEV 说。

2.10.5 动态助词 AS

暗示了动作发生状态的动词助词需要被标注为 AS。这一类只包括“了”，“着”，“过”，“的”。

2.10.6 句末助词 SP

SP 经常出现在句末。如，他 好 吧/SP?
有时候，也可以用来表示暂停，如，他 吧/SP，人 很 好。
例：了，呢，吧，啊，呀，吗。

2.10.7 ETC

ETC 用于“等”和“等等”两个词。ETC 有两种模式：
1.XP 等 NP：科技 、 文教 等/ETC 领域。
其中，XP 是任意短语，NP 是名词短语
2.XP 等/等等：科技 、 文教 等等/ETC。

2.10.8 其他助词 MSP

这一类包括“所”，“以”，“来”，“而”，且它们需要出现在动词短语前面。“所”出现在主语和动词短语之间，其他 3 个词都出现在介词短语/动词短语和另一个动词短语之间。虽然它们的句法分布不完全一致，但因为 MSP 的集合比较小，我们没有为“所”新增标签。
例：
所：他 所/MSP 需要 的/DEC
以、来：用 … 以/MSP（或 来） 维持
而：为 … 而/MSP 奋斗

2.11 其他词性 IJ，ON，LB，SB，BA，JJ，FW，PU

2.11.1 感叹词 IJ

感叹词总是出现在句首，如，啊。

2.11.2 拟声词 ON

拟声词是模仿声音的词语。ON 通常不能被副词等修饰。ON 的出现模式如下：
1. 在“ON 地 V”中修饰动词短语。如，雨 哗哗/ON 地/DEV 下 了/AS 一 夜。
2. 在“ON 的 N”中修饰名词短语。如，砰/ON 的/DEG 一声!
3. 独立成句。如，砰砰/ON！ 屋里 传出 两 声/M 枪响。
例：哗啦啦，咯吱。

2.11.3 长“被”结构 LB

这一类仅包括出现在长被结构中的“被”，“叫”，“给”（口语中），“为”（二声）。长被结构：名词短语 0+LB+ 名词短语 1+ 动词短语。如，他 被/LB 我 训 了/AS 一 顿/M。

注：“叫”有的时候会被标注成动词，如，他 叫/VV 你 去。

2.11.4 短“被”结构 SB

这一类仅包括出现在短被结构中的“被”和“给”（口语中）。短被结构：名词短语 0+SB+ 动词短语。如，他 被/SB 训 了/AS 一 顿/M。

注：“给”有其他词性，如 LB，VV 和 P。如，你 给/P 他 写 封/M 信。

2.11.5 把字结构 BA

这一类只包括出现在把字结构中的“把”和“将”，名词短语 0+BA+ 名词短语 1+ 动词短语。如，他 把/BA 你 骗 了/AS。

注：“将”还有别的词性，如 AD 和 VV。如，他 将/VV 了/AS 我 的/DEG 军。

2.11.6 其他名词修饰器 JJ

JJ 包括以下 3 类：

第一类，“区别词”（非谓形容词）：它们以“JJ+ 的/DEG+{N}”或者“JJ+N”的模式来修饰名词，但是它们没有“的”的帮助的话不能成为句子的谓语。它们不可以被程度副词修饰。

例：共同/JJ {的/DEG} 目标/NN，她 是/VC 女/JJ 的/DEG

第二类，带有连字符的复合词：这些词可以看做是定语从句和介词从句的简短形式。这些词通常有两个音节。其中一个或者两个是一个长单词的缩写形式。常见的这种类型的 JJ 有 V+N，P+N/LC，AD+VA 等。

模式：JJ+N，如，留美/JJ 学者/NN

第三类，形容词

模式：JJ+N，如，新/JJ 消息/NN

注：当形容词和名词之间加了“的/DEC”时，这个形容词应该被标注为 VA。

2.11.7 外文单词 FW

FW 是用来标注外语单词（foreign word）的。FW 不包括外文单词的翻译，也不包括那些中外混合的单词。例如“卡拉 OK/NN，A 型/NN”。它也不包括那些释义和词性标注在上下文中很明晰的词。我们应该尽量避免使用 FW，它只用来标注那些根据上下文不太能够确定词性的外文单词。

2.11.8 标点符号 PU

标点符号会被标注成 PU，但如果它们是某个词语的一部分的话，就不会被标注成 PU。

例：张三/NR ，/PU 李四/NR 和/CC 王五/NR。

3 词性标签难点讲解

3.1 易混淆的词性组合

词性标注在以下两种情形中可能会让人困扰：

1. 两种词性功能类似，比如 JJ 和 CD 都可以修饰名词。

2. 一些词语可以被标注为多种词性。例如，“政治”大部分时候是一个名词，但是在“只 能 政治/AD 解决/VV 这 个/M 问题”中，“政治”是一个副词 AD。

3.1.1 AD 和 AS

“在”出现在动词前面的时候，虽然也暗示了动作发生的状态，但还是应该被标注为 AD，而不是 AS。

注：其他副词可以出现在“在”和一个动词之间，但是它们不能出现在一个动词和动态助词之间。

3.1.2 AD 和 CC

在模式 X+ 主语/动词短语中，X 既可能是 AD，也可能是 CC。

CC 连接的两个等价的短语，而 AD 不是。一个连接副词经常和一个 CS 成对出现，它的功能是连接子句。

这两种词性都有的词语有：“又”，意思是“和”的时候是 CC（如“又/CC 唱歌 又/CC 跳舞”），意思是“又，再”的时候是 AD（“又/AD 下雨 了”）。“还是”，意思是“或者”的时候是 CC（如“喝 茶 还是/CC 喝 咖啡 都 可以”），意思是“仍然”的时候是 AD（如“你 还是/AD 做 错 了”）。

连接副词 AD 举例：否则，但是，但

3.1.3 AD 和 CS

在模式 X+ 主语/动词短语中，X 既可能是 AD，也可能是 CS。

CS 引导一个从句，它通常出现在从句中主语的前面。

可以被标注为 AD 的词语和可以被标注为 CS 的词语暂未发现重叠。

3.1.4 AD 和 JJ

在模式 X+ 名词短语中，X 既可能是 AD，也可能是 JJ。

一些副词可以修饰名词短语（如“又”，“又/AD 一 个/M 天/NN”，“才”，“才/AD 一 个/M 人”）。

一个比较好的区分 AD 和 JJ 的方法是在这个词和后面的名词短语之间加“的”。如果新的短语仍然通顺，那这个词就是 JJ，否则它就是 AD。

3.1.5 AD 和 NN

在模式 X+ 动词短语中，X 既可能是 AD，也可能是 NN。

时间名词和表示地点的词都可以直接修饰动词短语/主语，这些都是标注成名词的。对于其他名词来说，如果能在这个词前面加“在”并且不影响上下文的意思的话，就把这个词标注成 NN，否则标注成 AD。

例：政治/AD 解决/VV 这 个/M 问题；重点/AD 抓/VV 生产

3.1.6 AD 和 NT

在模式 X+ 动词短语中，X 既可能是 AD，也可能是 NT。

如果这个词 X 可以是一个名词短语的开头，那它就是 NT。例如，“昨天”就是一个 NT，而不是 AD。

AD：早日/AD 实现/VV

NT：目前/NT，今后/NT

3.1.7 AD 和 VA

VA 是不能直接修饰动词短语的，它需要“地”的帮助。但 VA 和 AD 都可以出现在“X+ 地 + 动词短语”结构中。

在模式 X+ 地 + 动词短语中，如果这个词作 AD 和作 VA 时意思一样，就把它标注成 VA，否则根据这个词在上下文中的释义进行标注。

在模式 X+ 动词短语中，X 只能被标注为 AD。

AD：大大/AD 提高

VA：高兴/VA 地/DEV，紧密/VA 地/DEV

3.1.8 AD 和 VV

在模式 X+ 动词短语/主语中，X 既可能是 AD，也可能是 VV。

VV 可以出现在 A-不-A 结构中，AD 不能。

像情态动词一样的 AD 有：大概，将，一定。

像副词的 VV 有：是否/VV。

3.1.9 AS 和 VV

在模式动词 +X 中，X 既可能是 AS，也可能是 VV。

目前，我们只认可四个动态助词 AS，“了”，“着”，“过”，“的”。像“写 完”里的“完”，“跑 起来”里的“起来”，我们都标注成 VV，而不是 AS。

3.1.10 CC 和 P

一些并列连词 CC（如“与”，“和”，“跟”，“同”）也可以看做介词。
在模式名词短语 0+X+ 名词短语 1 中，X 既可能是 CC，也可能是 P。

- 如果这两个名词短语可以交换，那么 X 就是 CC。
- 如果 X 前面有任何状语修饰的话，它就是介词 P。
- 如果名词短语 0 比名词短语 1 有更强的话题性或移情性，那么 X 就是 P。如，自由 比/P 生命 更 重要。
- 如果删去名词短语 0 后，必须删掉 X+ 名词短语 1 才能让语句继续保持通顺的话，X 是 CC。

3.1.11 CS 和 P

CS 的参数必须是主语，但是 P 的参数既可以是主语，也可以是名词短语。

CS 和 P 都可以接从句，但是“CS S1, S2”和“P S1, S2”之间是有区别的，S1 和 S2 是两个从句。我们假定 S1 和 S2 的主语是不同的。

- 在“P S1, S2”中，P+S1 可以被插入 S2 的主语和谓语之间。

  例：对/P 我 不能 按时 来， 他 感到 很 失望。

  他 对/P 我 不能 按时 来 感到 很 失望。

  虽然/CS 我 不能 去， 你 还是 应该 去。

  你 虽然/CS 我 不能 去 还是 应该 去。【不通顺，不成句】

- 在“CS S1, S2”中，CS 可以被插入 S1 的主语和 S1 的谓语之间。

  例：虽然/CS 他 不能 上场， 这 个/M 队 还是 会 赢。

  他 虽然/CS 不能 上场， 这 个/M 队 还是 会 赢。

  对/P 我 不能 按时 来， 他 感到 很 失望。

  我 对/P 不能 按时 来， 他 感到 很 失望。【不通顺，不成句】

- 在“P S1, S2”中，S1 可以用一个名词短语替代。

  例：对/P 我 不能 按时 来， 他 感到 很 失望。

  对/P 这 件/M 事， 他 感到 很 失望。

  虽然/CS 我 不能 去， 你 还是 应该 去。

  虽然/CS 这 件/M 事， 你 还是 应该 去。【语意不明，不成句】

例：
CS：虽然/CS。
P：对/P，由于/P，因为/P。

3.1.12 CD 和 DT

DT 包括限定词和不定量量词（如 “些”，“点儿”）。CD 包括定量量词（如 “对”，“打”）。当一个名词短语中同时出现 CD 和 DT 的时候，DT 会在 CD 的前面。CD 可以用于回答 “多少” 这样的问题。

词语诸如 “全体”，“全部”，“一切” 现在被标注为 DT。但如果我们发现别的词性更加合适也可以修改。

例：

CD：许多，若干，个把。

DT：各，全，某，这。

3.1.13 CD 和 JJ

CD 和 JJ 都可以修饰名词。JJ 用于描述某个属性来将某个个体区分开来，但是 CD 没有这个功能，CD 是数量词。

对于一些 JJ 而言，“JJ+ 的/DEG+N” 暗示着 “N 是 JJ 的”。如，好吃/JJ 的/DEG 苹果/NN。

例：

CD：一些，大批，好几，不少。

JJ：共同，女。

3.1.14 CD 和 NT

像 “一九九一”、“九一” 这样的词，如果它们的意思是 “一九九一年” 的话，它们应该被标注为 NT。如，“一九九一/NT 至 一九九五年/NT”。

3.1.15 DT 和 JJ

模式是 “X+N”。

DT 有可能是限定词，也可能是数量修饰词。而 JJ 是描述名词的属性。

“前” 有这两种词性：

JJ：美国 前/JJ 总统 尼克松

DT：前/DT 7 个/M 月

3.1.16 DT 和 OD

“首” 是 OD，不是 DT（如 “首/OD 届/M”）。

3.1.17 DT 和 PN

模式是 X 本身是一个名词短语，X+ 名词短语，X+ 的 + 名词短语。

只有指示代词（如“这”，“那”）既可能被标注为 PN，也可能被标注为 DT。在“X”和“X+的/DEG+ 名词短语”这两种模式中，指示代词应该被标注为 PN。在“X+ 名词短语”这种模式中，指示代词应该被标注为 DT。

例：

同时有 DT 和 PN 两种词性的词语：这，此。

这/DT 本/M 书 很 好。

这/PN 很 好。

只有 PN 词性的词语：他/PN 爸爸/NN，其/PN。

只有 DT 词性的词语：所有/DT 的/DEG 东西/NN。

3.1.18 JJ 和 NN

JJ 不能是名词短语的开头，但是 NN 可以。

为了一致性和简洁性，我们把以下几种类型处理成 NN，而不是 JJ：

N+ 形：椭圆形，V 形

N+ 状：带状，颗粒状

N+ 制：货币制，股份制

3.1.19 JJ 和 P

“有关”有 JJ 和 P 两种词性。当“有关”的意思是“相关的，涉及到的”的时候，应该被标注为 JJ。当“有关”是“关于”的意思的时候应该被标注为 P。

例：有关/JJ 单位，有关/P 撤军/NN 的/DEG 报告/NN。

3.1.20 LC 和 NN

LC 的一个功能是接在名词短语/主语后面，所以整个短语可以作为介词的参数或者直接修饰动词短语/主语。语料库中未发现这两个词性的重叠。

3.1.21 LC 和 MSP

“来”有 LC 和 MSP 两种词性。

LC：近年 来/LC

MSP：用 暴力 来/MSP 维持

3.1.22 M 和 NN

模式是 CD+X。

如果可以加一个量词在 CD 和 X 之间并且不影响句子含义的话，就把 X 标注成 NN，否则把 X 标注成 M。临时量词也要被标注为 M，临时量词指的是暂时充当量词的名词，并且可以被其他名词修饰。

例：一/CD 学生/NN，一/CD 年/M，一/CD 箱子/M 书/NN。

3.1.23 NN 和 NR

NR 指的是人名，组织名和国家名。它们通常都不能被 Det+M 结构修饰。

以下面这些字结尾的词语应该被标注成 NN，“方”，“军”，“人”（包括人种，血统），“会”（包括会议，委员会），“队”，“厅”。

3.1.24 NN 和 NT

以 NT 开头的名词短语直接可以修饰主语/动词短语，可以回答“什么时候”这类的问题。

除了“地点”名词以外的 NN 可以直接修饰主语/动词短语。

3.1.25 NN 和 VA

如果一个词语可以被“很”修饰，就把它标注为 VA/VV。

例：表示/VV 乐观/VA

3.1.26 NN 和 VV

如果一个词语 X 是名词短语的开头，就把它标注成 NN。如果 X 是动词短语的开头，就把它标注成 VV。详细点说就是：

- X 是名词短语的开头，也就是说 X 出现在 Det+M+X 结构中时，把 X 标注成 NN。
- 当 X 出现在 X+N 结构中修饰 N 时，如果 X 在没有“的/DEC”的定语从句或者补语从句中充当谓语，并且修饰名词的话，可以根据子句来对 X 进行标注，一般来说是 VA/VV。除了这种情况以外，把 X 标注成 NN。
- 状语短语可以修饰 X：如果 X 在上下文中被 YP+ 地/DEV 修饰（YP 是任何短语）或者被那种只能修饰动词的词语修饰，就把 X 标注成 VV。
- 当 X 是一个动词 V 的参数的开头时，需要再根据 V 来划分词性。如果 V 的参数是一个名词短语，就把 X 标注成 NN。如果 V 的参数是动词短语/主语，就把 X 标注成 VV。如果 V 的参数这两类都可以的话，就根据理解再定 X 的词性。

  区分动词的参数 X 是一个名词短语还是主语时，可以根据 X 后面是否跟了自己的参数来判断。如果 X 后面接着自己的参数，就认为 X 是一个名词短语。

  例如，“进行 改革 宗教”，“进行”的参数是一个名词短语，“改革”应该被标注为 NN。

  下列动词的参数是名词短语，而非动词短语：

  进行：进行/VV 密切/JJ 合作/NN

  加以：加以/VV 推进/NN

  受到：受到/VV 批评/NN

  给以：给以/VV 奖励/NN

予以：予以/VV 表彰/NN

注：“得以”后面可以接动词短语/主语，例如，“得以 提高/VV 教学 水平”。

3.1.27 NN 和 PN

在 DT+N 和 PN+N 结构中，它们都包含一个 PN 代词和一个名词短语。为了一致性考虑，我们把它们都标注成 NN 或者 NT。

NN（DT+N 结构）：本人/NN，本校/NN，全球/NN，当地/NN，我校/NN

NT（PN+N 结构）：当今/NT，何时/NT

3.1.28 P 和 VV

汉语中的很多介词都来源于动词，并且有些情况下，这些介词也能作为动词使用，所以要区分它们并不容易。

NP0 X NP1 YP，X 可能是 P，也可能是 VV：

- 如果整个句子中没有其他动词，那 X 是个动词。

  例：他/PN 在/VV 家/NN。

- 如果 NP1 缺失或者位置不对（比如在 VP-不-V 疑问结构中，VP 是动词短语，V 是动词），那 X 是个动词。

  例：他 在/VV 家 不 在/VV?

- 如果 X 后面可以接 AS，那 X 是个动词。

  注：有一些介词本身就以“着”或“了”结束（如“为了/P”），所以要注意。

- 一些词语这两种词性都可以，但是意义上有一些区别。

  例：靠，意思是“靠着，倚着”的时候应该被标注成 P，意思是“依赖”的时候应该被标注成 VV。

  他 靠/P 墙 站着。

  你 不 能 总 靠/VV 父母。

- 如果使用上述规则以后仍然有些不确定，就标注成 P。

3.1.29 VA 和 VV

VA 不能接宾语，几乎所有的 VA 都可以被“很”修饰，除了一些有“一定，确定”意思的词语。

一些词语有 VA 和 VV 两种词性，比如“丰富”，“少”，“多”。例如，“这儿 少/VV 了 一 本 书”，“这 书 很 少/VA”。

注：关于宾语还需要定义一下，如下。

现有一个动词和一个名词短语，P+NP+V 和 V+NP 结构都是符合语法的，这时候要把 NP 看做是动词的宾语，把动词标注成 VV。

例：他 对 我 很 关心/VV。

如果动词后面接了一个名词短语或 Det-M 结构表示程度，那它就不是这个动词的宾语。

例：鞋 大/VA 了 一号。

为了一致性，我们把起到动词短语功能的所有的四字成语都标注成 VV。

例：堂堂正正，欣欣向荣，有条不紊，扎扎实实，不得人心，熟视无睹，微不足道，危机重重，闻名于世，无人不晓，变幻无常，别具一格，并行不悖，波澜壮阔，出人意料，死灰复燃，腾空而起，同室操戈，万事如意，稳中求进，握手言和，息息相关，相依为伴，相映成辉，一分为二，因地制宜。

起到名词短语功能的四字成语应该被标注为 NN。

例：无名之辈，有识之士。

3.1.30 VE 和 VV

只有“有”，“没 [有]”，“无”作为主要动词时会被标注为 VE。把这些词语标注成 VE 的话可以很方便地从语料库中找到“存在句”。像其他动词，比如“具”，“具有”，“拥有”，“富有”这些都是正常标注成 VV。

3.2 特殊的词语和搭配

3.2.1 的

“的”有 4 种词性，DEC，DEG，AS 和 SP。

- DEC：我 买 的/DEC 书
- DEG：我 的/DEG 书
- SP：他 是/VC 一定 要 来 的/SP
- AS：他 是/VC 在 这里 下 的/AS 车

3.2.2 等，等等

“等”和“等等”在以下两种结构中应该被标注成 ETC：

- XP 等 NP：科技 、 文教 等/ETC 领域。

  其中，XP 是任意短语，NP 是名词短语
- XP 等/等等：科技、文教 等等/ETC。

3.2.3 来

“来”有好几种词性：

- VV：他 来/VV 了/AS。他 拿来/VV 了/AS 一 本/M 书。他 走/VV 上来/VV。
- SP：他 拿出 书 来/SP。
- MSP：他 用 这 个/M 来/MSP 证明 他 是/VC 无辜 的/SP。
- LC：3 年/M 来/LC。
- CD 的一部分：30 来/CD 个/M。

3.2.4 连

“连”在“连…都…”中被标注为 AD。
例：他 连/AD 我 都 不 认识。

3.2.5 有

- “有”是句中主要动词时应该被标注为 VE。
- “没有”可以被标注为 VE 或 VV。当“没有”是句中主要动词的时候被标注成 VE，如“他 没有/VE 书”。当“没有”修饰一个动词短语的时候被标注成 VV，如“他 没有/VV 来/VV”。
- “没”可以被标注为 VE 或 AD。当“没”单独作为句中主要动词的时候被标注成 VE，如“他 没/VE 书”。当“没”修饰一个动词短语的时候被标注成 AD，如“他 没/AD 来/VV”。当它出现在 A-没-A 结构中时也被标注成 AD，如“他 来 没/AD 来?”。
- “有没有”可以被标注成 VE 或 VV。当“有没有”作为句中主要动词的时候，应该被标注成“有/VE 没有/VE”。当“有没有”修饰一个动词短语的时候，应该被标注成“有/VV 没有/VV”。

3.2.6 这样

PN：这样/PN 的/DEG 伙伴/NN 关系/NN。
AD：你 这样/AD 做/VV 不 对。
VV：你 别 这样/VV。