

词性标注中难归类词语分析

邢红兵

北京语言文化大学语言信息处理研究所

北京师范大学心理系

Email: xinghb@blcu.edu.cn

摘要: 本文依据 200 万字经过人工校对的分词和词性标注的语料, 对其中的词性标记不一致但不属于兼类或同形的词语进行分析。文章根据这些标记不一致词语的产生原因将它们分成三大类, 并具体分析各类中的主要类型。

关键词: 词类, 词性标注, 难归类词语

Analysis on Words Hard to Be Classified in Part of Speech Tagging

Xing Hongbing

Language Information Processing Center, Beijing Language and Culture University

Psychology Department, Beijing Normal University

E-mail: xinghb@blcu.edu.cn

Abstract: This paper is based on a corpus which has 2 million Chinese characters and has been segmented and tagged. We analyze the words which have different part of speech tags except polysemy and homonymy. We classify these words which have different part of speech tags into three types, then we analyze the main words of each type in detail.

Keywords: parts of speech, part of speech tagging, words hard to be classified

一、引言

词性标注实际上涉及两方面的问题: 一是制定词性标注体系; 二是按照规范对具体语料进行自动标注和人工校对。目前各家标注体系虽有一定的差别, 但在大类上基本一致, 这方面应该说不会有太多问题。但是在第二个方面, 即给具体的词归类的时候, 会出现一些问

题。我们从语料库加工的经验中得到的体会是：规范只是基本原则，它不可能下辖每一个词语，在具体的词性标注过程中，大部分词语的词性是明确的，但有一部分词语常常存在难以找到合适的类、或者可以归入不同的类等问题。本文将以 200 万字经过人工校对的分词和词性标注的语料为依据，对词性标注的结果进行分析。文章分析的主要对象是非兼类、非同形的有多个标记的词语。

二、关于难归类词语

2.1 语料库中使用的词类标记

该语料库使用的标记共 119 个，其中词类标记 95 个，标点符号标记 24 个，这 95 个词类标记采用层级体系，最多有三个层级，如“npf”中“n”是最高层，表示名词，“p”是第二层，表示名词中的专有名词，“f”是第三层，表示名词中的专名中的外国人名（不包括日本人名及朝鲜、越南的汉式姓名）。最高层有 22 个大类，它们是：名词（n）、动词（v）、形容词（a）、状态词（z）、区别词（b）、时间词（t）、处所词（s）、方位词（f）、数词（m）、量词（q）、副词（d）、代词（r）、象声词（o）、叹词（e）、连词（c）、介词（p）、助词（u）、语气词（y）、插入语（l）、成语（i）、词缀（k）、阿拉伯数字、英文等（x）。本文分析难归类词语主要涉及大类，部分词语考虑到了小类。

2.2 什么是难归类词语

语料库中带标记的切分单位包括：各类词、各类语、词缀、语素、非汉字（包括阿拉伯数字、英文、符号、公式、段落标记、标点等）。本文分析的对象只是各类词、各类语。这些词语从词性标记角度大致可以分为以下几类：

A 类：意义固定，语法功能单一的词语。比如大部分名词、动词、短语（包括成语、惯用语等）、形容词等，它们无论是意义还是语法功能都比较单一，这类词语不会有兼类的问题，比较容易归类，这类词语在语料库中占有很高的比例。

B 类：具有明显不同的语法功能，意义上有很大差别的兼类词（包括同形词），这类词应该归入两个或多个类中，比如一部分词就兼动词和名词。

C 类：要不要看成兼类还存在不同看法。这类词往往具有不同类型的句法功能，但是意义上没有什么不同，因此，要不要看成兼类还值得研究。

D 类：语法功能和意义方面可以归入两个或多个类中，但是又不属于兼类的情况，比如形容词和不及物动词。

E类：语法功能不典型，找不到合适的类。

本文所研究的难归类的词语只包括C、D、E类三种情况。

2.3 从词表看难归类词语

2.3.1 语料库的标注结果

我们对语料库中出现的全部词语进行了统计。该语料库中标有不同词性标记的切分单位共有 75684 条（其中包括标点符号、阿拉伯数字、英文、篇章标记、段落标记、其他符号等），如果不考虑兼类和同形，共有 68937 条。每个单位的标记数从一个标记到 8 个标记不等，具体数据见下表。

	单标记	2 标记	3 标记	4 标记	5 标记	6 标记	7 标记	8 标记
数量	63655	4240	743	215	53	25	3	3
比例 (%)	92.3379	6.1505	1.0778	0.3119	0.0769	0.0363	0.0044	0.0044

2.3.2 难归类词语在语料库中的分布特点

从上表可以看出，语料库中的词语包括两类：单标记词语和多标记词语。经过分析，我们发现难归类词语在两类中都有分布，但主要分布在多标记词语中，因为在对这类词语进行标记时，不同的校对者看法可能不一致，即使是同一校对者，也会产生前后不一致。具体地说，单标记词语大部分属于 A 类，少数词语属于 E 类情况；难归类词语主要集中在多标记词语中。从分析结果看，难归类词语和易归类词语是混在一起的，因此，我们必须对全部的词语进行分类。由于时间的限制，本文分析的重点是两标记和三标记词语中的一些常见的类型。

三、 难归类词语分析

3.1 C类词语

名词和动词 这两类词大部分属于兼类，比如“工作”、“贡献”、“翻译”等，这些词充当名词和充当动词的句法功能完全不同，意义也有很大的变化。但是，有一部分这类词语，由于他们在句中所处的位置导致在是不是兼类的问题上也有分歧。这些词语主要有：

保障 供应 估计 观察 会谈 交流 交往 结束 竞争 渴望 控制 期待 期望 请求 审判 谈判 讨论

这些词语的特点是能够自由地出现在主宾语和定中结构的中心语位置上，很容易被标成名词，但是在意义上它们和处于谓语位置的动词没有任何区别。

副词 副词的问题主要集中在和时间词、动词、形容词及部分名词的区分上，因为这些词有一个重要的功能就是作状语，对它们的标注很容易出现不一致。

副词和时间词的问题比较突出，这是因为时间词的主要功能是充当状语，因此，很多时间词往往被标成了副词。应该说，副词和时间词的区分还是有一定标准的，那就是：时间词除了可以做状语以外，还可以作主语、宾语和定语。下列词语就是分歧比较大的词语。

不多久 不久 不一会 长期 常年 彻夜 初 从此 打小 当年 当时 到头来 短期 多咱 尔后 方才 好久 很久 回头 届时 近来 近年来 开始 开头 连日来 猛然间 平常 平生 平素 起初 起先 迄今 顷刻 日夜 事先 素来 突然间 下次 现 须臾 一开始 一刻 一时 一时间 一天到晚 一早 有生以来 有一天 原 原来 原先 早 早晚 至今 终 终年 终日 终身 终生 昼夜 转瞬间 转眼 自古 自古以来 自小 最终

一般来说，动词和副词的区别是比较明显的，因为它们往往是修饰语和中心语的关系，但下列两个方面存在问题：(1) 这些词语既可以作状语，也可以作谓语，但常常是出现在状语的位置上，修饰别的动词。因而对这类词语出现在状语位置上的标记常常不一致。(2) 少数词语存在分歧是由于句法分析的问题，如“继续前进”、“难免出现问题”、“开始出现”可能有动宾和状中两种不同的分析方法。这些词包括：

公开 几经 继续 加紧 加速 交错 交替 尽力 开始 联合 联手 埋头 没准儿 免费 难免 难以 破例 抢先 使劲 顺路 提前 统一 玩命 协作 照常 抓紧

副词和形容词的情况类似于副词和时间词，因为很多形容词都是可以作状语，下列形容词用作状语时，也常常产生分歧。

彻底 分明 疯狂 故意 过分 慌忙 即时 具体 绝对 没准 秘密 勉强 任意 适当 完全 无意 意外 主动 准时

区别词 语料库中标注的区别词共 1226 条，单标记的 585 条，多标记的 641 条，按照规范规定，区别词和副词可以兼类，那么，主要的分歧在于区别词和名词、区别词和动词的划分。例如：

bn: 本职 本质 常规 电子 分支 工商 固体 核心 客观 民间

bv: 常备 倒数 定向 独资 对外 辅助 复合 合资 假冒 模拟 涉外 特有 贴身 未婚 相关 自选

这些词语虽然具有名词或动词的语法特征，但常常作修饰成分，而且有一定的区别意义，容易出现分歧。

3.2 D类词语

形容词和不及物动词 形容词和不及物动词是汉语词类研究的难点之一，在语料库加工过程中，这类问题也比较突出。常见的这类难归类词语如下：

爱国 安心 饱和 不满 不适 惭愧 馋 缠绵 长寿 畅通 畅销 吵 沉默 沉 称职
成熟 吃惊 吃亏 迟疑 出名 出众 刺眼 错 担忧 胆怯 丢脸 懂事 懂 动情 恶心
饿 烦人 放心 费事 管用 害羞 后悔 昏 惊讶 渴 口渴 苦恼 满 忙 疲软 奇怪
失望 听话 头痛 惋惜 为难 委屈 闻名 吓人 小心 遗憾 有效 胀 知足 执着

这类词语有一些共同特点：可以受程度副词的修饰，可以受否定副词“不”的修饰，不能带宾语，可以带程度补语，因此区分它们有一定难度。

概数词和数量词 规范中规定了两个量词小类：概数词和数量词，前者的特点是表示大概的数目，可以带上量词来修饰名词，如：“无数个烈士”、“若干个代表”等，但是，从下面的标注结果来看，两类还是有很多交叉的词语，可见这个标准并不好把握。这类词语主要有：

概数词：不少 大半 多少 多数 多种 个把 好多 好几 好些 好些个 若干 少数 无数 许多 诸多

数量词：半拉 半数 不少 大半 大部分 大多 大多数 大量 大批 多 多少 多数 个把 个别 好些 好些个 若干 少量 少数 少许 无数 些许 许多 一点 一点点 一些 众多 诸多

这两类词语还有和代词、形容词交叉的问题，例如：“多少、多数、个别、有些”等常常被标为代词；“半拉、不老少、不少、大量、众多”等常常被标为形容词。

代词和名词 代词是个封闭的类，这类词语虽不多，但也存在一些问题，主要就是有一部分词语到底是代词还是名词的问题。例如：

本人 本身 敝人 别人 个人 各位 男的 女的 旁人 其人 全部 全体 人们 他人 有的 有人 自身 自我

应该说区分名代词和名词的标准不能是语法标准，因为他们的语法功能完全相同，所以主要依靠意义标准来进行区分。

3. 3 E 类词语

数词前置、中置、后置词 这类词语之所以难以归类，主要存在两方面的问题：第一个方面是它们本身不表示数，它们的功能是辅助数词表示一定的数量的，因此，把它们归入数词中是不是合适。这些词语有：

数词前置词：不到 不过 不满 不下 不止 初 第 好 将近 近 快来 上 摄氏 首 头 小 有 约 整 整整

数词后置词：把 倍 出头 多 多点 挂零 光景 见方 开外 来 强 上下 许 以上 以下 有 有余 余 整 正 之多 左右 啷当

数词中置词：点 分之 零 有

第二方面实际属于C类问题，即要不要看成兼类的问题。比如“快”、“约”、“整”、“整整”等词要不要看成前置词和副词的兼类；“不到、不过、不满、不下、不止、上、近、差不多”等词要不要看成前置词和动词的兼类；“以上、以下、上下、左右”等词，要不要看成是数词后置词和方位词的兼类。

其他助词 在助词中有一个小类是其他助词，这类词语常常放在句末，不充当句子成

分，意思也比较虚化，很多只表示语气，所以，给这类词归类比较困难。

罢了 不成 不过 不可 不说的话 等 等等 而论 而外 而言 而已 好不 好了 价 看来 来讲 来看 来说 来着 起见 说来 算啦 算了 为止 许 也罢 也成 也好 一来 一类 一气 一通 与否 云 云 再说 在内 之分 之际 之久 之类 之一 之余

从标注结果来看，一部分这类词语被标成语气词，例如：“罢了、不可、好了、而已”等。

另外，对下列词语的归类还存在很大的分歧，有的标记达到5种，这些词语有：

部分 (bmq) 长期 (abnt) 初 (bdfnt) 单个 (bm) 个个 (nqr) 开头 (dfnt) 每当 (dp)
每逢 (dpv) 其次 (cn) 全 (abdr) 任何 (br) 首先 (cd) 丝毫 (dmq) 同时 (cdnt) 系列 (bmq)
先 (dfntv) 一点点 (dmq) 整个 (abdr) 终年 (dntv) 最后 (bdft)

产生分歧的原因是在判别的时候考虑的因素不同，有的以句法功能为主，有的只考虑了意义。

四、 结论

通过对语料库中多标记的词语进行分析，我们可以看出，这些多标记的词语包括三个部分：(1) 应该有不同标记的词语，包括兼类和同形等。(2) 标注错误而产生的多标记词语，这类词语是由于校对者的失误或者没有严格按照规范执行。(3) 难以归类的词语。当然，本文还没有涵盖全部的难归类词语，要想使语料库的标注结果更加准确一致，除了尽量减少校对错误以外，最重要的还是要从这些难归类的词语着手，对汉语词类问题进行进一步的专门研究。

参考文献

- [1] 吕叔湘，汉语语法论文集（增订本），商务印书馆，1984年。
- [2] 朱德熙，语法讲义，商务印书馆，1997年。
- [3] 胡明扬主编，词类问题考察，北京语言学院出版社，1996年。
- [4] 胡裕树、范晓主编，动词研究综述，山西高校联合出版社，1996年。
- [5] 俞士汶等，现代汉语语法信息词典详解，清华大学出版社，1997年。
- [6] 邢红兵，现代汉语词类使用情况统计，浙江师大学报（社会科学版），1999年第3期。
- [7] 邢红兵，从分词的不一致性看汉语分词的难点，“机器翻译与计算机语言信息处理国际学术研讨会”论文，1999年6月，北京。