Modern Linguistics 现代语言学, 2021, 9(2), 524-529
Published Online April 2021 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2021.92072

Hans汉斯

兼类词概率分布计量考察及语法搭配模式在中文信息处理中的应用

王浩学，徐艳华*

鲁东大学文学院，山东 烟台
Email: 2226900937@qq.com, *ysyh0401@163.com

收稿日期：2021年3月25日；录用日期：2021年4月21日；发布日期：2021年4月29日

摘 要

在词性标注的过程中，汉语中兼类词的存在是影响词性标注准确率的主要原因。本研究以三部词典标注一致的78个形名兼类词为测试对象，基于规则和统计相结合的词性标注方法，将统计的兼类词分布概率与语法搭配规则结合起来，利用兼类词语法搭配模式构建规则库，对国家语委现代汉语通用平衡语料库标注的兼类词结果进行修正，准确率可以提高14.57%。

关键词

兼类词，语法搭配，语料库应用，词性标注

A Study of the Probability Distribution and Grammatical Collocation Patterns of Multi-Category Words in Chinese Information Processing

Haoxue Wang, Yanhua Xu*

College of Literature, Ludong University, Yantai Shandong
Email: 2226900937@qq.com, *ysyh0401@163.com

Received: Mar. 25th, 2021; accepted: Apr. 21st, 2021; published: Apr. 29th, 2021

*通讯作者。

文章引用：王浩学，徐艳华．兼类词概率分布计量考察及语法搭配模式在中文信息处理中的应用[J]．现代语言学，2021, 9(2): 524-529. DOI: 10.12677/ml.2021.92072

Abstract

In the process of part-of-speech tagging, the existence of multi-category words in Chinese is the main reason that affects the accuracy of part-of-speech tagging. In this study, 78 adjective-noun multi-category words of the same part-of-speech tagging in the three dictionaries are the test objects. The part-of-speech tagging method based on the combination of rules and statistics combines the statistical distribution probability of multi-category words with grammatical collocation rules, and builds a rule database using the grammatical collocation mode of multi-category words. The rule database corrects the results of the multi-category words tagged by the modern Chinese corpus of State Language Commission, and the accuracy rate can be increased by 14.57%.

Keywords

Multi-Category Words, Grammatical Collocation, Corpus Application, Part-of-Speech Tagging

1. 汉语中兼类词的现象

在汉语本体研究中，“词类问题可以说是一个老大难问题，长期以来众说纷纭，很难取得一致意见”[1]。这其中的原因之一是因为汉语中有大量的兼类词。兼类现象是语言中客观存在的，词的使用范围越广、使用频率越高，不同的用法就越多，产生兼类的可能性就越大。以《现代汉语八百词》为例，这是一部供非汉族人学习汉语使用的工具书，所收的都是高频常用词，而且以虚词为主，也收一部分实词，尽管所收实词不多，但该工具书中的“兼类词所占比例竟高达 22.5%”[2]，也就是说 800 个词里有 180 个词是兼类词。张虎(2004)等人对北京大学计算语言所公布的 200 万汉字语料进行统计，结果显示“兼类词占到 11%，但兼类词的词次却占到了 47%”[3]。兼类词往往集中在常用的基本词汇中，使用程度高，覆盖面广，兼类情况复杂。汉语兼类词的这些特点，如果还用传统的、面向人的语法信息来服务于中文信息处理是远远不够的，应该从中文信息处理的实际需要出发，细化语法描写的颗粒度，构建符合汉语特点较完备的语法规则库。

2. 词性标注的方法与规范

当今主流的词性标注方法可以分为三类：基于规则的词性标注方法、基于统计的词性标注方法、规则和统计相结合的词性标注方法。

1、基于规则的词性标注方法

基于规则的词性标注方法是一种理性主义的研究路线。该方法是利用语言学理论与知识，建立并利用规则库进行词性标注。这一方法的局限性主要表现在三个方面：1) 规则是由人工统计归纳出来的，具有较强的主观性，难以形成统一的规则标准，并且建立完备的规则库成本高昂，需要大量的人力。2) 对于大规模复杂语料标注效果不佳，很难挖掘到语言中较为边缘化的规则。3) 新旧规则的兼容性较差，前后规则容易相互冲突，增加规则时需要注意其与已存在规则的优先级与适用性。影响基于规则的词性标注准确率主要因素有两个：一是规则的准度，即规则本身是否正确严谨，二是规则的广度，即规则能否

覆盖语料的细微方面。这两个因素成反比关系，在实践层面很难协调，看重准确率的话可能覆盖率会受影响，看重覆盖率就又不能保证准确率。随着计算机技术的发展，出现了利用机器学习原理获取词性标注规则的方法，其主要研究成果有基于决策树(Decision Tree, DT)方法、基于转换的错误驱动(Transform-Based Error Driven, TBED)方法等。

2、基于统计的词性标注方法

基于统计的词性标注方法是一种经验主义的研究路线。随着计算机科学技术的迅速发展，基于大规模语料的计量统计研究成为可能，可以根据概率模型计算某个词所属某些词性的概率，并筛选其中最大概率的标记序列作为标注结果。概率计算是通过大规模语料来进行的，所以我们先要通过语料库对模型进行训练。其词性判断标准通过概率统计获得，客观性更强，减少了人工规则归纳的主观性，并且借助于计算机的性能，大幅提高了对大规模文本的标注速度。但是基于统计的方法也有不少的局限性。基于统计的方法即使不断扩大训练语料的规模，也会出现数据稀疏的情况，导致产生零概率问题，数据平滑算法是解决数据稀疏问题的方法之一。基于统计的词性标注方法是目前主流的研究方法。基于统计模型的词性标注方法的典型研究成果有 CLAWS 系统、最大熵(Maximum Entropy, ME)、隐马尔可夫(Hidden Markov Model, HMM)、条件随机场(Condition Random Field, CRF)、神经网络(Neural Networks, NN)等基于统计模型的词性标注方法。

3、规则和统计相结合的词性标注方法

基于规则和基于统计的两种方法各自具有优越性和局限性，结合上述两种方法的优缺点，规则和统计相结合的方法是大势所趋的，规则消歧与统计概率结合起来，互相克服各自的缺点，既改进统计方法中的数据稀疏等问题，也弥补规则方法中覆盖面小、规则不完备等缺陷，从而提高词性标注的准确率。典型的研究成果有张民等人“通过研究统计的可信度，引入了置信区间的方法，建立了一种基于置信区间的评价函数，统计和规则相结合进行词性标注”[4]。

纵观上述三种标注方法，它们都不同程度地忽略了汉语本体研究中的语法搭配模式，未能重视语法搭配知识在词性标注尤其是兼类词标注中的作用。传统的语法搭配模式虽然属于人工归纳，数量及效率上不如机器的性能强大，但语法搭配模式由人的认知模式归纳而出的，应当是最高优先级的规则，因为当我们利用标注结果去与正确答案比对时，其正确的标注答案也是人工标注的，人之所以能够标注出完全正确的答案，其根本依据还是人的语法思维，而不是机器的标注逻辑，所以依据人工归纳的语法搭配模式作为统计结果的辅助手段是具有价值的，且当标注矛盾时，依据语法搭配模式构建的规则应该具有比其他规则更高的优先级。

4、词性标注规范

词性是词的语法性质的分类，是按其在短语或句子中充当句法成分的能力来划分的，尽管汉语这一划分词类的标准已被学界认可，但由于汉语词类问题比较复杂，在“汉语中的词究竟包含哪些类”“每类所辖词汇范围”等问题上一直存在争议，没有完全统一的标准。汉语本体研究的这一问题在中文信息处理领域也同样存在分歧，分词与词性规范也没有完全统一的标准。目前，比较有影响的是国家语委制定的《信息处理用现代汉语词类标记规范》和北京大学制定的《北京大学现代汉语语料库基本加工规范》，两者的词性分类具有明显差异。本研究的研究对象是国家语委现代汉语通用平衡语料库中标注的兼类词，所以本研究以《信息处理用现代汉语词类标记规范》为词性标注规范。

3. 基于语法搭配模式的消歧策略设计思想

本研究的设计思想是基于规则和统计相结合的词性标注方法，利用兼类词语法搭配规则对基于概率统计模型的大规模词性标注加以辅助，也就是在概率统计后利用规则消歧的方法。汉语语法现象纷繁复

杂，语法规则数量庞大，我们曾对清华大学 100 万字的树库进行统计，结果发现语法规则竟多达 14000 多条。因此，想单纯利用语法规则实现大规模语料标注是不现实的。但我们知道，汉语词类划分的标准是语法功能，兼类词作为词类中的特殊现象，其判断标准也应该是语法功能。鉴于此，本研究将归纳的兼类词各词性搭配模式进行正则代码化，转换为计算机语言，用其对国家语委现代汉语语料库中的 78 个兼类词的标注结果进行校验。以“不幸”一词为例来阐述这种消歧策略的测试过程及结果。

1、语法搭配模式归纳

以朱德熙先生的“词组本位”为理论依据，分析了 7800 个实例(每个词抽取 100 个例句加以分析)，归纳语法搭配模式。以“不幸”为例，通过分析实例，归纳了它的主要语法搭配模式：

1) “不幸”作形容词的语法搭配模式

① 不幸 + 名词：意大利建筑界称这段历史为“[不幸]时期”。

② 副词 + 不幸：急促的电话铃声给伊丽莎白和丈夫佛兰克带来了一生中最[不幸]的消息。

③ 不幸 + 动词：2001 年，Y[不幸]发生意外，Z 向保险公司报案，保险公司决定给付 16 万元人身保险金。

④ 不幸 + 介宾结构：4 月 17 日，她[不幸]被感染。

2) “不幸”作名词的语法搭配模式

① 人称代词 + 的 + 不幸：聪明的人都能忘掉他们的[不幸]，适应他们的生活。

② 形容词 + 的 + 不幸：不在陆地上进行武装干涉，以拯救希腊的崩溃，这将是最大的[不幸]。

③ 动词 + 不幸：但，并未料到会有[不幸]。

④ 指量短语 + 不幸：在这个灾难或者说[不幸]面前，我觉得人们更需要理性，我渴望大家能够理性地面对。

2、转换规则代码

在得到“不幸”一词的语法搭配模式之后，以《信息处理用现代汉语词类标记规范》的词性标注规范为标准，将其转换为计算机语言的正则模式如下：

```
pattern_bxa1=r'不幸/[a-zA-Z]+[一-龥]+/n'
pattern_bxa2=r'[一-龥]+/d 不幸/[a-zA-Z]+'
pattern_bxa3=r'不幸/[a-zA-Z]+[一-龥]+/v'
pattern_bxa4=r'不幸/[a-zA-Z]+[一-龥]+/p[一-龥]+(/n|/r)'
pattern_bxa5=r'不幸/[a-zA-Z]+[一-龥]+/u'
pattern_bxn1=r'[一-龥]+/r 的/u 不幸/[a-zA-Z]+'
pattern_bxn2=r'[一-龥]+/a 的/u 不幸/[a-zA-Z]+'
pattern_bxn3=r'[一-龥]+/v 不幸/[a-zA-Z]+'
pattern_bxn4=r'[一-龥]+/r[一-龥]+/q[一-龥]*不幸/[a-zA-Z]+'
```

将正则表达式导入文本文件，逐行读取，匹配所输入的字符串或文本文件，导出所匹配的数量以及具体内容，并且将修正结果再次导出，对比前后标注文件的不同并计算其修正率。

3、测试结果及分析

仍以“不幸”一词为例，通过规则代码进行匹配，我们可以发现国家语委现代汉语通用平衡语料库中对“不幸”一词词性标注不当之处，如通过“不幸”作名词用的第一条规则“人称代词 + 的 + 不幸”，我们发现现代汉语通用平衡语料库将其全部判断为了形容词，如“各/r 的/u 不幸/a”“别人/r 的/u 不幸/a”。同样，“动词 + 不幸”的搭配模式中，“不幸”应判断为名词，现代汉语通用平衡语料库也都判断成了形容词，如“炫耀/v 不幸/a”“有/v 不幸/a”“忘记/v 不幸/a”“遭/v 不幸/a”“分担/v 不幸/a”等。这

种应当判断为名词却判断为形容词的情况比较普遍，仅在“不幸”一词的 100 条随机有效语料中，标注错误的句子就有 14 句，用我们的消岐策略全部可以修正过来，修正率为 14%。具体情况见表 1。

Table 1. Comparison table of part of speech tagging results of ‘不幸’
表 1. “不幸”词性标注结果前后对比表

词	语委自动标注		规则修正后		修正率
不幸	/a:100 条	/n:0 条	/a:86 条	/n:14 条	14%

通过仔细分析标注错误的实例，大都是忽略“不幸”的名词属。出现这种情况，并不是因为语委语料库词性标注的标准中把“不幸”只看作形容词而不看作名词。如果我们在语委语料库中搜索“不幸/n”，会出现 19 条检索结果，这说明语委语料库对于“不幸”的词类是包含名词的，承认“不幸”是形名兼类。在这 19 条检索结果中，“历史造成的不幸”“最大的不幸”等结构是符合上文两条“的 + 不幸”规则的；“种种不幸”“一段不幸”等结构是符合“指量短语 + 不幸”规则的；“带来不幸”“有过不幸”等是符合“动词 + 不幸”规则的。这说明语委的词类分类和我们的规则相一致，但是在进入自动标注后，就全然舍弃了名词的用法。我们将语料库中已经判断为“不幸/n”的例句再次进行自动标注，仍旧会判断为形容词。如“带来不幸和灾难”是检索“不幸/n”的例句之一，标注结果为“带来/v 不幸/a 和/c 灾难/n”，“不幸”仍判断为形容词。诸如此类的矛盾情况是较为普遍，这说明语委语料库的词性标注功能系统存在一定的缺陷，这也是基于概率模型的标注方法难以解决的问题，存在这一问题的原因可以归纳为以下两个方面：

1) 由于其采用的是“机助人校”的语料标注方法，而未对自动词性标注中的底层语法规则进行归纳总结，在机器标注后单纯依据人工校对修正部分兼类体系，并未针对大量的兼类现象进一步完善其消歧方法。

2) 由于训练样本规模不足而导致数据稀疏问题。语委的语料库相对于其他语料库规模偏小，根据最大似然估计，会将一些词的非常用词性的概率判断为零，但实际上其概率并不一定为零。对于这种情况，利用规则消歧辅助词性标注就显得尤为重要，能够大幅度提升因训练不足而忽略小概率的情况。

采用诸如“不幸”一词的标注方法，对 78 个形名兼类词的正则模式进行描写，最终共转换为 645 条规则代码，建立形名兼类词消歧规则库，然后对 7800 个形名兼类词的有效实例一一验证，发现有 1137 个句子词性标注错误，用我们的方法可以全部准确修正，修正率为 14.57%，这跟靳光瑾(2005)关于现代汉语通用平衡语料库标注质量的结论“词性标注错误率低于千分之五”不符，语委语料库的词性标注结果与正确结果之间存在不小的误差。我们只是作了这样一个小范围的测试，如果将本研究已经统计过的 2200 余个兼类词的搭配模式全部加入到语法规则库中，那么对标注结果准确率会更大的提高。

4. 基于语法搭配模式的消歧策略的局限性

本研究归纳的兼类词语法搭配规则，的确对提高词性标注尤其是兼类词标注结果在准确率有很大帮助，但不能据此否认这种方法的不足：

1) 汉语语法现象纷繁复杂，语法搭配模式的统计归纳只能统计出主要的搭配模式，而不能穷尽所有搭配模式。

2) 只关注有直接组合关系的搭配模式的提取，忽略了远距离的搭配模式。在代码化过程中，对于正则表达式的准度和广度提出了较高要求。

5. 结语

汉语具有不同于印欧语的语法特点，要进一步提高汉语词性标注质量，需要构建大量精确且适用的

规则。尤其在面对汉语的兼类词消歧这个“瓶颈”时，利用其语法搭配模式所构建的规则库对已标注语料加以修正，这不失为一种可靠的方法。这种运用兼类词语法搭配模式进行词性消岐的策略在基于概率的词性标注方法占据主流研究的今天仍具有很大的价值。在基于统计的方法进行大规模语料标注后，完全可以使用我们的设计思想对语料进行完善和修正，可以提高自动词性标注的准确率。

目前所有的机器自动标注始终不能达到百分之百的正确率，其本质原因是人与机器思维的差异，人能够依据自己的语法思维和认知模式标注出完全正确的语句，而机器利用与人类认知不同的算法逻辑只能接近完全正确的标注结果。如果想要无限接近完全正确的结果，需要让自动标注过程无限接近人类本身的语法思维和认知模式。本研究的语法搭配模式规则库的构建是一次小规模的实验，这种研究任重而道远，不仅需要先进的计算机技术如人工智能的运用，也需要前沿的汉语本体理论如认知语言学的研究，挖掘人类在语言中本质的认知模式，需要不断结合跨学科的先进理论与研究经验，从而推进中文信息处理不断取得理论创新和实践突破，推动汉语本体与应用研究不断完善与发展。

基金项目

国家级大学生创新训练项目“基于语料库的词典义项设置调查研究(S202010451033)”。

参考文献

[1] 胡明扬. 现代汉语的词类问题[C]//世界汉语教学学会. 第六届国际汉语教学讨论会论文选. 世界汉语教学学会: 世界汉语教学学会, 1999: 10.

[2] 张虎, 郑家恒, 刘江. 语料库词性标注一致性检查方法研究[J]. 中文信息学报, 2004(5): 11-16.

[3] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013.

[4] 张民, 李生, 赵铁军, 张艳风. 统计与规则并举的汉语词性自动标注算法[J]. 软件学报, 1998(2): 55-59.