文章编号：1003-0077(2020)09-0009-10

精细化的中文词性标注评测集的研制

唐乾桐[1,2]，常宝宝[1]，詹卫东[1,2,3]

（1. 北京大学 计算语言学教育部重点实验室，北京 100871；
2. 北京大学 中国语言文学系，北京 100871；
3. 北京大学 中国语言学研究中心，北京 100871）

摘 要：该文提出了一套精细化的中文词性标注评测体系。该文的工作重点在于确立其中的评测项目以及每个项目所对应的词例，提出了比对、归类、合取的方法；依此，该文初步建立了规模为 5 873 句、涵盖了 2 326 项词例和 70 个评测项目的评测试题集，并用这套试题集对几个常见的开源词性标注程序进行了评测。最后，该文指出了精细化评测体系将评测项目和评测语料联系起来的好处——在传统体系中，两者是分开的。该文从评测项目的价值和评测语料的组织性两个方面阐述了该文的评测体系相对于传统评测体系的优势，并指出了利用该文提出的评测体系改进被测程序的方法。

关键词：精细化评测；词性标注；语言资源

中图分类号：TP391 **文献标识码**：A

A Fine-grained Evaluation Set for Chinese POS Tagging

TANG Qiantong[1,2], CHANG Baobao[1], ZHAN Weidong[1,2,3]

(1. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China;
2. Department of Chinese Language and Literature, Peking University, Beijing 100871, China;
3. Center for Chinese Linguistics, Peking University, Beijing 100871, China)

Abstract: This paper proposes a fine-grained evaluation scheme on Chinese POS Tagging. The key to this task is to determine the evaluation items and the samples (words) for each item. This paper presents an evaluation set of 5 873 sentences, totaling 2 326 words for 70 evaluation items. Several common open source POS taggers are evaluated. Finally, this paper discusses the advantages of the merits of this evaluation approach, especially in contrast to the classical methods.

Keywords: fine-grained evaluation; POS tagging; language resource

0 引言

对于词性标注，现有的评测方式是直接用真实语料进行测试。这些评测语料通常是以符合语言总体的真实分布为前提的，但没有进行过进一步的筛选，也没有针对具体的评测项目做专门的设计。因而，其评测结果虽然可以反映标注程序处理真实语料的整体水平，但对于兼类词的处理能力，则缺乏更具针对性的测试，因而也难以准确反映程序在语言能力(competence)[①]方面的特征。

因此，需要一种更加语言学导向的精细化评测方法。这种评测方法强调评测项目的精细化和全面化。根据选定的评测项目，可以搜集评测语料，形成最终的评测集。相较于现有的评测方式，精细化评测专门挑出词性具有兼类情况的用例组成评测集，而不是像现有的评测集那样，包含大量没有评测价值的非兼类词。

① 语言能力(competence, I-language)：与“言语表现”(performance, E-language)相对，用来指说话人对自己语言的知识，即已经掌握的规则系统。据此人们可以生成句子、识别语法错误和歧义。

收稿日期：2019-09-09 **定稿日期**：2019-10-14

基金项目：教育部人文社科重点研究基地重大项目(15JJD740002)；国家自然科学基金(61876004)

进一步地，精细化的评测集不仅可以用于评测词性标注程序的性能，而且在发展完善之后，还可以用于评测那些依赖词性标注的下游 NLP 任务。

1 相关工作

精细化评测并非是近几年才兴起的风潮。早在 20 世纪 90 年代，研究者就已经对一些 NLP 任务设计了精细的评测项目[1-2]。但这些评测项目大都针对机器翻译等任务[2]或关注于语义方面的一些语言知识[3]，词性标注领域长期以来一直缺乏相应的精细化的评测工具。

这种情况一直持续到了近几年，Ali Elkahky 等开始关注名动歧义对词性标注任务的影响，进而构建了一个针对名动歧义的词性标注“挑战集”(Challenge Set)，可用于词性标注的评测。至此，英语的词性标注有了一个语言学导向的精细化的评测工具。[4]然而，这个评测集只针对“名动歧义”这一种语言知识，评测项目过于单一，评测体系不够全面[5]。

在中文词性标注领域，主流评测体系一直是粗糙而数据导向的，缺乏针对具体语言知识而设计的精细化评测体系，这使得词性标注对句法分析、中文分词[6]等相关任务的帮助大打折扣。有研究者也尝试改变现状，如杨尔弘等，将“相对标注精确率”和“兼类词标注精确率”作为词性标注的评测项目[7]，然而这两个评测项目一方面没有得到广泛使用，另一方面也依然比较粗糙、不够精细。

2 评测项目的选定

要确保评测体系的精细、全面和语言学导向，评测项目的选定至关重要。

2.1 选定步骤

本文采用比对、归类、合取①的方法，初步筛选出值得评测的项目；再对筛选出来的项目按照灵活性和一致性两个原则进行人工筛定，得到最后的评测项目。具体的步骤如图 1 所示。

2.1.1 标注

首先，准备一个适当规模的生语料，分别让专家和若干不同的词性标注程序对这个生语料进行标注，得到由专家标注的结果和由不同的词性标注程序标注的结果，并将专家的标注结果视为标准的标注结果。

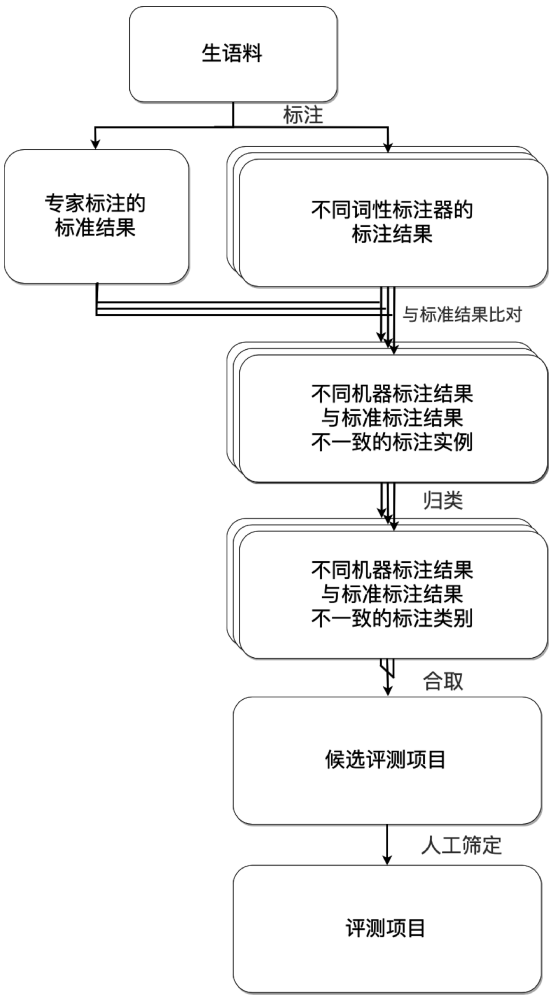

图 1 评测项目的选定步骤

具体地，本文从人民日报语料[8]中按篇抽取了 60 514 字的语料；并选取了 5 个常见的开源中文信息处理工具包中的分词和词性标注工具来对上述语料进行标注，这 5 个分词和词性标注工具分别记作 A、B、C、D、E，所基于的模型如表 1 所示[9-14]。

表 1 五家词性标注程序所基于的模型

程序	所基于的模型
A[9]	主要基于经典的 CRF 模型，辅以精调的特征
B[10]	主要基于最大熵模型，其中分词采用字序列标记方法，并充分发挥标点符号的分隔作用来寻找正反例

① 比对、归类、合取三个环节由机器自动实现，具体的实现程序会公开在项目网站上，以方便不同的使用者对评测项目再做修改和扩展。网址：https://github.com/hatriotism/A-Fine-grained-Evaluation-Set-on-Chinese-POS-Tagging

续表

程序	所基于的模型
C[11]	词性标注主要基于 HMM 模型，分词运用的则是基于二阶马尔科夫链的字序列生成模型(character-based generative model)
D[12-13]	主要基于 BiLSTM＋CRF 模型
E[14]	主要基于 HHMM 模型，即层次化隐马尔可夫模型

所有程序都保持各自默认的参数和设置。

2.1.2　比对

将不同机器标注的结果与标准的标注结果相比对，分别找出不同机器标注结果与标准标注结果不一致的标注实例。

值得一提的是，由于机器与专家之间的标注规范不同，不同标注程序之间的标注规范也不同，如果不对标注规范做统一，那么通过比对得到的不一致实例的数量就会特别庞大，且会直接影响之后合取的质量。为了解决这个问题，本文给每个标注程序分别建立了最低限度的映射规则，把它们的标注规范统一映射到专家的规范上来。例如，有的标注规范具体区分了 ule(助词“了”)、uzhi(助词“之”)等特殊助词，而专家的规范统一标注成 u(助词)，那么就将 ule、uzhi 等映射为 u。

具体地，A、B、C、D 和 E 分别找到了 2 977、3 219、2 487、12 023 和 2 880 例与标准结果不一致的标注实例。

2.1.3　归类

将不一致的标注实例按词类的不同进行归类。例如，如果标准标注结果是名词，而机器标注结果是动词，那么就是＜[n]，[v]＞类。

本文将形如＜[n]，[v]＞的式子称为“标注结果对应组”，其左边的中括号内是标准标注结果，右边的中括号内是机器标注结果。本文站在评测机器标注结果的视角，将前者称为标注不足项，将后者称为标注过度项。任何一例具体的词性错标，都会在宏观层面，致使一个词性被少标了，另一个词性被多标了。那么从标注程序对特定词性的敏感程度的角度而言，前者就是标注不足，后者是标注过度。

值得一提的是，由于词性标注涉及分词问题，本文在归类时也会考虑切分颗粒度的问题。例如，标准标注结果将“就此”切开，标注为“就/p 此/r”；而不少标注程序将其标注为“就此/d”，没有切开。本文在归类环节会保留这类不一致(比如，会用＜[p＋r]，[d]＞类来归纳上面“就是”的例子)，并会在之后的人工筛定环节中筛选出这类不一致中真正值得评测的项目。

具体地，A、B、C、D 和 E 分别得到了 432、335、298、1 015 和 470 个类错误类型。

2.1.4　合取

这一步要对上一步得到的各标注程序的错误类型做合取，得到候选评测项目。这里不做析取的原因是，虽然析取能得到更全面的候选评测项目，但是其中会掺杂过多的没有评测价值的项目，使评测失焦。

但合取的标准也不宜过高，要尽量在“全面”与“精炼”之间平衡，前者要求标准尽量低(最低就是完全的析取)，后者要求标准尽量高(最高就是 100％的合取)。本文选择了 51％这个折中值，将 51％以上的词性标注程序都标注错误的类型作为候选评测项目。这样一来，不仅可以有效过滤一些不显著的错误，以及在“比对”环节没有完全照顾到的由分词不一致、标注规范不一致等原因所导致的伪错误，保证了候选评测项目的精炼性；而且也可以使候选集尽可能全面，尽量不偏倚错误类型总是重合的标注程序，保证了全面性。

具体地，本文将上述五个词性标注程序之中的任意 3 个都标注错误的类型加入候选评测项目，选出了共 84 个候选评测项目。

2.1.5　人工筛定

最后，做人工筛选和总结，过滤掉不重要的候选评测项目，确定最终的评测项目，并记录下这些评测项目分别对应的词例。

本文在做人工筛定时，主要关注切分颗粒度问题。对于切分颗粒度问题，本文遵循两个原则。

(1) 灵活性原则。本文采取宋柔的“柔性”态度[15]，即分词单位的大小允许有较大的灵活性，切分结果在分词单位的上下界之间即可。分词单位的上界应允许以下结构成为一个分词单位：简单动宾、动补、形宾、形补、偏正结构，动词和形容词的各种变形结构，时间短语，处所短语，缩略语等。分词单位的下界应允许以下结构成为一个分词单位：动宾、动补、偏正结构中可扩展的成分，二字以上词语的前加成分、后加成分，表示儿化音的“儿”，二字以上地名的通名与专名，词重叠的重叠成分等。对于纯粹切分颗粒度不一致而不存在词性标注问题的项目，本文将予以筛除。

(2) 一致性原则。当机器标的比专家标的颗粒

度粗时，除了切分结果要求处于分词单位的上下界之间，两种切分的中心语的词性也要保持一致。不一致的话就视为标注错误，并保留为最终的评测项目；一致的话则会予以筛除。

如表2所示，本文通过观察每个项目的词例来进行判断，具体筛除了以下情况下的项目。

表2 人工筛除掉的评测项目

人工筛除的评测项目	部分词例	筛除原因
<[nr]，[nr+nr]>	邓小平、江泽民、兰红光	灵活性原则
<[n]，[nr+n]>	沈先生、哈桑国王	
<[d]，[d+v]>	先是、更是、常有	
<[n+n]，[n]>	中医理论、学术论文	一致性原则&灵活性原则
<[a+n]，[n]>	老干部、高性能、新国情	
<[b+n]，[n]>	金佛、金戒指、副院长	
<[m+n]，[n]>	七仙女、半月	
<[m+Ng]，[n]>	一体	
<[ns+n]，[n]>	广州人、纽约时报	
<[v+v]，[v]>	唱出、钻取、患有	
<[v+n]，[v]>	成团、到家、搭车	
<[v+Ng]，[v]>	降息、尽孝、投食	
<[Ag+v]，[v]>	私设、私分	
<[d+v]，[v]>	空守、初露、共建	

(1) 当机器标的比专家标的颗粒度细时，筛除了切分结果在切分上下界之间的项目。比如<[nr]，[nr+nr]>类(如邓小平、江泽民、兰红光)、<[n]，[nr+n]>类(如，沈先生、哈桑国王)、<[d]，[d+v]>类(如，先是、更是、常有)3类。

(2) 当机器标的比专家标的颗粒度粗时，筛除了不仅切分结果在分词上下界之间，而且两种切分的中心语的词性也保持一致的项目。比如<[n+n]，[n]>类(如，中医理论、学术论文)、<[v+n]，[v]>类(如，成团、到家、搭车)、<[ns+n]，[n]>类(如广州人、纽约时报)等11类。

最终，本文确定了70个评测项目，这些评测项目总共对应了2 326项词例。

2.2 评测项目的概况

2.2.1 不涉及分词问题的评测项目

在本文最终确定的70个评测项目中，共有61个不涉及分词问题，如图2所示。

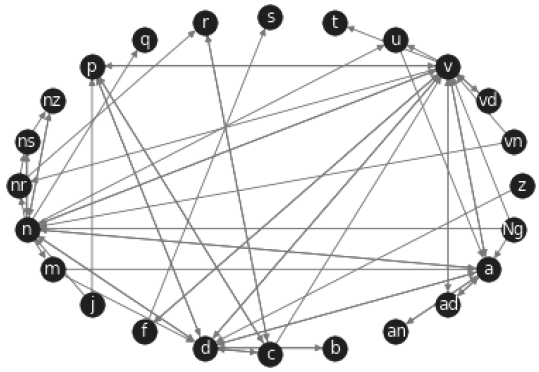

图2 不涉及分词问题的评测项目概况

图中的点表示词性，边表示两个词性之间产生了评测项目；边有方向，由标注不足项指向标注过度项；边可以是双向的，每个方向表示一个评测项目。例如，一条从z指向d的边，即表示<[z]，[d]>这个评测项目。

从图2中可以看出，不涉及分词问题的评测项目，大致可以分以下几种情况：

(1) 标注不足项和标注过度项属于同族的词性标签。例如，a(形容词)、ad(副形词)和an(名形词)之间；n(名词)、nr(人名)、ns(地名)和nz(其他专名)之间；等等。

(2) 标注不足项或标注过度项中的至少一个是比较开放的一级词性标签。这类词性标签有：a(形容词)、d(副词)、n(名词)、v(动词)。

(3) 其他。例如，<[c]，[p]>、<[c]，[r]>、<[f]，[s]>、<[j]，[p]>、<[p]，[c]>、<[r]，[c]>。

2.2.2 涉及分词问题的评测项目

在本文最终确定的70个评测项目中，共有9个涉及分词问题。如图3所示。

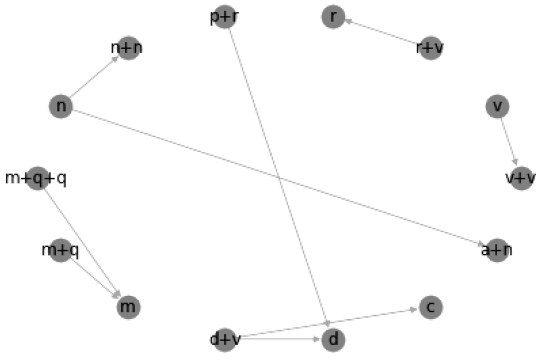

图3 涉及分词问题的评测项目概况

这类评测项目由于涉及分词问题，情况比较复杂，大部分词例数量比较少。

另外，这类评测项目较少，主要是因为在合取和人工筛定两个环节筛掉了大量的这类项目。鉴于这类评测项目涉及分词问题，而各个标注程序的分词颗粒度不尽一致，再加上有一些项目（如＜[v＋a]，[v]＞），虽然某一个词性标注程序在这个项目上表现不佳，但其他程序并没有不佳的表现，因此在合取环节就由于没有达到合取条件而被筛掉了。再加上在人工筛定的环节主要筛除的对象就是涉及分词问题的项目，因此这类项目最终留下来的就较少。

3　评测试题集的构建

3.1　构建方法

如前文所述，本文的工作重点是确立评测项目以及每个项目中的词例。在这个过程中，就会附带地找到每个词例所处的句子——这些句子就可以作为初始的评测试题集。

具体地，本文根据第 2 节确立的 70 个评测项目及其对应的 2 326 项具体词例，在人民日报语料中收集了规模为 5 873 句的评测试题集。

进一步地，根据词例，还可以到开放的真实语料中收集更多的句子，并辅以人工甄别，以此来扩展评测集的“题库”。①②

3.2　评测试题集的结构

评测试题集由 70 个评测章节构成，每个章节有一个一级评测项目，每个一级评测项目下面有若干个词例作为更加精细的二级评测项目，每个词例下面有 1～5 个包含这个词例的句子作为评测试题，如图 4 所示。

在评测试题集中，一级评测项目以“＃”开头，其词例即二级评测项目以“＃＃”开头。例如在图 4 中，＜[v]，[t]＞和＜[v]，[u]＞是一级评测项目，而“过去”“等”“了”“过”则分别是这两个一级评测项目下面的词例，即二级评测项目。

3.3　评测分数的计算

每道评测试题，仅关注目标词例是否被按照一级评测项目的模式错标。如果标对，则记对（Correct）；如果标错，且是按照一级评测项目的模式被标错，则记错（Wrong）；如果不是按照一级评测项目的模式被标错，则记录（Record）到后台。

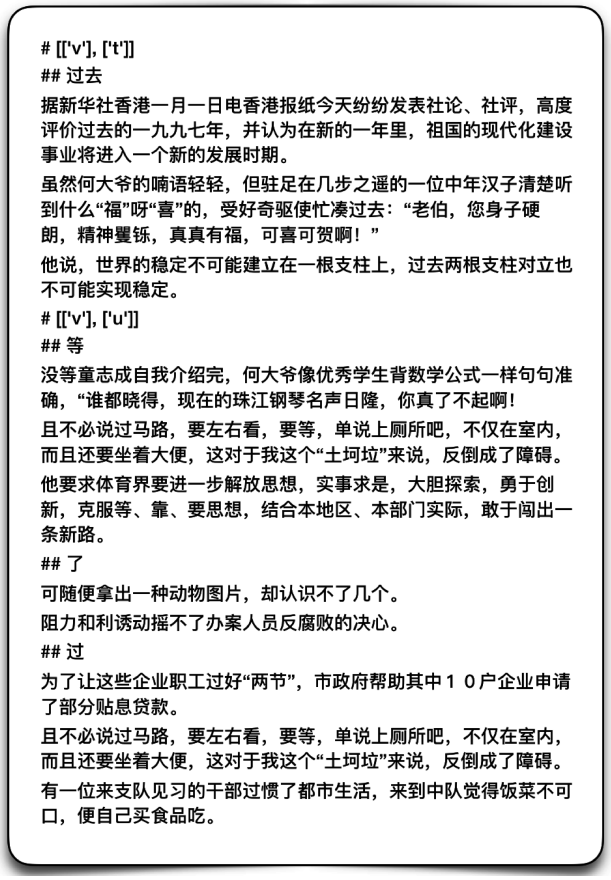
[['v'], ['t']]
过去
据新华社香港一月一日电香港报纸今天纷纷发表社论、社评，高度评价过去的一九九七年，并认为在新的一年里，祖国的现代化建设事业将进入一个新的发展时期。
虽然何大爷的喃语轻轻，但驻足在几步之遥的一位中年汉子清楚听到什么“福”呀“喜”的，受好奇驱使忙凑过去：“老伯，您身子硬朗，精神矍铄，真真有福，可喜可贺啊！”
他说，世界的稳定不可能建立在一根支柱上，过去两根支柱对立也不可能实现稳定。
[['v'], ['u']]
等
没等童志成自我介绍完，何大爷像优秀学生背数学公式一样句句准确，“谁都晓得，现在的珠江钢琴名声日隆，你真了不起啊！
且不必说过马路，要左右看，要等，单说上厕所吧，不仅在室内，而且还要坐着大便，这对于我这个“土坷垃”来说，反倒成了障碍。
他要求体育界要进一步解放思想，实事求是，大胆探索，勇于创新，克服等、靠、要思想，结合本地区、本部门实际，敢于闯出一条新路。
了
可随便拿出一种动物图片，却认识不了几个。
阻力和利诱动摇不了办案人员反腐败的决心。
过
为了让这些企业职工过好“两节”，市政府帮助其中１０户企业申请了部分贴息贷款。
且不必说过马路，要左右看，要等，单说上厕所吧，不仅在室内，而且还要坐着大便，这对于我这个“土坷垃”来说，反倒成了障碍。
有一位来支队见习的干部过惯了都市生活，来到中队觉得饭菜不可口，便自己买食品吃。

图 4　评测试题集节选示意

评测分数分为单项分数 D 和总分 Z，并且每个评测项目记录 R。计算方法如式(1)～式(3)所示。

$$D=\frac{\text{Correct}}{\text{Total-Record}}\times 100 \tag{1}$$

$$Z=\frac{\sum_{i=1}^{n}D_i}{n} \tag{2}$$

$$R=\frac{\text{Record}}{\text{Total}} \tag{3}$$

单项分数由该评测项目中标注正确的试题数量除以该评测项目中的试题总数得到，取值范围为 0 到 100。在某一句试题中，只要词例片段标注正确，即视为标注正确。

总分为了突出每个评测项目的平等地位，则是取的所有单项分数的平均数。

① 本文仅指出扩展“题库”的可能性，但暂时还没有做相关的工作。

② 本文开发的试题集将公布在项目网站上。网址：https://github.com/hatriotism/A-Fine-grained-Evaluation-Set-on-Chinese-POS-Tagging。

4 评测结果与分析[①]

本文利用上述的评测试题集对A、B、C、D和E五个常见的开源中文信息处理工具包中的分词与词性标注程序进行了评测，整体结果如附表1所示。

五个词性标注程序分别得到了44.15、53.43、55.20、35.85、43.23的总分，可见上述评测试题集的效度和难度。其中C总体表现最佳，从表3可以看出，C在涉及分词问题的评测项目上，有优于其他词性标注程序的表现，并最终反映到了总分上。

本文认为，C在涉及分词问题的评测项目上得分较高(见表3)，原因应该在于C在分词模型上的独特设计。不同于A、B、D所使用的基于字序列的判别模型(CRF、最大熵等)和E所使用的基于词表的生成模型，C在分词上运用了基于字序列的生成模型(character-based generative model)。根据Wang等的研究成果，基于字序列的模型擅长处理未登录词，而生成模型又擅长处理已登录词，基于字序列的生成模型在未登录词和已登录词两方面都会有较好的表现[12]，因此C在涉及分词问题的测试项目上得分较高。

表3 五个程序在涉及分词问题的项目中的表现

	A	B	C	D	E
<[d+v], [c]>	0.00	83.33	100.00	0.00	87.50
<[d+v], [d]>	0.00	85.71	92.86	11.90	71.43
<[m+q+q], [m]>	0.00	0.00	0.00	0.00	0.00
<[m+q], [m]>	0.00	16.33	17.35	15.31	14.29
<[p+r], [d]>	0.00	37.50	100.00	12.50	12.50
<[r+v], [r]>	0.00	0.00	100.00	0.00	0.00
<[n], [a+n]>	84.62	80.77	96.15	23.08	76.92
<[n], [n+n]>	84.62	84.02	95.27	9.47	71.01
<[v], [v+v]>	84.62	82.84	93.52	11.27	61.54
平均	28.21	52.28	77.24	9.28	43.91

另一个值得注意的地方是，B在以形容词或动词作为标注不足项的评测项目中(即左边中括号为[a]或[v]的项目中)，普遍有优于其他词性标注程序的表现，如表4所示。

表4 五个程序在以[a]或[v]作为标注不足项的项目中的表现

	A	B	C	D	E
<[a], [ad]>	54.17	90.28	33.33	52.78	29.17
<[a], [an]>	68.35	98.73	60.76	69.62	10.13
<[a], [d]>	44.44	77.78	66.67	66.67	66.67
<[a], [n]>	25.40	66.67	68.25	42.86	50.79
<[a], [v]>	27.03	64.86	68.92	63.51	51.35
<[v], [a]>	28.24	50.59	32.94	31.76	37.65
<[v], [ad]>	36.36	63.64	27.27	18.18	18.18
<[v], [d]>	18.75	43.75	37.50	39.58	35.42
<[v], [f]>	72.88	84.75	15.25	67.80	64.41
<[v], [n]>	42.92	80.97	39.53	12.68	30.23
<[v], [nr]>	56.00	92.00	64.00	28.00	84.00
<[v], [p]>	43.79	60.36	31.95	51.48	21.89
<[v], [t]>	0.00	33.33	0.00	66.67	33.33
<[v], [u]>	66.67	76.19	4.76	19.05	9.52
<[v], [vd]>	62.50	91.67	58.33	29.17	8.33
平均	43.17	71.70	40.63	43.99	36.74

5 精细化评测体系的优势

5.1 评测项目的价值

传统的词性标注评测体系所关注的评测项目沿袭的是分类器的评测项目，即各类别(各词性)的精确率、召回率和F_1分数。这类评测项目善于反映各词性的标注过度和标注不足的总体情况。例如，某个词性的精确率越低，表示其标注过度的情况越严重；而召回率越低，则表示其标注不足的情况越严重。

而本文的精细化评测体系将评测项目精细到了具体的“标注结果对应组”，从而可以对评测语料的组织扩展和被测程序的改进产生很强的启发价值，并反映了程序在语言能力方面的特征，这是现有的评测项目所不具备的。

首先，精细化的评测项目可以将评测项目和评测语料联系起来。在传统的评测方式中，一般是直

① 鉴于本文工作重心在于确立评测项目及其对应词例，不在于对具体的模型评测和分析上，因此本文对评测结果的分析将比较浅显。

接用无组织的自然语料进行测试，这样一来，评测语料和评测项目就是分开的。而具体的“标注结果对应组”及其词例可以将原本无序的评测语料组织起来，并可以启发评测语料依循这些评测项目做扩展。换言之，精细化的评测项目搭好了骨架，向其中填充评测语料就变得方便了。关于将评测项目和评测语料联系起来的优势，会在 5.2 节中详细讨论。

其次，在本文的评测体系下，当发现被测程序在某个评测项目上表现不佳时，可以针对性地扩充训练语料，或者针对性地引入规则，从而使针对性地改进被测程序成为可能。而反观现有的评测体系，就不能启发被测程序做针对性的改进。

最后，由于本文研制的评测体系避开了过于简单的“标注结果对应组”，把评测的焦点放在更加困难的项目上，并针对这些项目总结了一系列的词例，从而提高了评测的效度。另外，分词问题对词性标注具有重要影响，而传统体系的评测项目不涉及分词问题，本文体系包含了涉及分词问题的评测项目，也提高了评测的效度。

5.2　评测语料的组织性

在传统的评测方式中，一般是直接用无组织的自然语料进行测试，这样一来，评测语料和评测项目就是分开的，从而造成评测语料缺乏组织性。

而在本文的精细化评测体系中，根据选定的评测项目，可以搜集评测语料，形成最终的评测集。相较于现有的评测方式，具有以下优势：

首先，专门挑出词性具有兼类情况的用例组成评测集，而不是像传统的评测集那样，包含大量没有评测意义的非兼类词。

其次，方便定位错误，从而指导被测程序的改进。而在传统的评测体系中，由于评测语料没有组织化，即使评测结果反映了被测程序的一些不足之处，也不方便定位具体的错误、指导被测程序的改进。

最后，方便语言学家或相关专家对评测语料做扩展，从而递进式地测试被测程序的改进效果。例如，当被测程序在针对某个评测项目做了改进之后，需要进一步评测这个改进的效果，那么就需要针对这个评测项目扩充评测语料，这时就可以依据具体的词例做扩充。

6　总结与展望

本文提出，中文词性标注需要建立一套精细化的评测方法，以评测被测程序内在的“语言能力”，指导被测程序的改进。

本文认为，精细化评测方法的重点，在于评测项目的精细和全面。因此，本文把工作重心放在评测项目及其词例的选定上，提出了比对、归类、合取的方法，并依此初步建立了规模为 5 873 句的、涵盖了 2 326 项词例和 70 个评测项目的评测试题集。按照这套评测方法，本文尝试性地对 A、B、C、D 和 E 五个常见的开源词性标注程序进行了评测。最后，本文从评测项目的价值和评测语料的组织性两个方面阐述了本文精细化评测体系相对于现有的传统评测体系的优势，并指出了根据本文提出的精细化评测体系改进被测程序的方法。在传统体系中，评测项目与评测语料被割裂开来，而本文重点指出了将两者联系起来的优势。

不过由于时间的有限性和问题的复杂性，本文的工作尚有许多不足之处，留待将来补足，例如，

(1) 本文指出了根据本文提出的精细化评测体系改进被测程序的方法，但没有做充分的实践。我们将在之后的工作中补充这一点。

(2) 本文提出的评测试题集，不仅可以用于评测词性标注程序的性能，而且可以进一步用于评测那些依赖词性标注的下游 NLP 任务，但本文没有涉及下游任务的评测。我们会在将来进一步考虑下游任务，并针对下游任务对评测试题集做优化。

(3) 鉴于本文的主要目的在于研制一个精细化的中文词性标注评测集，因此工作重心在于确立评测项目及其对应词例，而不在于对具体的模型评测和分析上，因此本文对评测结果的分析尚比较浅显。

参考文献

[1] Lehmann S, Oepen S, Regnier-Prost S, et al. Tsnlp: Test suites for natural language processing[C]//Proceedings of the 16th Conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 1996: 711-716.

[2] King M, Falkedal K. Using test suites in evaluation of machine translation systems [G]. COLNG 1990 Volume 2: Papers Presented to the 13th International Conference on Computational Linguistics, 1990: 211-216.

[3] Cooper R, Crouch D, VanEijck J, et al. Using the framework[R]. Technical Report LRE 62-051 D-16, The FraCaS Consortium, 1996.

[4] Elkahky A, Webster K, Andor D, et al. A challenge set and methods for moun-verb ambiguity[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2562-2572.

[5] Belinkov Y, Glass J. Analysis methods in neural language processing: A survey[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 49-72.

[6] 刘金宁. 词性标注体系对中文分词的影响[D].大连：大连理工大学硕士学位论文,2010.

[7] 杨尔弘,方莹,刘冬明,等.汉语自动分词和词性标注评测[J].中文信息学报,2006,20(01): 44-49,97.

[8] 俞士汶,段慧明,朱学锋,等. 北大语料库加工规范：切分·词性标注·注音[J]. 汉语语言与计算学报, 2003,13(2): 121-158.

[9] Luo R, Xu J, Zhang Y, et al. PKUSEG: A toolkit for multi-domain Chinese word segmentation [J]. arXiv preprint arXiv: 1906.11455, 2019.

[10] Sun M, Chen X, Zhang K, et al.Thulac: An efficient lexical analyzer for Chinese[CP/OL]. 2016-01-10. https://git.hub.com/thanlp/THUL AL.

[11] Brants T. TnT: a statistical part-of-speech tagger [C]//Proceedings of the 6th Conference on Applied Natural Language Processing. Association for Computational Linguistics, 2000: 224-231.

[12] Wang K,Zong C, Su K Y. Which is more suitable for Chinese word segmentation, the generative model or the discriminative one? [C]//Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2,2009: 827-834.

[13] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv: 1603.01360, 2016.

[14] Zhang H P, Yu H K,Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing-Volume 17. Association for Computational Linguistics, 2003: 184-187.

[15] 宋柔.关于分词规范的探讨[J].语言文字应用,1997(03): 113-114.

附录

附表 1 五个程序在本文评测系统上的评测结果

	A	B	C	D	E
<[Ng], [a]>	0.00	0.00	0.00	0.00	0.00
<[Ng], [n]>	0.00	0.00	67.46	0.00	0.00
<[Ng], [v]>	0.00	0.00	15.38	0.00	0.00
<[a], [ad]>	54.17	90.28	33.33	52.78	29.17
<[a], [an]>	68.35	98.73	60.76	69.62	10.13
<[a], [d]>	44.44	77.78	66.67	66.67	66.67
<[a], [n]>	25.40	66.67	68.25	42.86	50.79
<[a], [v]>	27.03	64.86	68.92	63.51	51.35
<[ad], [a]>	81.90	0.00	77.62	80.95	80.00
<[ad], [v]>	0.00	0.00	100.00	33.33	100.00
<[an], [a]>	54.63	0.00	75.93	50.93	74.07
<[b], [d]>	45.00	0.00	35.00	55.00	50.00
<[c], [d]>	38.24	52.94	17.65	47.06	26.47
<[c], [p]>	68.89	62.22	42.22	68.89	53.33
<[c], [r]>	50.00	0.00	0.00	0.00	0.00
<[c], [v]>	8.33	58.33	0.00	8.33	16.67
<[d], [a]>	54.17	66.67	25.00	33.33	41.67

续表

	A	B	C	D	E
<[d], [b]>	94.44	88.89	66.67	72.22	16.67
<[d], [c]>	46.43	67.86	60.71	64.29	46.43
<[d], [n]>	50.00	37.50	37.50	50.00	12.50
<[d], [p]>	55.56	70.37	37.04	29.63	25.93
<[d], [v]>	25.00	63.89	50.00	38.89	30.56
<[f], [s]>	50.00	75.00	50.00	50.00	12.50
<[f], [v]>	3.91	96.09	98.44	88.28	96.09
<[j], [n]>	74.85	83.83	92.81	0.00	0.00
<[j], [p]>	70.00	60.00	20.00	0.00	0.00
<[m], [a]>	62.50	100.00	37.50	50.00	75.00
<[m], [d]>	10.71	71.43	78.57	71.43	75.00
<[n], [a]>	41.18	66.67	70.59	84.31	68.63
<[n], [d]>	33.33	66.67	16.67	16.67	16.67
<[n], [m]>	80.00	60.00	60.00	40.00	40.00
<[n], [nr]>	68.09	76.60	74.47	14.89	57.45
<[n], [ns]>	64.52	77.42	93.55	22.58	74.19
<[n], [nz]>	50.00	50.00	92.86	57.14	42.86
<[n], [q]>	54.90	64.71	29.41	56.86	19.61
<[n], [v]>	42.72	54.37	45.63	58.25	52.43
<[nr], [ns]>	71.43	0.00	42.86	11.43	94.29
<[nr], [r]>	61.54	0.00	84.62	0.00	0.00
<[ns], [n]>	41.18	70.59	76.47	35.29	47.06
<[ns], [nz]>	73.33	73.33	66.67	93.33	66.67
<[nz], [n]>	63.16	31.58	89.47	10.53	57.89
<[p], [c]>	65.38	69.23	53.85	36.54	42.31
<[p], [d]>	75.00	70.00	40.00	60.00	20.00
<[p], [v]>	28.86	82.55	71.14	63.76	61.07
<[r], [c]>	0.00	92.00	88.00	84.00	84.00
<[u], [a]>	33.33	66.67	66.67	0.00	88.89
<[v], [a]>	28.24	50.59	32.94	31.76	37.65
<[v], [ad]>	36.36	63.64	27.27	18.18	18.18
<[v], [d]>	18.75	43.75	37.50	39.58	35.42
<[v], [f]>	72.88	84.75	15.25	67.80	64.41
<[v], [n]>	42.92	80.97	39.53	12.68	30.23
<[v], [nr]>	56.00	92.00	64.00	28.00	84.00
<[v], [p]>	43.79	60.36	31.95	51.48	21.89

续表

	A	B	C	D	E
<[v], [t]>	0.00	33.33	0.00	66.67	33.33
<[v], [u]>	66.67	76.19	4.76	19.05	9.52
<[v], [vd]>	62.50	91.67	58.33	29.17	8.33
<[vd], [v]>	55.56	0.00	72.22	55.56	66.67
<[vn], [n]>	74.55	0.00	79.39	0.00	80.41
<[vn], [v]>	72.08	0.00	76.24	0.00	76.73
<[z], [d]>	50.00	0.00	83.33	16.67	33.33
<[n], [u]>	44.44	66.67	0.00	55.56	55.56
<[d+v], [c]>	0.00	83.33	100.00	0.00	87.50
<[d+v], [d]>	0.00	85.71	92.86	11.90	71.43
<[m+q+q], [m]>	0.00	0.00	0.00	0.00	0.00
<[m+q], [m]>	0.00	16.33	17.35	15.31	14.29
<[p+r], [d]>	0.00	37.50	100.00	12.50	12.50
<[r+v], [r]>	0.00	0.00	100.00	0.00	0.00
<[n], [a+n]>	84.62	80.77	96.15	23.08	76.92
<[n], [n+n]>	84.62	84.02	95.27	9.47	71.01
<[v], [v+v]>	84.62	82.84	93.52	11.27	61.54
总分	44.15	53.43	55.20	35.85	43.23

唐乾桐(1996—),硕士研究生,主要研究领域为中文信息处理。

常宝宝(1971—),博士,副教授,主要研究领域为自然语言处理。

詹卫东(1972—),通信作者,博士,教授,主要研究领域为现代汉语形式语法、中文信息处理、汉语语言知识工程。