基于本体标准的汉语词法分析评测方法研究*

张永伟　刘　婷　王玉莹

提要　现代汉语的自动词法分析存在多种参考标准,不同的词法分析工具基于不同的语言学理论和应用场景,采用不同的操作规范。在语言教学与研究领域,学界普遍期待词法分析工具能够更符合语言学本体标准。然而,传统词法分析评测方法专注于对算法性能的评测,在对语言学本体标准的契合度评估方面存在局限性。本文提出一种新的词法分析评价方法:通过精选现代汉语基本词汇和能充分反映汉语本体标准的例句,建立专门的评测语料库,将核心词汇在评测语料库中的分词和词性标注准确率作为评估指标,对词法分析工具的性能进行评估。这种本体标准的汉语分析评测方法,为语言教学与研究领域选择和优化词法分析工具提供了新的参考。基于本体标准的评测结果显示,现有词法分析工具的性能指标普遍低于传统评测方法的评估结果,在方言词、口语词、文言词及低频词的切分与标注方面仍有较大的优化空间。

关键词　汉语本体标准　词法分析　辞书　语料库　评测

1. 引言

在人工智能快速发展的背景下,汉语词法分析作为自然语言处理(natural language processing)的基础任务,一直是计算语言学重要的研究课题。虽然基于深度学习的大规模预训练语言模型已经展现了出色的语言理解和生成能力,但在具体的语言分析任务上仍面临挑战。为了服务语言信息处理,人们研发了许多词法分析工具,这些工具在标准测试集上的表现也较为优异。然而,当这些工具被应用于语言教学与研究并按语言学本体标准进行检查时发现,自动分析结果的错误显著,难以直接应用。这使我们不得不思考:现有的词法分析工具是否真正理解了汉语词汇系统,能否按照语言学家的标准来分析文本?

汉语词法分析主要由分词和词性标注两个任务组成。传统上,汉语词法分析工具的效果主要利用精确率(precision)、召回率(recall)、F1 值(F1-score)和准确率(accuracy)等指标在特定测试集上进行评测。由于语言数据存在长尾分布,工具可以针对高频语言现象进行较好处理,从而获得较高的基础性能。典型案例如标点符号的切分标注,几乎所有工具都能达到满分的评测结果。然而,在评测时,将此类易于处理的高频语言现象赋予大的权重,容易导致对工具整体切分标注效果的乐观估计。相反,工具在处理低频语言现象时表现往往不尽如人意,而这些低频现象在整体评测中所占权重较小,使得传统评测方法难以准确反映工具在长尾分布

* 本研究得到国家社科基金一般项目"融合句法信息的大规模汉语语料库分析工具研制研究"(22BYY086)的资助。俞敬松教授、化柏林副教授曾对本文修改提出重要建议,谨此致谢。文中错误由作者负责。

下的真实性能,更侧重于算法的整体表现,而非语言分析结果的最终质量。

汉语词法分析的根本困难在于词语切分和词性标注缺乏普遍认可的、统一的标准,不同标准间存在着显著差异。服务语言教学与研究的词法分析工具需要遵循语言本体的标准,但目前的评测方法仅关注工具在特定测试集上的表现,而并未评估结果是否符合语言学本体标准。因此,需要建立新的评估方法,专门针对语言本体标准进行评价。

本研究精选汉语基本词和能充分体现汉语本体标准的例句构建评测语料库,评估词法分析工具对例句中特定词语切分和词性标注的准确率,从而考察工具是否符合语言本体标准。实验使用权威语文辞书收录的词目作为汉语基本词,辞书例句作为体现汉语本体标准的例句。尽管语文辞书内容不完全等同于语言本体研究成果,但其编纂和历版修订过程中均会吸收大量最新的语言学研究成果,是反映语言本体标准的重要典范。这种评测方法旨在发现更加契合语言本体标准的词法分析工具,为语言教学与研究中词法分析工具的改进和选择提供新的参考。

2. 相关研究

2.1 词法分析算法与工具

汉语分词和词性标注算法分为基于规则的算法、传统机器学习算法和深度学习算法(唐琳等,2020)。就分词而言,基于规则的算法是一种机械式算法,需要借助词典实现,实用性较弱。机器学习算法又分为基于词和基于字的两种,后者是主流。分词和词性标注通常被视为序列标注问题,由人工选择特征,使用 N 元模型、隐马尔可夫、支持向量机、最大熵、条件随机场等算法确定标签序列的概率分布,选出标签序列得分最高的作为分词和词性标注结果。机器学习算法实现的分词和词性标注工具如中国科学院计算技术研究所 ICTCLAS①(刘群等,2004),哈尔滨工业大学 LTP 初期版本(刘挺等,2011),斯坦福大学 CoreNLP(Toutanvoa 和 Manning,2000;Tseng 等,2005)、HanLP 1.x(何晗,2019)等。深度学习方法无需人工选择特征,常使用循环网络模型、卷积网络模型、Transformer 模型等编码器对输入文本进行编码,并使用 Softmax 或者 CRF 进行解码预测。深度学习算法实现的分词和词性标注工具如哈尔滨工业大学 LTP 4.x(Che 等,2021)、HanLP 2.x(He 和 Choi,2021)、斯坦福大学 Stanza(Qi 等,2020)、俄勒冈大学 Trankit(Van Nguyen 等,2021)等。

本研究在知网上选择近 10 年(2015—2024 年)中国语言文字类的期刊为调查对象,统计了这些期刊近 10 年所刊登论文中常见汉语词法分析工具的使用情况②。统计后发现 NLPIR(同 ICTCLAS 结果合并)、jieba③、LTP、HanLP④ 是应用最多的 4 个工具,它们的使用总次数分别为 143、98、62、25,历年使用详情如图 1 所示。

图 1 显示,10 年来 NLPIR 的使用在整体趋势上先升后降,其他工具大致逐年递增。在 2019 年及以前,NLPIR 的使用优势明显,单个工具的使用次数超过了其他 3 个工具总和。从

① ICTCLAS 后期发展为 NLPIR,主页为:http://ictclas.nlpir.org/。2023 年 08 月 03 日最后访问。

② 许多论文直接使用分词和词性标注的语料库,也有些论文没有明确说明使用的工具名称。本文统计时忽略以上两种情况,仅统计明确指出分词和词性标注工具的论文。

③ jieba 工具的主页为:https://github.com/fxsjy/jieba。2023 年 08 月 03 日最后访问。

④ HanLP 工具的主页为:https://www.hanlp.com/。2025 年 03 月 08 日最后访问。

2018 年开始,jieba 的使用逐渐增多,并在 2021 年成为使用最多的汉语词法分析工具。此外,图 1 也显示早期分词和词性标注工具的选择较为单一,近年来选择逐渐丰富,一家独大的情况不复存在。值得注意的是,大部分论文未给出选择某个词法分析工具的缘由。

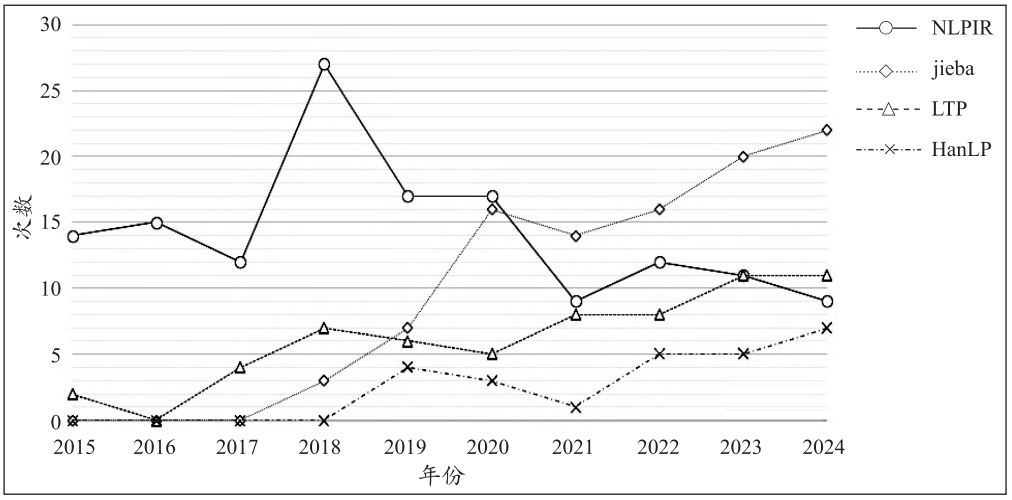

图 1　4 个汉语分词和词性标注工具的使用情况

2.2 工具性能评测

SIGHAN 和 CIPS-SIGHAN 举办过多次汉语词法分析评测,以分词和词性标注评测为主。通常做法是参与者对评测语料原始文本进行切分标注,组织者将全部切分标注结果同标准答案文本进行对比,计算评测指标。分词效果通常由精确率(简称 P)、召回率(简称 R)和 F1 值来评价(张奇等,2023),值越大越好,计算公式如下:

$$P=\frac{\text{算法输出的正确分词结果个数}}{\text{算法输出的全部分词结果个数}}\times 100\% \quad (1)$$

$$R=\frac{\text{算法输出的正确分词结果个数}}{\text{测试集中全部答案个数}}\times 100\% \quad (2)$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \quad (3)$$

F1 是精确率和召回率的调和平均值,兼顾了精确率和召回率。词性标注效果通常由准确率(简称 A)和宏 F1 值(Macro-F1)来评价(张奇等,2023),值越大越好,准确率计算公式如下:

$$A=\frac{\text{算法输出的正确结果个数}}{\text{算法输出的全部结果总数}}\times 100\% \quad (4)$$

参考分词评测中 F1 值的计算方法,也可以计算特定词类对应的词性标注 F1 值。将所有词类的 F1 值进行算术平均,即可得到整个词性标注的宏 F1 值。评测通常分为封闭评测和开放评测,前者的评测文本和训练文本原本属于同一个数据集,后者情况不同,评测文本与训练文本毫无关系。同一个工具在封闭评测中的得分通常高于开放评测。

综上所述,随着汉语词法分析算法的不断改进,词法分析工具也越来越多。传统的评测方法使用大规模真实语料,统计所有词的切分标注结果,兼顾切分标注的精确率和召回率,具有一定的合理性,但也存在不足。

第一,不同工具采用的训练语料不同,切分标注标准不同,哪些工具对汉语基本词的切分

标注更符合汉语基本词本体研究的需求和期望，即更符合汉语本体研究标准，传统评测语料和评测方法缺乏有针对性的评价。

第二，从评测语料质量上看，一方面，评测语料词数较多，导致难以投入巨大精力反复校对，容易出现切分标注错误或不一致现象；另一方面，评测语料发布后鲜有更新，缺乏有序的版本迭代，难以持续改进评测语料质量。

第三，从词汇分布上看，评测语料源于真实文本，字词分布服从齐普夫（Zipf）定律，这也意味着评测时正确切分标注高频词比正确切分标注低频词可以为评测结果带来更高的分数，即性能指标容易被高频词“遮蔽”。

因此，为了弥补上述不足，既需要有能为参与评测的所有字词提供公平机会的评测方法，又需要有能体现汉语本体研究最新成果的高质量评测语料。

3. 评测方法与评测语料

3.1 评测方法

本文提出了一种新的汉语词法分析评测方法，旨在评估词法分析工具是否充分反映了语言本体研究的最新成果。具体而言，通过精选汉语基本词及能够充分体现汉语语言本体标准的例句，构建标准化的评测语料库，以词法分析工具对这些核心词切分标注的准确率（简称$\mathrm{Acc}_{\mathrm{core}}$）作为评估指标，考察其与语言本体标准的契合程度。准确率的计算公式为：

$$\mathrm{Acc}_{\mathrm{core}}=\frac{\text{核心词被正确切分标注的句子数}}{\text{测试集中全部句子数}}\times 100\% \quad (5)$$

核心词切分标注准确率的最小值为0%，最大值为100%，值越大，表明切分标注越符合语言学本体标准。与传统评测方法不同，本评测方法不统计句中全部词语的切分标注结果，只统计核心词的切分标注结果，让评测结果聚焦在例句中的核心词（每个例句仅包含一个核心词），为所有评测词提供公平的评测机会。本评测方法的评测语料和训练语料来源不同，因此属于开放评测。本评测方法特点如下：

1）评测语料中每个核心词的例句数有限，近似服从均匀分布，低频词获得了同高频词相同的权重，为低频词的切分标注提供了公平的评测机会。

2）当期望评测的字词没有被语文辞书收录或者缺少例句时，可以使用其他同类语文辞书进行扩展，语料质量优于传统抽样方法。

3）评测词汇有更多的分类标记（如语体标记），可以针对不同类别的词汇分别评估切分标注效果。

3.2 评测语料

为评估汉语词法分析工具是否采纳了汉语本体研究的最新成果，应选取汉语本体研究的典范成果作为评测语料。《现代汉语词典》（以下简称《现汉》）以确定词汇规范为目的，以推广普通话、促进汉语规范化为宗旨，具有权威性。本研究将《现汉》第7版的例句作为评测语料，和传统评测语料相比，优势在于：1）从切分标注标准上看，《现汉》的收词和词性标注代表了当前汉语本体研究的最新成果；2）从评测语料质量上看，内容经过语言学家、辞书学家反复推敲，并通过广大读者的反馈不断完善，定期系统性修订，更权威，质量更高；3）从评测对象上看，《现汉》对汉语基本词汇系统进行了精选，评测更有针对性。

《现汉》不仅收录词，还收录不成词的语素、非语素字、词组、成语和其他熟语等。为使研

究结果更有针对性,本研究只选择词的例句作为评测句子⑤。《现汉》的配例包括多种形式,比如词例、短语例、例句等。《现汉》未对不同类型的配例加以区分,因此需要从不同形式的配例中识别例句。根据例句字数普遍比短语例、词例更多的特点,本研究将字数多于5(统计时忽略句末点号)的配例视为例句。评测语料样例如表1所示。

表1　评测语料样例⑥

序号	词目	词性(简称)	语体标记	例句
1	吃劲	形	方	这出戏不怎么样,看不看不**吃劲**。
2	尺寸	名	口	说话要掌握好**尺寸**。
3	低回	动	书	使人**低回**不忍离去。
4	宝贵	形	—	这是一些十分**宝贵**的出土文物。

表1中,行1~3分别是方言词、口语词、文言词的样例,行4是普通汉语词的样例。经统计,字数多于5的例句合计25460个,涉及17611个词目。其中,方言词例句1141个,口语词例句666个,文言词例句571个,普通词例句23082个。

为探讨词频对词法分析性能的影响,本研究基于《现代汉语常用词表》第2版(李行健和苏新春,2021)对高频词与低频词的词法分析性能进行了分析。《现代汉语常用词表》第2版共收录词语56790个,频序号较小的词语,其词频不低于频序号较大的词语。经统计,本研究所涉及的17611个词目中,有14841个在该表中出现,占比84.27%。为分析高频词与低频词切分标注的性能差异,本研究将词频序号最小和最大的10%词目(各1484个)视为高频词和低频词代表,并以此为基础对例句切分标注的性能进行统计分析。

4. 性能评测与分析

4.1 评测工具介绍

由于分词工具大多同时支持词性标注,本研究选择的分词和词性标注工具完全相同,具体信息如下:

1) HanLP:由一系列模型和算法组成的自然语言处理工具包。评测选择基于深度学习方法实现的2.1.0-beta.50版。该工具提供了多种分词和词性标注模型,评测选择性能最佳的分词模型(CTB9_TOK_ELECTRA_BASE)和词性标注模型(C863_POS_ELECTRA_SMALL),命名为HanLP。

2) Jiagu:包含分词和词性标注功能在内的自然语言处理工具,评测选择0.2.3版默认模型。

3) jieba:"结巴"工具,分词支持精确切分、全切分、搜索引擎、paddle等四种模式。其中精确切分和paddle模式以更高的F1值为目标,前者采用传统的机器学习方法,后者采用深度学习方法。评测选择这两种模式,分别命名为jieba[1]和jieba[2],版本为0.42.1版,模型均为默认。

4) LTP:哈尔滨工业大学语言技术平台,支持深度学习算法和感知机算法,其中深度学习方法提供了5个模型。评测选择4.2.13版性能最佳的深度学习模型(Base1)以及感知机算法

⑤　根据《现汉》第7版凡例,词均有词类标记。因此,是否标记了词类可以作为词目是否为词的依据(王惠,2009)。

⑥　语体标记"方、口、书"说明词目在例句中分别为方言词、口语词和文言词,"—"说明词目在例句中为普通汉语词(简称普通词)。

默认模型,分别命名为 LTP[1]和 LTP[2]。

5) NLPIR:中国科学院计算技术研究所最早推出的分词和词性标注工具。该工具采用传统的机器学习算法进行切分标注一体化分析,评测选择 Python 语言封装的 PyNLPIR 0.6.0 版默认模型。

6) pkuseg:北京大学推出的多领域中文分词工具包,基于传统的机器学习方法研制。评测选择 Python 0.0.25 版由默认的混合数据集训练的通用模型。

7) SnowNLP:一套专注简体汉语文本处理的工具包,基于传统机器学习方法研制,评测选择 0.12.3 版默认模型。

8) Stanza:斯坦福大学推出的多语言处理工具包,基于深度学习算法研制,支持 60 余种语言的分析。评测选择 1.5.0 版默认的简体汉语模型。

9) THULAC:清华大学推出的一套中文词法分析工具包,基于机器学习方法研制。评测选择其 Python 语言版本 0.2.2 版,模型为《人民日报》分词和词性标注语料库训练的模型 Model_2。

10) Trankit:俄勒冈大学推出的轻量级多语言处理工具包,基于深度学习算法研制,支持 56 种语言的分析。评测选择 1.1.1 版默认的简体汉语模型。

4.2 分词性能评测与分析

实验使用 12 款工具或模型⑦对评测语料进行分词,并对核心词的切分结果进行统计。除分词总准确率外,还分别对不同语体词(方言词、口语词、文言词和普通词)和不同词频词(高频词和低频词)的分词准确率进行了统计,详情如表 2 所示。

表 2 分词准确率(%)

工具	不同语体词分词准确率				不同频次词分词准确率		总准确率
	方言词	口语词	文言词	普通词	高频词	低频词	
HanLP	77.04	81.08	86.51	89.65	87.47	88.31	88.79
LTP[1]	73.53	75.23	80.04	87.83	90.63	71.87	86.69
pkuseg	70.55	74.92	66.73	86.60	82.43	80.97	85.13
LTP[2]	65.03	69.82	70.93	84.94	87.17	72.31	83.34
Trankit	68.62	66.52	81.26	83.97	75.34	82.18	82.77
jieba[1]	73.18	79.43	72.15	83.48	74.81	86.21	82.66
Jiagu	64.33	62.91	69.53	84.07	79.30	67.13	82.31
THULAC	63.80	60.51	62.87	81.27	84.91	67.13	79.53
NLPIR	52.67	58.11	52.54	79.37	92.38	40.26	77.01
Stanza	52.50	48.65	64.62	73.09	76.04	61.94	71.34
jieba[2]	50.31	54.05	45.71	73.34	77.54	49.20	71.19
SnowNLP	51.53	50.00	52.54	72.53	85.24	42.64	70.55

表 2 显示,12 款工具分词的总准确率均未达到 90%。其中,HanLP 性能显著高于其他工具,不但总准确率最高(88.79%),对不同语体词的分词准确率也是最高,说明 HanLP 具有良好的分词性能和语体适应性。

从语体角度看,所有工具切分普通词的准确率均比切分其他语体词的高,HanLP、LTP[1]和

⑦ 下文不再区分是工具还是模型,统称为工具。

pkuseg 分别达到 89.65%、87.83%和 86.60%，其他工具亦普遍在 72.53%及以上。然而，在切分其他语体词时，准确率下降明显。比如，切分方言词时仅 HanLP（77.04%）和 LTP[1]（73.53%）维持较高水平，大多数工具的准确率都在 71%以下。再比如，在切分文言词时，尽管 HanLP（86.51%）和 Trankit（81.26%）达到了较高的准确率，但依然低于同工具在普通词上的 89.65%和 83.97%。

从词频角度看，各工具切分高频词和低频词的准确率存在较大差异。大多数工具切分高频词的准确率较高，如 NLPIR 和 LTP[1]切分高频词的准确率分别为 92.38%和 90.63%，仅 3 个工具（HanLP、Trankit、jieba[1]）切分低频词的准确率较高。在切分低频词时，各工具的表现差异较大。部分工具如 HanLP 和 pkuseg 在切分低频词和高频词时表现出相近的准确率（HanLP 的相差 0.84%，pkuseg 的相差 1.46%），显示出较好的稳定性和通用性，总的切分准确率也比较高。部分工具切分高频词和低频词时表现出较大的性能差异。例如，NLPIR 切分高频词的准确率高达 92.38%，但切分低频词的准确率仅为 40.26%，相差 52.12%。类似的，SnowNLP 切分高频词的准确率为 85.24%，而切分低频词的准确率仅为 42.64%，显示出较差的稳定性和通用性，总的切分准确率也较低。

4.3 分词与词性标注评测与分析

实验使用 12 款工具对评测语料同时进行分词与词性标注，对核心词的切分标注结果进行统计。由于不同工具采用的词性标签集不同，同《现汉》的 12 种词性的标签体系也有区别，因此评测时需先将不同标签体系的词性统一转换为《现汉》词类标签，然后再计算准确率。分词与词性标注准确率详情如表 3 所示。

表 3　分词与词性标注准确率（%）

工具	不同语体词分词与词性标注准确率				不同频次词分词与词性标注准确率		总准确率
	方言词	口语词	文言词	普通词	高频词	低频词	
LTP[1]	55.92	63.96	62.00	80.24	80.60	63.60	78.32
HanLP	55.74	64.71	61.30	79.68	74.86	74.08	77.80
LTP[2]	44.52	54.80	48.16	74.33	72.61	58.08	71.89
pkuseg	46.28	55.41	38.88	72.26	66.44	61.89	69.91
Jiagu	42.51	48.65	42.56	71.67	63.36	50.08	69.11
THULAC	43.30	47.90	41.16	70.94	70.23	54.83	68.43
Trankit	50.92	53.90	50.44	69.91	58.65	66.46	68.21
NLPIR	29.97	47.00	36.25	69.22	71.78	35.14	66.14
jieba[2]	36.46	42.49	30.47	63.35	62.16	38.39	60.86
jieba[1]	36.81	46.55	40.81	58.16	50.95	55.87	56.51
Stanza	34.71	35.89	34.68	57.50	55.34	47.21	55.40
SnowNLP	19.37	29.13	23.12	53.22	66.34	11.31	50.40

分词错误会进一步导致词性标注错误。本评测中，只有分词和词性标注结果均正确时才视为正确。表 3 显示，在分词与词性标注一体化任务中，LTP[1]、HanLP 的总准确率最高。但是和分词准确率相比，分词与词性标注总准确率偏低，最高的为 78.32%。增加词性标注后，Jiagu 和 jieba[2]的整体性能由分词时的第七和第十一提升至第五和第九，而 Trankit 由分词时的第五下降至第七，其余工具的总准确率浮动均比较小。由此可见，词性标注性能在不同工具间存在差异，且与分词性能并不总是正相关。

从语体角度看,所有工具切分标注普通词的准确率均高于切分标注方言词、口语词和文言词。其中,LTP[1]切分标注普通词的准确率最高,为 80.24%。然而,在切分标注方言词、口语词和文言词时,所有工具的准确率均下降明显。jieba[1]切分标注口语词的准确率为46.55%,相较于其在普通词分词标注中的准确率(58.16%)下降了 11.61%,为准确率下降幅度最小的情形。换言之,当切分标注方言词、口语词和文言词时,所有工具的分词与词性标注准确率相比切分标注普通词的最低降幅为 11.61%。由此可见,词法分析时,普通语文词的准确率有近 20%(19.76%)的提升空间,而方言词、口语词和文言词的提升空间更大。

从词频角度看,LTP[1]在高频词的切分标注上表现最好,准确率达到 80.60%,而 HanLP 则在高低频词的切分标注上表现最为均衡,高低频词的切分标注准确率分别为 74.86%和 74.08%,相差 0.78%。值得注意的是,大多数工具在切分标注低频词时性能明显下降,如 SnowNLP 切分标注低频词的准确率仅为 11.31%,与高频词的 66.34%形成鲜明对比。准确率下降超过 5%的有 8 款工具,这表明低频词的切分标注仍是当前自动词法分析的重要难点之一。

4.4 方法对比分析

对于传统的公开评测数据集,不排除有词法分析工具将评测数据也加入模型训练。因此,不同工具的性能难以通过这些公开的数据集进行评测对比。为了将本文方法与传统评测方法进行对比,需选择发布了多个模型的工具,对同一工具的不同模型进行评测。

LTP 发布了 6 个分词模型,并公布了它们的性能得分。使用本文方法对这 6 个模型进行评测,不同方法评测的性能结果如表 4 所示。

表 4　两种评测方法对 LTP 不同模型的分词评测性能(%)

LTP 模型	Acc_{core}	F1
Base1	86.69	99.22
Base2	86.09	99.18
Base	85.46	98.70
Small	84.37	98.40
Legacy	83.34	97.93
Tiny	82.44	96.80

表 4 显示,在分词任务上,本文方法对 6 个模型评测的性能排序与 LTP 公布的排序完全一致,说明了本文方法同传统评测方法一样有效。但是,本文方法得到的性能普遍低于传统方法,且不同模型之间的分差更大。传统评测方法下 Base1 的性能为 99.22%,而本文方法下 Base1 的性能只有 86.69%,传统评测方法下最高性能的与最低性能相差 2.42%,而本文方法相差 4.25%。

本文方法评测得到的性能较传统评测方法更低,主要原因有:1)评测语料的切分标准不同,本文方法以语文辞书例句为评测语料,更体现语言学本体研究的最新成果;2)评测对象不同,本文方法忽略了标点、成语等容易切分准确的语言现象,而传统评测方法将标点、成语等均统一作为评测对象;3)低频词在训练语料中出现少,容易出错,在传统评测方法中对整体性能的影响也小,而本文方法为低频词的评测赋予了同样权重;4)本文方法是一种完全开放的评测。本文方法评测的性能不如 F1 值反映出来的乐观,这也说明汉语的词法分析还存在很大的改进空间,比传统评测方法体现出的改进空间更大。

4.5 分词与词性标注错误分析

12 款词法分析工具对 25460 个例句中的核心词进行分词与词性标注的错误详情见表 5。

表 5　12 款词法分析工具分词与词性标注的错误详情

分析错的工具数	分词任务		分词与词性标注任务	
	核心词例句数	百分比(%)	核心词例句数	百分比(%)
0	8579	33.70	4607	18.10
1	4903	19.26	3814	14.98
2	3210	12.61	3121	12.26
3	2287	8.98	2342	9.20
4	1592	6.25	1987	7.80
5	1231	4.84	1645	6.46
6	910	3.57	1372	5.39
7	677	2.66	1205	4.73
8	575	2.26	1018	4.00
9	502	1.97	1148	4.51
10	435	1.71	1093	4.29
11	311	1.22	996	3.91
12	248	0.97	1112	4.37
合计	25460	100.00	25460	100.00

从分词结果看,8579 个核心词的例句(33.70%)被所有工具正确切分,16881 个核心词的例句(66.30%)存在被错误切分的情况,248 个例句(0.97%)被所有工具错误切分。从分词与词性标注结果看,4607 个核心词的例句(18.10%)能被所有工具正确切分标注,81.90%的核心词例句存在被错误切分标注的情况,有 1112 个核心词的例句(4.37%)被所有工具切分标注错误。这充分说明现有词法分析工具与语言学本体标准存在显著差距。

我们以“我男、女乒乓球队双双获得冠军。”为例(副词“双双”的例句)说明 12 款词法分析工具的结果。12 款工具中,1 款将“双双”错误切分为两个不同的词,11 款将其正确切分为一个词但全部将其词性标错,其中 2 款工具(pkuseg、Stanza)将其标注为名词,2 款工具(HanLP、Jiagu)将其标注为形容词,其余工具将其标注为数词。

4.6 意见建议

本文方法可以为评测词法分析工具是否契合语言学本体标准提供新的手段。在词法分析工具改进方面,一方面建议建设质量更高、更契合汉语本体研究成果的语料;另一方面,建议研制对不同频次词、不同语体词赋予同等权重的算法模型。在词法分析工具选择方面,自动词法分析结果可以满足研究需求时,词法分析工具可参考本文评测结果直接选择。但是,如果词法分析结果用于语言教学与研究等对性能要求较高的领域,依然有必要对自动词法分析的结果进行人工校对。

5. 结语

本文提出了一种新的汉语词法分析评测方法,仅对精选例句核心词切分标注的准确性进行评测。具体实验时,本文将《现汉》第 7 版词目视为汉语基本词汇系统的代表,将词目例句作为评测语料。评测结果更能说明在切分标注标准尚存分歧的前提下,不同的词法分析工具是否契合当下权威的语言学本体标准,是对传统评测方法的有效补充。

本文主要贡献有:1)提出了一种区别于传统评测方法的新方法,专注于对句子中核心词的切分标注结果进行评价,评价结果更能体现与汉语本体研究成果的一致性;2)格式化《现

汉》第7版词条释义,提取词目、词性、例句等信息,制作了一个新的反映汉语本体标准的评测语料库,为面向汉语本体标准的词法分析工具评测提供了数据支持;3)本文研究成果不仅可以指导词法分析工具的改进,也可以为汉语本体研究及相关应用研究中选择合适的词法分析工具提供建议。

本文不足在于评测语料规模较小,不同语体词数量分布不均衡,未考虑未登录词(Out-of-Vocabulary,OOV)和新词的切分标注,且未对工具的分析速度进行测试。后续将整理更多反映汉语本体最新研究成果的例句,扩大评测语料库规模,尤其是有针对性地提高文言词、口语词、方言词的例句比例,以便获得更准确、更全面的评测结果。

本文评测结果显示,现有词法分析工具的性能依然有较大的改进空间,如何研制更优的词法分析工具,更好地服务学界和社会需求,依然是值得持续深入研究的重要课题。

参考文献

何　晗　2019　《自然语言处理入门》,人民邮电出版社。

李行健　苏新春(主编)　2021　《现代汉语常用词表》(第2版),商务印书馆。

刘　群　张华平　俞鸿魁　程学旗　2004　《基于层叠隐马模型的汉语词法分析》,《计算机研究与发展》第8期。

刘　挺　车万翔　李正华　2011　《语言技术平台》,《中文信息学报》第6期。

唐　琳　郭崇慧　陈静锋　2020　《中文分词技术研究综述》,《数据分析与知识发现》Z1期。

王　惠　2009　《词义·词长·词频——〈现代汉语词典〉(第5版)多义词计量分析》,《中国语文》第2期。

张　奇　桂　韬　黄萱菁　2023　《自然语言处理导论》,电子工业出版社。

中国社会科学院语言研究所词典编辑室(编)　2016　《现代汉语词典》(第7版),商务印书馆。

Che, Wanxiang, Yunlong Feng, Libo Qin, and Ting Liu　2021　N-LTP: An open-source neural language technology platform for Chinese. ArXiv. https://doi.org/10.48550/arXiv.2009.11616.

He, Han, and Jinho D. Choi　2021　The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. ArXiv. https://doi.org/10.48550/arXiv.2109.06939.

Van Nguyen, Minh, Viet Dac Lai, Amir Pouran Ben Veyseh, Thien Huu Nguyen　2021　Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. ArXiv. https://doi.org/10.48550/arXiv.2101.03289.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning　2020　Stanza: A Python natural language processing toolkit for many human languages. ArXiv. https://doi.org/10.48550/arXiv.2003.07082.

Toutanvoa, Kristina, and Christopher D. Manning　2000　Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the* 2000 *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 63–70, Hong Kong, China.

Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning　2005　A conditional random field word segmenter for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 168–171, Jeju Island, Korea.

张永伟　北京　中国社会科学院大学文学院/中国社会科学院语言学重点实验室　zhangyw@cass.org.cn;

刘　婷　北京　中国社会科学院大学文学院　2621146148@qq.com;

王玉莹　北京　中国科学院大学计算技术研究所　wangyuying23@mails.ucas.ac.cn

ZHONGGUO YUWEN
STUDIES OF THE CHINESE LANGUAGE
May, 2025

Abstracts of Major Papers in This Issue

CAO Zhiyun, The documentation and study of Chinese dialectal oral culture

Chinese dialectal oral culture can be embodied by the proverbs, folklores, ballads, folk performing arts and short plays preserved in various Chinese dialects. In history, the collection and study of Chinese oral literature have long been valued. But little attention has been given to scientific research on the recording, transcribing and annotating of these materials from the perspective of dialectology. This paper proposes comprehensive guidelines on the documentation of Chinese dialectal oral culture. It also discusses issues concerning the selection of survey subjects, investigation items, recording methods, audio-visual recording, transcribing methods, the application of IPA, text processing, video editing and text compilation etc. In contrast with existing disciplines, the guidelines emphasize the significance of dialectological viewpoints and academic standards in the study of Chinese dialectal oral culture.
Keywords: dialectal oral culture; folk literature; transcription; The Chinese Language Resources Protection Project

YING Xuefeng and LU Bingfu, Principle of elastic iconicity and the position of localizers in Chinese street names

By analyzing the street names in Shanghai, Beijing and Shenzhen, this paper propses that the position of monosyllabic localizers such as *dong* (东, "east") and *shang* (上, "upper") in street names demonstrates certain regularities. For street names given in modern times, monosyllabic localizers are strictly posited in the middle as in *zhongshan nan lu* (中山南路, "Zhongshan south road"). In street names which already existed in earlier times, monosyllabic localizers are often posited at the beginning. These ancient street names are replaced by new street names in a rather gradual and slow manner. The sequence of words in street names reflects the elastic iconicity regarding rhythm, grammar, semantics and register. In general, modern street names tend to be more formal whereas those passed down from earlier times are more colloquial.
Keywords: elastic iconicity; street names; monosyllabic modifiers; modifier-head structure; register

ZHANG Yongwei, LIU Ting and WANG Yuying, Evaluation scheme for Chinese lexical analysis toolkits from the perspective of theoretical linguistics

For lexical analysis of modern Chinese, no consensus has been reached on the standards of segmentation and part-of-speech tagging. This has led to the adoption of different practices when processing Chinese. Since traditional evaluation methods of lexical analysis toolkits have come across many limitations, this paper proposes a novel evaluation scheme. A specialized corpus of basic words and sentences is build up according to the theories of Chinese grammar. The performance of lexical analysis toolkits is then systematically assessed on their accuracy of segmentation and part-of-speech tagging in this specialized corpus as evaluation metrics. The results indicate that, recent lexical analysis toolkits cannot fully align with linguistic facts; Optimization potential still remains in the processing of

dialects, colloquial expressions, Classical Chinese, and low-frequency words.

Keywords: segmentation standards; lexical analysis; dictionary; corpus; evaluation

ZHOU Minli, LI Xiaojun and QUAN Yanping, Contraction-induced grammaticalization and constructional changes: On degree adverb *ɕiau*^{45}*te*^{0} (晓得) and related forms in southern Hunan dialects

In southern Hunan dialects, *ɕiau*^{54}*liau*^{54} (晓了) and *ɕiau*^{45}(*te*^{0}) (晓(得)) can function as subjective degree adverbs. Their grammaticalization is closely related to the contraction of the construction "*ɕiau*^{54}*pu*^{33}*te*^{33} (晓不得) + *xau*^{54} (好) + X". Such constructions with the negation of a verb meaning "to know" (i.e. *ɕiau*^{54}) and a degree interrogative (i.e. *xau*^{54}) have undergone various degrees of contraction in Chinese dialects, which can be roughly categorized into two main types (direct omission and phonetic fusion) and twelve subcategories with four reinforcement strategies. The original construction and its contracted form present three main distribution types synchronically. Influential factors include the occurrence of set phrases resulted from high frequency of use, constraints of Chinese degree expressions, as well as semantic and phonological limitations. Comparing the similar types of constructions in different dialects thus building up a network of constructional changes would not only help to explore the origin of certain function words in Chinese dialects but also deepen our understanding of the diachronic change of grammatical constructions.

Keywords: degree adverbs; contraction; grammaticalization; constructional change; construction network

WANG Yili, The *sy*^{33}(处) and *dou*^{22}(度) for the expression of location in Guangzhou dialect

In Guangzhou dialect, there are two words for the expression of location, *sy*^{33}(处) and *dou*^{22}(度). In early literature of Guangzhou dialect, only *sy*^{33} was attested and *dou*^{22} is believed to have appeared no earlier than the end of the 19th century. After the coexistence for about half a century, *dou*^{22} has finally replaced *sy*^{33} as the common word for location in Guangzhou dialect, which is reflected in the choice of words between generations. On the origin of the two words, *sy*^{33} already existed in Classical Chinese and has undergone certain diachronic phonetic weakening; *dou*^{22}, on the other hand, has been derived from *dao* (道, "path"). Their functions and variation in modern Yue dialects suggest that, *sy*^{33} represents an early layer whereas the gradual spread of *dou*^{22} into surrounding dialects started after its occurrence in the authoritative Guangzhou dialect. Hence, variation between different generations and geographical areas can reveal phases of diachronic evolution.

Keywords: expressions of location; diachronic replacement; geographical distribution; generation difference; grammaticalization

WU Yonghuan, The origin of the demonstrative pronoun *nie* (乜) in Shandong dialects revisited

Among Shandong dialects, the demonstrative pronoun *nie* (乜) has different referential functions (i.e. proximal, middle and distal) in different regions. By analyzing the phonetic and semantic features of *nie* in comparison with other historical materials and contemporary Chinese dialects, this paper proposes that the origin of *nie* could be traced back to the demonstrative pronoun *er* (尔) in Archaic Chinese.

Keywords: *nie* (乜); *er* (尔); demonstrative pronoun; Shandong dialects

ZHANG Shifang, Aspiration-conditioned final split in Chinese dialects

In modern Chinese dialects, "aspiration-conditioned final split" refers to the phenomenon that, depending on the [±aspirated] feature of the initial, words which are supposed to share the same final under phonological evolution rules have diverged into words with different finals. This phenomenon mainly concerns three sets of stops, i.e. p-p^{h}, t-t^{h}, and k-k^{h}. The paper proposes that, aspiration-conditioned final split is not caused by articulation mechanisms. It represents a systematic "bidirectional orderliness" formed during the merge/split or historical sound change of MC