文章编号：1003-0077（2017）00-0000-00

基于众包标注的语文教材句子难易度评估研究

于东[1]，吴思远[1,2]，耿朝阳[1]，唐玉玲[1]
（1.北京语言大学 信息科学学院，北京 100083；2.北京语言大学 汉语国际教育研究院，北京 100083）

摘要：该文提出了一种基于成对比较的众包标注方法，该方法可以通过非专业人士的简单判断获取标准统一的句子难度标注结果。基于该方法，构建了基于语文教材的汉语句子难度语料库。面向单句绝对难度评估和句对相对难度评估两项基本的句子难易度评估任务，使用机器学习方法训练汉语句子难度评估模型，并进一步探讨了不同层面语言特征对模型性能的影响。实验结果显示，基于机器学习的分类模型可以有效预测句子的绝对难度和相对难度，最高准确率分别为 63.37%和 67.95%。语言特征可以帮助提升模型的性能，相比于词汇和句法层面的特征，加入汉字层面特征的模型在两项任务上的准确率最高，说明汉字特征对句子难度的预测作用最强。

关键词：句子难易度评估；可读性研究；众包标注；语文教材语料库

中图分类号：TP391 **文献标识码：**A

Assessing Sentence Difficulty in Chinese Textbooks Based on Crowdsourcing

YU Dong[1] ,WU Siyuan[1,2], GENG Zhaoyang[1] ,TANG Yuling[1]
(1. College of Information Science, Beijing Language and Culture University, Beijing 100083, China;
2.Research Institute of International Chinese Language Education, Beijing Language and Culture University, Beijing 100083, China)

Abstract: We propose a crowdsourcing annotation approach based on pairwise comparison, which use hundreds of non-experts to determine the difficulty of amounts of sentences according to a uniformed standard. Using this annotation approach, we construct a textbook-based sentence corpus which contains 18,411 Chinese sentences. Based on this annotated corpus, machine learning models are trained to predict the absolute difficulty of single sentences and the relative difficulty of sentence-pairs. We further study the impact of multi-level linguistic features. In the first task, we effectively predict the difficulty level of sentences with 63.37% accuracy. In the second task, our model distinguishes the easy and difficulty sentences with 67.95% accuracy. We find that the Chinese character level features are the strongest predictive features on the two tasks compared with the lexical and syntactic level features.

Key words: Sentence Difficulty Assessment; Readability Research; Crowdsourcing; Textbook Corpus;

0 引言

阅读是人类获取信息、认识世界和思维发展的重要活动，也是语言学习的重要内容。难度合适的阅读文本可以促进阅读过程的顺利进行，难度不合适则会阻碍阅读的进行，甚至损害读者的阅读兴趣。因此，评估阅读材料的难度并根据语言水平进行针对性、个性化的阅读逐渐成为社会各界的共识。其中，评估阅读文本的难易程度，即文本可读性研究扮演着关键而基础的角色[1]。

文本可读性的自动评估是文本可读性研究的

收稿收稿日期：xxxx-xx-xx；**定稿日期**：xxxx-xx-xx
基金项目：国家社会科学基金（17ZDA305）；教育部人文社会科学研究青年基金项目（19YJCZH230）；北京语言大学中青年学术骨干支持计划

核心，也是语言学、心理学与自然语言处理领域共同探讨的课题之一。自动评估文本可读性，就是将影响阅读难度的、可以量化的文本因素综合起来，构建一个自动评估模型，通过模型评估文本的可读性[2]。由于文本的可读性可以用连续的难度值或者离散的难度级别（如年级）表示，可读性自动评估任务通常被转化为回归或分类问题。基于多层面语言特征的机器学习方法是可读性自动评估的主流方法，其核心是从字、词、句和篇章等层面分析和筛选可以预测文本难度的有效语言特征[3-4]。语言特征的选择与文本的语言属性有关，其他语言研究中的有效特征对汉语特征选择具有启发意义，但不能直接应用于汉语可读性评估[1-2]。

按照文本粒度的不同，可读性自动评估任务主要分为文档级的可读性评估和句子级的可读性评估[5]。现有研究多以文档级为主，但文档级的评估模型在短文本上表现不佳，也无法满足特定任务的需求[6]。句子级的难易度评估拥有更加切实的应用场景。例如，根据句子难易度评估结果，教师和图书出版商可以有针对性地修改困难句子[6]。作为一项语言评价技术，句子难易度评估在试题研制、翻译质量评估上也有广泛的需求[7]。同时，句子难度评估方法的研究可以为文档级的可读性研究奠定基础。

目前的汉语可读性研究集中在文档级的可读性评估上[8-10]。一些句子难易度评估研究对影响句子难度的语言特征进行了探讨，但缺乏具体的量化方式和实验证据，在语言特征的选择上也存在不足[11-13]。没有发现使用机器学习方法进行汉语句子难易度评估的研究。因此，汉语句子难易度自动评估有很大的研究空间。

本文首先提出了一种基于众包标注的成对比较方法来标注句子的难度级别。基于该方法，我们构建了基于语文教材的汉语句子难易度语料库，把句子难易度评估转化成分类问题，探究了机器学习方法在两种句子难度评估任务（句对相对难度评估和单句绝对难度评估）上的表现。本文还对比分析了汉字、词汇和句法特征对句子难度评估的作用。实验结果表明，机器学习方法可以有效地评估汉语句子的难度。在预测单句难易度的五分类任务上，模型的准确率达到了63.37%。在句对相对难度评估任务上，模型有效区分了简单句和困难句，最高准确率为67.95%。

本研究的主要贡献包括以下三个方面：

（1）提出了基于众包标注的句子难度标注方法，这种方法通过非专家的简单判断任务就可以获取标准统一的难度标注结果，可适用于大规模的句子难度标注语料库的构建。

（2）构建了基于语文教材的汉语句子难度标注语料库。该语料库包含 18,411 个具有五个难度级别标注的汉语句子，为汉语可读性研究提供了数据支持。

（3）使用机器学习方法进行单句绝对难度评估和句对相对难度评估两项任务，验证了机器学习模型在汉语句子难易度自动评估上的有效性。

（4）选取并分析了多层面语言特征，并对语言特征在难度评估上的预测作用进行了验证。

1 相关研究

评估文本的难易程度一直是教育学、语言学和自然语言处理领域所关心的问题。从 20 世纪 20 年代以来，各个语言的研究者根据自身语言特点，通过量化不同层面、不同维度的语言特征，构建线性或者非线性的模型进行自动评估[1,3]。传统的可读性研究通过量化文本的表层特征（如词长、词频等），构建多元线性回归公式来评估文本的阅读难度。最具代表性的可读性公式有 Flesch-Kincaid 可读性公式[14]和 Smog 公式[15]等。随着计算机和自然语言处理技术的发展，越来越多的复杂模型被构建出来应用于文本可读性评估工作[16-18]。可读性自动评估拥有广泛的应用场景，不仅可以帮助教师选择合适的阅读材料，为教材编写、阅读测试提供参考，而且也可以应用于一些自然语言处理任务如智能改编、作文自动评分上[19]。

有监督的机器学习方法是自动评估文本可读性的主流方法。相关研究包括构建统计语言模型评估网页文本的阅读难度[16]，或者把可读性评估任务视为分类任务，构建分类模型预测文本的可读性级别[3,8]。这些基于特征工程的方法发现，语言特征的选择对于可读性评估起着重要的作用[20]。但有效特征的预测能力与语言特点有关[20-21]。这些研究中预测能力高的语言特征是否适用于汉语，还有待进一步探究。

句子是语言学习中常用的语言单位，也是多

项自然语言处理任务的基本处理单元。句子级的可读性研究受到越来越多的关注，按照任务的不同可以把句子级可读性评估分为单句绝对难度评估和句对相对难度评估两项。

Pilán 等[5]从第二语言学习角度探讨了影响瑞典语句子难易度的语言因素。该研究将句子可读性评估抽象为二分类问题，支持向量机分类器在该任务上达到了 71%的准确率。Dell’Orletta 等[22]对比了表层特征、词汇特征、形态句法特征和句法特征在意大利语文本可读性评估中的作用。他们的研究表明，无论是句子级还是文档级的可读性评估，句法特征都是预测意大利语文本可读性最重要的预测指标。Brunato 等[23]发现，在表层特征、形态句法特征和句法特征中，与句子结构相关的句法特征与英语文本的阅读难度高度相关。

Inui 和 Yamamoto[24]首次提出了句对相对难度评估的任务，通过收集原句与手工简化句之间的相对难度判断，该研究使用基于支持向量机的比较器评估了听力障碍人士对句子难度的感知。Vajjala 和 Meures[25]提出基于配对排序的句子可读性评估方法。该任务是对句对的相对难度进行判断，具体来说，给定包含一个简化句-原句的句子对，判断哪个句子更难。Schumacher 等[26]评估了一组句子在有上下文和无上下文条件下的相对阅读难度。该研究使用众包标注的方法收集了人类对句对相对难度的判断，然后使用词法和句法特征训练了逻辑回归模型预测句子对的相对难度。研究发现，词汇相关特征可以帮助预测句对相对难度，句子在文本中的上下文信息会影响人类对句子难度的判断。

国内句子难易度自动评估的研究仍处于起步阶段。江少敏[11]采用调查问卷和对比分析的方法，从字、词和句法层面收集了被试对语言特征预测能力的主观评价，并建立了句子难易度测量公式。庞成[13]把影响句子难度的因素分为内部结构、外部结构和意义形式三个范畴。郭望皓[12]对字层面和词层面的特征进行了量化，并使用 CRITIC 加权赋值法计算了各指标在预测句子难度上的权重，构建了线性公式。上述研究在影响句子难度因素的选择上缺乏系统性和结构性，还没有学者使用机器学习的方法进行汉语句子难易度评估工作，也没有对语言特征的预测作用进行系统的考察。汉语句子的难易度自动评估的难点在于缺乏一定规模的难度标注句子语料库。

2 众包标注方法

基于机器学习的文本难易度自动评估方法需要一定规模的标注数据。然而作为一种缺乏形式标记的信息，文本难度标注的困难之处在于难度无法界定、标注标准无法统一。

主观量表法与成对比较法是主要句子难度标注方法。主观量表是一种包含若干有序级别的量表，用以测量个人对文本的主观难度评价，按照级别可分为 5 点量表、7 点量表、9 点量表等，标注人员多为有经验的专家、教师等[27]。主观量表法可以有效地确定文本的阅读难度[28]，但数据规模较小。标注者面对大量的待标注数据时，很难保证统一的标注标准。成对比较法需要标注者比较给定的两个句子并判定哪个句子更难[24-25]。成对比较法是一种相对简单的标注任务，具有正常语言能力的标注者都可以进行句子相对难度的判断。但该方法只能得到句子的相对难度，无法给出精确的难度级别或难度值[17]。现有的两种难度标注方法只能标注小规模的数据，面对大规模数据无法给出标准统一且具有具体难度值的标注结果。

本文提出了一种基于成对比较的众包标注方法，该方法首先通过主观量表标注小规模句子的难度。量表的使用不仅为量化难易度提供统一的坐标，量表中的量点还可以把连续的难度划分成若干个难度区间。然后通过基于成对比较的众包方法把未标注句归类到某个难度区间上，达到标注句子难度的目的。这种方法把难度标注转化为简单的难易判断任务，便于非专业人士使用统一的标注标准对句子难度进行标注。该标注方法的原理如图 1 所示。

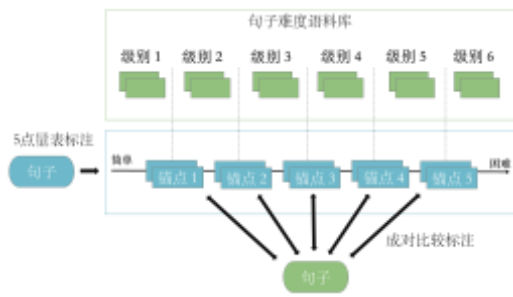

图 1 基于成对比较的句子难度众包标注方法

具体来说，该方法主要包括两个步骤：

1）在小规模数据上使用主观量表量化句子的难易程度。根据主观量表的评定结果，在量表的每个点上选择部分句子作为锚点，如 5 点量表，则选择可以代表 5 个难度点的句子作为锚点，5 个锚点把连续的难度划分为 6 个区域；若使用 7

点量表，则可以选择 7 个难度点上的句子作为锚点，把连续的难度划分为 8 个区域。

2）使用众包标注的方法，通过成对比较任务判断锚点句和未标注句的相对难度，根据判断结果把未标注句划分到特定的难度区间中，该难度区间即为该句子的难度级别。主要流程如图 2 所示。

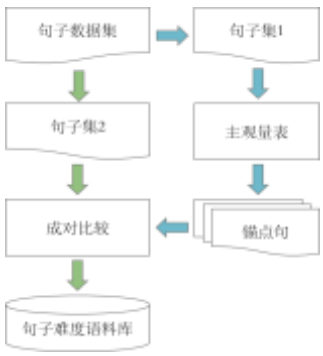

图 2 基于成对比较的句子难度众包标注流程

3 语料库构建

3.1 数据收集

语料库中的汉语句子来源于汉语语文教材中的课文文本。语文教材中的课文属于权威典范的文本，体裁丰富，来源广泛。我们收集了人教版、苏教版和北师大版三个版本 1-12 年级的语文课文，剔除了特殊语言使用的和不完整的文本，如文言文、诗歌、诗词、剧本、识字文本等。对句子进行去重后，1392 篇课文共产生 51,298 个句子，句子的平均长度为 24.6（MD=16.19）。

3.2 基于五点量表的专家标注

我们从原始句子集中随机选择 250 个句子。3 名小学教师和 2 名教育领域研究生被要求认真阅读这些句子，并在五点量表上对句子的难度进行评分，1 表示非常简单，5 表示非常难。包含 250 个句子的五点量表问卷需要大约 20 分钟完成。最终收集了 1250 个标注数据。5 位专家之间的肯德尔一致性系数（Kendall's coefficient of concordance）为 0.712（P<0.001），说明 5 位专家的标注一致性较高。

对于每一句话，我们使用多数投票原则确定句子的最终难度。为了保证作为锚点的句子难易度一致，计算了每个句子被标注为最终难度的概率，具体来说，如果 5 位专家都把句子标注为 5，则被标注为 5 的概率为 1.0，如果有 4 位专家把句子标注为 5，1 位专家标注为 4，则被标注为最终难度 5 的概率为 0.8。我们选择概率大于等于 0.8 的句子作为锚点句。

最终，62 个句子被选择为锚点句，四组句子代表四个难度锚点（没有难度为 5 的句子）。为了保证四组锚点句之间在难度上具有较高的差异，对四组锚点句的难度差异进行了测量。单因素方差分析结果显示，四组句子的难度差异显著（F=469，P<0.01）。更多信息和示例如表 1 所示。

3.3 基于成对比较的众包标注

我们使用成对比较的标注任务，通过众包标注确定大规模句子的难度级别。本研究的标注过程如下：

标注平台 为了发布众包任务，我们在微信开放平台上开发了众包标注的微信小程序。

标注人员 共有 110 名标注人员参与了众包任务。在参与标注之前，他们被要求报告自己的年龄、性别、教育程度等个人信息。标注者年龄在 19 至 27 岁之间，男女比例为 1：5，大多数人接受大学教育。

表 1 锚点句的基本统计信息及部分锚点句示例

锚点	数量	平均分	方差	例句
1	33	1.1	0.01	小河唱起了快乐的歌。
2	16	2.0	0.03	早晨，雾从山谷里升起来，整个森林浸在乳白色的浓雾里。
3	10	3.0	0.08	瑰丽的朝霞倾泻在戈壁滩上，裸露在黄沙上的石头闪着珠光玉彩。
4	3	3.8	0.00	语言跟着思想情感走，你不肯用俗滥的语言，自然也就不肯用俗滥的思想情感，你遇事就会朝深一层去想，你的文章也就真正是“作”出来的，不致落入下乘。

表 2 句子难度标注语料库的部分例句

难度级别	例句
1	四个班共有四十人。
2	地上的雪厚厚的，又松又软，常常没过膝盖。
3	建筑是社会的缩影，民族的象征，但绝不是某一民族的，而是全人类的结晶。
4	我们要同香港各界人士广泛交换意见，制定我们在十五年中的方针政策以及十五年后的方针政策。
5	周恩来、聂荣臻也很快注意到这种情况，他们接到钱学森的请辞报告后，果断决定，配备强有力的行政领导，把钱学森从这些繁杂事务中解脱出来，让他集中精力思考和解决重大技术问题。

图 3 基于成对比较的众包标注平台界面

标注流程 登录标注平台后，屏幕上会显示一条标注指导语和一对句子，一个是锚点句，一个是待标注的句子，如图 3 所示。标注者被要求认真阅读这两条句子并选择相对简单的那条。每个待标注句会随机与特定锚点中的句子进行匹配。为了减少标注工作量，我们在匹配过程中使用了折半插入策略。例如，一个待标注句首先与锚点 2 的某个句子进行匹配，根据标注结果，该句子与锚点 1 或者锚点 3 的某个句子进行配对。重复这个过程直至确定一个句子的难度级别。每个句子由至少三个标注者进行标注，即每个句子至少被标注三次。平均说来，每个待标注句需要经过两次成对比较得到最终的难度标签，每个句子平均需要 30s 的时间进行判断。

数据处理 4 周的标注共收集了 378,183 个成对判断。对于每个句子，我们删除了标注时间小于 15 秒（1%）和标注次数小于 3 次（28%）的句子。我们使用多数投票原则决定单个句子的难度级别。

数据集构建 最终我们构建了一个基于汉语语文教材的句子难度语料库。该语料库共包含 18,411 个汉语句子，每个句子被标注为 1 到 5 共 5 个难度级别，级别 1 表示很简单，级别 5 表示很难。表 2 给出了每个难度级别上的示例句子。语料库中 5 个难度级别的统计信息如表 3 所示。表中除了包含每个级别中句子的数量信息，还包括了每个级别上句子的平均长度（以字为单位）和句子的平均难度值。句子的难度值的计算方式来自于江少敏[11]，值越大则难度越高。

表 3 句子难度标注语料库的基本统计信息

等级	数量	平均句长	Jiang(2009)
1	2068	8.10	112.86
2	6103	16.58	220.69
3	6485	27.56	353.07
4	2939	42.83	530.36
5	816	65.59	790.50
均值	——	25.86	401.50

在单句绝对难度评估任务上，我们使用基于语文教材的句子难度标注语料库作为实验数据。

在句对相对难度评估任务上，基于句子难度标注语料库，我们使用随机配对的方法构建了句对数据集。具体来说，对于句子 S_i，从语料库中随机选择句子 S_j 组成<S_i,S_j>句对。为了保证数据不重复出现在训练集或者测试集中，句对数据集中的每个句子在整个数据集中仅出现一次，因此，在随机匹配的过程中，每个句子 S 只能匹配或被匹配一次。最终，18,411 个句子共组成 9,205 个句对。我们把相对难度定义为两个句子难度级别的关系。例如，如果句对中两个句子的难度级别相等，则这两句话的相对难度标签为 0。每类难度关系在句对数据集上的分布如表 4 所示。

表 4 句对数据集中类别的分布信息

标签	数量	平均难度差	平均句长差
1	3324	1.53	21,20
-1	3374	1.52	21.13
0	2507	0.0	9.53

4 特征及模型

本文在基于语文教材的汉语句子难度语料库基础上进行两项句子难易度评估任务，分别是：单句绝对难度评估和句对相对难度评估。我们把这两项任务抽象为有监督的机器学习任务，通过构建模型评估句子的绝对难易度和相对难易度。为了提高模型的准确率，并探讨不同层面语言特征在汉语句子难易度评估任务上的作用，我们加入了汉字、词汇和句法层面的语言特征。

本小节将会对所用语言特征、模型和实验设置进行介绍。

4.1 特征抽取

特征体系的设计参考了吴思远[29]的特征框架，该研究从汉字、词汇、句法和篇章四个层面构建了指标体系来进行文档级的汉语可读性评估。本文从汉字、词汇和句法三个层面实现句子语言特征的量化。下面是三个层面语言特征的简要说明。

汉字层面 汉字是汉语的书写符号，汉字的识别难度影响句子的阅读难度。汉字层面的语言特征可以从字形复杂度、汉字熟悉度和汉字多样性三个角度进行量化。

汉字字形复杂度的量化主要考虑了汉字笔画数、汉字对称性，共计 6 个指标。考虑到笔画数效应的大小与汉字频率有关，相比于低频字，笔画数效应在高频字上作用更小[30]。因此在量化笔画数时，对笔画数进行了频率加权，加权方式参考了吴建国等[31]的研究。熟悉度表现为汉字的使用频率，包括句子中只出现一次的单次字信息，汉字字频和常见字信息，共计 7 个指标。汉字字频信息来源于国家语委现代话语语料库提供的《现代汉语语料库字频表》和《汉语字幕字频表》，常见字信息来源于《现代汉语常用字表（3500 字）》。汉字多样性的量化主要使用类符-形符比（Type-Token Ratio,TTR），即文本中出现的不重复汉字数和汉字总数的比值。TTR 有多种计算方式，共计 8 个指标。

词汇层面 词是语言中最基本的造句单位，词汇复杂性在句子理解中起着关键作用。影响词汇难度的特征主要包括词长、词汇熟悉度、词汇多样性和词汇语义难度四个维度。

词长是拼音文字中预测可读性的主要指标，该维度主要量化了 6 个指标，考虑到词长与词频的协同作用，对词长进行了频率加权。词汇熟悉度的量化主要计算词频和单次词，共 5 个指标。词频的信息来自于国家语委现代话语语料库的《现代汉语语料库词频表》和《汉语字幕词频表》。词汇多样性上计算了句子的总词数、句子中不重复的词数和 6 个词的 TTR 值，共计 8 个指标。词汇语义难度是汉语可读性研究中由于技术限制没有纳入的维度，但词义的理解是句子理解的重要内容。本文关注句中五类具有特殊语义作用的词的使用情况，包括实词、虚词、否定词、命名实体和成语。此外，词汇语义难度还包括词在词典中的义项数。共有 8 个指标来量化句子的词汇语义难度。

句法层面 句子结构层面共包括 3 个维度的句法特征：句子表层的复杂度、词性复杂度、句法结构复杂度，共计 25 个指标。

表层复杂度包括句子的长度信息和单句复句信息。句长是影响句子难度判断的重要标准之一，同时长句会倾向于包含更复杂的句法结构，因此句长可以反映句法的结构复杂性。词性层面包括句子中五种主要词性（动词、名词、形容词、副词和介词）的使用情况。句法结构复杂度分别量化自基于短语结构的句法分析结果和基于依存结构的句法分析结果，计算了句子中名词短语、动词短语、形容词短语、副词短语和介词短语的使用情况，统计了句法树的树高作为句法复杂性的指标。主要动词和依存距离[32]被认为可以反映句子加工的难度，因此，句法结构复杂度还对主要动词前的词数和依存距离进行了计算。句法结构复杂度维度共计 10 个指标。

特征计算 首先，我们对文本进行了一系列的分析，使用哈工大研发的语言技术平台（Language Technology Platform，LTP）对文本进行分词、词性标注、命名实体识别和依存句法树构建[33]，使用斯坦福大学研发的斯坦福句法分析工具（The Stanford Parser）构建了短语句法树[34]。在文本分析的基础上，我们通过 python 编程计算得到了汉字、词汇和句法层面的特征指标。这些特征指标将被用作文本的表示，以应用于机器学习模型。

4.2 模型与实验设计

4.2.1 任务一：单句绝对难度评估

任务 单句绝对难度评估任务是句子可读性研究中的典型任务，其目标是，给定任意一个句子，评估该句的难度水平。我们把单句绝对难度评估任务定义为五分类问题。

模型 我们对比了支持向量分类（Support Vector Machine, SVM）和逻辑回归（Logistic Regression, LogR）两种模型的表现。

我们把基于 tf-idf 的词袋向量作为输入构建了基线模型，词袋向量的维度是 200 维；然后把不同层面的语言特征作为句子的向量表示，构建了特征模型。在训练过程中采用了 5 折交叉验证。我们在 Python 中使用 scikit-learn 实现了模型。

评估标准 任务一使用准确率（Accuracy）作为分类模型的评估指标。在句子难度分类任务中，难度级别之间并不是相互独立的，而是有序的。难度为 1 的句子比难度为 2 的句子简单，如果模型把难度为 2 的句子判定为 5 比判定为 3 误差更大。因此，任务一还使用邻近准确率（±Accuracy）和皮尔逊相关系数（Pearson）作为模型的评估指标。

准确率（Accuracy）：被预测正确的句子占所有句子的比例；

邻近准确率（±Accuracy）：句子的预测级别与标注级别的误差在 1 个级别内的句子占所有句子的比例；

皮尔逊相关系数（Pearson）：句子预测级别与实际级别的相关程度。

4.2.2 任务二：句对相对难度评估

任务 句对相对难度评估任务的内容是评估两个给定句子之间的相对难度关系[16-17]。具体来说，给定一个随机句子对<S_i, S_j>，句对的相对难度关系为[-1，0，1]，其中-1 表示比容易，其中 0 表示比难度相等，1 表示比难，这种关系可以形式化地表示为：

$$relative = \begin{cases} 1 & if \quad D(S_i) > D(S_j) \\ 0 & if \quad D(S_i) = D(S_j) \\ -1 & if \quad D(S_i) < D(S_j) \end{cases}$$

其中，$D(S_i)$为句子 S_i 的难度，*relative* 为句对<S_i, S_j>之间的难度关系。我们把句对相对难度评估任务抽象为三分类问题。

模型 本任务使用了 SVM 和 LogR 两个经典的分类模型。我们把两个句子的 tf-idf 向量拼接起来，组成 400 维的向量作为句对表示，构建了基线模型；把两个句子的特征向量拼接起来作为句对表示，构建了特征模型。在训练过程中，采用了 5 折交叉验证。

评估指标 任务二采用准确率（Accuracy）作为模型的评估指标。

5 实验结果与分析

5.1 任务一

单句绝对难度评估的实验结果如表 5 所示，该表展示了仅使用词袋特征的基线模型和加入不同层面语言特征的模型在准确率、邻近准确率和皮尔逊相关系数上的表现。我们对比了 SVM 和 LogR 两种不同的分类模型在该任务上的表现。可以看出，LogR 的准确率高于 SVM，两个模型的准确率整体相差不大。

由表可知，基于 tf-idf 的词袋模型在该任务上可以达到 43-46%的准确率，说明词的使用可以在一定程度上评估单句的难度。基于语言特征的模型高于仅基于词袋的模型，说明语言特征的使用可以提升模型的准确率，帮助预测句子的难易度。

表 5 句子绝对难度评估的实验结果

模型	SVM			LogR		
	Acc	±Acc	Perason	Acc	±Acc	Pearson
tf-idf	46.62%	93.32%	0.55	43.25%	92.13%	0.47
汉字	62.56%	**98.05%**	**0.76**	**63.37%**	**98.18%**	0.76
词汇	61.72%	98.18%	0.75	61.25%	98.05%	0.75
句法	59.41%	97.96%	0.73	58.35%	97.86%	0.72
所有	**62.86%**	97.56%	**0.76**	63.21%	98.10%	**0.77**

SVM 模型中，基于所有语言特征模型的准确率和相关系数最高，但汉字的相关系数与基于所有语言特征的模型相当。LogR 中，基于汉字层面特征的模型达到最高的准确率和邻近准确率，皮尔逊相关系数比基于所有语言特征模型下降了 0.01。基于所有特征的模型可以达到较高的准确率，只基于汉字层面特征模型可以达到与基于所有特征的模型相当的水平，说明语言特征的使用并不是越多越好。基于汉字、词汇和句法三个层面特征的模型准确率都高于基线模型，说明加入语言特征有助于提升句子难度预测模型的性能。

在汉字、词汇和句法三个层面的语言特征中，基于汉字层面特征的模型准确率较高，基于词汇层面特征的模型次之，基于句法层面特征的模型准确率最低，说明汉字特征对于单句难易度评估的预测能力更强。该结果和江少敏[11]的结论不一致，江少敏把句子的难度分为句法、短语和字词三个层面，调查问卷的结果发现，对于小学生和留学生来说，句法层面的因素要难于短语层面和字词层面的因素。本研究把影响句子难易程度的因素分为汉字、词汇和句法三个层面，发现相比于词汇和句法层面的特征，汉字层面特征拥有最好的预测能力。这可能是由于，江少敏的目标群体为小学生和留学生，而本研究的标注人员为汉语水平较高的大学生，大学生的句法知识已经较为丰富，句法因素在判断句子难易程度的时候影响较小。

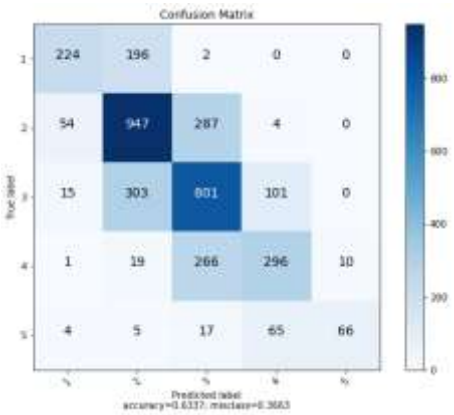

图 4 单句绝对难度预测的混淆矩阵

图 4 显示了基于汉字层面特征的 LogR 模型的混淆矩阵。可以看到，大部分被错误分类的句子都被分到了邻近的难度级别。这说明，句子的难易程度实际上是连续的，分类模型可以把句子分为某几个难度级别，但是各等级之间的边界较为模糊。即使我们使用两两比较的方法，通过锚点句把句子按难度级别划分开来，但标注者在识别具有微小难度误差的句子时仍比较困难。同时，由于我们的标注者的教育水平为大学以上，语言水平较高，在区分低难度级别的句子时不敏感，导致句子的难度被高估或者低估。

模型在难度级别为 5 的句子上分类效果不佳，仅有约 42%的句子被正确地分类。通过分析混淆实例，我们发现，被标注为难度级别为 5 的句子不仅包括现代白话文的句子，还包括语境依赖度较高的对话、非白话文的句子等。这些句子由于使用了特殊的文体或表达方式，理解时需要依靠上下文信息或者背景知识，因此标注者把这些句子认定为难度较高的句子。我们的模型只依靠句子的语言特征区分难易，不能考虑文体和语境依赖程度的影响，所以在这些句子上模型的判断与人工标注的结果产生了偏差。这也说明，语境和文体是影响篇章中句子理解难度的重要因素。

5.2 任务二

表 6 句对相对难度评估的实验结果

模型	SVM	LogR
tf-idf	36.45%	36.94%
汉字	**67.95%**	**66.87%**
词汇	66.27%	66.43%
句法	64.69%	65.67%
所有	67.08%	66.39%

我们把词袋模型作为基线模型，表 6 对比了基线模型与加入不同语言特征的模型在预测句对相对难度任务上的实验结果。SVM 与 LogR 的对比显示，SVM 的预测准确率略高于 LogR。词袋模型的准确率只能达到 36%左右，基于语言特征的模型比词袋模型准确率高 30%左右，说明语言特征可以提升句对相对难度预测模型的性能。从整体上看，基于汉字特征的模型准确率最高，分别为 67.95%和 66.87%，基于句法特征的模型准确率最低，分别为 64.69%和 65.67%，比最高的基于汉字特征的模型降低了 2%左右。说明汉字特征在句对相对难度评估中的预测作用最强。

图 5 显示了基于所有特征的支持向量机模型的混淆矩阵。从表中可以看出，标签 1 和标签-1 之间的混淆最小，模型在原标签为 0 的实例上没有达到较好的分类结果。标签为 0 的实例是难度级别一致的句子，我们的难度级别只分为 5 个级别，但语言难度是一个连续体，即使在一个级别内部，句子与句子之间也有难度的差距。在数据集构建过程中，我们把两句话的难度相等定义为两个句子的难度级别相等，这种做法忽略了级别内部的句子难度差异。句对相对难度的预测任务

实际上是在学习句子之间的难度关系，也说明相比于五分类任务，句对相对难度评估任务可以关注到更小的难度差距。

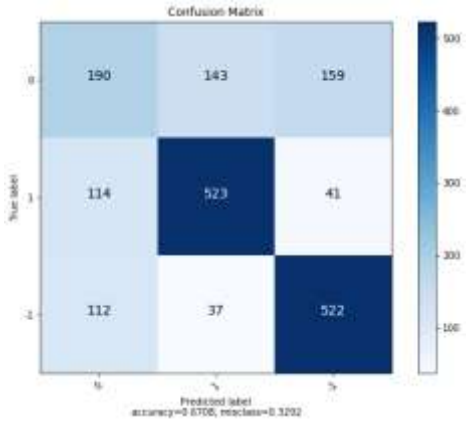

图 5 句对相对难度预测的混淆矩阵

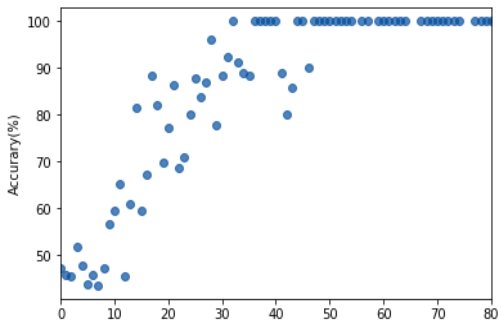

图 6 模型准确率在不同长度差句对上的分布

由于标签为 0 的句对，其平均长度差小于标签为 1 和标签为-1 的句对，这可能是导致标签为 0 句对的相对难度难以预测的原因。因此，我们绘制了模型准确率在不同长度差的句对上的分布，见图 6。从图 6 可以看出，在两个句子的长度差小于 30 时，句子长度差与模型的准确率成正比关系，长度差越大，模型预测句对相对难易度的准确率越高。当句对的长度差大于 50 字时，模型可以达到 100%的准确率。这说明当句对中两个句子的长度差大于一定的阈值时，句子的长度差可以准确预测两个句子的相对难度。在成对比较的标注过程中，我们同样发现了句子长度差对句对相对难易度的影响。当呈现两个句子，标注者被要求判断两个句子的相对难易度时，句子长度是标注者考虑的首要因素，只有在句子长度相近或者字义和词义的理解难度过大，标注者才会考虑从其他因素评估句子难度。

6 总结

本研究提出基于成对比较的众包标注方法来标注大规模句子的难度级别。并使用该方法构建了基于语文教材的汉语句子难度语料库，该语料库中包含 18,411 个被标注为 5 个难度级别的句子。基于该语料库，本研究探讨了有监督的机器学习方法在单句绝对难度评估和句对相对难度评估两项句子难易度评估任务上的表现。为了提升模型的性能，本研究量化并提取了汉字、词汇和句法层面的句子特征，并对比了这些特征对汉语句子难易度评估的作用。

实验结果显示，机器学习模型可以有效预测汉语句子的难度级别，加入语言特征可以提升模型的预测准确率，尤其是，相比于词汇和句法特征，基于汉字层面特征模型的预测准确率最高，说明汉字特征对句子难易度的预测作用最强。实验结果还显示，在单句绝对难度评估中，句子的语境依赖程度和表达方式影响句子的理解难度，在句对相对难度评估中，句对中两个句子的长度差影响模型的预测性能。

未来的研究会考虑扩大句子语料的规模，以期实现更复杂的模型。同时，本研究仅使用语言水平较高的大学生作为标注人员，未来的标注会面向年龄跨度更大，教育背景更丰富的广泛群体。

参考文献

[1] 吴思远，蔡建永，于东，等. 文本可读性的自动分析研究综述[J]. 中文信息学报，2018，32(12)：1-25.

[2] 王蕾. 可读性公式的内涵及研究范式——兼议对外汉语可读性公式的研究任务[J]. 语言教学与研究，2008(6)：46-53.

[3] Collins-Thompson，Kevyn. Computational assessment of text readability：A survey of current and future research[J]. ITL - International Journal of Applied Linguistics，2014，165(2)：97-135.

[4] Pilán I，Vajjala S，Volodina E. A readable read：Automatic assessment of language learning materials based on linguistic complexity[J]. 2016，arXiv preprint arXiv：1603.08868.

[5] Pilán I，Volodina E，Johansson R. Rule-based and machine learning approaches for second language sentence-level readability[C]//Proceedings of the ninth workshop on innovative use of NLP for building educational applications. 2014：174-184.

[6] Leal S E，Duran M S，Aluísio S M. A Nontrivial

Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 401-413.
[7] Vajjala S, Meurers D. Readability-based sentence ranking for evaluating text simplification[J]. 2016, arXiv preprint arXiv: 1603.06009.
[8] Sung Y T, Chen J L, Cha J H, et al. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning[J]. Behavior research methods, 2015, 47(2): 340-354.
[9] 孙刚. 基于线性回归的中文文本可读性预测方法研究[D]. 南京大学 硕士学位论文, 2015.
[10] 蒋智威. 面向可读性评估的文本表示技术研究[D]. 南京大学 博士学位论文, 2018.
[11] 江少敏. 句子难度度量研究[D]. 厦门大学 硕士学位论文, 2009.
[12] 郭望皓. 基于 CRITIC 加权赋值的汉语句子难度测定[J]. 语文学刊(教育版), 2016(12): 10-12.
[13] 庞成. 汉语句子难易度影响因素分析[J]. 语文学刊(教育版), 2016(1): 18-19.
[14] Kincaid J P, Fishburn R P, Chisson B S. Derivation of new readability formulas for navy enlisted personnel[J]. Adult Basic Education, 1975: 49.
[15] Laughlin G H M. SMOG Grading-a New Readability Formula[J]. Journal of Reading, 1969, 12(8): 639-646.
[16] Luo S, Callan J. A statistical model for scientific readability[C]//Tenth International Conference on Information and Knowledge Management. ACM, 2001: 574-576.
[17] Tanaka-Ishii K, Tezuka S, Terada H. Sorting texts by readability[J]. Computational Linguistics, 2010, 36(2):203-227.
[18] Kate R J, Luo X, Patwardhan S, et al. Learning to Predict Readability using Diverse Linguistic Features[C]//Coling 2010 - 23rd International Conference on Computational Linguistics , Proceedings of the Conference. COLING, 2010: 546-554.
[19] Attali Y, Burstein J. Automated essay scoring with e-rater® V. 2[J]. The Journal of Technology, Learning and Assessment, 2006, 4(3): 3-29.
[20] Feng L, Huenerfauth M. Cognitively motivated features for readability assessment[C]// Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 229-237.
[21] Karpov N, Baranova J, Vitugin F. Single-sentence readability prediction in Russian[C]// International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2014: 91-100.
[22] Dell'Orletta F, Montemagni S, Venturi G. Read-it: Assessing readability of italian texts with a view to text simplification[C]//Proceedings of the second workshop on speech and language processing for assistive technologies. Association for Computational Linguistics, 2011: 73-83.
[23] Brunato, D., De Mattei, L., Dell' orletta, F., Iavarone, B., & Venturi, G. (2018). Is this Sentence Difficult? Do you Agree? [C]// Conference on Empirical Methods in Natural Language Processing. 2018: 2690-2699.
[24] Inui K, Yamamoto S, Inui H. Corpus-Based Acquisition of Sentence Readability Ranking Models for Deaf People[C]// NLPRS. 2001: 159-166.
[25] Vajjala S, Meurers D. Assessing the relative reading level of sentence pairs for text simplification[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014: 288-297.
[26] Schumacher E, Eskenazi M, Frishkoff G, et al. Predicting the Relative Difficulty of Single Sentences With and Without Surrounding Context[C]//Conference on Empirical Methods in Natural Language Processing. 2016: 1871-1881.
[27] Pitler E, Nenkova A. Revisiting readability: a unifed framework for predicting text quality[C]//Conference on Empirical Methods in Natural Language Processing. 2008: 186 - 195.
[28] Klare G R. Readability[J]. Handbook of reading research, 1984, 1: 681-744.
[29] 吴思远, 于东, 江新. 汉语文本可读性特征体系构建及其效度验证.
[30] 沈烈敏, 朱晓平. 汉字识别中笔画数与字频效应的研究[J]. 心理科学, 1994, 4: 245-247.
[31] 吴建国, 俞庆英, 吴海辉. 汉字笔画若干数据的统计方法研究与应用[J]. 安徽大學學報 (自然科學版), 2005, 29(3): 14-20.
[32] Liu, Haitao. Probability distribution of dependency distance[J]. Glottometrics, 2007, 15: 1 - 12.
[33] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 13-16.
[34] Levy R, Manning C. Is it harder to parse Chinese, or the Chinese Treebank?[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 439-446.

于东(1982—),通讯作者,博士,副教授,主要研究领域为自然语言处理。

E-mail: yudong_blcu@126.com

吴思远（1998—）,硕士研究生,主要研究领域为第二语言习得,自然语言处理。

E-mail: wusiyuan2401@163.com

耿朝阳(1996-),硕士研究生，主要研究领域为自然语言处理。

E-mail:yangican@163.com