结合深度学习和语言难度特征的句子可读性计算方法

唐玉玲
北京语言大学信息科学学院
blcutyling@163.com

于东*
北京语言大学信息科学学院
yudong_blcu@126.com

摘要

本文提出了可读性语料库构建的改进方法，基于该方法，构建了规模更大的汉语句子可读性语料库。该语料库在句子绝对难度评估任务上的准确率达到0.7869，相对前人工作提升了0.15以上，证明了改进方法的有效性。将深度学习方法应用于汉语可读性评估，探究了不同深度学习方法自动捕获难度特征的能力，并进一步探究了向深度学习特征中融入不同层面的语言难度特征对模型整体性能的影响。实验结果显示，不同深度学习模型的难度特征捕获能力不尽相同，语言难度特征可以不同程度地提高深度学习模型的难度表征能力。

关键词： 深度学习 ； 语言难度特征 ； 句子可读性

The method of calculating sentence readability combined with deep learning and language difficulty characteristics

Yuling Tang
College of Information Science
Beijing Language and Culture
University, Beijing 100083, China
blcutyling@163.com

Dong Yu *
College of Information Science
Beijing Language and Culture
University, Beijing 100083, China
yudong_blcu@126.com

Abstract

In this paper, an improved method for the construction of readable corpus is proposed, and a larger Chinese sentence readability corpus is constructed based on this method. The accuracy rate of the corpus in the task of evaluating the absolute difficulty of sentences can reach 0.7869, which is 0.15 higher than the previous work, which proves the effectiveness of the improvement method. Applying the deep learning method to the evaluation of the readability of Chinese language, the ability of different deep learning methods to automatically capture difficulty characteristics was explored, and the influence of incorporating different levels of language difficulty characteristics into the deep learning features on the overall performance of the model was further explored. The experimental results show that the difficulty features of different deep learning models are different, and language difficulty characteristics can improve the difficulty characterization ability of deep learning models to varying degrees.

Keywords: deep learning , language difficulty characteristics , sentence readability

*为通讯作者

第十九届中国计算语言学大会论文集，第732页-第743页，海南，中国，2020年10月30日至11月1日。

1 引言

作为衡量阅读难度的标准之一，文本可读性对于阅读教学、教材编排有重要意义。可读性体现了给定文本与读者理解文本的认知负荷的关系。这种复杂的关系受到很多因素的影响，如词汇与句法复杂程度、语境和背景知识(Crossley et al, 2017)。传统的可读性研究通过量化不同层面、不同维度的语言特征，如句子长度和单词难度(Davison and Kantor, 1982)，构建多元线性回归公式来评估文本的阅读难度。这些方法因其薄弱的统计基础而受到诟病(Crossley et al, 2017)。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(Luo and Callan, 2001; Tanaka et al, 2010; Kate et al, 2010)。有监督的机器学习方法是现行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难度(Luo and Callan, 2001)，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(Collins and Kevyn, 2014; Sung et al, 2015)。从20世纪20年代以来，各个语言的研究者根据自身语言的特点，构建线性或者非线性的模型进行自动评估(Collins and Kevyn, 2014; Wu Siyuan et al, 2020)。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(Feng and Huenerfauth, 2009)。但有效特征的预测能力与语言特点有关(Feng and Huenerfauth, 2009; Karpov et al, 2014)。这些研究中预测能力强的语言特征是否适用于汉语，已在于东等(2020)的工作中得到验证。

到目前为止，深度学习方法(Goodfellow et al, 2016)在很多自然语言处理任务中都有很好的表现，尤其是与语义相关的任务(Collobert et al, 2011; Zhang,Zhao and LeCun, 2015)，但是只有很少的学者将深度学习方法用于可读性研究，Matrinc等(2019)在几大公开文本级可读性数据集如WeeBit(Vajjala and Meurers, 2012)，OneStopEnglish(Vajjala and Lucic, 2018)，Newsela(Xu, Callison-Burch, and Napoles, 2015)以及Slovenian SB(Matrinc et al, 2019)上分别用HAN(Yang et al, 2016)，BiLSTM(Zhou et al, 2016)和Bert(Devlin et al, 2018)模型进行了有监督的可读性自动评估研究。这项研究是使用现有的深度学习方法在可读性问题上的初尝试，探究了不同深度学习模型在不同数据集上的表现。深度学习模型自动学习到的特征在多大程度上表征难度没有得到验证，这种自动学习获取的特征与人工抽取的语言难度特征的差别体现在何处？现阶段还没有工作使用深度学习方法对汉语可读性问题进行研究，本研究主要是在汉语可读性问题上，结合深度学习方法与外部语言难度特征，探究深度学习方法自动学习获取的特征的表征能力以及与外部语言难度特征表征能力的是否互补的问题。

本文首先参考于东等(2020)基于五点量表和锚点对比构建可读性语料库的方法，提出改进思路，构建了新的句子可读性语料库。基于于东等(2020)构建的语料库（下称set1）和本次构建的语料库（下称set2），探究了机器学习方法在句子绝对难度评估任务上的表现，本次工作使用的语言特征为吴思远等(2020)构造的汉语可读性语言特征体系，包含汉字层面、词汇层面、句法层面。实验结果表明，通过固定标注人员进行标注构造的语料库set2能达到0.7869的准确率。同时，探究了深度学习方法在句子绝对难度评估任务上的表现。实验结果表明，深度学习方法通过自动学习提取特征，能达到比机器学习方法略胜一筹的效果，说明深度学习方法获得的特征可以很好地表征难度。本文试图通过向深度学习特征中加入外部语言难度特征来提高模型的难度表征能力。实验结果表明，外部语言难度特征能不同程度地提高深度学习特征向量的难度表征能力。

本研究的主要贡献包含以下三个方面：第一，构建了一个规模更大、噪点更低、质量更高的句子级可读性标注语料库。该语料库包含37247条汉语句子，具有五个难度等级，为汉语可读性研究提供了数据支持。第二，将深度学习方法应用于汉语句子可读性等级评估任务，验证了深度学习方法在汉语句子可读性等级评估任务上的有效性。第三，通过向深度学习特征中融入外部语言难度特征进行实验，结果表明语言难度特征能不同程度地提高模型整体性能(Tovly Deutsch et al, 2020)。

2 相关研究

自动评估可读性的方法试图发现和利用与可读性感知密切相关的因素来达到自动评估的目的。传统的可读性公式试图建立一个简单的人类可理解公式，其与人类认为的可读性程度有良好的相关性，它们考虑各种统计因素，如词长、句长等。这些公式最初是用于英语的可读性

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

研究，后来也被借鉴用于其他语言的相关研究。目前，大多数的文本可读性公式都将句子长度和词数纳入计算，如针对成人的Flesh公式(Kincaid et al, 1975)，便于个人使用的SMOG公式(Laughlin, 1969)，估计文本等级的The Gunning Fog公式(Gunning and Robert, 1952)以及用于评价书本的Dale-Chall公式(Dale and Chall, 1948)等。较新的方法是在人工标注的可读性数据集上训练机器学习模型，通过一系列的语言难度特征用来预测给定的无难度标签的文本的难度。这些方法通常依赖于广泛的特征工程，构造许多人类易于理解的特征。

现行的测量可读性的新方法是将其视为一项分类任务，并构建自动预测模型，根据多种特征属性自动预测文本的可读性得分(Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012; Petersen et al, 2009)。更复杂和适应性更强的方法通常能达到更好的效果，但是需要大量额外的数据资源作为支撑，特征的选择大多依赖专家进行判定，人工抽取特征耗时耗力，所以这些方法在不同语言，不同数据集之间的可迁移性比较差。目前几乎还没有工作涉及到跨语言，多语言，甚至多体裁，多数据库的有监督可读性自动评估。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，其核心是从词汇、句法和篇章等层面分析和筛选可以预测文本难度的有效特征(Collins and Kevyn, 2014; Pilan et al, 2016)。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(Wu Siyuan et al, 2018; Wang Lei, 2008)。

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。Pilan等(2016)从第二语言学习的角度探讨了影响瑞典句子难易度的语言因素。该研究将句子可读性评估抽象为多分类问题，支持向量机分类器在该任务上达到了71%的准确率。Dell'Orletta等(2011)对比了表层特征、词汇特征、形态句法特征与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别的可读性评估，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(2018)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。Schumacher等(2016)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子相对难度的判断，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子对相对难度，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。句子级的可读性研究受到越来越多的关注，于东等(2020)按照任务的不同把句子级可读性评估分为单句绝对难度评估和句子对相对难度评估两项，通过抽取一系列难度特征训练机器学习模型用于句子可读性自动评估。

国内句子难易度自动评估的研究仍处于起步阶段。江少敏(2009)采用调查问卷和对比分析的方法，从汉字，词汇和句法层面收集了被试者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(2016)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(2016)对汉字层面和词汇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。于东等(2020)等通过机器学习的方法进行了语文教材句子的难易度评估工作，也对语言特征的预测作用进行了系统的考察。深度学习方法在英语文本可读性上的应用研究使可读性研究有了进一步的突破(Matrinc et al, 2019)，然而深度学习方法在汉语可读性上的研究工作甚少，本次工作希望句子级可读性问题在深度学习方法上取得突破，并探究融入语言特征的深度学习模型是否具有更好的性能。

3 数据集构建

于东等(2020)的工作中已经构建了一个包含18411条句子的开放的可读性数据集set1，难度标签为5个等级。这个数据集的优点是数据集中的数据来源于权威的语文教材，五点量表的标注方法和锚点对比标注流程的科学性也是不容置疑的。不足之处在于：第一，构建这个数据集的锚点集的数据量太少，各等级占比不均衡；第二，虽然采用众包标注可以节省成本，但是众包标注也意味着难度衡量标准的稳定性更差，噪点更多；第三，标注数据集（标3次）与锚点集（标5次）的标注次数不一致，很大程度上会影响句子的最终标签，偏差更大。针对这些问题，本文提出了相对应的数据集构建改进方法，首先扩充锚点集，然后采取固定标注人员的方式进行标注，每条数据标注5次。基于以上改进方法，我们重新构建了一个句子级可读性数据集set2，我们的句子数据集也是来源于具有权威性的北师大版、人教版和苏教版的汉语语文教材。我们在处理数据的过程中去掉了使用特殊体裁的文本和不完整的文本，如诗歌、词赋、识字文本等，经过句子去重，最后得到的句子数据集包含40192个句子，句子的平均长度为29。

3.1 基于专家标注的锚点句扩充

我们采用锚点比较法进行数据标注，在正式标注之前，首先要构建锚点数据集，我们在于东等(2020)工作的锚点数据集的基础上进行了扩充。首先在原始数据集中选取500条没有进行任何标注的句子集，邀请5名小学语文教师认真阅读句子，并根据五点量表对句子进行等级评定，1表示非常简单，5表示非常难。完成500个句子的难度评定工作大约需要一个小时。最终收集到每个句子被标注5次的数据，5位专家之间的肯德尔一致性系数为0.723（p <0.001）,说明5位专家的标注一致性较高。

对于每一个句子，我们采用多数投票原则确定锚点句的难度等级。为了保证作为锚点句的句子难易程度一致，我们计算了每个句子被标注为最终难度的概率。如句子A被标注了5次，其中三位专家标注A的难度为等级3，一位专家标注A的难度为等级1，一位专家标注A的难度为等级4，那么该句子A最终难度为等级1的概率为0%，等级2的概率为20%，等级3的概率为60%，等级4的概率为20%，等级5的概率为0%。我们选取概率大于或者等于80%的难度等级作为该句子的最终等级，并确定该句子为锚点句。

经过概率筛选后，我们确定了205个句子的最终难度等级，除去难度为5的10个句子，剩余的195个句子为最终的锚点句，其中，等级一的锚点句数量为60句，等级二的锚点句数量为48句，等级三的锚点句数量为75句，等级四的锚点句数量为12句。为了保证四组锚点句之间在难度上具有较高的差异性，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=580，p <0.01）。更多统计信息如表1所示。

锚点等级	于东(2020)	比例	本文锚点集	比例	相对差值
等级一	33	53%	60	31%	+23
等级二	16	26%	48	25%	+32
等级三	10	16%	75	38%	+65
等级四	3	5%	12	6%	+9

表 1: 锚点句对比详情统计

3.2 基于锚点比较的数据标注

3.2.1 标注流程

我们共招募了20名标注员对数据进行标注，标准规则为与锚点句成对比较，每个句子将在2-3次比较后被划分到最终难度等级。我们收集了标注员的年龄，性别，教育程度等个人信息，标注者年龄在19至27岁之间，学历为本科到博士，男女比例为1:5。在正式标注之前，对标注人员进行简单培训，明确标注任务和规则，然后客观负责地完成任务。我们每天定时在微信标注小程序上发布标注任务，并定期抽查，以监控标注质量。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 的某个句子再次进行配对。重复这个过程直至确定该句子的难度级别。每个句子由至少五位标注员进行标注，即每个句子至少被标注五次。我们的标注周期为4周，每周会对标注员的工作进行检查。

3.2.2 数据集构建

标注周期结束后，我们收集了40192条数据，每条数据都被标注了5次，删除了标注时间小于15 秒（1%）的句子。我们使用多数投票原则决定单个句子的难度级别，3名以上标注员（包含3名）意见一致则确定最终难度标签。最终我们构建了一个基于语文教材的句子难度语料库。该语料库共包含37427 个汉语句子，每个句子被标注为1 至5 的某个难度级别，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(2009)，值越大则难度越高。

3.3 数据集比较

我们将于东等人(2020)构造的set1与之进行对比。set1来源为汉语语文教材，基于五点量

难度等级	例句
等级一	大家都觉得很不方便。
等级二	一只塘鹅闭紧嘴巴，她急急地走在小路上。
等级三	这里“你是”含有假定语气，也带“你不是”一点讥刺的意味。
等级四	克莱谛不时用眼睛瞟我，从他的眼里表示出来的不是愤怒，而是悲哀。
等级五	根据英国南部物候的一种长期记录，拿1741到1750年10年平均的春初7种乔木油青和开花日期相比较，可以看出后者比前者早9天。

表 2: 句子难度标注语料库示例

难度等级	数量	比例	平均句长	Jiang(2009)
等级一	3158	0.08	8.08	112.84
等级二	9235	0.25	16.57	220.65
等级三	14627	0.39	28.02	353.37
等级四	7371	0.20	42.83	530.36
等级五	2856	0.08	65.40	790.42
平均值	–	–	29.29	272.128

表 3: 标注数据集详情统计

表，1表示很简单，5表示很难，通过专家标注获得锚点句，然后通过众包的方式进行大规模标注，标注的过程中采用目标句与锚点句对比的方式进行。每个句子被标注三次，通过投票原则确定句子的难度等级。最后经过数据处理得到的语料库包含18411个汉语句子。set1与set2是两个既有相同之处也存在差异的句子可读性数据集。其中的相同点包括：句子数据均来源于权威的苏教版，北师大版和人教版的汉语语文教材，在数据标注的过程中首先基于五点量表，通过专家标注获得锚点数据集，然后通过目标句与锚点句的对比来得到目标句的难度等级。标注过程中的数据处理方式为投票原则，评判标准为肯德尔系数和方差分析。

不同点之一则在于set1是通过众包的方式进行标注，而set2是通过招募固定的优秀标注员进行标注，每个人由于受教育水平和文化背景的差异，对于难度的评判标准是不一致的，那么众包标注就意味着在数据集构建过程中的评判标准差异性更大，在数据处理的时候会引入更多的噪声，从而降低数据集的质量，那么对比而言，固定的标注员会使整个数据集的评判标注趋于统一，会提高数据集的质量；不同点之二在于构建set2的锚点句的数量为195，而构建set1的锚点数据集包含62条锚点句，其中锚点一的数据量为33，锚点二的数据量为16，锚点三的数据量为10，锚点四的数据量为3。可以发现锚点句的总量相对较少，且各锚点句数量的比例相差较大，不同等级的句子在字数，句式和结构等方面都存在很大差异，若可对比的锚点句的数量过少，则可作为评判指导的依据就少，这对于整个数据集的质量会产生一定的负面影响；不同点之三在于set1中的数据内容与set2中的数据完全不重合，set1中每条数据被标注3次，set2中每条数据被标注5次，标注次数越多则产生偏差的概率越低，数据质量越高。在之后的实验中，我们分别基于这两个数据集进行实验。

4 特征及模型

4.1 特征选择

可读性特征体系的设计参考了吴思远等(2020)的特征框架，该研究把评估文本可读性的指标划分为四个层面，分别是汉字、词汇、句法和篇章结构。于东等(2020)从汉字、词汇和句法三个层面实现句子语言特征的量化，达到了较好的分类结果。

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是从字形复杂度、汉字熟悉度和汉字多样性三个角度进行量化，共22个指标，如汉字笔画数、字频等。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，共25个指标，如词频、词长等。句法结构层面共包括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。

在深度学习特征向量融合外部语言难度特征的实验部分，本文采用的外部语言难度特征即为汉字、词汇和句法层面的特征以及三个层面的组合特征。深度学习特征向量的抽取则根据模型的不同而不同，对于双向循环神经网络，则抽取模型最后一层的第一个神经元和最后一个神经元输出的特征向量组合。对于卷积神经网络，则抽取最后一个卷积层经过不同卷积核卷积之后输出的特征向量的组合，对于基于transformer的神经网络模型Bert，则使用肖涵博士开发的bert-as-service默认抽取倒数第二个transformer层的输出向量。

4.2 模型介绍

正如前文中提到，近年来文本分类任务的趋势表明，采用自动特征构建的深度学习方法占主导地位。Matrinc等(2019)在首次将深度学习方法用于英文可读性研究，为了确定不同模型在可读性研究中的性能和不足，评估了大量模型的复杂性。在此之前的可读性研究依赖人工构造特征和机器学习分类器(Vajjala and Lucic, 2018; Xia Kochmar Briscoe, 2016)。在汉语可读性研究中，即使是最新的中文可读性分类方法也依赖于人工构造的特征和传统的机器学习分类器(Wu Siyuan et al, 2020; Yu Dong et al, 2020)。在这一部分中，我们将重点介绍三大特征提取器RNN，CNN和Transformer，以及语言特征融入实验的框架流程，如图1所示。

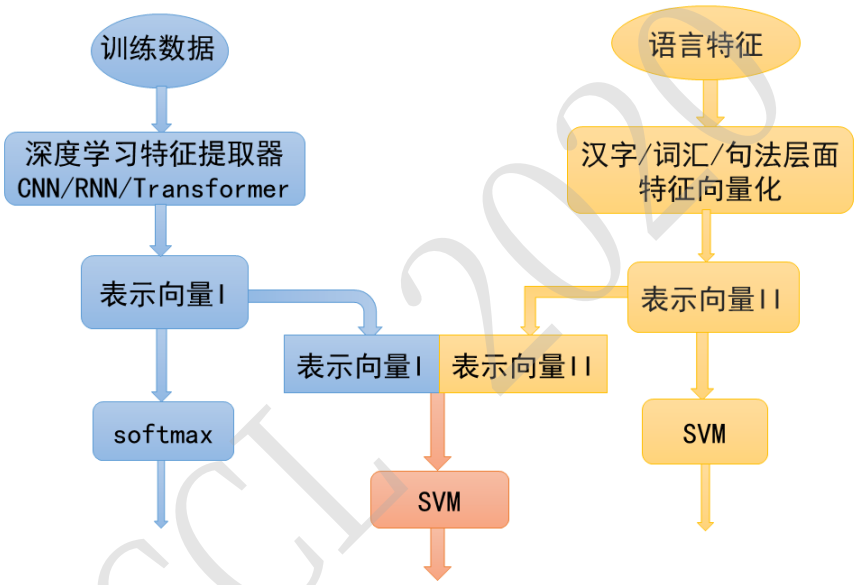

图 1: 语言特征融入流程图

- RNN：在一些自然语言处理任务中，当对序列进行处理时，我们一般会采用循环神经网络RNN，尤其是它的一些变种，如BiLSTM(Zhou et al, 2016)，GRU(Cho et al, 2014)等。循环神经网络善于捕捉更长的序列信息。LSTM因在文本分类任务上的效果非常好而受到重视。LSTM在每个时间步上的输入有两部分信息，一部分是前一个时间步的保留信息，一部分是当前时间步对应的原始信息，由此，LSTM可以在最后一个时间步获取到整个序列的信息，并且丢弃掉模型认为没有用的信息。句子作为序列，其长度是相对篇章来说较短的长度，所以LSTM模型可以很好地胜任句子的特征抽取任务。在实验中，我们采用基于双向LSTM的textRNN模型(Liu et al, 2016)，在该模型上，则抽取模型最后一层的第一个神经元和最后一个神经元输出的特征向量组合。

- CNN：卷积神经网络因其在句子分类任务上的突出表现而被选中(Kim, 2014)。本文使用的CNN模型基于Kim(2014)描述的textCNN模型。将卷积神经网络CNN应用到文本分类任务，利用多个不同尺寸的卷积核进行一维卷积来提取句子中的关键信息，每次能处理不同尺寸长度个词的完整的词向量，从上往下依次滑动卷积，这个过程中的输出就成了我们需要的特征向量，类似于多窗口大小的ngram，可以能够更好地捕捉局部相关性。CNN的并行计算能力非常强，可以快速实现特征提取。在textCNN模型上，我们抽取最后一个卷积池化层后输出的特征向量组合。

- Transformer：Transformer(Ashish Vaswani et al, 2017)是使用自注意力的Encoder-Decoder模型，其在包括可读性评估在内的众多自然语言处理任务上取得了最新的结果(Matrinc et al, 2019)。Transformer利用注意力机制，使得模型在构造输出向量时能够注意到输入的特定部分。尽管它们被表示为序列到序列模型，但是可以通过在网络的末端放置一个额外的线性层并训练该层以产生所需的输出来修改它们以完成各种NLP任务。这种方法在与预训练模型相结合时通常会获得最好的结果。在本文中，我们使用了基于Transformer的Bert中文模型(Devlin et al, 2018)，该模型是在图书语料库（800M words）(Zhu et al, 2015)和中文维基百科上预先训练的，然后在特定的可读性语料库上对模型进行微调。预训练的Bert模型来源于Huggingface的Transformer库(Thomas Wolf et al, 2019)，由12个隐藏层组成，每个隐藏层的大小为768和12个自注意力头。Transformer模型突破了RNN 模型不能并行计算的限制。相比CNN模型，计算两个位置之间的关联所需的操作次数不随距离增长。自注意力可以产生更具可解释性的模型。我们可以从模型中检查注意力分布。各个注意力头可以学会执行不同的任务。在语言特征融入的实验中，本文使用肖涵博士开发的bert-as-service[0]默认抽取倒数第二个transformer层的输出向量。

5 实验设计与结果分析

5.1 在机器学习方法上验证set2改进思路的有效性

为了验证本文提出的数据集构建改进方法的有效性，我们在set1与set2上对比了支持向量机（Support Vector Machine, SVM）和逻辑回归（Logistic Regression, LogR）两种模型的表现，我们以于东等(2020)基于tf-idf的词袋向量作为输入构建的模型作为基线模型，然后分别把汉字，词汇，句法层面的语言特征以及三个层面的组合特征作为句子的向量表示，构建特征模型。在实验过程中训练集与测试集的比例为8:2，采用五折交叉验证，评价指标为准确率。我们使用Python语言，在scikit-learn库(Pedregosa et al, 2011)中实现了模型。

Models	set1		set2		set1+set2	
	SVM	LogR	SVM	LogR	SVM	LogR
汉字	**0.6303**	**0.6308**	0.7779	**0.7740**	0.6928	**0.6901**
词汇	**0.6303**	0.6212	**0.7819**	0.7544	**0.6947**	0.6777
句法	0.6242	0.6242	0.7662	0.7461	0.6779	0.6552
all	0.6179	0.6301	0.7817	0.7597	0.6911	0.6895
tf-idf	0.4720	0.4549	0.5213	0.4740	0.4935	0.4772

表 4: 两个数据集在机器学习上的表现

我们对比分析了SVM和LogR两种不同分类模型在set1与set2上的表现，实验结果如表4所示.相对于基线模型，各个层面的语言难度特征都表现出了较好的效果，验证了语言难度特征的使用可以提升模型的效果，基于单一层面的语言特征的模型甚至比基于字词句组合特征的模型的效果更好，说明特征的使用并不是越多越好，汉字层面和词汇层面的特征的效度要比单一句法层面和三个层面的组合特征的效度更好(Yu Dong et al, 2020)。同时验证了机器学习方法在set2上的有效性。可以看出set1的整体效果在0.6179到0.6308之间，set2的整体效果在0.7461到0.7819之间，set1的最优结果是在LogR上以汉字层面特征作为特征向量的模型，准确率达到0.6308，set2的最优结果是在SVM上以词汇层面特征作为输入的模型，准确率为0.7819。set2的整体效果高出set1约0.15，说明set2数据集更加优质，噪点更低，证明了本文提出的数据集改进方法的有效性。set2在可读性评估上的效果更好，以set1作为可读性评估数据集具有更高的挑战性。

为了进一步对比两个数据集，以set1作为训练集，set2作为测试集进行实验，混淆矩阵的结果显示set2会更多地被分到比原等级更低的等级，如图2 所示。以set2作为训练集，set1作为测试集，混淆矩阵的结果显示set1会更多地被分到比原等级更高的等级，如图3所示。由此可以得出，set1与set2两个数据集的难度中心点不一致，set1的难度中心点更高，set2的难度中心点较set1偏低，set1的整体难度比set2高。这大约是因为固定的优秀标注员都有着相对较高

[0]https://github.com/hanxiao/bert-as-service

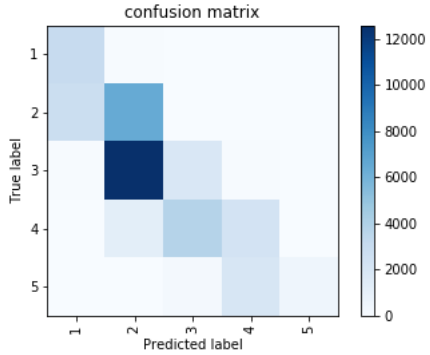

图 2: train set1,test set2.

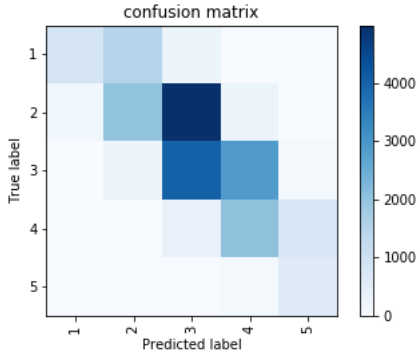

图 3: train set2,test set1.

的语言水平，对难度的感知相对较低，得到的句子难度等级更低，而众包标注过程中，参与标注的人语言水平门槛较低，使得参与标注的所有人的语言水平差异较大，低语言水平的标注员对难度的感知会更高，得到的句子难度等级更高，所以会使得set1的整体难度会比set2的高。合并set1与set2，随机混合数据，以8:2的比例切分训练集与测试集，以各个层面的特征作为输入。结果如表4所示，最优结果为0.6947。对比以上set1最高准确率0.6308，set2最高准确率0.7819，合并数据集得到的是一个折中的实验效果。通常来说，合并数据集之后，数据量更大，会得到更高的结果，但是这里合并数据集之后，得到的结果是以set1与set2分别进行实验的结果的折中效果，并且两个数据集的难度中心点不一致，所以本文认为，set1与set2是两个独立的数据集，合并set1与set2会得到一个更具挑战性的可读性数据集。

5.2 深度学习模型在汉语可读性评估上的有效性验证

深度学习方法在汉语可读性研究中的应用是可读性研究的更进一步，通过深度学习模型征提取器来自动学习获取特征,可以有效避免大规模的人工特征抽取。对于基于卷积神经网络和基于循环神经网络的模型，利用预训练的百度百科静态词向量sgns.baidubaike.bigram-char(Shen Li et al, 2018)，词嵌入维度为128，对于基于transformer的模型则在Bert上进行分类任务的微调。训练集，验证集，测试集的占比为8:1:1，以两个数据集在机器学习模型上的最优结果作为基线，评价指标为准确率和F1值。整个模型基于pytorch深度学习框架(Paszke et al, 2019)。

Models—Datasets	set1		set2	
	ACC	F1	ACC	F1
SVM	0.6303	0.6315	0.7819	0.7784
LogR	0.6308	0.6298	0.7740	0.7756
TextCNN	0.6020	0.6018	0.7199	0.7187
TextRNN	0.6071	0.6051	0.7293	0.7304
TextCNN(Bert词向量)	0.6324	0.6313	0.7752	0.7745
TextCNN(Bert字向量)	**0.6357**	**0.6345**	0.7773	0.7782
TextRNN(Bert词向量)	0.6326	0.6327	0.7830	0.7828
TextRNN(Bert字向量)	0.6345	0.6336	0.7850	0.7851
Bert	0.6209	0.6316	**0.7869**	**0.7943**

表 5: 两个数据集在深度学习上的表现

我们以SVM和LogR的实验结果作为基线，对比分析了TextCNN模型，TextRNN模型以及基于Transformer的Bert模型在句子级汉语可读性评估任务上的表现，实验结果如表5所示。实验结果证明了深度学习方法在汉语可读性评估任务上的有效性。从以上结果可以看出，以百度百科词向量作为辅助的TextCNN和TextRNN模型的效果不及机器学习模型的效果，这在set2上表现得更加突出。以Bert-as-service生成的的预训练Bert字向量和Bert词向量作为TextCNN与TextRNN的输入的模型，则在之前实验的基础上表现出了显著的提升，整体

效果比机器学习模型的效果更好，说明普通的静态词向量对这两个模型的难度特征的捕获没有起到很好的辅助作用。以普通静态词向量作为输入的TextCNN和TextRNN的难度表征能力不如人工抽取的语言特征的难度表征能力强。同时对比以Bert向量作为输入的实验结果，可以发现以Bert字特征向量为输入的模型具有更优的性能，说明相比Bert词向量，Bert字向量在该数据集上能更大程度上地表征难度信息。同时，在Bert模型上进行微调的实验结果在set1和set2上的准确率分别为0.6209和0.7869，在set2上达到了最好的结果。说明Bert预训练语言模型自动捕获的特征向量在很大程度上代表了难度信息(Tovly Deutsch et al, 2020)。对比TextRNN和TextCNN的所有实验结果可以发现TextRNN的效果总是比TextCNN模型的效果更好，说明在汉语句子级可读性评估任务上TextRNN的难度表征能力比TextCNN更强。

5.3 探究语言特征能否提升深度学习模型的整体性能

将人工抽取的语言特征应用于机器学习模型，在可读性难度判别任务上达到了不错的效果。为了探究人工抽取的语言难度特征能否提升深度学习模型的整体性能，我们以深度学习模型作为特征提取器，向深度学习特征向量中融入不同层面的语言难度特征进行实验，以期这种组合的特征向量可以更好地表征难度。本实验中的特征提取器分别是TextCNN模型、TextRNN模型和基于Transformer的Bert预训练语言模型，TextCNN和TextRNN不使用中文静态词向量。语言难度特征采用吴思远(2020)的可读性特征体系，在实验中分别加入各个层面的语言难度特征进行实验。

Models—Datasets	set1		set2	
	ACC	F1	ACC	F1
TextCNN	0.6021	0.6018	0.7087	0.7027
TextCNN+汉字层面	**0.6135**	0.6119	**0.7592**	**0.7564**
TextCNN+词汇层面	0.6134	**0.6148**	0.7541	0.7542
TextCNN+句法层面	0.6128	0.6114	0.7347	0.7282
TextCNN+三个层面	0.6133	0.6123	0.7501	0.7479
TextRNN	0.6027	0.6013	0.7183	0.7084
TextRNN+汉字层面	0.6142	0.6147	**0.7643**	**0.7622**
TextRNN+词汇层面	**0.6168**	**0.6202**	0.7579	0.7615
TextRNN+句法层面	0.6086	0.6056	0.7303	0.7337
TextRNN+三个层面	0.6125	0.6144	0.7499	0.7528
Bert	0.6209	0.6226	0.7812	0.7724
Bert+汉字层面	0.6292	0.6240	0.7819	0.7830
Bert+词汇层面	0.6234	0.6212	0.7814	0.7789
Bert+句法层面	0.6341	0.6328	0.7848	0.7851
Bert+三个层面	**0.6390**	**0.6361**	**0.7874**	**0.7919**

表 6: 深度学习融合外部语言特征的表现

以三种深度学习特征提取器抽取的特征向量单独作为输入的模型作为基线，通过向TextCNN，TextRNN和Bert的特征向量中融入不同层面的语言特征，可以发现语言特征能不同程度地提高模型的效果(Tovly Deutsch et al, 2020)，实验结果如表6所示。横向对比，在set1上的提升不显著，在set2上的提升则更加显著。纵向对比，在Bert上的提升不显著，在TextCNN和TextRNN上的提升则更加显著。在TextCNN和TextRNN上，融入汉字层面的特征和融入词汇层面的特征得到的提升更多，而在融入句法特征的模型上提升更少，说明TextCNN和TextRNN模型捕获的难度特征更偏向于类似句法层面的特征，对于整句信息的保留能力更强，则与汉字和词汇层面特征的互补性更强。在Bert的一些列组合特征模型中，在融合句法层面特征的模型和融合所有层面特征的模型上提升更多，说明Bert模型自动捕获的难度特征类型更偏向于汉字和词汇层面的细粒度特征，所以与句法层面的特征的互补性更强。在整个实验中，TextCNN和TextRNN的基线效果比Bert的基线效果相差许多，其融入特征的最优模型尚且没有达到Bert的基线效果，说明Transformer作为特征提取器，其难度特征捕获能力在各个层面都优于CNN和RNN特征提取器。

6 总结

在本文中，我们首先提出了改进语料库构建的方法，基于改进的方法思路，我们构建了一个规模更大、噪点更低、质量更高的句子级可读性语料库，该数语料库包含37247条数据。通过在机器学习模型中加入汉字层面、词汇层面、句法层面以及三个层面的组合语言特征来探究在set2上的表现，并且与set1的结果进行对比。实验结果显示，set2的准确率比set1的准确率高出约0.15，验证了本文中改进方法的有效性，以及该数据集的有效性。说明扩充锚点句对提高数据集质量有直接影响，固定标注人员可以保证难度衡量标准的一致性和稳定性，标注次数越多，难度等级偏差越小。

将深度学习方法应用于汉语可读性评估，验证了深度学习方法在该任务上的有效性，并且得到比机器学习略胜一筹的效果，说明深度学习自动捕获的特征在很大程度上能够代表难度信息。在这两个非平行语义的数据集中，效果最好的模型都与Bert相关，说明非平行语义的可读性评估与语义不可分割。TextRNN的难度特征捕获能力比TextCNN更胜一筹。深度学习模型自动捕获特征的能力可以有效减少人工抽取语言特征的成本，促进可读性研究。我们探讨了语言难度特征的融入能否提升深度学习模型的难度表征能力，深度学习特征与语言特征的互补性关系。实验结果表明，在汉语可读性评估中，语言难度特征可以不同程度地提升深度学习模型的表征能力。TextCNN和TextRNN模型捕获的特征与汉字和词汇层面的语言特征互补性更强，Bert预训练语言模型捕获的特征与句法层面的语言特征互补性更强。

总的来说，深度学习方法在可读性评估上的应用是一条必由之路，语言特征对深度学习模型的难度表征能力的提升不显著，那么我们之后的工作似乎更应该关注到深度学习特征在多大程度上代表了难度信息，深度学习模型如何能捕获到更多区别于语言特征的难度信息。未来，我们的数据集会进行开放，便于更多学者的研究。同时，我们的研究也将在更具挑战性的两个数据集的合并集上进行。

参考文献

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, and Illia Polosukhin. 2017 *Attention is all you need.* In Advances in neural information processing systems 30, pages 5998–6008. Curran Associates, Inc.

Brunato,De Mattei,Dell'orletta,Iavarone,Venturi. 2018 *Is this Sentence Difficult? Do you Agree?* Conference on Empirical Methods in Natural Language Processing. pages：2690-2699.

Cho, K. , Van Merrienboer, B. , Gulcehre, C. , Bahdanau, D. , Bougares, Schwenk, H. 2014 *Learning phrase representations using rnn encoder-decoder for statistical machine translation.* Computer Science.

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

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011 *Natural language processing (almost) from scratch.* Journal of Machine Learning Research, 12(Aug):2493–2537.

Crossley, Scott A, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017 *Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas.* Discourse Processes, 54(5-6):340–359.

Dale, Edgar and Jeanne S Chall. 1948 *A formula for predicting readability: Instructions.* Educational research bulletin,pages 37–54.

Davison, Alice and Robert N Kantor. 1982 *On the failure of readability formulas to define readable texts: A case study from adaptations.* Reading research quarterly,pages 187–209.

Dell'Orletta F,Montemagni S,Venturi G. 2011 *Read-it*：*Assessing readability of italian texts with a view to text simplification.* Proceedings of the second workshop on speech and language processing for assistive technologies. Association for Computational Linguistics,2011：73-83.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee,and Kristina Toutanova. 2018 *Bert:Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher,Matthieu Perrot, and Edouard Duchesnay. 2011 *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12:2825-2830.

Feng L, Huenerfauth M. 2009 *Cognitively motivated features for readability assessment.* Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,2009:229-237.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016 *Deep Learning.* MIT Press.

Gunning, Robert. 1952 *The technique of clear writing.* McGraw-Hill, New York.

Karpov N,Baranova J,Vitugin F. 2014 *Single-sentence readability prediction in Russian.* International Conference on Analysis of Images,Social Networks and Texts. Springer,Cham,2014:91-100.

Kate R J,Luo X,Patwardhan S,et al. 2010 *Learning to Predict Readability using Diverse Linguistic Features.* Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference. 546-556.

Kim Y . 2014 *Convolutional Neural Networks for Sentence Classification.* Eprint Arxiv, 2014.

Kincaid J P， Fishburn R P， Chisson B S. 1975 *Derivation of new readability formulas for navy enlisted personnel.* Adult Basic Education，1975：49.

Laughlin G H M. 1969 *SMOG Grading-a New Readability Formula.* Journal of Reading,12(8)：639-646.

Liu P , Qiu X , Huang X . 2016 *Recurrent Neural Network for Text Classification with Multi-Task Learning.*

Luo S,Callan J. 2001 *A statistical model for scientific readability.* Tenth International Conference on Information and Knowledge Man-agement. ACM，2001：574-576.

Matej Martinc, Senja Pollak, Marko Robnik-Šikonja. 2019 *Supervised and Unsupervised Nerual Approches to text readability.* Computational Linguistic journal.

Paszke A, Gross S , Massa F . 2019 *PyTorch: An Imperative Style, High-Performance Deep Learning Library.*

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi , Bingchen Li, Hongwei Hao, Bo Xu. 2016 *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification* Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,pages:229-237.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao. 2016 *Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification.* Meeting of the Association for Computational Linguistics.

Petersen, Sarah E and Mari Ostendorf. 2009 *A machine learning approach to reading level assessment.* Computer speech and language, 23(1):89–106.

Pilan I，Vajjala S，Volodina E. 2016 *A readable read：Automatic assessment of language learning materials based on linguistic complexity.*, arXiv preprint arXiv：1603.08868.

Schumacher E，Eskenazi M，Frishkoff G，et al. 2016 *Predicting the Relative Difficulty of Single Sentences With and Without Surrounding Context.* Conference on Empirical Methods in Natural Language Processing. pages：1871-1881.

Schwarm, Sarah E and Mari Ostendorf. 2005 *Reading level assessment using support vector machines and statistical language models.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 523–530. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du. 2018 *Analogical Reasoning on Chinese Morphological and Semantic Relations.* Meeting of the Association for Computational Linguistics.

Sung Y T,Chen J L,Cha J H,et al. 2015 *Constructing and validating readability models：the method of integrating multilevel linguistic features with machine learning[J].* Behavior research methods,47(2)：340-354.

Tanaka-Ishii K, Tezuka S, Terada H. 2010 *Sorting texts by readability.* Computational Linguistics, 36(2):203-227.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Morgan Funtowicz, and Jamie Brew. 2019 *HuggingFace's transformers: state-of-the-art natural language processing.* Technical report.

Tovly Deutsch ,Masoud Jasbi, Stuart Shieber. 2020 *Linguistic Features for Readability Assessment* arXiv preprint arXiv:2006.00377.

Vajjala, Sowmya and Detmar Meurers. 2012 *On improving the accuracy of readability classification using insights from second language acquisition.* In Proceedings of the seventh workshop on building educational applications using NLP, pages 163–173. Association for Computational Linguistics.

Vajjala, Sowmya and Ivana Lucic. 2018 *Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification.* In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications,pages 297–304. Association for Computational Linguistics.

Xia, Kochmar, Briscoe. 2016 *Text Readability Assessment for Second Language Learners.* Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015 *Problems in current text simplification research: New data can help.* Transactions of the Association of Computational Linguistics, 3(1):283–297.

Yang, Zichao, Diyi Yang, Chris Dyer,Xiao dong He, Alex Smola, and Eduard Hovy. 2016 *Hierarchical attention networks for document classification.* In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489.

Yoon Kim. 2014 *Convolutional Neural Networks for Sentence Classification.* In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015 *Aligning Books and Movies:Towards Story-like Visual Explanations by Watching Movies and Reading Books.* Computing Research Repository, arXiv:1506.06724.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015 *Character-level convolutional networks for text classification.* In Advances in neural information processing systems, pages 649–657.

郭望皓. 2016 基于*CRITIC* 加权赋值的汉语句子难度测定. 语文学刊(教育版),2016(12)：10-12.

江少敏. 2009 句子难度度量研究. 厦门大学硕士学位论文.

庞成. 2016 汉语句子难易度影响因素分析. 语文学刊(教育版),2016(1)：18-19.

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

吴思远，于东，江新. 2020 汉语文本可读性特征体系构建及其效度验证. 世界汉语教学，2020(1):81-97.

王蕾. 2008 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. 语言教学与研究，2008(6)：46-53.

于东,吴思远,耿朝阳,唐玉玲. 2020 基于众包标注的语文教材句子难易度评估研究. 中文信息学报34(2):16-26.