文章编号：1003-0077（2019）00-0000-00

一个面向中文古诗词理解难易度的人工标注数据集

刘磊[1,2] 何苯[1,2] 孙乐[2]
（1. 中国科学院大学 计算机科学与技术学院，北京 100049；
2. 中国科学院软件研究所 中文信息处理实验室，北京 100190）

摘要：向读者推荐阅读难度合适的古诗词有助于提升读者的诗词鉴赏能力。现阶段，围绕古诗词可读性自动化分析的相关研究的突出局限之一是缺乏大规模高质量的数据集。针对该问题，本文研究面向古诗词可读性自动化分析的数据集构建。我们对外开放包含 1915 篇古诗词的标注阅读理解难度的数据集①。我们首先将数据集划分成易中难三级，构建数据集 APRD；然后进一步细化标注构建六级分类数据集 APRD+。我们抽取教材中的诗词组成标准集，以年级为标准难度级别，计算标准集与 APRD、APRD+之间的 Spearman 相关性分别为 0.786 与 0.804，表明该数据集标记结果与标准集具有较高一致性。本文提取了字频、注释数等古诗词特征，采用 SVM、随机森林等算法进行了初步古诗词阅读理解难易度分类测试。本文提出的古诗词可读性数据集与实验结果可作为后续研究的测试基准。

关键词：中文古诗词；可读性分析

中图分类号：TP391 **文献标识码：**A

An Annotated Dataset for Ancient Chinese Poetry Readability

LIU Lei[1,2], HE Ben[1,2], and SUN Le[2]
(1. School of Computer Science and Technology, University of Chinese Academy of Science, Beijing, 100049,China ;
2. Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, 100190,China)

Abstract：Reading Chinese ancient poems with appropriate difficulty is beneficial to improving reader' literacy appreciation skills. One of the major limitations of the current research on the automatic analysis of ancient Chinese poetry readability is the lack of large-scale high-quality corpus. To bridge the gap, this work provides a collection of 1,915 ancient Chinese poetry with manually annotated readability levels. We provide two kinds of granularities of the readability classification. The initially annotated APRD dataset contains three readability classes, which is further refined into the APRD+ dataset with six readability classes. The Spearman correlation of the readability labels in APRD or APRD with the textbook grades is 0.786 or 0.804, indicating a fair strong correlation of our annotations with the golden standard. The SVM and random forest algorithms based on statistical text features are applied to classify the difficulty levels of poetries. The annotated poetry readability dataset and the obtained results on it can sever as a benchmark for further studies.

Key words: Ancient Chinese poetry; readability analysis

0 引言

古诗词分级阅读是一种按照读者的阅读能力为读者提供难度系数匹配古诗词的阅读计划，在阅读能力培养、个性化诗词推荐等方面具有积极的现实意义。不同古诗词在字词使用、语言风格、写作手法等方面存有差异，使得相应的阅读理解难度不尽相同，因而适合于不同阅读能力水平的人群。阅读难度适当的诗词不仅使得读者易于深

① https://github.com/lailoo/APRD

收稿收稿日期： ；**定稿日期**：
基金项目：基金名（基金号）；基金名(基金号)

刻理解诗词主旨大意，有效提取内容信息，而且可以提升阅读体验，激发读者阅读兴趣，培养读者对古典文学的阅读与鉴赏能力，同时可以帮助读者根据当前自身阅读能力水平制定科学的诗词阅读计划，从而循序渐进提升自身阅读能力。

然而当下具有难度级别标注的古诗词匮乏。教材与诗词分级读物中提供了由专家标注按照年级划分难度的诗词，但由于这类诗词数量较少且所涵盖的内容范围有限，难以满足读者广泛的阅读需求。现存古诗词规模超过 80 万首，大部分处于未分级状态，如果通过专家对全部诗词的可读性评估与分级，需要投入大量时间与人物力成本，代价高昂且难以扩展。如何快速准确的对现存大量古诗词进行难度分级是一个亟待解决的问题，应用文本可读性分析技术可望自动评估诗词的阅读难度。但目前具有标注难度等级数据集的缺乏成为限制古诗词可读性研究的一个重要瓶颈。

本文旨在提供一个大规模人工标注的诗词可读性数据集。该数据集包含 1915 首诗词，具有三级与六级两种不同粒度的难度等级。

整个标注过程分成两阶段进行。在第一阶段，我们将古诗词按照阅读理解难度分为三级；在第二阶段将已有的三级分类进一步细化成六级分类。

我们首先将诗词的难度被划分成三级：简单、中等、困难，由三名在读研究生作为标记员，通过阅读原文、注释、译文以及赏析等多种材料，在词汇、句子、篇章以及情感感知四个维度对诗词的难易度进行分析，进而综合这些信息对诗词的难度分级，得到古诗词可读性基础数据集（Ancient Poetry Readability Dataset，APRD）。

为了验证标记结果的可靠性，我们抽取教材中的 136 首诗词组成标准集，以诗词所属年级作为标准难度级别，共划分为 12 级。用 Spearman 系数作为标准难度级别与人工标注难度级别之间的一致性衡量指标，一致性为 0.786。

由于三级难度划分粒度过粗，难以满足不同层次用户群体的阅读需求，因此我们在 APRD 数据集三级分类基础上，对已分级的每类数据进一步细化成偏难与偏易两类。由于难度细化后相邻难度级别的差异变小，对标注工作带来困扰，标注者难以准确评估诗词所属的具体难度级别，因此我们采用成对比较法[1]通过比较诗词之间的相对难度得到整个数据集中诗词两两之间的难度关系，计算每首诗词的难度评分值，根据诗词难度的全局排序构建古诗词六级难度划分的古诗词数据集（Ancient Poetry Readability Dataset Plus， APRD+）。最后我们运用基于特征工程的随机森林、SVM 等机器学习模型对诗词难度进行分类。本文提出的古诗词可读性数据集与实验结果可作为后续研究的测试基准。

1 相关工作

1.1 古诗词相关研究

应用人工智能技术对各类文学作品进行自动化分析早已引起学者的广泛兴趣，古诗词因作为中国古典文化的核心组成部分受到重点关注，围绕古诗词已诞生许多有意义的工作。

诗词分类是常见的研究内容。MuY 等人[2][3][4][5]运用机器学习分类模型识别宋词的风格类型。Hu 和 Zhu[6]按照内容类型的对古诗词主题分类。Wang[7] 则关注如何根据格式不同对宋词的词牌名进行自动识别。此外，Zhao[8]对古诗词做情感分析，挖掘潜藏在诗词中的情感。

另一类研究热点则是古诗词自动生成与创作。人们尝试以关键字或者图片作为输入提示，让机器创作对应主题的诗词，基于统计学习[9]以及深度学习的诗词生成模型[10][11][12][13]在诗词创作任务中得到广泛的应用并取得了显著的成果。

在上述诸多诗词应用研究中，核心目标之一是实现让机器理解诗词，从而辅助读者阅读，增加诗词阅读的趣味性与游戏性，降低诗词学习与阅读的难度系数，激发读者的兴趣。然而这些研究均没有考虑到诗词固有的可读性因素，当读者阅读晦涩难懂或者过于简单等与自身阅读能力不匹配的诗词时易于失去阅读兴趣。

1.2 文本可读性相关研究

文本可读性研究主要目的是应用自然语言处理等技术对文本的难度进行定量分析与自动化评估。由于评估工作基本上由计算机自动完成，

减少对人工标注的依赖，具有高效、经济、客观的优势。但与人工评估相比，自动化评估结果的效度通常遭到人们的质疑。如何提升评估有效性与探讨影响难度的关键因素是自动化分析的核心问题，围绕这些问题开展的研究被统称为可读性研究。

高质量大规模的数据集是研究可读性自动化分析的基石，由于英文可读性研究起步较早，数据集较为丰富。其中，Common Core State Standards②[14]（CCSS）是一套指导教材制定的规范，由美国教育部官方发布，旨在统一全国各州的教育大纲标准，对各年级的学生阅读能力与教材难度范围有着详细规定，其附录所提供的例文常被作为可读性分析的训练数据。CCSS 对我们分析与标注古诗词的难度等级具有借鉴意义。

E Schumacher [1]通过众包的方式构建句子对相对难度评估数据集，应用 Trueskill 算法基于句子对的偏序关系计算每个句子的难度系数，获取句子难度的全局排序。

中文可读性数据集比较匮乏，按年级划分的教材是训练数据的主要来源[15][16]。但对于古诗词而言，课文中提供的样本数量过少，且涉及的内容种类、体裁类别有限，难以构建大规模的语料。

在模型方面，公式法与分类法是主流解决思路。在早期研究者通过构建文本特征与难度系数之间的映射公式预测文本可读性。常见的特征有词频、音节数、词长、句长等因素，主要的可读性公式有 ATOS、DRP、FK 公式等[14]。

之后自然语言处理技术发展成熟，可以有效分析文本语法、句法特征，为研究者提供更丰富的灵感来源。Collins-Thompson[17]发现字词分布对预测文本难度具有非常强的导向性，提出基于大规模语料词频表分类模型。Si 等[18]应用 unigram 语言模型对科技 Web 文本进行难度分类。Schwarm[19]基于 SVM 结合 N-gram 语言模型与词法句法特征实现对可读性的分析。E Schumacher [1]考虑句子层次可读性的预测，使用可解释的线性回归模型训练数据。

综上所述，缺乏大规模高质量数据集是当前文本可读性研究的主要瓶颈，在接下来两章中我们将详细描述古诗词可读性数据集的标注过程并对标记结果的可靠性进行验证。在第 2 章中我们描述如何在第一阶段中完成将古诗词划分成“易中难”三个难度等级，创建 3 级基础数据集 APRD。在第 3 章中我们给出如何在第二阶段完成古诗词难度粒度的细化。基于 APRD 数据集的三级分类结果，我们将每一级中的古诗词进一步细化成偏难与偏易两个子级难度，实现六级难度数据集 APRD+的构建。

2 构建三级难度数据集 APRD

2.1 数据收集与预处理

我们利用古诗文网③作为数据源收集待标记的候选诗词样本。古诗文网是一个内容丰富多样、结构清晰的诗词阅读网站。除了提供完整准确的诗词正文外，网站会尽可能提供注释、译文、赏析等辅助资料。充足且高质量的数据源是标注结果可用性的基础。

古诗文网对诗词按照创作年代、风格类型、格式等特征归纳分类，方便我们对样本进行有效过滤与筛选。网站中文本规范清晰的组织结构便于特征的抽取与清洗等预处理工作。网站提供的点赞功能帮助我们获取读者的点评信息，点赞数量能够反映出作品的流行度，为模型提供额外的训练特征。最后，古诗文网的数据对外公开，易于收集与获取。

虽然收集的原始数据集规模庞大，但具有注释、译文等全部辅助资料的样本较少，且含有文言文、近现代诗词等噪声样本。因此我们按照创作年代、辅助信息的完整度对数据进行自动化过滤删除。然后对样本逐个人工检查，剔除乱码、存有质量问题样本。最后在清理后的数据集中，我们随机采样 2000 条样本准备进行难易度标记。

2.2 标注者人选

专家标注是获取高质量可靠数据集的有效途径，能够对诗词的难度进行精准的标记，但由于专家标注成本较高，适合于标记规模小但分级精细的数据集，如教材或分级读物。

对于大规模古诗词难易度标注工作而言，当难度分级粒度较细、如六类分类，对标记者的诗词鉴赏能力要求较高。而对于较粗粒度的类别标

② http://www.corestandards.org/ELA-Literacy/

③ https://www.gushiwen.org/

注如三类分类，对标记者的鉴赏能力的依赖性则相应较低。

因此我们邀请三名具有数学、计算机与语言学专业背景的在读研究生作为数据的标记者。三名标记者均在中国高考语文考试中取得优异的成绩，具备一定的诗词阅读与鉴赏能力。同时三名非专家标注者能够一定程度上代表普通读者，因而能够从普通读者角度标注诗词的阅读理解难易度，借助全面的诗词赏析辅助资料实现对数据精确的标注。

2.3 标注方法

2.3.1 三级分类体系

已有可读性工作中，根据年级将文本分成 12 级是最常见的分级模式[14][15]。但这种模式的缺陷在于难度级别的粒度过于细化，导致不同级别之间界限模糊，标记者难以有把握的的确定文本的具体难度级别，可实践性较低[16]。因此在构建 APRD 数据集时，我们将难度粗粒度的划分为易中难三级，以降低标注工作的复杂性，提高标注结果的准确性。

“简单”表示诗词的词汇与语法简单易于理解而且诗词中蕴涵的语义单一而且表达情感明显；“中等”代表诗词中的用词与语法更复杂而且情感或语义相对丰富，但仍旧可以部分理解；“困难”不仅用词生僻语法复杂，而且语义晦涩、情感复杂，理解困难。

2.3.2 诗词难度标注考虑的两个问题

在阅读诗词中，一般是通过原文获取其表层基本释义，在此基础上通过对诗词的分析并结合相关背景知识来感知作者的潜在语义的表达，从而得到真正的理解。因此分析诗词的可读性也将从影响理解诗词两层含义的角度入手。

考虑原文与译文的语义距离

衡量诗词难度首先需要判断诗词原文与译文之间的语义距离。译文是对诗词语义的通俗化表达。在语义上，译文与原文对等，但由于大部分读者接受的是现代文的语法训练，在阅读诗词原文时首先应用现代文的语法规则对诗词进行解析，导致译文理解难度远远低于诗词原文。这种现象说明在诗词标注工作中应考虑原文与译文的语义距离，若诗词原文中字词含义以及语法规则与现代文越接近，二者之间的语义距离越小，则该首诗词的难易度越低。

考虑译文与赏析的语义距离

在诗词中借助典故或意象等实体对象来寄托情感或表达主题是一种常见的现象，比如在诗句“但愿人长久，千里共婵娟。”中作者用“婵娟”形容月中嫦娥，嫦娥指代月亮，月亮代表思念，如果读者不了解意象“月亮”所寓意的情感，仅仅理解为“赏月”便无法理解诗词中所表达的真正诗意。

译文是诗词内容的直接描述，缺乏对诗词中实体对象所蕴涵潜在含义的解析，若要实现对诗词真正的理解需要借助赏析以补充相关的语言与背景知识。赏析对诗词的语义解析，是译文的延伸与扩展。衡量赏析与译文的语义距离以及考察诗词主题含义的复杂度与情感表达的隐晦程度，会进一步增加诗词难度标注的准确性。

2.3.3 具体难度因素分析

在具体评估中，我们要求标记者从多个层次分析影响诗词的难度因素。

词汇层次：考虑字形的生僻度，分析字词的古今释义是否一致，衡量理解诗词中典故、意象等实体对象潜在含义的难易度等。

句子层次：考虑句子中各种语法结构及修辞技巧的理解难度，即在词汇理解的基础上理解整个句子含义的难度。

篇章层次：衡量将句子含义串联起来翻译成诗词译文的难度，考虑句子之间的关系，语义清晰度与流畅度等因素。

情感感知层次：评估诗词中语义的复杂性与主旨大意的理解难度，进而考虑情感表达的隐晦程度。

2.4 标注过程

对于每一首诗词，我们要求标记者完整浏览原文、注释、译文以及赏析，分别考虑理解字词、句子、篇章以及情感感知的四个维度的难度，最

后确定诗词整体的难度级别。

在标记过程中，我们随机选择两个标记者独自评估样本。将得到的标记结果进行汇总，评估二者之间的一致性。保留两人达成一致的结果作为该样本的最终难度类别。对于存有争议的样本交由第三人进行裁决，以决定最终难度级别。最后汇总所有人的结果采用投票机制决定有争议诗词的最终难度。

经进一步整理后，最终标记结果可分成表 1 中列出的四种类型。类型 A：难度类别由前两人达成一致决定；类型 B：前两人存有争议但通过第三人裁决决定；类型 C：三人均有争议；类型 D：重复标记或残余的近现代诗等无效数据。我们在表 2 中给出不同类型样本的统计数据。

表 1：不同类型标记结果的分布

结果类型	A	B	C	D	总计
数量	1116	862	63	22	2000
比例	0.56	0.43	0.03	0.01	1

对于三人均存有争议的样本，经过分析，发现主要有以下两点原因：一是诗词本身流行度比较高，对标记者产生干扰；二则是诗词本身存有难度存有争议，不同标记者关注点不同。

我们剔除掉类型 C 与 D 的样本后获取最终数据集 APRD，整个数据集包含 1915 首样本。

我们以表 2 中列出三首难易度不同的诗词做为案例，阐述标记者如何通过分析各层次的难度因素来实现对诗词难易度进行分级的标注过程。

对于第一首诗词《三衢道中》，其在词汇层次主要包含简单字词诸如“日日晴、小溪、黄鹂”，且均为生活中普通对象，且无隐藏含义，对背景知识要求低，易于理解；在句子层次，语法与现代文较为接近，语义通俗,以句子“梅子黄时日日晴”为例，其译文为“梅子黄透了的时候，天天都是晴朗的好天气”，原文与译文内容差异教材小，理解难度小；在篇章层次，上下文语义衔接自然，上文道出山行原因，下文描述山间见闻，前后连贯流畅；最后在情感感知层次上，全诗明快自然，诗人游玩欢愉之情跃与纸上，情感表达明显，且该诗为写景，主题平凡常见,也降低了诗词理解的难度。基于以上四点原因，我们将该首诗词标记为“简单”。

第二首诗《咏蝉》的难点主要在于文本中含有“蝉、南冠、玄鬓、白头吟”等多个意象典故，语义双关，对读者背景知识有要求，导致诗词难度的提升；此外该诗使用托物言志的手法，主题虽为咏物，但旨在明志，情感表达较为含蓄，理解诗词要求对作者生平与创作背景有了解。但由于使用语法简单，句子结构清晰，上下文语义流畅，因此该诗的难度为“中等”。

第三首《狼跋》含有包含“疐、跋、舄”诸多生僻字词，语言凝练，要求读者具有古汉语语法积累。同时情感表达隐晦且主题存有争议。因此该诗的难度为“困难”。

表 2：诗词难度标注样例

难度	样例
易	三衢道中 梅子黄时日日晴，小溪泛尽却山行。 绿阴不减来时路，添得黄鹂四五声。
中	咏蝉 / 在狱咏蝉 西陆蝉声唱，南冠客思深。 不堪玄鬓影，来对白头吟。 露重飞难进，风多响易沉。 无人信高洁，谁为表予心？
难	狼跋 狼跋其胡，载疐其尾。 公孙硕肤，赤舄几几。 狼疐其尾，载跋其胡。 公孙硕肤，德音不瑕？

2.5 标注结果统计分析

数据集包含诗与词以及极少量的文言文，每个样本附带注释、译文、赏析辅助资料。

表 3 列出 APRD 数据集的统计信息。三个级别的样本总体分布较为均匀，但困难类别诗词占据比例较小。同时可以发现诗词难易度越高，样本平均长度越大，样本长度与难易度之间可能存有一定关系。

表 3：APRD 数据集的统计信息

	数量	样本均长	字典大小
简单	806(42%)	37.15	5138
中等	815(42.5%)	76.03	
困难	294(13.5%)	179.46	

对于标记结果的可靠性我们选择人教版教材

中的诗词进行验证。提取教材与 APRD 中共有的 136 首诗词构建标准集，以教材所属的年级作为诗词的实际难度级别即标准难度。标准集中所包含各个年级的诗词样本数如表 4 所示。

表 4：标准集各个年级所含诗词数

年级	1	2	3	4	5	6
诗词数	14	15	16	8	7	12
年级	7	8	9	10	11	12
诗词数	11	17	11	4	13	8

计算得到标准集与 APRD 数据集之间的 Spearman 相关性为 0.786，表明了人工标注的难易度与诗词的实际难度之间具有较高的一致性。

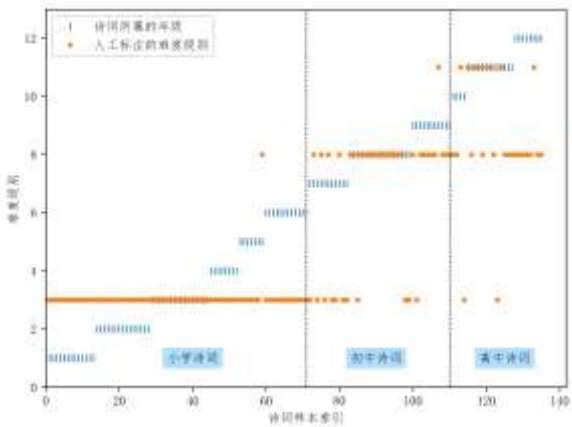

图 1 APRD 与标准集难度等级对比图

图 1 中给出标准集中诗词的标准难度级别与人工标注难度的对比结果。横轴代表诗词样本，纵轴为每首诗词对应的难度级别。为了便于比较，我们将“易、中、难”三级难度数字化为“3、8、11”。从图中可以发现我们的标记结果能够有体现出诗词的难易度，对于实际难度越高的诗词，我们同样对其难度等级评估的越高。比如对于低年级的小学诗词，基本全被标记为简单级，而对于初中诗词会被标记为中等级而较少分类为简单级。同时我们发现，对于难度较高的高中诗词我们标记难度为中等的诗词比例较大，对于这一现象可能的原因是标记者已学习过该首诗词而对其熟悉导致其标记为较低的难度等级，即诗词的流行度对标记者产生干扰。

3 细化标注构建六类数据集 APRD+

3.1 标记方法

经过第一阶段的人工标注，我们得到三级分类的 APRD 数据集。在第二阶段中，APRD 数据集的每一类别的诗词被进一步细化成偏难与偏易两类。由于划分到同一类的诗词之间的难度差异较小，导致标注者判断诗词准确类别的难度增大，为了降低数据标注的复杂度以及提升标注结果准确度，我们采取成对比较法通过对同一类别中的诗词样本进行两两比较，通过不同诗词之间的相对难度关系实现对诗词难易度的分级。

根据样本之间的偏序关系可以构建相对难度有向图，由于相对难度关系具有可传递性，因此任意两个样本之间的相对难度关系即使未被人工标记，但若两者之间存有一条有向路径，则可以根据已标记样本对推导出未标记样本对之间的难度关系。

为了降低标记成本，避免不必要的样本对比较，在生成具体的诗词匹配对时，我们仅从相对难度未知即二者之间不存在一条有向路径的样本对中进行采样。

诗词间的相对难度关系共有三类：偏难、偏易、相等。每一对诗词样本之间的相对难度都由三个标记者进行判断，投票决定最终的相对难度关系。每个样本对的类别都至少有两人达成一致，若三人结果均不同则该样本对直接被舍弃，重新采样新的样本对进行标记，而该样本对之间相对难度关系则通过后续再采样中标注或者根据已有样本对推导得到。

3.2 标记过程

如图 2 所示，整个数据标注流程分成两步：第一步是不断生成匹配对，由标记者标注样本间的相对难度关系，迭代多次直到数据集中任意两个样本之间可相互比较；第二步是基于样本对的偏序关系应用 Trueskill 算法计算样本的难度评分，基于该评分细化 APRD 中的每类诗词的难度级别，从而获取六级分类的 APRD+数据集。

图 2：APRD+标注方法

3.2.1 生成匹配对

在实际标记过程中因无法保证标记者实时在线标记，因此我们采用多轮迭代的方式逐步获取所有样本之间的偏序关系。在每一轮标记中，遍历 APRD 数据集，对于每首诗词从同一类别的的样本子集中找到相对难度关系未知的样本构成采样池，从中随机采样出一条样本生成匹配对，如果采样池为空则代表该诗词可与数据集中任意诗词之间比较相对难度，因此该首诗词不需要进行重复标记，可以直接跳过。每轮生成多条匹配对，分别交由三个标记者独立标记。

每轮迭代可以获取部分样本对之间的相对难度关系，当数据集中任意两个样本之间可比较时，迭代过程终止。

表 5：每轮标记样本对数

迭代轮数	1	2	3	4	5
样本对数	1915	985	369	261	137
迭代轮数	6	7	8	9	10
样本对数	69	66	7	4	2

实际标记过程共进行十轮，共标记 3815 对样本，表 5 列出每轮标记样本对数。

3.2.2 计算难度评分

为了获取诗词的整体难度排序，我们借鉴文献[1]提出处理方案，采用微软开发的 Trueskill 算法基于偏序对样本的难度进行排序。

Trueskill[20]系统是基于贝叶斯推断的评分系统，是对传统 Elo 评分系统的衍伸与推广，擅长根据样本对之间的偏序关系中计算样本的质量评分以产生数据的全局排序。

Trueskill 主要用于游戏中玩家能力值的排名，根据玩家每局的对战结果计算玩家的能力值。在具体应用中，我们将每首诗词视作一个玩家，在匹配对中标记为更难的样本视作赢家，因此相对难度三个类别“偏难、偏易与难度相同”分别对应为“赢、输与平局”。

由于 Trueskill 算法依赖于输入样本对的先后顺序，因此我们对外报告 Trueskill 算法运行 500 轮后的评分结果。在计算评分过程中，每轮输出的诗词难度评分将作为下一轮算法运行时该诗词难度评分的初始值，通过多轮计算消除样本对顺序的影响。最后一轮的输出被作为诗词最终的难度评分，评分值越大代表该诗词难度更高。

在 Trueskill 算法中，玩家的初始评分值为 25，之后根据对局结果调整评分值，战败的一方评分值下调，而战胜的玩家评分值得到提升。由于我们希望对 APRD 中已有的每个级别样本子集中的数据集划分成偏难与偏易两个子级别。因此采用 25 作为子级别划分的界限，评分高于 25 的样本标记为偏难，低于 25 标记为偏易，对三个类别的数据分别进行相同处理，得到诗词的六级分类数据集 APRD+。

3.3 标记结果统计分析

APRD+各类数据分布参考表 6,可以发现各级别的数据分布较为均衡。

表 6：数据集的统计信息

级别	1	2	3	4	5	6
数量	465	341	439	377	138	155

与验证 APRD 标记结果有效性类似，我们计算出 APRD+数据集六级分类结果与标准集的 Spearman 一致性为 0.804，相比于 APRD 数据集，APRD+数据集的一致性得到轻微的提升。

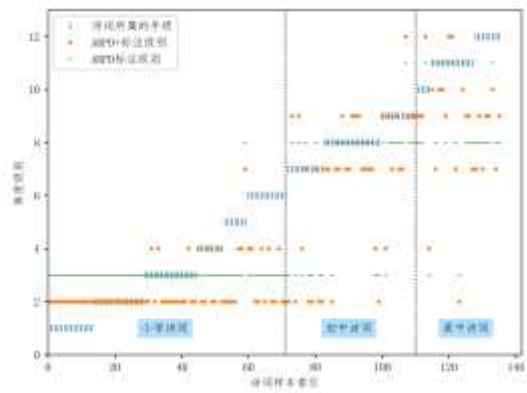

图 3 标准集与 APRD+标记难度对比图

图 3 给出标准集与 APRD+标记难度的对比结果。类似于图 1，但 APRD+的六级难度被数字化成“2/4/7/9/10/12”。

4 基准实验

如同已有的中文可读性分析[21]工作，我们将诗词难度评测作为一个文本分类任务。将 APRD 与 APRD+数据集作为训练语料，我们提取常见文本特征并应用 SVM 和随机森林模型对古诗词难易度进行分类，测试模型分别在古诗词的三级与六级难易度自动评估的作用效果。

4.1 实验设置

预处理中，我们利用正则表达式过滤掉文本中的特殊字符与混合在原文中的注解拼音，并进行分字处理，将诗词文本视作若干字的集合。

我们提取若干蕴涵诗词难度信息的特征作为诗词文本的特征表示。直觉上，字使用频率越易被读者熟悉，理解难度越低，因此字频蕴涵字的难易度信息，在具体应用中我们 *tf-idf* 公式对字频进行归一化转换，将其结果作为衡量字的难度系数，同时将每首诗的平均字 *tf-idf* 值也作为文本特征，代表该首诗词所包含字的平均难度。

此外，在标注数据过程中发现，辅助资料的文本特征与诗词难度有一定相关性。越难理解的诗词其解释会愈加详尽，因此译文长度越长。而注释则代表难词的个数。最后点赞数是对诗词质量的直观反映，我们考察诗词质量与可读性之间的关联。表 7 中列出本文用于古诗词可读性分析的全部特征。

表 7：特征集

特征	缩写
字 tf-idf	tf
平均字 tf-idf	atf
诗词长度	lt
句子数量	ns
译文长度	ltt
注释数量	na
点赞数	nu

实验结果采用十折交叉验证，每个模型运行十次，每次运行时随机选取 80%的样本作为训练集，10%作为验证集，剩下的 10%作为测试集。记录每次实验在测试集的结果，求平均值作为该模型的分类性能。

由于数据集的规模有限，字在数据集中的文档频率不能有效代表字的实际值，因此我们选择语料库在线网站④提供的古代汉语语料库字频表来近似字的文档频率。

由于直接使用 *tf-idf* 值构造文档表示矩阵时会导致严重的特征稀疏性问题，因此我们选择 PCA 对字的 *tf-idf* 矩阵进行降维，之后加入其他文本特征来训练模型。

4.2 实验结果

我们遵循可读性的评价方法，将准确率（ACC）、精确率（PREC）、召回率(PREC)以及 F1 作为评估分类器性能的指标。

表 8 列出 SVM 与随机森林模型在 APRD 与 APRD+上的分类效果。从结果可以看出，模型对于古诗词难易度分类的效果影响较小，两个模型分别作用于三级分类任务与六级分类任务时有着近似的分类结果。在模型中加入 PCA 对数据进行降维均能够给分类效果带来一定程度的提升。

此外，相比于三级分类，六级分类效果发生下降，由于 APRD 与 APRD+样本数据完全一样，仅是类别标签粒度不同，因此类别粒度的细化是导致分类效果的直接原因，这也说明如何提升细粒度的诗词可读性分类效果将是一件具有挑战性的任务。

表 8：不同机器学习模型分类结果对比

数据集	模型	ACC	PREC	REC	F1
APRD	RF	0.74	0.65	0.56	0.56
	RF+PCA	0.75	**0.73**	**0.71**	**0.72**
	SVM	0.75	0.70	0.66	0.67
	SVM+PCA	**0.76**	0.72	0.68	0.70
APRD+	RF	0.51	0.47	0.45	0.42
	RF+PCA	**0.52**	**0.51**	**0.48**	**0.46**
	SVM	0.46	0.43	0.42	0.40
	SVM+PCA	0.48	0.44	0.43	0.41

SVM：SupportVectorMachine，支持向量机模型；

RF：RandomForest，随机森林模型；

4.3 烧蚀分析

4.3.1 字频特征有效性验证

特征字的频率值是一类特殊的特征，高频词的理解难度往往低于低频词，因此我们将字频作为衡量字难易度的一种指标。数据集中出现的每个特征字都代表一个特征维度，但由于部分特征字出现频率低，使得该部分字作为特征输入会带来特征维度的扩增以及特征表示稀疏问题，同时可能给模型引入新的噪声，因此我们需要验证不同频段的字对模型分类效果的影响。

④ http://corpus.zhonghuayuwen.org/

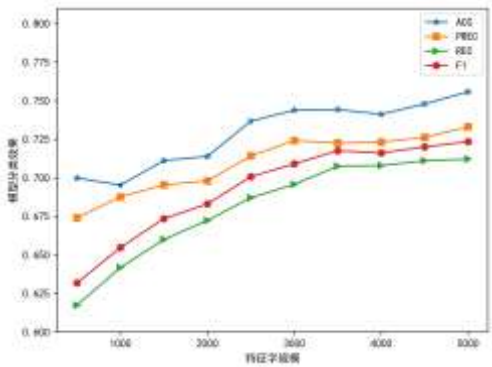

图 4：特征词数量对分类效果的影响

我们将特征字按照字频排序，优先选择高频字，逐步扩大特征字的规模。由于 APRD 与 APRD+仅类别标签不同，但数据完全一样，因此我们仅给出基于 APRD 三级分类的测试结果。我们选择平均分类效果最好的 RF 模型进行测试，分类效果变化趋势如图 4 所示。

从图 4 中可以看出，随着低频特征字的逐渐加入，模型分类效果不断得到提升。这一结果验证了各个频段的字对诗词的难易度均具有区分性，字频能够反映出字的复杂度信息，使用字频特征有助于对诗词的难易度分级。

4.3.2 其他文本特征有效性验证

我们已经验证字频信息的有效性，对于其他文本特征的有效性，我们采用以字频特征为基础，每次仅将由字频信息与待验证特征组合得到的特征子集作为输入的方式进行验证，即每次试验选择两类特征：根据字频计算得到 *tf-idf* 值与待验证的文本特征。为了显著突出不同子特征对性能的影响程度，我们将特征子集的分类结果与仅用 *tf-idf* 特征的效果的差值表示该特征的有效性的衡量。

图 5 列出不同特征效度对比结果，结果显示在这些文本特征中，单个特征的增加分类性能的影响较小。但这些特征之间的相对效度有较大差异，其中对性能带来显著提升的特征主要有注释数量(na)、句子个数(ns)、平均字频(awf)、译文长度 (ltt)，而诗词长度(lt)、点赞数(nu)、赏析文长度(lat)对分类器的提升甚微，甚至会降低分类器的效果。

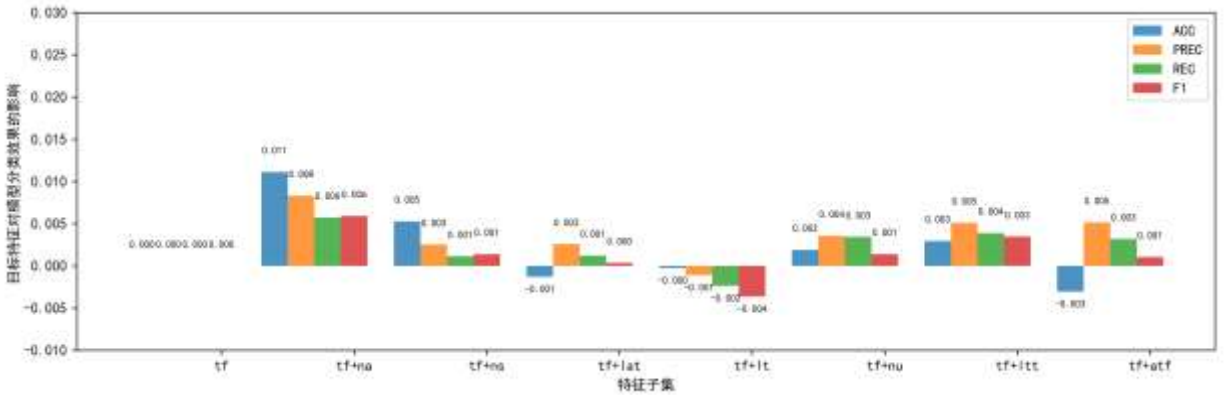

图 5：不同特征的效度比较

5 结论

在本文中工作中，我们构建了一个包含 1915 首的人工标注难度的中文古诗词可读性语料，并提供三级与六级两种粒度的难易度标记。数据集

中每个样本都具有原文、注释、译文赏析等辅助数据，以便标记者全面了解诗词。每首诗词的难度级别至少被两人评估，从字词、句子、篇章以及情感感知四个维度对诗词中难点进行分析，综合考虑影响难度的因素对诗词的难度类别以及诗词间的相对难度进行分类，并将标注结果与专家标注的难易度类别进行比较，较高的一致性值验证我们提供的数据集的可靠性。

最后我们基于获取的两种粒度的数据集进行基准实验，当数据集的难度设置为三级时，模型能够有效实现对古诗词难度的分级，但将难度设置为六级时，模型分级效果会发生大幅下降，分类效果不佳。同时当前数据集的规模有待继续扩展，以满足深度学习文本分类模型对数据量的要求。如何提升六级难易度分类的效果以及继续扩展数据集的规模将成为我们未来工作的主要内容。

参考文献

[1] Schumacher E, Eskenazi M, Frishkoff G, et al. Predicting the relative difficulty of single sentences with and without surrounding context[J]. arXiv preprint arXiv:1606.08425, 2016.

[2] Mu Y. Using keyword features to automatically classify genre of Song Ci poem[C]//Workshop on Chinese Lexical Semantics. Springer, Cham, 2015: 478-485.

[3] Yong, Y.: A Study on Style Identification and Chinese Couplet Responses Oriented Computer Aided Poetry Composing. Ph.D. Dissertation of Chongqing University (2005). (in Chinese)

[4] Chunlong, W.: The Research of Computer Assistant Analysis on Chinese Song Poems' Style. The master degree thesis of Xiamen University (2008). (in Chinese)

[5] Z. He, W. Liang, L. Li and Y. Tian, "SVM-Based Classification Method for Poetry Style," ICMLC, Hong Kong, 2007, pp. 2936-2940.

[6] Hu R, Zhu Y. Automatic classification of tang poetry themes[J]. Journal of Peking University (Science and Technology), 2015, 51(2): 262-68.

[7] B. Wang, J. Zheng, Y. Du and L. Yang, "Automatic Recognition of Tune Names of Song Ci-Poetry," IALP, Bandung, Indonesia, 2018, pp. 189-192.

[8] H. Zhao, B. Wu, H. Wang and C. Shi, "Sentiment analysis based on transfer learning for Chinese ancient literature," BESC2014, Shanghai, 2014, pp. 1-7.

[9] He J, Zhou M, Jiang L. Generating chinese classical poems with statistical machine translation models[C]//AAAI, 2012.

[10] Zhang X, Lapata M. Chinese poetry generation with recurrent neural networks[C]//EMNLP. 2014: 670-680.

[11] Yan R. i, Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema[C]//IJCAI. 2016: 2238-2244.

[12] Yi X, Li R, Sun M. Generating chinese classical poems with rnn encoder-decoder[M]//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham, 2017: 211-223.

[13] Cheng W F, Wu C C, Song R, et al. Image inspired poetry generation in xiaoice[J]. arXiv preprint arXiv:1808.03090, 2018.

[14] **吴思远，蔡建永，于东**, et al. **文本可读性的自动分析研究综述**[J]. **中文信息学报**, 2018, 32(12).

[15] 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.

[16] Chen Y T, Chen Y H, Cheng Y C. Assessing chinese readability using term frequency and lexical chain[J]. International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 2, June 2013-Special Issue on Chinese Lexical Resources: Theories and Applications, 2013, 18(2).

[17] Heilman M, Collins-Thompson K, Eskenazi M. An analysis of statistical models and features for reading difficulty prediction[C]//ACL 2008: 71-79.

[18] Si L, Callan J. A statistical model for scientific readability[C]//CIKM. 2001, 1: 574-576.

[19] Schwarm S E, Ostendorf M. Reading level assessment using support vector machines and statistical language models[C]//ACL, 2005: 523-530.

[20] Graepel T, Minka T, Herbrich R T S. A Bayesian skill rating system[J]. Advances in Neural Information Processing Systems, 2007, 19: 569-576.

[21] Liu H , Li S , Zhao J , et al. Chinese teaching material readability assessment with contextual information[C]// IALP IEEE, 2018