

面向数字人文的古诗文本情感术语抽取与应用研究^{*}

张 卫 王 昊 邓三鸿 张宝隆

摘 要 在跨学科知识范式下,数字人文的研究范畴随着自身学科体系的拓展而不断泛化,采取关键语义技术解析文化对象中的人文内涵与情感知识对于重拾学科“人文性”与“计算性”特质具有重要意义。本文以古诗文本为例,面向汉语诗文及其鉴赏实现大规模人文情感术语的自动化抽取与分析。首先在无标注集环境下提出一种基于“冷启动”的字序列自动标引方法来获取学习语料,随后在字向量(Char2Vec)指导下将汉字特征(部首、拼音等)和BERT语言学模型分别引入机器学习与深度学习模型,并从知识发现的角度定义新术语识别规则。研究发现,将现代鉴赏融入古诗原文显著优化了情感知识的广度与深度,领域术语能够被有效标引。训练的BERT-BiLSTM-CRFs深度学习模型的效果明显优于CRFs机器学习,最佳F1与F1_distinct可分别达到95.63%和85.43%;同时汉字特征的引入也有效提升了传统CRFs效果,以领域特征和基于“竖心旁”“心字底”部首约束特征为最优。相较于机器学习抽取出的长篇幅新术语,深度学习能够拓展出更多寄托情感知识的新意象词。源于诗文与鉴赏的情感术语为文学信息资源的情感分析与知识服务提供了参考(人文性),基于汉字语言特征的抽取方案为中文领域自然语言处理技术的深化提供了启迪(计算性)。图11。表6。参考文献30。

关键词 数字人文 古诗 情感术语抽取 汉字语言特征 Char2Vec BERT

分类号 G254

Sentiment Term Extraction and Application of Chinese Ancient Poetry Text for Digital Humanities

ZHANG Wei, WANG Hao, DENG Sanhong & ZHANG Baolong

ABSTRACT

In recent years, the application of digital technologies such as digital library, information visualization, multimedia publishing, and geographic information system in the field of humanities has made the scope of digital humanities more extensive. It is of great significance to regain the “humanity” and “computability” characteristics of discipline by taking key technology to parse sentiment knowledge in humanistic objects.

^{*} 本文系国家自然科学基金面上项目“关联数据驱动下我国非遗文本的语义解析与人文计算研究”(编号:72074108)和中央高校基本科研业务费项目“面向人文计算的方志文本的语义分析和知识图谱研究”(编号:010814370113)的研究成果之一。(This article is an outcome of the project “Research on Semantic Parsing and Humanities Computing of Chinese Intangible Cultural Heritage Text Driven by Linked Data” (No. 72074108) supported by National Natural Science Foundation of China and the project “The Semantic Analysis and Knowledge Graph Research of Local Chronicle Text Oriented to Humanistic Computing” (No. 010814370113) supported by the Fundamental Research Funds for the Central Universities.)

通信作者:王昊,Email:ywhaowang@nju.edu.cn,ORCID:0000-0002-0131-0823 (Correspondence should be addressed to WANG Hao,Email:ywhaowang@nju.edu.cn,ORCID:0000-0002-0131-0823)

Ancient poetry contains the sentiment knowledge about the political background, historical events, folk customs, and other things, and occupies an extremely important literary position. At present, existing studies mainly automatically classified the sentiment trend of ancient poetry, while the mining of fine-grained sentiment knowledge is still insufficient. In order to achieve a more accurate sentiment analysis of ancient poetry, this paper focuses on the sentiment term automatic extraction in the domain text and its application.

Firstly, this paper introduces the modern appreciation to extend the sentiment knowledge and human connotations in the target ancient poetry for the first time to solve the problems of limited term number, coarse sentiment granularity and insufficient learning of text feature caused by the condensed language characteristic of ancient poetry; secondly, a “cold start” automatic citation method for character sequences is proposed to obtain learning corpus; thirdly, based on a character vector mapping (Char2Vec) of the BERT model, we focus on exploring the extraction effect of sentiment terms under the introduction of Chinese character linguistics feature on the benchmark of CRFs model, and compare it with the BERT-BiLSTM-CRFs model, and then we define a new term recognition rule from the view of knowledge discovery; finally, based on the term set of humanistic sentiment in the field of ancient poetry, this paper explores the digital application of term retrieval, granularity mining and poet portrait. It was found that: 1) The integration of modern appreciation into ancient poetry significantly optimizes the breadth and depth of sentiment knowledge, and the domain terms were effectively labeled by the method proposed in this paper. 2) The trained BERT-BiLSTM-CRFs model outperformed the CRFs model, and the best F1 and F1_{distinct} can reach 95.63% and 85.43%. At the same time, the introduction of Chinese character features also improves the effect of traditional CRFs, among which the field feature and the constraint radical feature (“shuxinpang” and “xinzidi”) are optimal. 3) Compared with the long new terms extracted by machine learning, deep learning expands more new imagery words that deliver sentiment. The latter integrates with domain terms to form 14 599 distinctive terms and lays the foundation for the construction of sentiment dictionary and sentiment analysis in the field of ancient poetry.

The contribution of this paper lies in two aspects. The sentiment term derived from poetry and appreciation provides a reference for sentiment analysis and knowledge service of literary information resources (humanity), and the extraction scheme based on the linguistic knowledge provides inspiration for the deepening of natural language processing technology in the Chinese domain (computability). 11 figs. 6 tabs. 30 refs.

KEY WORDS

Digital humanities. Ancient poetry. Sentiment term extraction. Chinese character linguistics feature. Char2Vec. BERT.

0 引言

数字人文(也称“人文计算”),是由人文知识、计算机基础设施、数据挖掘技术等多方理论与实践融合发展形成的跨学科研究领域^[1]。近年来,多维数字化技术在人文领域的应用大大

拓展了数字人文的知识体系^[2],然而宽泛的研究范畴稍许模糊了其学科特点^[3]——“人文性”与“计算性”,前者强调文化资源中所传承的人文情感与思维内涵,后者倡导通过数据的深度计算来挖掘人文素材中隐含的核心知识。由于文本是人文知识的根本载体,故而以语义解析技术为驱动、以人文情感为对象的文本知识挖

掘愈发重要。

在我国的文化遗产中,古诗囊括了古人对政治背景、历史事件、民俗风情等事物的认知,涵盖了丰富的情感知识,占据着极为重要的文学地位。目前,领域文本的情感内涵尚未得到有效发掘和利用,如何从语义关联视角对多源异构的人文情感信息进行知识组织更是重中之重。对此,有学者试图对古诗的情感极性进行自动分类^[4],但篇章级的极性分析对细粒度情感知识的挖掘尚显不足。目前我国古诗领域尚无完善的情感词典,而且现有中文情感词典构建技术难以保障情感术语的完整性与准确性^[5]。为了实现更精准的古诗情感分析,需要先实现领域文本中情感术语的自动化抽取及其性能优化,其中所涉情感术语是指古诗领域中具有人文内涵的情感词语,包括汉语单字、词语与短语。存在的关键问题包括:①古诗文本多为基于单汉字行文的古汉语原文,高度凝练的文体限制了术语的抽取规模、情感粒度以及文本特征的学习效果,因此有必要对古诗文本的内容进行扩展;②大规模术语抽取目前主要采取机器学习技术,但古诗领域标注语料的匮乏使得学习任务难以启动;③BERT的兴起使得以CRFs为代表的传统机器学习算法受到挑战,但其字向量映射(Char2Vec)启迪后者引入汉字结构与内容上的语言特征来优化情感术语的抽取效果。

本研究基于古诗文本内的核心情感内涵(人文性),采用信息抽取中的关键语义技术(计算性),一方面引入现代鉴赏的方法,将古诗的鉴赏文本融入其所评鉴的诗文中,通过鉴赏文本扩展古诗原文中的情感知识,形成本文研究范畴内的古诗文本;另一方面,提出一种冷环境下(无学习语料)的字序列自动标注方法以获取学习语料。在此基础上,将古诗领域的语言学知识融入CRFs机器学习模型,重点探索汉字语言特征对情感术语抽取的有效性,并与BERT-BiLSTM-CRFs深度学习模型进行实验对比,最终归纳出最优模型以实现古诗文本中大规模情

感术语的自动化抽取及应用。

1 相关研究成果

1.1 古诗的知识组织和人文内涵解析

古诗文本是我国诗词文脉传承的重要载体,蕴含着丰富的人文知识。目前对古诗文本的知识组织多集中于自动化分类,其维度包括诗文风格^[6]、诗文题材^[7]以及诗文韵律^[8]等。但是,这些研究大多从诗文外部结构的角度出发,而古诗内涵的挖掘与情感分析则更能体现人文性。对此,吴斌等基于迁移学习模型将诗文判定为正、中、负三种情感极性^[9];为了提升情感分类的精度,Tang等提出了一种基于卷积神经网络和门控循环单元的唐诗情感分析方案,采取多通道处理模型来提取语义特征,并通过与主流深度学习模型的对比验证了研究方案的有效性^[10]。然而上述研究均属于篇章级别的情感极性分析,难以洞悉诗文中细粒度的人文知识。相较之下,情感术语能更为精细地解析古诗的人文内涵,但要先解决因古诗凝练性所引发的术语规模受限、情感粒度较粗以及文本特征学习不充分等问题。而诗文鉴赏^[11]是一种对古诗内涵进行分析的现代文学文本,能够显著地拓展领域术语的数量并细化情感知识粒度。基于上述考虑,本文采取现代鉴赏的方法,将古诗的鉴赏文本融入相对应的诗文中,致力于古诗文本中词语粒度的人文内涵解析。

1.2 领域情感词典的构建方法

情感词典是一种词语粒度的知识组织与情感分析工具^[12]。目前,我国古诗领域情感词典的空白使得自动构建工作势在必行,核心构建方法包括:①语义知识库,即在通用情感词典(如HowNet)的基础上通过词语关系(同反义、上下位)的挖掘来构建情感词典^[13],此方法虽快捷实用,却存在领域适用性差的问题;②领域语料库,旨在从领域文本中自动抽取出情感词汇,主要包括连接关系法^[14]、统计量(点间互信

息、卡方统计量等)^[15]以及表示学习法(如 Word2Vec)^[16]。其中,连接关系法所诉诸的语言规则容易使情感词汇的覆盖面受到限制^[17];统计量则难以避免对低频术语识别的固有缺陷^[18];表示学习法通过词向量计算术语间的相似度来确定情感词,但中文分词过程往往会将领域未登录词错误切分^[19]。综合来看,现有技术在面向具有显著领域特性的古诗文本时,很难实现较为完整、准确的情感术语抽取效果,这将大大制约情感词典的质量以及后续情感分析的精度。为此,本文重点探讨古诗文本中情感术语抽取方法的有效性及其优化方案。

1.3 情感术语抽取的方法与技术

目前,有关情感术语抽取有效性的研究多与观点挖掘^[20]、属性抽取^[21]、情感标签抽取^[22]等细粒度情感分析任务相结合。其中,具有指导意义的是郁圣卫等^[23]利用序列标注模型抽取观点对象与情感术语等要素,并通过扩展文本特征取得了良好的识别效果。不过,该研究更侧重于观点对象的挖掘,对情感术语的探索尚不深入。因此,采取有监督学习模型来抽取古诗领域的情感术语,能够通过模型对过往数据的经验学习实现序列的自动标注,在克服语言规则限制的同时,亦能弥补统计量无法识别低频术语的缺陷。

CRFs 是目前实现序列标注任务最为经典的算法^[24],不足之处在于古诗领域目前尚缺乏成熟的学习语料。此外,如何有效地提取领域文本特征是模型优化的关键问题。近年来,BERT^[25]的兴起使传统的机器学习算法受到了挑战,其双向语言编码、遮罩及注意力机制可获得更高质量的语义向量^[26],且适用于主流的深度学习模型(如 BiLSTM-CRFs)^[27],已在多个领域的实体识别中取得优异表现。此外,由于 BERT 在中文领域的标注单元为汉字^[28],能够避免中文分词技术所引发的术语碎片化问题,可保障识别术语的完整性。更为重要的是,字标注模型可通过字向量映射(Char2Vec)充分发挥领域文本中汉字的语言

学知识,这为传统的机器学习提供了一种基于汉字特征的模型优化思路。

综上,本文将构建基于字标注的机器学习和深度学习模型来实现古诗及其鉴赏文本中情感术语的抽取与应用,并在 CRFs 算法基础上重点探索汉字语言特征对情感术语抽取的影响效果,为领域情感词典的构建与术语级的古诗情感解析打下基础。

2 数据及方法

2.1 研究框架

本研究基于字标注模型提出了一种“冷启动”式学习语料自动获取方法,旨在解决古诗语料中情感术语的标注问题,在此基础上将汉字语言知识分别融入机器和深度学习算法,以实现古诗文本中情感术语的自动化抽取及应用,如图 1 所示。

由图 1 可知,研究框架主要包括三个模块。①冷启动。首先搜集并整理多源情感词汇列表,融合形成更具知识扩展性的领域情感词集;随后将现代鉴赏文本(以下简称“鉴赏”)融入对应的古诗原文(以下简称“诗文”)中,形成古诗文本语料,并利用词集对其进行术语匹配以将其切分为 0—1 文本片段;通过字标注模型将文本映射为字—角色空间以探析汉字特征;最后划分标注语料以启动后续学习任务。②机器学习和深度学习模型。就机器学习来说,本文将汉字语言特征引入特征空间训练 CRFs 算法,以期在字特征基准上实现模型改进;就深度学习而言,旨在融入 BERT 语言知识对神经网络进行优化,包括 Char2Vec 映射领域汉字深层特征、BiLSTM 上下文信息编码与网络参数训练、CRFs 解码约束标签顺序等过程。最后,利用模型预测测试集获得标注序列。③数字人文应用。将预测序列映射至角色—字空间,抽取出领域术语和新术语,融合为古诗领域人文情感术语集,进而探索情感术语集在术语检索、粒度挖掘以及诗人画像等多维角度的数字化应用。

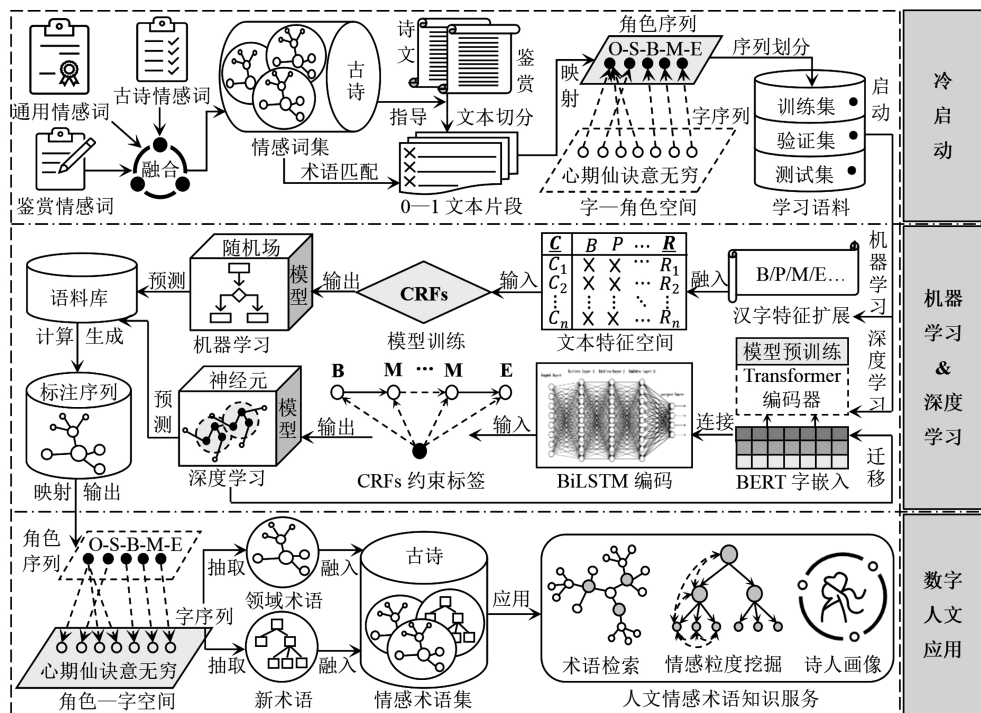


图1 古诗文本情感术语抽取与应用研究框架

2.2 数据来源与预处理

本文的古诗文本语料主要由诗文及其鉴赏组成。在我国诗文化中,唐代诗歌对于后世的政治、民情、风俗、文化有着重要的影响^[29]。本文以《唐诗鉴赏辞典》为素材进行分析,该书是一本包括古诗原文和现代鉴赏解析的文学研究成果,能够较为全面、准确地映射诗文中隐含的情感知识。该书的数字文本来自“文学100”网站^①,如图2所示。

通过数据清洗并对文本进行结构化组织,共得到有效诗文1374篇,选择“篇名”“诗文”“鉴赏”三个层次的字段作为语料。进一步以“。”“?”“!”“;”为断句标准切割,得到51545个段落,并将段落初始类别设为“0”,形成待(未)标注语料,本文称之为冷语料。此外,由于本文

语料中诗文的篇幅相对较小,为验证其有效性,笔者特从外部开放资源“GitHub”网站^②引入本文诗歌以外的唐诗语料,构建基于诗文的新训练集,使新训练集的文本规模与本文训练集规模一致,进而对比出不同训练集上的模型在术语抽取表现中的差异。

2.3 基于冷启动的古诗情感学习语料自动生成

冷启动是指在无或少学习语料的情况下,通过相关域知识对冷语料进行自动标注来生成学习语料,进而实现目标域模型训练的方法。在本文中,冷启动的核心机理是通过情感词集匹配冷语料中的术语,在借助字标注模型映射文本片段为字序列的同时标注角色,以实现学习语料的自动生成。

① <http://www.wenxue100.com>

② <https://github.com/Werneror/Poetry>



图2 古诗文本语料与情感知识术语(部分)

(1) 情感词集,用于领域文本中术语的匹配及其角色的自动化标注,主要源于通用情感词典及领域专有情感词汇。其中,中文领域通用的情感词典包括知网 HowNet 情感词典、清华大学李军情感词典、大连理工大学情感词汇本体库、台湾大学情感词典等^[30];领域专有情感词汇囊括古诗中的常见情感词汇与鉴赏中的评价词汇,由笔者从 Web 资源中搜索获得。将二者相融合后去除重复的词语,最后形成具有知识扩展性与延伸性的古诗领域情感词集。

(2) 字标注模型。定义字标注模型为特征映射的考虑有三:其一,中文以字作为最小知识单元,相比于词,采取字标注能够挖掘更细粒度的语义特征;其二,由于汉语文本的连续性特点,采用词角色标注需要先对领域文本进行分词,然而现有中文分词技术尚不成熟,尤其在面向古汉语文本时,极易引发文本序列错误切割、长术语碎片化(如“岂不/惮/艰险”)等问题;其三,BERT 是一种字嵌入的语言模型,其 Char2Vec 映射也为 CRFs 模型中汉字特征的引入提供了思路。为此笔者采取字角色标注 {OSBME} 映射文本序列空间,如 {心/O 期/S 仙/O 诀/O 意/B 无/M 穷/E},旨在通过词首、词中、词尾、词外等外部特征序化术语的汉字组成结构。

(3) 基于冷启动的字序列自动标注,通过自动化标注逻辑将非结构化文本解析为字序列与

角色序列,包括术语匹配与序列映射两个方面。前者主要以情感词集与冷语料文本为实验输入,在遍历词集中情感词汇、文本内段落及其初始类别(全0)的基础上,通过文本段落内情感词汇的索引实现术语匹配,并将成功匹配到词汇的文本类别标记为1,进而输出文本片段及其类别;后者以文本片段及其类别为实验输入,通过文本片段中单字区间的截取实现字序列映射,进而输出字序列与角色序列,最终在冷环境下自动获取古诗领域中具备情感标注集的学习语料。

2.4 Char2Vec 驱动下汉字语言特征的引入与 CRFs 模型构建

初始条件下学习语料包含字序列(C)与标注角色序列(R),而 BERT 语言学模型正是依凭 Char2Vec 的单字向量解析充分挖掘中文语言知识。受此启发,以 CRFs 为代表的传统机器学习是否基于汉字结构和汉语知识扩展语义特征空间以提升模型表现效果,还有待深入探索。对此,笔者重点发掘了汉字的核心语言特征,如图3所示。

由图3可知,笔者在语言学视角下为汉字定义了部首特征、拼音特征、情感特征、领域特征等八种汉字属性作为扩展以优化模型学习,其特征标记、取值如表1所示。

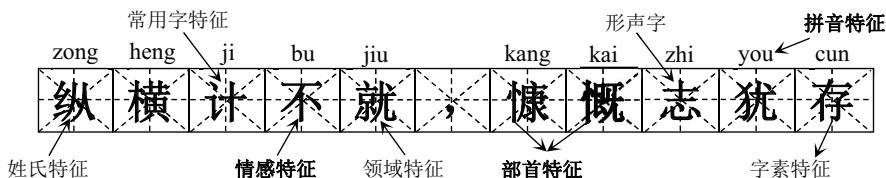


图3 古诗文本汉字语言特征解析

表1 基于古诗文本的汉字语言特征归纳

汉字特征		部首特征(B)		拼音特征(P)			字素特征(M)			情感特征(E)																
描述		汉字字形组件		汉字发音组件			字素的词性倾向			一级情感		二级情感		其他												
取值		部首字符串		拼音字符串			词性字符串			X		Y		N												
汉字特征		领域特征(F)		常用字特征(U)			姓氏特征(N)			汉字分类特征(T)																
描述		常用字	其他	一级字	二级字	其他	常用姓	罕用姓	其他	指事	象形	形声	会意	其他												
取值		Y	N	X	Y	N	X	Y	N	X	Y	Z	U	N												
特征空间样例	<u>C</u>		委	<u>婉</u>	<u>地</u>	<u>透</u>	<u>露</u>	了	诗	人	有	苏	秦	张	仪	之	雄	才	而	不	逢	明	主	的	苦	恼
	B		女	女	土	<u>之</u>	雨	乙	人	月	禾	弓	彳	彳	彳	佳	才	而	一	<u>之</u>	日	、	白	彳	彳	
	P		wei	wan	di	<u>lou</u>	lu	le	shi	ren	you	su	qin	zhang	yi	zhi	xiong	cai	er	bu	feng	ming	zhu	de	ku	nao
	M		zg	a	uv	<u>v</u>	v	ul	n	n	v	j	nr	q	ng	u	n	d	c	d	zg	a	b	uj	a	zg
	E		N	N	X	<u>N</u>	N	Y	N	X	X	N	N	N	N	X	N	X	X	X	N	X	Y	X	X	N
	F		Y	Y	Y	<u>Y</u>	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
	U		X	Y	X	<u>X</u>	X	X	X	X	X	X	Y	X	X	X	X	X	X	X	X	X	X	X	X	X
	N		Y	N	N	<u>N</u>	N	N	Y	X	X	X	X	X	Y	Y	N	Y	N	Y	Y	X	N	N	Y	N
	T		U	Z	Z	<u>Z</u>	Z	Y	Z	Y	U	Z	U	Z	Z	X	Z	X	X	X	Z	U	Y	Z	Z	Z
	<u>R</u>		B	E	O	<u>B</u>	E	O	O	O	O	O	O	O	O	O	S	O	O	O	O	S	O	O	B	E

(1) 汉字特征扩展。①部首是汉语字典中汉字字形和意义分类的依据,相对于汉字,部首是一种更细粒度的中文语义表征方式;②拼音是汉字的发音组件,其中读音也具备较强情感色彩;③情感特征是汉字的情感倾向,可以以本文构建的情感词集为知识元,区分出常用情感字级别,作为汉字情感倾向的判断依据;④字素特征是指汉字在特定语境中的词性倾向;⑤领域特征是指领域语料中常用汉字的类别区分;⑥常用字特征是指汉语通用常用字类别区分;⑦姓氏特征用来判断汉字与姓氏字的关联;⑧分类特征是汉语造字的抽象依据。笔者在本文语料基础上引入上述汉字的特征作为实验对比,探讨汉字语言特征扩展的鲁棒性。需要说

明的是,特征 B/P/M/U/N/T 均通过外部字典匹配获得,而 E/F 则依靠字频统计与后续实验调参选优获得。

(2) 特征空间约束。将所有汉字特征加入初始观察序列便形成基于字序列的基准模型,在机器学习过程中存在横向与纵向两个方面的约束。在表1的特征空间样例中,前者(横向虚线)指字长窗口,常见三字与五字两种方案;后者(纵向虚线)旨在对观察序列特征组合进行约束。对此笔者在单特征基础上采取正向叠加的方式获取多特征序列,将超过基准模型结果的单特征记为正向特征,并对所获正向特征集合进行组合再训练,如组合测试结果超过正向特征,则继续扩展组合特征维度,如此多层次递

进。横向与纵向序列是基于上下文信息与语言特征约束模型的训练结果,而模型的具体成效则有待后续实验论证。

2.5 基于机器学习和深度学习模型的情感术语抽取算法

在 BERT 模型的 Char2Vec 指导下,本文构建融入汉字语言特征的 CRFs 机器学习模型和 BERT-BiLSTM-CRFs 深度学习模型,其核心算法如下。

(1) CRFs 算法。对于输入的字序列,输出角色序列与序列分数;然后,将所有序列路径归一化得到角色序列的概率分布,并使用序列级似然函数求解其最大后验概率,进而提升标签预测的准确性;最终,以解码阶段预测总得分最高的序列作为最优输出序列。

(2) BERT-BiLSTM-CRFs 算法,将 BERT 语言学模型嵌入 BiLSTM-CRFs 深度学习模型,实现深层特征的自动提取。其中,BERT 以汉字、句子和位置作为文本嵌入,通过 Char2Vec 驱动神经网络更好地捕获汉字语义特征;BiLSTM 层用于对 BERT 层输入的字向量进行双向编码,以表征上下文相关的语义信息;CRFs 层能够对神经网络所训练的语义向量进行解码以输出概率最大的标签序列,最终获得字—角色空间以抽取情感术语。

2.6 情感术语抽取评价指标与新术语鉴定规则

(1) 情感术语抽取的评价指标,主要包括召回率(R)、精确率(P)和调和平均值($F_{measure}$)。

原始术语计算方式为: $R = \frac{TP}{TP + FN}$, $P = \frac{TP}{TP + FP}$, $F_{measure} = \frac{(n2 + 1) RP}{R + n2P}$, FN 、 TP 、 FP 分别代表未识别、正确识别与错误识别的术语数,本研究认为 R 与 P 同等重要,故设 n 为 1。然而,由于原始术语中的重复项会使得识别率虚高,又额外引入区分性(distinct)术语指标,即非

重复性术语,用于考察模型对不同术语的辨别能力,可通过对原始术语去重后获得,再计算各评价标准。

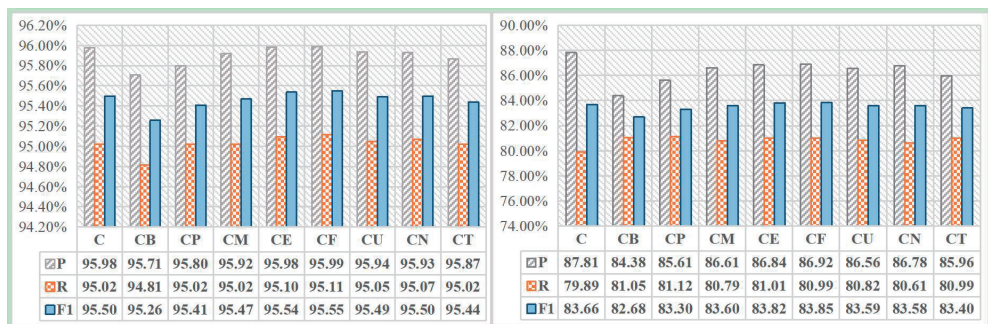
(2) 新术语识别规则。区别于利用词集进行术语匹配,采用机器学习模型能够进一步抽取出领域新术语,为情感知识的发现奠定基础。笔者通过预实验对标注和预测序列展开相对性观察,发现:①标注序列“悼/O/伤/O”的预测序列为{B/E},可定义为新发现的情感术语;②标注序列“凄/S/戚/O”的预测序列为{B/E},可定义为增量新术语;③标注序列“深/B/情/E/厚/S/志/O”的情感单元包括“深情”“厚”,而预测序列为{B/M/M/E},可融合为复合型术语“深情厚志”。因此,本文归纳出新术语识别的规律:对于预测序列抽取的术语,若标注序列首字角色为 B/S/O,尾字角色为 E/S/O,并且标注序列不等同于预测序列,则将其作为领域新术语候选。

3 实验结果与分析

本文以古诗领域构建的情感词集为术语集(40 173 个),以诗文与鉴赏相融合(诗文:鉴赏=1:14)的古诗文本为语料,由于本文抽取的术语旨在应用于已有古诗的情感分析,故在主体实验中不再引入外部诗文语料。采取冷启动的方法匹配情感术语共 298 400 个,将学习语料按照 4:1 的比例分配形成训练集和测试集,运用 Python 3.7 工具、CRFsuite 模块、TensorFlow 1.15 框架分别构建机器学习与深度学习模型,以实现古诗领域大规模人文情感术语的自动化抽取。

3.1 基于汉字特征扩展的人文情感术语抽取结果

(1) 单特征扩展。首先基于三字与五字长窗口进行预实验,前者 P、R、F1 分别为 95.98%、95.02%、95.50%,均高于后者的 95.21%、94.01%、94.60%,说明更大字长会引入一定噪音特征,故选取三字编码,并根据表 1 中的汉字特征训练 CRFs 算法,如图 4 所示。



(a) 原始术语抽取结果计算

(b) 区分性术语抽取结果计算

图4 基于单特征扩展的情感术语抽取结果计算

在图4中,C与B/P/M/E/F/U/N/T中任一特征的集合代表了以字序列为基准的单特征扩展的实验方案。①baseline(C)的原始术语F1值高达95.50%,已达到十分不错的水准,不过区分性术语的F1值仅有83.66%,说明重复性术语优化了模型表现效果。②从P值来看,两种评价标准下特征集合的准确率大多低于baseline。③从R值来看,领域(F)、情感(E)、常用字(U)以及姓氏(N)特征能够较为明显地提升原始术语召回率;在区分性术语识别中,所有特征均能有效提升模型的召回效果,此时效果最好的为拼音(P)与部首(B)特征。④从F1值来看,F/E的扩展在不同指标下均优于baseline。综上,特征扩展使模型捕获了丰富的汉字特征,有效提升了术语的识别效果。其中,F/E对重复性术语识别的稳定性更好,P/B对区分性术语识别的增量性更强。此外,结合②、③可以发现,F1的表现取决于R和P的变化幅度,这说明当引入外部特征时,P值的降低往往能够带来R值的提高,在本实验中也反映出F和E鲁棒性的特征。

(2)约束特征扩展。考虑到③中所述B与P特征在召回上的优异表现,故对其进行进一步探索。细究发现,导致F1劣于baseline的主要原因是错误术语大量增多,这是由于候选新术语在实验中往往被直接计入错误识别之列,极大地弱化了模型表现效果。此外笔者推断,由

于二者的特征取值由字符串组成,而过多变量容易降低模型的准确性,故需要对特征取值进行合理的统计约束。又因为④中已证实特征F和E的有效性,故本文基于F/E对B与P建立约束,统计古诗语料以及情感词集中常用的拼音与部首,构建新特征进行训练测试,结果如图5所示。

由图5可知,对B/P建立汉字的领域和情感约束后,模型表现效果均上升到与baseline相当的水平。就P而言,经过实验调参对比后,将领域汉字拼音中频率前25名作为领域常用拼音(F_P),情感汉字中频率前60名作为情感常用拼音(E_P),模型F1值均超过baseline,这验证了拼音特征的正向作用;另外,统计领域文本及情感词集中的汉字部首,在参数调整后将领域部首前115名作为领域常用部首(F_B),情感部首前20名作为情感常用部首(E_B),得到的最优F1值依旧略逊于baseline。对此笔者分析术语抽取数目以探究其缘由,如表2所示。

由表2可知,利用F/E对P/B特征约束后的模型在术语抽取的召回上产生了较为明显的提升效果。①两种评估指标下,F_B和E_B正确识别与新识别数总体上优于F_P及E_P,表明部首特征能够更好地召回情感术语并发掘新术语。②E_B在原始术语召回和新术语识别上均优于F_B,说明情感部首的约束更为有效,故笔者观察了E_B集合内的二十个汉字部首,并启发

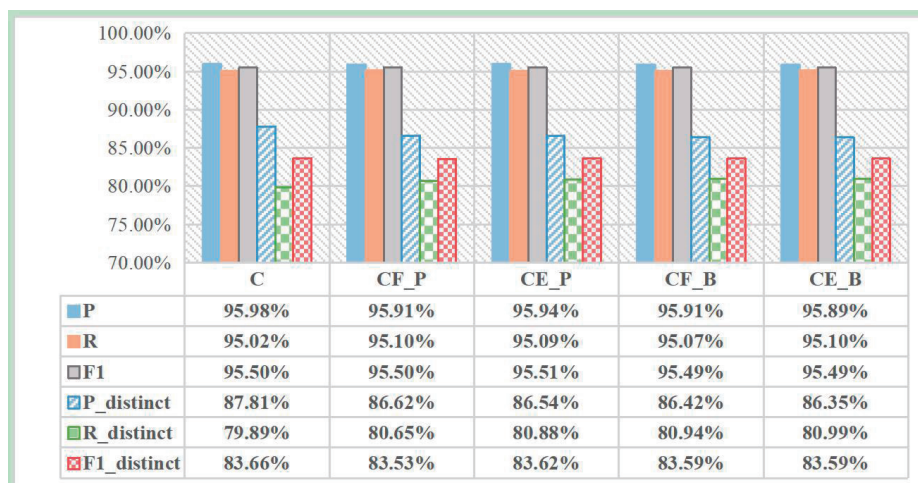


图5 基于约束特征扩展的情感术语抽取结果计算

表2 基于约束特征扩展的情感术语数目抽取结果

约束 特征扩展	原始评估标准				区分性评估标准			
	原有术语	识别术语	正确识别	新识别数	原有术语	识别术语	正确识别	新识别数
C	63 151	62 519	60 005	796	7 045	6 409	5 628	647
F_P	63 151	62 618	60 057	865	7 045	6 560	5 682	742
E_P	63 151	62 592	60 051	893	7 045	6 584	5 698	747
F_B	63 151	62 601	60 039	894	7 045	6 598	5 702	752
E_B	63 151	62 632	60 057	912	7 045	6 608	5 706	761

性地以其中{“↑”,“心”}为约束特征展开实验,F1值高达95.54%,跃升至整体实验第二,仅次于领域特征;F1_distinct也达到83.69%,显著高于baseline、B、F_B和E_B,验证了部首特征的有效性。在此基础上,笔者以{“↑”,“心”}为基准,将另外18个部首分别加入集合进行增量实验,结果如图6所示。

图6为以{“↑”,“心”}为基准部首集合的计算结果,“+部首”为其增量扩展的实验结果,图中F1值以顺时针方向倒序排列。经过分析可知:①所有部首特征的增量扩展均可使R与R_distinct优于baseline;②从F1来看,增量部首的作用排序前五为:“一”>“大”>“目”>“宀”>“讠”,其中优于baseline的部首特征为“一”

“大”“目”;③F1_distinct前五位为:“目”>“人”>“宀”>“讠”>“大”,其中部首“目”优于baseline;④综合来看,使用{“↑”,“心”}为部首集合进行情感抽取实验的整体表现依旧保持最好。

(3)多特征扩展。对实验结果优于baseline的单特征采取正向叠加的方式获取多特征序列,包括E/F/P/B。其中,拼音特征选择约束特征中表现更优的E_P,部首特征则以{“↑”,“心”}为集合约束特征。基于上述特征形成六组多特征方案进行实验,结果显示无一超过baseline。对此笔者推断,多特征的引入并不一定利于模型的优化,需由具体领域语料特征而定。

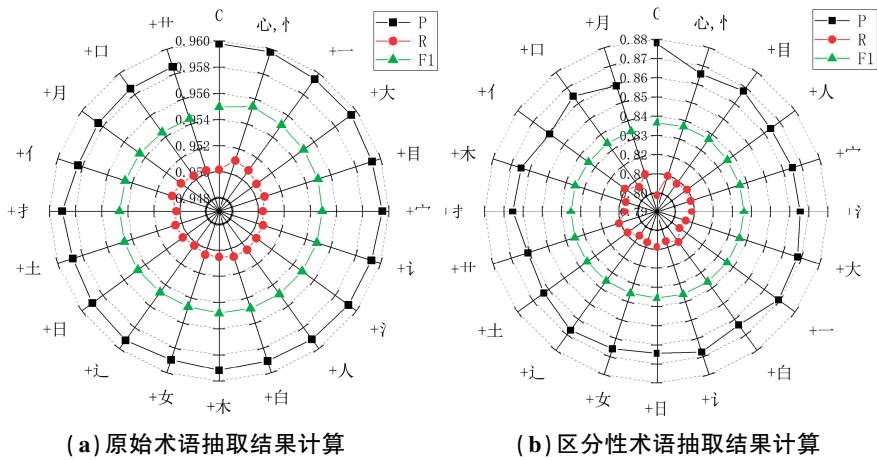


图6 基于{“心”, “心”}基准扩展的情感术语抽取结果计算

3.2 基于迁移学习的人文情感术语抽取结果

汉字特征扩展的提升效果相对于 baseline 较为有限。对此本文进一步搭建 BERT-BiLSTM-CRFs 深度学习模型。考虑到 BERT 模型的可迁移特性,为防止训练轮数不足或过量所

引发的模型欠拟合或过拟合,笔者将单个周期模型训练次数设置为 10,通过多周期训练方式对模型进行多重迁移并计算每次迁移的识别效果,进而归纳出领域最优模型,具体超参数配置如表 3 所示。

表3 BERT-BiLSTM-CRFs 模型训练的超参数配置

超参数名称	超参数取值	超参数名称	超参数取值
batch_size	64	dropout_rate	0.5
max_seq_length	128	lstm_size	128
learning_rate	2.00E-05	num_train_epochs	10

(1) 基于 BERT 模型的 Char2Vec 训练。首先对学习语料进一步分割,从训练集中抽取整体语料的 1% 作为验证集。接着,采用谷歌提

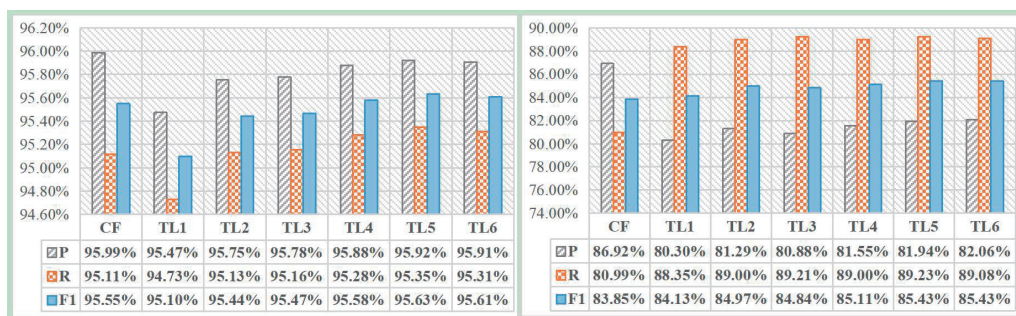
供的预训练模型作为初始迁移模型。最后,通过本文语料的字嵌入以及预训练模型的微调实现 Char2Vec,结果如表 4 所示。

表4 基于 BERT 的古诗文本语料字嵌入(示例)

tokens	纵	横	计	不	就	,	慷	慨	志	犹	存	...
input_ids	101	5 288	3 566	6 369	679	2 218	117	2 724	2 717	2 562	4 310	...
input_mask	1	1	1	1	1	1	1	1	1	1	1	...
segment_ids	0	0	0	0	0	0	0	0	0	0	0	...
label_ids	8	7	2	4	7	2	4	7	2	4	4	...

(2) 情感术语抽取结果计算。以机器学习最优模型 CF 为 baseline,训练 BERT-BiLSTM-

CRFs 深度学习模型,并利用每次迁移训练的分类器进行测试,结果如图 7 所示。



(a) 原始术语抽取结果计算

(b) 区分性术语抽取结果计算

图7 基于BERT-BiLSTM-CRFs的情感术语抽取结果计算

由图7可知:①从整体上看,随着迁移次数的增加,模型在情感术语抽取中的表现愈发优秀,在不同标准下均超过了baseline的效果。②对原始术语抽取结果分析可知:就P值而言,baseline始终保持最高;就R值而言,从TL1之后便一直高于baseline;就F1值而言,模型从TL4开始超越baseline,随后一直稳定在较高水平(95.63%)。自此,深度学习模型对原始术语抽取的水准已明显优于机器学习模型。③对区分性术语抽取结果分析可知:就P值而言,其最高值仅为82.06%,远逊色于CF(86.92%);就R值而言,模型的TL1结果大大优于CF,达到88.35%,同比上升7.36%,刷新了本文实验中的

所有记录;就F1值而言,在P值与R值抵消下模型仍然在TL1处超越CF,并稳步提升至TL6(85.43%)。④TL5在不同标准下的F1值均达到了峰值,而TL6结果相较TL5稍逊,这说明模型有过拟合趋势,因此TL5即为最优模型。

(3)不同古诗文本语料术语抽取差异性分析。本文对最优模型TL5的测试结果进行拆分,从中分别抽取诗文和鉴赏序列后计算各类指标;此外,为了验证本文语料的有效性,笔者在外部17365首诗文本所构建的新训练集上重新训练深度学习模型,并在拆分的诗文序列上进行测试,通过二十轮迭代达到最优抽取结果,如表5所示。

表5 不同古诗文本语料的情感术语抽取结果计算

训练集	测试集	原始评估标准				区分性评估标准			
		P(%)	R(%)	F1(%)	新识别数	P(%)	R(%)	F1(%)	新识别数
诗文+鉴赏	诗文+鉴赏	95.92	95.35	95.63	941	81.94%	89.23%	85.43%	836
诗文+鉴赏	鉴赏	95.85	95.26	95.56	888	82.32%	89.34%	85.68%	805
诗文+鉴赏	诗文	96.69	96.25	96.46	53	90.74%	92.61%	91.66%	38
外部诗文	诗文	98.47	98.14	98.31	27	95.81%	94.18%	94.99%	25

由表5可知:①诗文和鉴赏文本中原始术语的测试效果相近,均保持良好的识别效率;而对于区分性术语而言,诗文的效果则要明显优于鉴赏,差距主要体现在准确率上(相差8.42%),这说明诗文语料的凝练特性对于区分

性术语更具优势。②相比本文语料,外部诗文的模型训练达到了更优精度,可见更精准的语料环境更契合诗文中情感术语的抽取;不足之处在于新术语识别效果稍逊,这说明鉴赏中更丰富的文本特征更利于情感知识发现。此外,

经笔者统计,综合训练集中的区分性术语数量(12 151)是外部诗文训练集(6 259)的两倍,故前者更有助于模型充分学习特征。综上所述,采取“诗文+鉴赏”的古诗文本语料能够保证较高水准的情感术语抽取效果,鉴赏文本的引入不仅可以有效优化情感术语的数量与粒度,更有益于情感新词的知识发现。

3.3 古诗文本的新情感术语发现

机器学习技术在术语抽取中的更深层价值在于未登录词汇的知识发现。笔者基于本文对

新术语鉴别的规则,分别利用最优机器学习(CF)与深度学习(TL5)模型抽取新情感术语,分别获得1 057和1 531个候选新词。其中,机器学习识别的新词最大长度为8,深度学习为4。另外,后者识别的两字新词数量(1 101)是前者(481)的两倍有余,识别的四字新词数量(339)也大幅领先前者(190)。根据上述情况,笔者率先统计出机器学习抽取的长篇幅新词,随后从深度学习识别的新词集合中抽取出不包含的高频术语,并在此基础上对术语的来源文本进行区分,如表6所示。

表6 古诗文本新情感术语候选(部分)

基于机器学习的长篇幅新术语				基于深度学习的拓展性新术语			
候选新词(诗文)	词长	候选新词(鉴赏)	词长	候选新词(诗文)	词频	候选新词(鉴赏)	词频
有才过屈宋	5	王孙公子肆无忌惮	8	彩凤	2	沦谪	11
可笑不自量	5	不甘于无所作为	7	肺腑无言	1	秀艳	7
挺立不动膝	5	惶惶若丧家无主	7	东风日暖	1	离席	5
自献自为酬	5	花心如痴如醉	6	千里山河	1	末世	5
艰极泰循	4	依依惜别之情	6	薄俗	1	欢情	3
举目凄凉	4	风格清新俊逸	6	浮世	1	轻扬	3
恨不胜	3	不期然而然的	6	多雨	1	亡灵	3
自作为	3	变得模糊不清	6	惜金	1	春光易逝	2
惊满座	3	决胜千里之外	6	春泉	1	慷慨不平	2
神通力	3	明珠交相辉耀	6	娉婷	1	忍辱饮恨	2
辛苦力	3	放荡无不拘的	6	濡染	1	名盛一时	2
奔亡	2	寄情山水自然	6	惨澹	1	征人思妇	2
忍耻	2	美人迟暮之恨	6	愬武	1	有志难伸	2
惜惧	2	文采风流而生	6	私债	1	阶下囚	2
冤谪	2	鸣声有如不如	6	角逐	1	被贬	2
抚膺	2	痛感无能为力	6	霸图	1	唐军	2
威怒	2	活生生感染力	6	丧逝	1	宦官	2
恳苦	2	年华迟暮之悲	6	兵荒	1	哀怆	2

由表6可知,机器学习与深度学习模型所抽取的领域新词总体上具备较好的情感倾向。

其中,前者能够在诗文中识别包含情感内涵的诗句,术语凝练性更强;同时亦可在鉴赏中抽

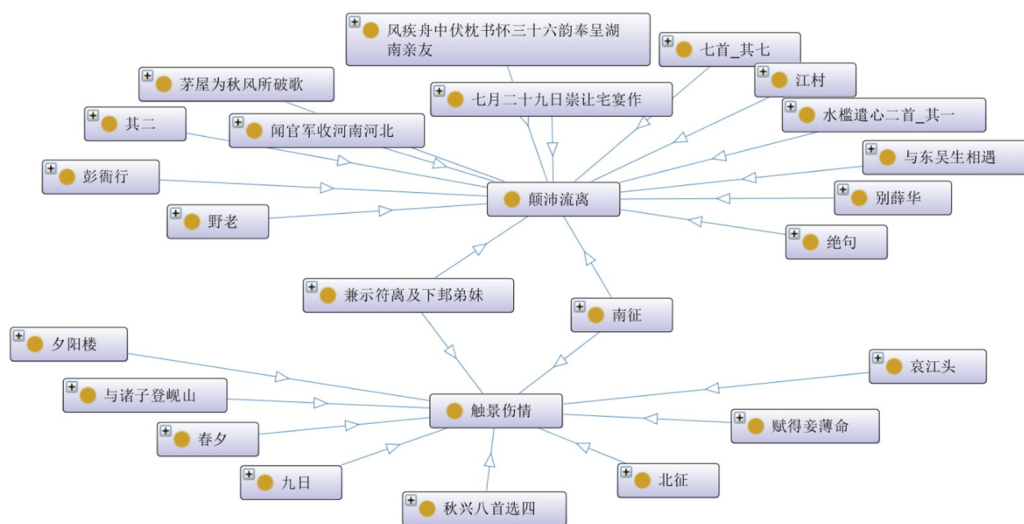


图9 基于“以词查诗”的人文情感术语检索(示例)

双术语“颠沛流离”与“触景伤情”进行缩检,能够精炼出融合二者情感的诗篇《兼示符离及下邳弟妹》和《南征》。其中,前者旨在通过“雁行”描绘诗人与失散兄弟如离群的孤雁那样孤苦无依、颠沛流离,触发对亲人的思念之情;后者描绘诗人面对眼前美景叹息自身光景不多、前途渺茫,表达长年颠沛流离、远适南国的羁旅悲愁。

4.2 人文情感粒度挖掘

人文情感粒度挖掘是指从不同层次文本中解析多粒度的情感内涵。古诗情感知识粒度表现为“诗题”>“诗文”>“鉴赏”并逐层细化,笔者以诗文《述怀》为例抽取三种文本中的情感术语。随后,以此三种文本为元数据标签、以情感元为属性、以情感术语为属性值绘制知识图谱,如图10所示。

图10“以小见大”地展示了诗文《述怀》中情感知识元的细化轨迹。①诗题中的情感知识元为术语单字“怀”,说明该诗旨在陈述诗人的胸襟与抱负。②诗文中的情感知识元大致可分为三类:其一为“纵横”“慷慨”,表现了诗人以天下为己任的博大胸怀;其二为“艰险”“寒”

“啼”,阐述了任务的艰难以及征程环境的恶劣;其三为“重”“意气”,反映了诗人重视信义,纵使前路坎坷也绝不负使命。③鉴赏内的情感知识元是对上述三种情感和思想的映射与细化。就志向而言,“从军”“济世”“安邦定国”“建功立业”等术语揭示出诗人怀揣弃文从武、为国效力、成就一番丰功伟绩的抱负;就环境而言,术语“荒凉”“崎岖不平”“巍峨”“哀鸣”突出了征途的艰险,“凶”“急迫”反映了任务的凶险与紧急,二者反映了在恶劣环境下诗人内心的此起彼伏;就信义而言,术语“义无反顾”“不畏艰险”“百折不挠”“季布一诺”表明诗人在艰难险阻面前依旧恪守信义,表达出不负使命的决心。

基于上述分析,实际应用中便可对鉴赏、诗文、诗题内多层次情感术语进行自底向上的解析。其中,鉴赏中的细粒度情感术语用以辅助诗文中情感词汇的判读,诗文内中等粒度术语支撑诗题内粗粒度情感术语的泛化。如《述怀》中的“怀”字,在经过粒度映射后可归纳为诗人矢志报国、不畏艰险、慷慨重义的胸襟抱负,如此可为古诗领域的人文研究提供有效的知识服务。

然”(108)、“友人”(37)、“权贵”(37)、“思妇”(37)、“理想”(37)、“苦闷”(37)、“黑暗”(37)、“送别”(36)等,折射其情感色彩多集中于对亲友的思念以及对统治集团荒淫腐败的批判。②孟浩然诗文中共抽取4 775个情感术语,高频情感词为“自然”(36)、“景物”(17)、“朋友”(12)、“隐居”(11)、“垂钓”(9)、“平淡”(9)、“隐逸”(9)、“名山”(8)等,多反映诗人钟情于山水且不媚世俗、修道归隐的高洁之风。③杜甫诗文中共抽取32 483个情感术语,其中高频情感术语主要为“自然”(102)、“人民”(68)、“战乱”(43)、“遭遇”(35)、“凄凉”(31)、“沉郁”(25)、“思念”(25)、“多病”(23)等,说明其诗多涉笔社会动荡、政治黑暗、人民疾苦,承载了忧国忧民的现实主义色彩。④李商隐的诗文中共抽取17 684个情感术语,高频情感词汇为“身世”(54)、“爱情”(52)、“自然”(48)、“相思”(31)、“遭遇”(23)、“痛苦”(23)、“寂寞”(21)、“凄凉”(21)等,表明其诗文主要聚焦于男女爱情相思,多表达诗人形单影只的寂寞和对妻子的深切思念之情。⑤四位诗人的高频情感术语中均出现了“自然”一词,体现出汉语古诗文善于融情于景的显著特征。

5 结语

本文立足于数字人文的“人文性”与“计算性”特质,针对古诗及鉴赏文本缺少学习语料的

现状,提出一种基于“冷启动”的字序列自动标注方法,将汉字语言知识引入机器学习和深度学习模型中,实现了大规模情感术语的自动抽取与人文应用。研究发现,机器学习中领域、情感、部首和拼音等汉字特征的引入有效提升了模型效果,以领域特征和基于“竖心旁”“心字底”的部首约束特征最优,而深度学习的表现则更为出色,以区分性术语的召回效果最为明显。此外,相较于机器学习抽取出的长篇幅新术语,深度学习拓展出更多寄托情感知识的新意象词,与领域术语融合后形成14 599个区分性术语,为古诗领域情感词典的构建与情感分析奠定了基础。

本研究还存在可深化之处。其一,自动标注语料采用的情感词集是一种有限规模的知识库,难以避免标注不准确、不充分等问题,后续将引入增量迭代和远程知识库检验进行优化;其二,本文对部首特征的扩展仅为初步探索,未来将尝试对部首进行知识组织,基于不同的部首语义内涵来验证其在特征扩展上的有效性;其三,情感术语含有多种类别,如“喜”“怒”“哀”“思”“赞”“讽”“惊”等,且在古诗领域会更为丰富,而本文主要关注在“冷环境”下对古诗文本整体情感术语抽取从0至1的突破,并未细分情感类别,后续研究将聚焦于情感术语的自动化分类,致力于古诗领域情感词典的构建及其情感分析应用。

参考文献

- [1] Poole A H, Garwood D A. Interdisciplinary scholarly collaboration in data-intensive, public-funded, international digital humanities project work[J]. Library & Information Science Research, 2018, 40(3-4): 184-193.
- [2] 刘伟, 叶鹰. 数字人文的技术体系与理论结构探讨[J]. 中国图书馆学报, 2017, 43(5): 32-41. (Liu Wei, Ye F Y. Exploring technical system and theoretical structure of digital humanities[J]. Journal of Library Science in China, 2017, 43(5): 32-41.)
- [3] 黄水清. 人文计算与数字人文: 概念、问题、范式及关键环节[J]. 图书馆建设, 2019(5): 68-78. (Huang Shuiqing. Humanity computing and digital humanities: concept, problem, paradigm and key step[J]. Library Development, 2019(5): 68-78.)

- [4] Li G, Li J. Research on sentiment classification for Tang poetry based on TF-IDF and FP-Growth[C]//2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2018:630-634.
- [5] 梅莉莉, 黄河燕, 周新宇, 等. 情感词典构建综述[J]. 中文信息学报, 2016, 30(5):19-27. (Mei Lili, Huang Heyan, Zhou Xinyu, et al. A survey on sentiment lexicon construction[J]. Journal of Chinese Information Processing, 2016, 30(5):19-27.)
- [6] 易勇, 何中市, 李良炎, 等. 基于遗传算法改进诗词风格判别的研究[J]. 计算机科学, 2005(7):156-158. (Yi Yong, He Zhongshi, Li Liangyan, et al. A traditional Chinese poetry style identification calculation improvement model[J]. Computer Science, 2005(7):156-158.)
- [7] 胡韧奋, 诸雨辰. 唐诗题材自动分类研究[J]. 北京大学学报(自然科学版), 2015, 51(2):262-268. (Hu Renfen, Zhu Yuchen. Automatic classification of Tang poetry themes[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2):262-268.)
- [8] 罗凤珠, 李元萍, 曹伟政. 中国古代诗词格律自动检索与教学系统[J]. 中文信息学报, 1999, 13(1):36-43. (Luo Fengzhu, Li Yuanping, Cao Weizheng. The format auto-checking and database indexing teaching system of Chinese poetry and lyrics[J]. Journal of Chinese Information Processing, 1999, 13(1):36-43.)
- [9] 吴斌, 吉佳, 孟琳, 等. 基于迁移学习的唐诗宋词情感分析[J]. 电子学报, 2016, 44(11):2780-2787. (Wu Bin, Ji Jia, Meng Lin, et al. Transfer learning based sentiment analysis for poetry of the Tang Dynasty and Song Dynasty[J]. Acta Electronica Sinica, 2016, 44(11):2780-2787.)
- [10] Tang Y, Wang X, Qi P, et al. A neural network-based sentiment analysis scheme for Tang poetry[C]//2020 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2020:1783-1788.
- [11] 陶幼琴. 从“书友沙龙”探索用户知识资源的开发应用[J]. 图书馆杂志, 2019, 38(12):67-71. (Tao Youqin. Exploring the development and application of readers' knowledge from the "readers' salon"[J]. Library Journal, 2019, 38(12):67-71.)
- [12] Xing F Z, Pallucchini F, Cambria E. Cognitive-inspired domain adaptation of sentiment lexicons[J]. Information Processing & Management, 2019, 56(3):554-564.
- [13] Huang S, Niu Z, Shi C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation[J]. Knowledge-Based Systems, 2014, 56:191-200.
- [14] Wu S, Wu F, Chang Y, et al. Automatic construction of target-specific sentiment lexicon[J]. Expert Systems with Applications, 2018, 116:285-298.
- [15] 李枫林, 范雅娴. 领域情感词典构建方法研究[J]. 图书馆理论与实践, 2019(12):60-65. (Li Fenglin, Fan Yaxian. Research on construction method of domain sentiment lexicon[J]. Library Theory and Practice, 2019(12):60-65.)
- [16] Viegas F, Alvim M S, Canuto S, et al. Exploiting semantic relationships for unsupervised expansion of sentiment lexicons[J]. Information Systems, 2020, 94:101606.
- [17] Wu F, Huang Y, Song Y, et al. Towards building a high-quality microblog-specific Chinese sentiment lexicon[J]. Decision Support Systems, 2016, 87:39-49.
- [18] 王昊, 王密平, 苏新宁. 面向本体学习的中文专利术语抽取研究[J]. 情报学报, 2016, 35(6):573-585. (Wang Hao, Wang Miping, Su Xinning. A study on Chinese patent terms extraction for ontology learning[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(6):573-585.)
- [19] 王科, 夏睿. 情感词典自动构建方法综述[J]. 自动化学报, 2016, 42(4):495-511. (Wang Ke, Xia Rui. A

- survey on automatical construction methods of sentiment lexicons[J]. Acta Automatica Sinica, 2016, 42(4): 495-511.)
- [20] Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network [J]. Knowledge-Based Systems, 2016, 108: 42-49.
- [21] Yang H, Zeng B, Yang J H, et al. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction[J]. Neurocomputing, 2021, 419: 344-356.
- [22] 李纲, 刘广兴, 毛进, 等. 一种基于句法分析的情感标签抽取方法[J]. 图书情报工作, 2014, 58(14): 12-20. (Li Gang, Liu Guangxing, Mao Jin, et al. A sentiment label extraction method based on dependency parsing [J]. Library and Information Service, 2014, 58(14): 12-20.)
- [23] 郁圣卫, 卢奇, 陈文亮. 基于领域情感词典特征表示的细粒度意见挖掘[J]. 中文信息学报, 2019, 33(2): 112-121. (Yu Shengwei, Lu Qi, Chen Wenliang. Fine-grained opinion mining based on feature representation of domain sentiment lexicon[J]. Journal of Chinese Information Processing, 2019, 33(2): 112-121.)
- [24] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报, 2016, 42(2): 66-80. (Ouyang Jian. Visual analysis and exploration of ancient texts for digital humanities research[J]. Journal of Library Science in China, 2016, 42(2): 66-80.)
- [25] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 4171-4186.
- [26] López-Santillán R, Montes-Y-Gómez M, González-Gurrola L C, et al. Richer document embeddings for author profiling tasks based on a heuristic search[J]. Information Processing & Management, 2020, 57(4): 102227.
- [27] 肖连杰, 孟涛, 王伟, 等. 基于深度学习的情报分析方法识别研究——以安全情报领域为例[J]. 数据分析与知识发现, 2019, 3(10): 20-28. (Xiao Lianjie, Meng Tao, Wang Wei, et al. Entity recognition of intelligence method based on deep learning: taking area of security intelligence for example[J]. Data Analysis and Knowledge Discovery, 2019, 3(10): 20-28.)
- [28] 吴俊, 程垚, 郝瀚, 等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究[J]. 情报学报, 2020, 39(4): 409-418. (Wu Jun, Cheng Yao, Hao Han, et al. Automatic extraction of Chinese terminology based on BERT embedding and BiLSTM-CRF model[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(4): 409-418.)
- [29] Hong L, Hou W, Wu Z, et al. A cooperative crowdsourcing framework for knowledge extraction in digital humanities-cases on Tang poetry[J]. Aslib Journal of Information Management, 2020, 72(2): 243-261.
- [30] 胡家珩, 岑咏华, 吴承尧. 基于深度学习的领域情感词典自动构建——以金融领域为例[J]. 数据分析与知识发现, 2018, 2(10): 95-102. (Hu Jiaheng, Cen Yonghua, Wu Chengyao. Constructing sentiment dictionary with deep learning: case study of financial data[J]. Data Analysis and Knowledge Discovery, 2018, 2(10): 95-102.)

张 卫 南京大学信息管理学院博士研究生。江苏 南京 210023。

王 昊 南京大学信息管理学院教授, 博士生导师。江苏 南京 210023。

邓三鸿 南京大学信息管理学院教授, 博士生导师。江苏 南京 210023。

张宝隆 南京大学信息管理学院博士研究生。江苏 南京 210023。

(收稿日期: 2020-08-02; 修回日期: 2020-11-07)