面向情感分析的汉语构式语料库构建与应用研究
—对汉语构式情感分析问题的思考

吴尹清
国防科技大学国际关系学院/江苏南京
wu.yinqing@outlook.com

李德俊
国防科技大学国际关系学院/江苏南京
njlide@sina.cn

摘要

文本情感分析又称为意见挖掘，是基于网络大数据对评价主体倾向性的研究。由于其在舆情监控、市场营销、金融等应用领域的特殊意义，近年来受到了越来越广泛的关注。本文关注情感分析面临的语义隐匿性问题，通过构建一个汉语构式语料库，对语料库中的汉语构式进行量化统计，讨论汉语构式与情感分析之间的关系。文章对语料库中表达量级和态度义的构式与词汇进行了标注，并基于该语料库对相关构式和词汇进行了计量分析，按照构式类型、语义类别、常项变项个数等标准统计了语料库中量级和态度义构式的信息，并与量级和态度义词汇的统计信息进行了比对，通过分析构式表义比重和词汇表义比重这两个指标，发现语料库中词汇承载了大部分态度和量级语义信息，构式所承载的态度和量级语义信息较少。虽然构式不是主要的表义单位，但其承载的态度语义信息仍占一定比例。文章为构式语法应用于汉语情感分析提供了实证数据，为后续该类研究提供了一种方法，也为汉语构式研究提供了基于汉语真实文本的数据。文章还专门探讨了目前构式语法应用于汉语情感分析乃至自然语言处理所面临的困难，对后续研究提出了展望。

关键词： 构式语法；汉语情感分析；构式语料库；态度义；量级义

A Study of Chinese Construction Corpus Compilation and Application for Sentiment Analysis: A Discussion of Sentiment Analysis Problems of Chinese Constructions

Wu Yinqing
College of International Studies,
National University of Defense
Technology / Nanjing, Jiangsu
wu.yinqing@outlook.com

Li Dejun
College of International Studies,
National University of Defense
Technology / Nanjing, Jiangsu
njlide@sina.cn

Abstract

Sentiment analysis, also called opinion mining, is a field that studies the sentiment orientation of the evaluation subjects based on Web data. Due to its significance in such fields as public opinion monitoring, marketing and finance, sentiment analysis has received increasing attention in recent years. The present study focuses on the problem of latent meaning faced by sentiment analysis, builds a Chinese construction corpus and studies the relations between Chinese constructions and sentiment analysis by quantifying the Chinese constructions within the corpus. We annotate the constructions and words expressing attitudinal and gradational meaning in the corpus. A qualitative

作者简介：吴尹清(第一作者/通讯作者)，博士生。李德俊，博士，教授，博士生导师。

analysis is conducted on these constructions and words. Statistics of attitudinal and gradational constructions are computed according to such standards as construction type, semantic type, constant number and variable number. By comparing the statistics of the constructions and words in the corpus, we find that most of the gradational and attitudinal semantic information is loaded by words rather constructions. In spite of this, there still is a certain portion of attitudinal and gradational information expressed by constructions. The present study provides empirical data and a research method to the study of applying construction grammar to Chinese sentiment analysis. Chinese construction studies are also be supported by these data computed from Chinese authentic texts. The present study also discusses the difficulities of applying construction grammar to Chinese sentiment analysis and natural language processing and looks forward to future studies.

Keywords: Construction grammar , Chinese sentiment analysis , Construction corpus , Attitudinal meaning , Gradational meaning

1 研究必要性：构式与情感分析的语义隐匿性问题

情感分析作为重要的自然语言处理任务，其目标是“从文本中分析出人们对于实体及其属性所表达的观点、情感、评价、态度和情绪”(刘兵, 2017: 1)。近年来互联网评价文本的爆炸性增长催生了对海量文本的情感语义进行自动化分析的强烈需求，使得情感分析成为了学术界的研究热点，在舆情监测、企业与政府决策、社会计算等方面均有应用。

汉语情感分析面对的语义隐匿性问题是由语义隐匿性现象造成的，即汉语中部分情感意义表达机制不确定性较强的语义现象，如反讽、非现实性、构式等。一般的情感分析方法难以应对这些语义现象的复杂性，不能准确解析它们的意义，导致它们的意义对于情感分析系统具有某种“隐匿性”。由于汉语的复杂性，语义隐匿性现象较为普遍，对语义隐匿性问题进行研究对于未来汉语情感分析系统性能的提升将有较大意义。

态度和量级义构式是上述语义隐匿性现象的重要一部分。“（典型的）构式是无递归性的非平凡的短语结构”(詹卫东, 2017: 232)，表达态度和量级意义的构式称为量级和态度义构式。态度意义是情感倾向，也即情感极性，包括正面、负面、中性等，量级意义描述的是情感极性的强度，包括低量、高量、极量等，也可使用数值来描述。态度和量级意义是情感意义计算的最重要的两个语义变量。由于构式的形式与意义的可推导性较弱(Goldberg, 1995:4)，其情感意义和形式难以通过短语结构语法等一般的组合性规则推导得到，意义表达机制的不确定性较大，自动解析难度较大，而汉语中的许多构式又具有明显的情感意义(詹卫东等, 2020)。而且汉语情感分析主要关注词汇的情感色彩，对构式这一特殊语言单位承载的语义信息缺少关注。因此态度和量级义构式的语义解析问题就属于汉语情感分析的语义隐匿性问题。

当前情感分析研究对语义隐匿性现象的研究不足，尚不能很好地应对语义隐匿性问题，满足于将词汇视为情感意义的承载单位，在理论构建、实证研究、资源建设等方面都以词汇为重点，未能参考借鉴当前构式语言学的理论与实证成果，对汉语中具有语义隐匿性的边缘性结构缺乏关注。因此情感分析研究可以考虑借鉴构式研究的成果，对构式这一层级的语言单位给予一定重视。文章基于真实评价语料，通过考察汉语态度和量级构式的分布情况，探索构式与情感意义之间的关系，为后续实证研究提供一种方法，并对汉语构式的情感分析乃至语义计算的研究现状进行分析、展望未来的研究方向。

2 相关研究回顾

本文研究重点是构式语法对于应对情感分析语义隐匿性问题的意义，因此主要回顾情感分析研究对情感表达的认识与处理方式以及汉语语言学界对构式语法的相关研究。

2.1 汉语构式语法研究

构式语法理论兴起于上世纪90年代，是借鉴认知心理学格式塔(Gestalt)完形理论所创立的一种新兴语法研究理论，集形式、意义、用法于一体来认识和分析语言(陆俭明、吴海波, 2018:

1)。构式的权威定义最早由Goldberg(1995: 4)给出：C是一个构式当且仅当C是一个形式—意义的配对，且C的形式或意义的某些方面不能从C的构成成分或其他先前已有的构式中得到完全预测。此后语言学界对“构式”概念范畴的界定出现了狭义和广义的区分。汉语语言学界多主张有选择性地吸收构式语言观的创新性视角，以狭义定义研究构式，促进对汉语边缘性结构的描写和解释。陆俭明、吴海波(2018: 2-3)主张“构式”应当是“自由的边缘性句法结构，既有象征关系，内部又有结构性的组成关系”。詹卫东(2017: 232)指出：“（典型的）构式是无递归性的非平凡的短语结构”。部分学者采用以上狭义定义对一部分具有态度意义的汉语构式进行了实证研究(刘宗保, 2011; 郑娟曼, 2012; 李劲荣, 2015; 刘晨阳, 2016; 胡习之, 2017)，进行了详尽的句法语义分析。总的来说，目前语言学界对于构式范畴的认定仍存在争论，但无论在心理认知还是句法语义接口研究等领域，构式作为一类在形式和意义方面都具较强自足性的语言单位的理据性正逐渐确立，而且在应用研究方面成果不菲，较好地描写和解释了对汉语中部分边缘性结构的句法语义特征，很好地补充了传统句法语义研究对非常规结构关注的不足。

2.2 情感分析研究对情感表达的认识与处理

情感表达是表达了情感意义或立场的语言单位。随着万维网的快速发展以及人文社科领域的“情感转向”(Hardt, 2007)，情感表达已成为目前语言学和计算机科学共同关注的热点研究领域。情感表达研究主要有两种路径：一是理论驱动的语言学路径，即情感表达的语言学研究，注重分析文本中特定语言单位与情感意义的关系；二是应用驱动的量化研究路径，即情感分析，以实现对情感意义的自动计算处理和量化分析为目标。

情感表达的量化研究以情感语义的量化计算以及情感表达的自动抽取为目标，属于应用研究，近年来该领域被统称为情感分析研究。情感分析方法主要可分为三种方法：基于知识库的方法；机器学习方法；融合知识库和机器学习的混合方法(Cambria et al., 2017)。知识库是情感分析的基础资源之一，专门用于情感分析的知识库又称为情感知识库，是语义知识库的一种，汉语情感分析领域已经产生了情感词典(王科、夏睿, 2016; 赵妍妍等, 2017)、情感语义规则库(万岩、杜振中, 2020)、情感常识库(杨亮等, 2019)、情感句子模式库(陈涛等, 2013)等多种形式的情感知识库的构建和应用研究。目前汉语情感知识库建设和应用研究过于偏重表达态度和量级意义的词汇以及习语、谚语、惯用语等短语的语义知识，情感词典作为汉语情感知识库资源的主体，收录的也是这部分词汇和短语。但对于同样表达态度和量级意义的边缘性结构，也就是文章所说的“构式”，则在情感分析研究中受到了忽视。汉语中存在许多这类构式。在网络评论等真实的汉语评价语料中，由于语料的非正式性、去中心化等特征，这些构式可能较高频地出现。

构式研究已经得到语言学界的重视，态度和量级义构式研究在汉语本体研究中也有较多成果，但这类较为特殊、边缘化的语言单位尚未得到情感分析研究的重视，除了少量研究，如黄思思、詹卫东(2018)以42条量级义构式和185条态度义构式为例，探讨了适用于情感分析的汉语量级和态度义构式的语义分析和形式化表征问题。目前既缺乏面向汉语情感分析的构式知识库，而且在实证研究方面，也未有研究基于汉语真实评价语料，分析构式的分布情况，探讨情感意义表达与构式之类的关系。

3 研究设计

3.1 研究问题

文章的研究对象为汉语量级与态度义构式，它们既可能包含常项也可能包含变项，它们的意义包含量级或态度这两类在情感分析中最为关键的语义成分，但其整体的量级或态度意义又无法通过其组成成分如字、词等的意义推出，情感词库等已有的知识库资源又难以分析该类结构，从而可能导致情感分析结果准确性的降低。文章主要探讨的三个研究问题为：(1) 汉语真实网络评论文本中态度和量级义构式的分布情况是怎样的？(2) 对于汉语网络评论文本的情感分析，是否有必要考虑计算构式所承载的态度和量级语义信息？(3) 汉语构式的情感分析乃至语义计算目前存在哪些需要突破的难点？

3.2 语料选取

构式语料库选取的语料来源为热门话题微博的评论，话题均具有较大争议性，评论中的立场和态度具有高度不一致性，每个话题下的评论都为一万条以上。我们从热门话题的微博评论

中随机抽取4000余条评论，形成语料库，总字符数为13.3万。选取微博评论作为语料，是因为微博评论属于网络短文本，具有语体非正式、口语化、去中心化、数据噪声较多、情感表达丰富等特征，是典型的网络评价文本，是情感分析的主要处理对象之一。另外，文章采用的语料都属于话题型微博的评论，根据侯敏等(2013：136-138)指出，具有以下特点：(1) 句子简短，单句多；(2) 观点负面倾向多；(3) 表达情感强烈，理性评价淡化；(4) 口语色彩浓重，情感因子颗粒度加大，往往不再是词，而是短语甚至短句；(5) 隐晦表达观点，常用习语、反讽等方式表达观点，而不使用直接表达态度意义的词汇；(6) 评价对象省略；(7) 语言不规范。这些特点给情感分析带来了很大困难。同时第二至第五个特点可能意味着该类语料中存在大量构式，契合文章的研究目的。

3.3 语料库标注

由于目前尚无准确率较高的自动标注汉语构式的方法，我们需要先对语料中的态度和量级构式分别进行人工标注，目前学界尚未形成在真实语料中标注汉语构式的标准，我们尝试使用一种基于XML(Extensible Markup Language)的构式标注方法，并根据汉语的特点，将詹卫东(2017: 232)的狭义构式观作为文章对构式的定义：“(典型的)构式是无递归性的非平凡的短语结构”。文章不采用将语素、词、常规句法结构等纳入构式范畴的广义构式观，而只是将构式视为“对常规短语结构语法组合的必要补充”(ibid.)，原因是该观点较契合文章的研究对象及研究问题，也有助于增加操作的可行性。在标注过程中，只要某个结构符合文章的构式定义及判定标准，就将其标注为态度或量级构式，判定标准采用Hilpert(2014: 14-23)基于Goldberg(2006: 5)的构式定义所提出的四条标准：(1) 形式特殊性：该结构的整体或部分在形式上有别于一般的语法结构；(2) 语义不可预测性：该结构的意义具有非组合性，不严格等于其组成成分的意义加和；(3) 特殊限制性：该结构的整体或部分是不完全自由的，受到某些条件制约；(4) 搭配倾向性：该结构倾向于与某些成分或结构共现。

构式的标注方法是在文本中使用开始标记和结束标记包围整个构式，开始标记为<C TYPE="" FORM="" CAT="">，结束标记为</C>。开始标记包括三个属性，TYPE属性表示构式的语义，对应六个值：“att/neg”(负面态度义)；“att/pos”(正面态度义)；“att/bipolar”(双极性态度义)；“gradational/low”(低量义)；“gradational/high”(高量义)；“gradational/veryhigh”(极量义)。前三个值分别对应态度义构式的三类语义，后三个值分别对应量级构式的三类语义。FORM属性表示构式的形式，参照詹卫东(2017; 2018: 35-36)提出的构式形式表示法，值为由实例化的常项、变项以及加号构成的表达式。CAT属性表示构式的类别，参考詹卫东(2017; 2018: 18-20)提出的构式分类标准，对应四个值“frozen”(凝固型构式)；“semi-frozen”(半凝固型构式)；“phrasal”(短语型构式)；“compound-sentential”(复句型构式)。凝固型构式为完全由常项组成的构式；半凝固型构式的变项不超过2个，长度较短；短语型构式的变项数量可为1个及以上，长度可短可长；复句型构式由两个相对独立的部分组成，变项数至少为2个。评论文本中的构式标注示例如下(选自构式语料库)：

进门抢狗，这是<C TYPE=”att/neg” FORM=”什么+np/vp” CAT=”semi-frozen”>什么行为</C>。
<C TYPE=”att/neg” FORM=”不+v+就+不+v” CAT=”phrasal”>不批就不批</C>咯，少布置就少布置咯。
我让他<C TYPE=”gradational/veryhigh” FORM=”要+多+a+就+(有)+多+a” CAT=”phrasal”>要多快乐就多快乐</C>。
哪些该做不该做，还用规定吗？<C TYPE=”att/neg” FORM=”真+是/的+够+够+的” CAT=”frozen”>真是够够的</C>。
说真的杯子也<C TYPE=”att/neg” FORM=”没+(x)+几+q+vp/ap” CAT=”phrasal”>没几家干净的</C>，从来不用酒店杯子喝水。

Table 1: 构式标注示例

另外，态度和量级义词汇的标注，也使用XML标记，开始标记为<W TYPE=“”>，结束标记为</W>。属性TYPE代表词汇的语义类别，对应五个值：“att/pos”(正面态度义)；“att/neg”(负面态度义)；“gradational/low”(低量义)；“gradational/high”(高量

义)；“gradational/veryhigh”(极量义)。

4 基于构式语料库的数据分析

4.1 构式分布基本情况

经过人工标注，在语料库中共发现119个量级和态度义构式，构式出现的总频次为207次。按照上文提到的构式分类标准：凝固型构式共出现36个，频次为58次；半凝固型构式共出现18个，频次为49次；短语型构式共出现57个，频次为92次；复句型构式共出现8个，频次为8次。四类构式在语料库中的分布情况如下：

构式类型	凝固型	半凝固型	短语型	复句型
频次	58	49	92	8
总频次中占比	28.0%	23.7%	44.4%	3.9%
个数	36	18	57	8
总个数中占比	30.3%	15.1%	47.9%	6.7%

Table 2: 四类构式在语料库中的分布情况

“现代汉语构式数据库”(Chinese Construction Grammar Database, CCGD)(詹卫东, 2021)是目前唯一的较大规模的汉语构式知识库，统计数据显示，在该知识库收录的1108个汉语构式中，以上四类构式的分布情况如下：

构式类型	凝固型	半凝固型	短语型	复句型
构式条数	224	238	545	98
构式占比	20.2%	21.5%	49.2%	8.8%

Table 3: 四类构式在“现代汉语构式数据库”中的分布情况

对比数据可知，评论文本中四类构式的频次占比和个数占比都与CCGD中的占比相近，其频次占比和个数占比分布比较接近2：2：5：1的比例，这一比例在一定程度上反映了汉语真实评论文本中量级和态度义构式的分布情况，也为CCGD的统计数字提供了真实语料的佐证。至于所有汉语构式在真实文本中的分布是否仍然遵循这一比例，尚需进一步研究的验证。

构式语料库中共发现9个量级构式，出现频次共为11次，其中，4个为极量构式，5个为高量构式；6个为短语型构式，3个为半凝固型构式。量级构式中没有出现低量构式，只出现了高量和极量构式，这说明在汉语真实评论文本中，量级构式的分布以高量和极量构式为主，低量构式可能占比偏小。另外，语料库中共发现111个态度义构式，占总个数的93.2%，出现频次共为197次，占总频次的95%。与态度义构式相比，量级构式在构式出现总频次和总个数中占比都很低，仅占总个数的7.6%和总频次的5.3%，这说明量级构式在汉语真实评论文本中的分布可能较少，远低于表达主要情感意义的态度义构式。

态度义构式在构式语料库中的分布情况如下：

态度义类型	负面	正面	双极性
构式个数	101	6	4
个数占比	91.0%	5.4%	3.6%
构式频次	181	8	8
频次占比	91.9%	4.1%	4.1%

Table 4: 态度义构式在语料库中的分布情况

构式义为负面或正面态度的构式又称负面或正面态度义构式。而双极性构式则较为特殊，它也表达态度意义，但其极性的正负向具有不确定性，较为典型的双极性构式如“a+的+是”，其极性由形容词a来决定，具有不确定性，可能是正面也可能是负面。

语料库中出现的负面态度义构式无论是个数还是频次都占90%以上，占所发现的态度义构式的绝大多数，正面态度义构式和双极性构式出现都较少，占比相近，都在5%左右。这种

构式义分布的显著不对称性现象，在一定程度上说明汉语的负面态度义更多通过构式义来表达，为研究汉语负面评价表达规约化(方梅, 2017)现象提供了真实语料的佐证，在一定程度上说明了汉语在构式层面正负面态度意义的不对称性以及心理认知层面的“负面偏见”(Negativity Bias)(Rozin and Royzman, 2001)在汉语中的表现。

通过对语料库中构式的分析，我们发现，态度义构式的常项中很少包含出现表达态度意义的词汇，而且对于量级和态度义构式而言，仅分析其常项与变项各自的语义，无法推出其整体的构式义。

经过统计，语料库中所有构式的常项个数和变项个数情况如表5和表6所示：

常项个数	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
构式条数	12	33	27	29	13	3	0	1	1
占比	10.1%	27.7%	22.7%	24.4%	10.9%	2.5%	0%	0.8%	0.8%

Table 5: 语料库中所有构式的常项个数统计

由表5可知，语料库中，常项个数为1的构式占10.1%，常项个数为2至4个的构式占主要比重，达74.8%，与此相近的是，CCGD(詹卫东, 2021)中，常项个数为1的构式占该数据库所有构式的13.4%，常项个数为2到4的构式占79%。但在“构式库”(ibid.)中，常项个数为0的构式占9%，语料库中却没有发现常项个数为0的构式，这可能与样本数据规模的限制有关；语料库中出现了常项个数多达9个的构式，但“构式库”中没有此类构式。这说明，在真实评论文本中，常项个数为1至4的构式出现最多，占主要比重，常项个数大于或等于5的构式则分布较少。

变项个数	**0**	**1**	**2**	**3**
构式条数	36	59	18	6
占比	30.3%	49.6%	15.1%	5.0%

Table 6: 语料库中所有构式的变项个数统计

由表6可知，语料库中不存在变项个数为3以上的构式，变项个数为0至2的构式占主要比重，达到95%。与此相似的是，在CCGD(ibid.)中，变项个数为0至2的构式占90.4%，同样所占比重相近。这说明变项个数为0到2的构式在真实评论文本中占主要比重，分布最多，变项个数大于或等于3的构式则分布较少。

以上数据表明，语料库中的构式主要是常项个数为1-4、变项个数为0-2的构式，与CCGD（ibid.）的情况基本一致。

4.2 构式在语料库量级和态度意义表达中的比重分析

为了分析汉语真实文本中构式所承载的态度和量级语义信息的多寡，我们对语料库中表达态度和量级语义的构式与词汇进行了标注和统计。结果显示，态度义词汇共出现2453次，量级义词汇出现523次，总计2976次，以下为相关数据：

语义类型	**正面**	**负面**	**双极性**	**总计**
词汇频次	643	1810	0	2453
在态度义词汇总频次中占比	26.2%	73.8%	0%	100%
构式频次	8	181	8	197
在态度义构式总频次中占比	4.1%	91.9%	4.1%	100%
构式表义比重	0.40%	22.00%	100%	7.40%
词汇表义比重	99.6%	78.0%	0%	92.6%
注：构式表义比重=构式频次/(词汇频次+构式频次)；词汇表义比重=词汇频次/(词汇频次+构式频次)				

Table 7: 语料库中态度义词汇与态度义构式频次数据对比

态度和量级意义的表义单位主要是构式和词汇。句子往往也表达态度和量级意义，但通常句子语义的可推导性较强，可以拆分为更小的“原子”，如构式和词汇，因为句子不适宜作为表

义单位的进行统计。因此文章只统计构式和词汇的表义比重，表义比重是指构式或词汇这两类语言单位在表达相应类型语义中所占的比重大小。

语料库中共出现态度义词汇2453次，正面态度义词汇643次，负面态度义词汇1810次，正负态度义词汇比例约为1:3，远高于正负态度义构式的1:9。从表义比重上看，正面态度语义的构式表义比重很小，仅为0.4%，词汇表义比重高达99.6%；负面态度语义的构式表义比重则相对较高，达到了22%，词汇表义比重为78%。正负态度构式表义比重的比例为1:55，该悬殊的比例也进一步佐证了汉语负面评价表达规约化的现象。而更为特殊的双极性态度语义方面，构式表义比重则为100%，语料库中未发现表达此类语义的词汇，这也说明，更加复杂和模糊的语义，可能由往往构式表达，而不是词汇。

从整体上看，态度语义的词汇表义比重为92.6%，这表明词汇是评价文本中态度语义表达的主要语义单位，构式表义比重为7.4%，这说明评价文本中的构式只承载了比例较小的态度语义，但7.4%的比例也不能算小，这表明评论文本中的构式也承载了相当比重的态度语义，尤其是负面态度义。如果在态度语义计算的过程中，忽视构式所承载的态度义，足以影响语义计算的准确性，造成结果的偏差。

语义类型	低量	高量	极量	总计
词汇频次	18	404	101	523
在量级义词汇总频次中占比	3.40%	77.20%	19.30%	100%
构式频次	0	6	5	11
在量级义构式总频次中占比	0%	54.5%	45.5%	100%
构式表义比重	0%	1.5%	4.7%	2.1%
词汇表义比重	100%	98.5%	95.3%	97.9%

Table 8: 语料库中量级义词汇与量级义构式频次数据对比

语料库中共出现量级义词汇523次，高量词汇404次，低量词汇18次，极量词汇101次，表达高量语义的词汇占大多数，为77.2%，表达低量和极量语义的词汇都相对较少，分别为3.4%和19.3%。

表义比重方面，低量语义的构式表义比重为0%，词汇表义比重为100%，其语义全部由词汇表达；高量语义的构式表义比重很小，仅为1.5%，词汇表义比重为98.5%；极量语义的构式表义比重略高于高量语义，为4.7%，而词汇表义比重为95.3%。整体而言，量级语义的词汇表义比重为97.9%，相比之下，量级语义的构式表义比重仅为2.1%，远低于词汇表义比重，说明评论文本中的量级语义绝大部分由词汇来承载，构式在量级语义表达中起到的作用很小。

4.3 构式语义对于情感分析的意义

总的来说，上述语料库数据的分析显示：在汉语真实评论文本中，词汇承载了大部分态度和量级语义信息，是主要的表义单位；构式所承载的态度和量级语义信息较少，构式虽然不是主要的情感意义单位，但作为边缘性结构，其承载的态度语义信息仍占一定比例，为7-8%，其承载的语义信息不应完全被忽视。而构式所表达的量级语义仅占2.1%，比重小。至于在情感分析中是否需要考虑计算构式所承载的态度和量级语义，需要根据研究要求的计算精确度和研究目的来具体确定，不宜一概而论，若研究对计算精度要求较高，则可以考虑将构式的语义纳入计算范围。但以上数据是仅基于13万字左右的汉语语料库得出的，要进一步验证上述结论，尚需后续基于更大规模、包含更丰富主题的语料库进行研究。

从数据上看，如果汉语构式的情感语义能够得到汉语情感分析系统的准确解析，最多可在情感倾向计算上提升7-8%的准确率，最多可在情感强度计算上提升2%左右的准确率。但要实现以上准确率的提升需要较好地处理汉语构式的情感分析问题，也即汉语构式的隐匿性语义计算的问题，在算法、知识库资源、知识表示等方面都需要做出较多努力和探索。因此，对于汉语情感分析是否应当考虑计算构式的语义，需要考虑投入成本与产出的比率，如果追求较高的情感计算准确率，则有必要计算构式的语义。如何实现对汉语构式的情感分析本身也是一个较复杂的语义计算问题，下文将对相关难点、研究现状、发展方向进行讨论。

5 对汉语构式情感分析问题的思考：问题与展望

在汉语情感分析中融入构式的语义知识，有助于应对汉语情感分析的部分语义隐匿性问题。汉语构式的情感分析属于构式计算和语义计算的研究范畴，该领域产生了一些研究成果，但也面临诸多难题，主要是四方面问题。文章对相关研究难点进行了梳理，提出了相应的对策，展望了后续研究。

5.1 探索面向计算的构式定义

第一个问题是构式的定义问题，即如何面向自然语言处理的需求界定构式的内涵和外延，给出一个可操作性较强、适用于计算机处理的汉语构式定义。此处论及定义问题，主要原因是，目前语言学界对构式这一重要概念界定的含糊不清、各自为政的现象，不能适应句法语义自动分析的要求。国外构式研究倾向于接受宽泛的“构式”定义，如“激进”构式学派把语素、词汇、短语等各个层级都涵盖到构式的范畴内(Croft and Cruse, 2004: 225-257)，Goldberg(1995: 4; 2003; 2006: 5)也一直不断修正其关于构式的定义，使得构式的概念同时涵盖了不可预测性的边缘结构以及出现频率高的可预测性常规结构(比如核心句式)。甚至有将构式延伸到语篇层面的观点(Ostman, 2005)。而国内汉语学界则倾向于认同狭义的“构式”定义，比如陆俭明与吴海波(2018: 3)认为构式应当是具有不可预测性的形义配对结构体，这比较接近詹卫东(2017: 232)及文章采用的定义——“无递归性的非平凡的短语结构”。但无论是国内汉语学界还是国际语言学界，对于构式应该包含哪些语言单位，仍无定论。虽然构式理论的提出和深入发展的确给句法-语义界面研究带来了新的理论视角，但此种理论层面的割裂现象阻碍了构式理论在句法语义分析中的应用。在句法语义研究中，某一基础概念内涵的确定是重要的研究工作，比如汉语学界对汉语词类划分标准这一基础问题的争论和探讨(詹卫东, 2013; 叶脉清、聂仁发, 2015; 杨丽姣等, 2021)，对词典编纂、汉语词汇知识库构建、自然语言处理等都产生了较大的影响。如果构式理论要更好地应用于语言工程领域，也需要这样一番专门的争论和探讨。语言学界需要加强对面向计算的构式定义的探讨，结合语义计算的实例进行分析，探索适用于计算的“构式”概念操作界定。

5.2 加强汉语构式知识库构建研究

第二个问题是汉语构式知识库的构建问题。构式作为意义与形式的配对体(Goldberg, 1995: 4)，其语义具有非组合性和不可预测性(Hilpert, 2014:14-23)。那么对于情感分析而言，计算机获取构式的先验语言知识有两种途径，一是真实文本中人工标注的构式知识，二是专门的量级和态度义构式知识库。在真实文本中人工标注构式也在某种程度需要依靠知识库提供的知识。因此有必要建设标注有句法和语义等信息的构式知识库。目前，成熟的汉语构式知识库资源中仅有CCGD，其收录的构式为1108条，主要收录的是已经出现在汉语本体研究论文中的构式，多为学界讨论较多、较典型的汉语构式(詹卫东, 2021)。目前对于汉语中到底有多少构式这一问题，学界尚无定论，就目前CCGD的规模而言，其应该远未能覆盖所有汉语语言使用中的构式，其收录的量级和态度义构式可能也不全面。因此，要为汉语情感分析提供构式先验知识的支持，一方面需要构建对量级和态度义构式收录全面的知识库，另一方面也需要进行相应研究来测算汉语构式总数的数量级，才能为构式知识库的构建提供参考。

此外汉语构式知识库的收录范围也是值得探讨的问题，主要是凝固型构式的边界问题，即成语、俗语、谚语以及新兴网络流行语等是否应该收录为凝固型构式。我们认为成语、俗语、谚语虽然也符合文章对构式的定义，但考虑到它们数量较大，如“汉语习语知识库”(Chinese Idiom Knowledge Base, CIKB)(Lei and Yu, 2010)就收录了多达38117条汉语成语、俗语、谚语等，而且它们出现时间长，很多辞书都已收录，语义确定性强，适合由情感词典收录。而新兴网络流行语，如“醉了”、“打脸了”、“凉凉”等网络流行语，往往有特殊的修辞效果，从字面义无法推知其意义。它们符合文章的构式定义，我们可以将其视为凝固型构式。同时由于网络流行语出现时间较短，语体非正式和偏口语化，传统语文词典和情感词典等知识库基本都没有收录，其构式义又往往具有情感意义，因此有必要由构式知识库来对其进行收录，未来应注意对这部分网络流行语的整理、追踪、收录、知识化。

5.3 探索科学的构式形式化表示方法

第三个问题是构式的形式化表示问题，即应该采用何种标准，来呈现单个构式的句法语

义知识或者汉语文本中关于构式的句法和语义知识，以便更好地服务于构式的识别。由于学界对汉语构式的计算处理研究还在早期阶段，构式的形式化表示问题尚缺乏系统研究。前人研究主要探索了结构化的知识库中单个构式的形式化表示问题：CCGD中使用10余种特征-值对(feature-value pair)来表示单个构式的句法语义特点，如变体、义项、释义模板、实例等(詹卫东, 2017; 2021)。而对于如何在语料库这一类非结构化数据库中表征汉语构式的语言学知识，研究则很少，少数相关研究如黄彤等(2020)尝试采用中文抽象语义表示(Chinese Abstract Meaning Representation, CAMR)这一面向自然语言处理的句法语义表示体系，对CCGD中所有条目的例句进行了构式的形式化标注，发现CAMR可以表示出61.2%的基本符合组合原则的构式，同时发现在文本中标注构式存在语义省略、凝固型构式难以拆分表示、语义范围难以确定、释义需要语境和语用推导等困难，并对标注策略进行了总结。

总体来看，学界在汉语构式的语料库标注方面的研究较为缺乏，也没有经过科学标注的汉语构式语料库作为参考标准和研究材料；而由于CCGD的实践，学界在汉语构式的知识库表示方面已经积累了一定经验，但仍需进一步研究和探索：从描写全面性的角度，探索应当用多少类属性来表示构式所承载的语言知识；探索在实际语义计算中，最常调用的构式知识属性是哪些；进一步探索构式的句法和语义是否有更合理、高效的形式化表示方法。

5.4 加强构式识别研究

第四个问题是构式的识别问题，即采用何种算法能够更好地从文本中自动识别构式，这是构式计算处理过程的最后一步。由于汉语句法语义规则本身以及汉语构式的复杂性，以及汉语句法语义分析研究对非常规、句法组合规则难以分析的边缘性结构重视相对不足，系统性研究较少(黄彤等, 2020)，目前尚未摸索出成熟高效的算法。同时我们也缺乏面向汉语构式识别的标注数据集，无法训练出基于有监督学习的构式标注器。黄海斌等(2020)在没有训练语料的情况下，探索了融合高斯混合模型、正则表达式以及词性匹配的无监督汉语构式自动识别算法，并指出，构式识别算法面临的最大困难是在无训练语料的情况下确定句子中构式的边界信息。除此之外，汉语学界缺少相关研究积累，这给构式知识在情感分析乃至自然语言处理中的应用带来了困难。后续需要加强汉语构式数据集的标注以及识别算法实验性研究。

6 结语

文章为构式语法应用于情感分析提供了实证数据，以量化的方式揭示了汉语构式与情感意义的关系，为汉语情感分析是否需要考虑计算构式的语义信息这一问题的解答提供了数据支持，也为情感分析提供了一种新的研究视角，即基于真实文本进行实证分析，弥补了汉语情感分析研究对以构式为代表的语义隐匿性现象关注的不足，还较为系统地梳理了汉语构式情感分析的相关研究及现实困难，推动了汉语情感分析研究的进展。同时，文章还为汉语构式研究提供了实证数据支持，在构式语料库标注方面进行了探索。因此文章对于汉语情感分析、构式计算、构式语法理论等研究均有一定价值，但这方面的研究还需更大规模语料库的支持，期望文章能够为后续研究抛砖引玉，提供一种研究思路和方法。

参考文献

Cambria, E., Das, D., Bandyopadhyay, S. & Feraco, A. Affective Computing and Sentiment Analysis. In Cambria, E., Das, D., Bandyopadhyay, S. & Feraco, A. (eds.). *A Practical Guide to Sentiment Analysis*. Berlin: Springer. 2017: 1-10.

Croft, W. & Cruse, D. A. *Cognitive Linguistics*. Cambridge: Cambridge University Press. 2004.

Goldberg, A. *A Construction Grammar Approach to Argument Structure*. Chicago/London: The University of Chicago Press. 1995.

Goldberg, A. Constructions: A New Theoretical Approach to Language. *Trends in Cognitive Sciences*, 2003, 7(5): 219-224.

Goldberg, A. *Construction at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press. 2006.

Hardt, M. Foreword: What Affects are Good for. In Clough P. & Halley J. (eds.). *The Affective Turn: Theorizing the Social*. Durham/London: Duke University Press. 2007: 1-5.

Hilpert, M. *Construction Grammar and Its Application to English.* Edinburgh: Edinburgh University Press. 2014.

Lei Wang & Yu Shiwen. Construction of Chinese Idiom Knowledge-base and Its Applications. In Laporte E., Nakov, P., Ramisch, C. & Villavicencio, A. (eds.). *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications.* Beijing: Chinese Information Processing Society of China. 2010: 11-18.

Ostman, J.-O. Construction Discourse: a Prolegomenon. In Ostman, J.-O. & Fried, M. (eds.). *Construction Grammars: Cognitive Grounding and Theoretical Extensions.* Amsterdam: John Benjamins. 2005: 121-144.

Rozin, P. & Royzman, E. B. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, 2001, 5(4): 296-320.

陈涛, 徐睿峰, 吴明芬, 刘滨. 一种基于情感句模的文本情感分类方法.中文信息学报, 2013(5): 67-74.

方梅. 负面评价表达的规约化.中国语文, 2017(2): 131-147.

侯敏, 滕永林, 李雪燕, 陈毓麟, 郑双美, 侯明午, 周红照. 话题型微博语言特点及其情感分析策略研究.语言文字应用, 2013(2): 135-143.

胡习之. 构式“你才X呢”再探.当代修辞学, 2017(6): 73-81.

黄海斌, 常宝宝, 詹卫东. 基于高斯混合模型的现代汉语构式成分自动标注方法.中文信息学报, 2020(9): 1-8.

黄思思, 詹卫东. 面向情感分析的构式主观态度义初探.外语教学, 2018(6): 27-33.

黄彤, 李斌, 闫培艺, 戴玉玲, 曲维光. 基于抽象语义表示的汉语构式标注与分析.中文信息学报, 2020(10): 1-10.

李劲荣. 列举形式“什么X”与“X什么的”的语义偏向.汉语学习, 2015(5): 40-48.

刘兵著, 刘康, 赵军译. 情感分析：挖掘观点、情感和情绪.北京: 机械工业出版社. 2017.

刘晨阳. 警告义“再VP”构式探析.语言科学, 2016(4): 412-421.

刘宗保. 警告义构式“叫/让”句探析.汉语学习, 2011(2): 60-67.

陆俭明, 吴海波. 构式语法理论研究中需要澄清的一些问题.外语研究, 2018(2): 1-5.

万岩, 杜振中. 融合情感词典和语义规则的微博评论细粒度情感分析.情报探索, 2020(11): 34-41.

王科, 夏睿. 情感词典自动构建方法综述.自动化学报, 2016(4): 495-511.

杨丽姣, 肖航, 刘智颖. 《信息处理用现代汉语词类标记规范》修订方案.语言文字应用, 2021(3): 111-120.

杨亮, 周逢清, 林鸿飞, 殷福亮, 张一鸣.基于情感常识的情感分析.中文信息学报, 2019(6): 94-99.

叶脉清, 聂仁发. 新世纪以来现代汉语词类研究综述.现代语文, 2015(8): 7-10.

詹卫东. 计算机句法结构分析需要什么样的词类知识——兼评近年来汉语词类研究的新进展.中国语文, 2013(2): 178-190.

詹卫东. 从短语到构式: 构式知识库建设的若干理论问题探析.中文信息学报, 2017(1): 230-238.

詹卫东. “现代汉语构式知识库”填写规范. http://ccl.pku.edu.cn/ccgd/downloaddoc/,2018-12-01.

詹卫东. 现代汉语构式知识库. http://ccl.pku.edu.cn/ccgd/,2021-12-01.

詹卫东, 王佳骏, 黄海斌, 陈龙. 构式的表征与语料标注——现代汉语构式数据资源建设中的基本问题. 第21届汉语词汇语义学国际研讨会, 中国香港, 2020年5月.

赵妍妍, 秦兵, 石秋慧, 刘挺. 大规模情感词典的构建及其在情感分类中的应用.中文信息学报, 2017(2): 187-193.

郑娟曼. 从贬抑性习语构式看构式化的机制——以“真是(的)”与“整个一个X”为例.世界汉语教学, 2012(4): 520-530.