

文章编号:

细颗粒度汽车评论语料库的构建和分析*

曹紫琰¹ 毛雪芬¹ 程宁¹ 宋阳² 李斌^{1,3} 冯敏萱¹

(1.南京师范大学 文学院, 江苏省 南京市 210097;

2.南京大学 计算机科学与技术学院, 江苏省 南京市 210023;

3.哈佛大学 计量社会科学研究, 美国 剑桥市 02138)

摘要: 产品评论文本是情感分析的重要研究对象, 目前已有的产品评论语料库大都较为粗疏, 没有完整地标注出对象、属性、极性“三要素”, 影响自动分析的应用场景。对此, 本文构建了细颗粒度情感语料库, 共包含 9343 句汽车评论短文本, 不仅人工标注了“三要素”的具体词语, 而且将其对应到产品和属性的知识本体树上。此外, 对无情感词的隐含表达、特殊文本(如建议文本、比较句等)也标注出对应的三元组并予以特殊标签。语料统计表明, 对象和属性要素的共现率高达 77.54%, 验证了构建“三要素”齐全标注体系的必要性; 基于本语料库的“三要素”自动标注实验 F1 值可达 69.25%, 验证了细颗粒度标注体系的可计算性以及语料库构建的规范性和应用价值。该语料库可以为细颗粒度情感分析研究提供基础数据。

关键词: 情感分析; 细颗粒度; 语料库

中图分类号: TP391

文献标识码: A

Construction and Analysis of Fine-grained Car Reviews Corpus

CAO Ziyan¹, MAO Xuefen¹, CHENG Ning¹, SONG Yang¹,
LI Bin¹, FENG Minxuan¹

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;

2. School of Computer Science and Technology, Nanjing University, Nanjing, Jiangsu 210023, China;

3. Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA)

Abstract: The product reviews text is an important research object of sentiment analysis. At present, most of the existing product reviews corpus are relatively coarse, and the three elements of the target, attribute and polarity are not completely annotated, which affects the application scenarios of automatic analysis. The paper constructed a fine-grained emotional corpus, including a total of 9343 short texts of car reviews. It not only annotated the specific words of “three elements”, but also corresponded them to the ontology knowledge tree of the products and attributes, and the implicit expressions of non-sentimental words and special texts (such as suggested texts, comparative sentences, etc.) were also annotated with corresponding triples and given specific labels. The statistics shows the co-occurrence rate of the target and attribute is as high as 77.54%, indicating it is necessary to construct a complete annotation system on three elements; and based on it, the experiment on three elements automatic annotation achieves a F1 value up to 69.25%, which verifies the application value of the fine-grained annotation system and the accuracy of the corpus’ construction. The corpus can provide the underlying corpus for fine-grained sentiment analysis.

Key words: sentiment analysis; fine-grained; corpus

1 引言

开放共享的 Web2.0 时代, 互联网用户发布的海量信息呈爆炸式增长, 其中围绕时事热点、产品服务主观评论文本是 UGC (用户原创内容) 的重要类型, 在大数据背景下具有很强的应用前景。由于这类文本难以借助人工方法处理分类, 因此, 面向主观文本的情感分析研究(以下简称“情感分析”)应运而生了。情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程^[1], 不仅包括对文本主观情绪的分析(如哀伤、惊讶等), 更侧重于对评论文本的意见挖掘。细颗粒度意见挖掘往往以短文本为信息抽取对象, 目的是获取情感要素并据此判断观点在文本中表达的情感倾向^[2]。其中, 常见的情感要素包括<观

* 收稿日期: 定稿日期:

基金项目: 国家社科基金项目(18BYY127); 江苏高校哲学社会科学优秀创新团队建设项目(2017STD006); 国家语委科研项目(YB135-61)。

点持有人, 评论对象, 评论属性, 情感极性^[3-4]四种, 其中评论属性(如口味、质量、性价比等)也被称为主题词或者评论方面(角度)。目前, 有监督的学习方法在各类意见挖掘任务中具有较好的任务独立性, 且需要借助人工标注的语料作为训练数据^[5-6]。因此, 语料库构建工作是情感分析研究的重要方面。

目前, 国内外现有的语料库标注大多较为粗略, 有的仅标注整个句子或语篇的极性, 有的没有完整标注出评论对象和属性信息, 对情感要素省缺情况的处理也较少提及。但事实上评论文本具有简洁灵活、表达丰富等特点, 现有语料库的标注方法往往会导致关键信息的遗漏。以如下句子为例:

当四驱系统介入时还是特别给力的, 但是后备箱空间也太小了。

该句包括了不同的评论对象(四驱系统和后备箱)、不同属性(动力和空间)和不同极性的评价词(“特别给力”和“太小”), 且“当四驱系统介入时还是特别给力的”尽管省略了评论属性的具体表述, 但从语义上仍然可获知评论的属性是“动力”。结合上述现象, 本文主要针对产品评论短文本的对象、属性、情感极性“三要素”, 进行细颗粒度语料库的构建, 不仅标注出涉及要素的具体词语, 同时标注出其在产品知识本体树上的对应类别, 最终形成评论六元组。由于产品评论观点持有人默认为消费者, 因此不予标注。

产品评论文本领域众多, 不少国内的评测竞赛选择汽车评论文本作为数据集, 如 CCF 大数据与计算智能大赛^②及 COAE(Chinese Opinion Analysis Evaluation)等, 因此, 本文也选取汽车线上评论文本, 最终构建了包含 9343 句汽车评论短文本的语料库。在标注方面, 除对“三要素”的标注, 对于不含情感词的隐含表达^[7]、对情感词进行修饰的程度副词和否定副词、主观情感词(例如喜欢、满意等)、建议文本等其他特殊文本赋予了额外的特殊标签。本文将在第 2 节重点阐述语料库的标注体系。

2 相关研究

现有的情感语料库按照题材分为新闻、对话、用户评论语料库^[8], 语料库来源主要为国内外评测训练集、高校等科研机构和个人提供等。近年来随着细颗粒度的中文情感分析任务逐渐引起关注, 一系列的评测和竞赛围绕不同的子任务提供了较为丰富的数据集。下面主要介绍现有的中英文情感分析用语料库资源及评测数据集:

2.1 英文情感语料库及国际评测

现有的英文情感语料库方面, MPQA (MultiPerspective Question Answering corpus) 是最著名的英文情感语料库之一^[2], 其对 535 篇新闻评论进行观点持有者、评论对象、主观表达式及其极性与强度的标注; 影评语料库^[9]选取了 11 部不同电影的评论, 每个电影抓取 100 条评论, 对含有{评论对象、观点}二元组的句子进行标注, 语料来源为 IMDB (Internet Movie Database)。该语料库针对评论对象人工制定 20 个目标种类标签(主要为电影评论涉及的若干元素)并在文本中提取出评论对象和情感词, 标注出这些词所属的种类标签。伊利诺伊大学芝加哥分校的 Hu 和 Liu 提供的产品领域的评论语料库^[4]主要为从亚马逊等网站下载的 5 种电子产品的网络评论, 并运用 [u] (attribute) 标签区分了评论对象为具体商品还是某一具体属性(如尺码、外观等)。国际评测方面, 国际文本检索会议 TREC 率先围绕博客的观点检索开展了一系列评测任务; NTCIR 的情感分析评测(multilingual opinion analysis task, MOAT) 从 2006 年起每年举行一次, 并拥有中、英、日 3 种语言的标准语料库, 主要针对新闻报道提取主观性信息。总体来看, 国外情感语料库构建的研究成果为面向中文文本的情感语料库提供了基本思路和参考。

^② <https://www.datafountain.cn/projects/2018CCF/index.html>

2.2 中文情感语料库及国内评测

面向中文的情感语料库资源相对较少。“情感语料库”由徐琳宏等构建^[10]，语料来源为小学教材、电影剧本、童话故事等共计4万句，从词汇、语句和篇章三个层面分别进行标注，以文本表达的情感类型为重点标注对象；情感信息抽取语料库由苏州大学戴敏等人构建^[7]，对情感极性、评论对象、情感词进行了标注，同时重点针对评论对象的省略、无情感词情感句的表示及极性转移等进行了处理，较为细致地关注到中文评论文本中错综复杂的语言现象，语料为笔记本、宾馆和化妆品的评论共计5000篇。

另一方面，由学术组织和企业举办的评测竞赛不乏各类情感分析任务，因此也提供了丰富的数据集资源。COAE是国内第一个情感分析方面的评测，其早期主要围绕评论对象和情感词的抽取、倾向性分析等较为基础的任务组织评测^[11]。近年来，随着对情感分析技术的应用需求不断扩大，更细颗粒度的情感分析评测和竞赛逐渐引起重视。其中，CCF大数据与计算智能大赛(CCF Big Data& Computational Intelligence Contest)[®]围绕汽车产品评论文本逐年举办竞赛，2017年为“基于主题的文本情感分析”，任务是从文本中抽取主题词和情感词并判断情感极性；2018年竞赛为“汽车行业用户观点主题及情感识别”，并规定主题词根据上下文提出或归纳，且必须在动力、价格、内饰等10个预设主题词内进行选择，最高得分为70.85%^④。此外，2016年的COAE评测[®]任务2也关注了属性等较细的抽取颗粒度，主要针对影评文本的属性词、对应情感词及极性的判断，最高系统准确率为63%^[12]。也有不少评测主要面向产品评论中的特殊表达类型，这些特殊表达类型往往无法运用情感自动分析常用技术手段处理，具体如COAE2013^[13]针对否定句、比较句等问题的信息抽取进行了评测；2016年的CCF竞赛针对不同视角下的情感分析，在语篇中抽取汽车品牌词并识别与之对应情感极性；CCL2018[®]评测任务主要针对隐喻识别及情感极性判别等。

综上，现有中文语料库大多以为情感自动分析提供训练集为主要目的，往往存在不同的标注侧重点，这导致了“三要素”标注不够全面，标注也往往较为粗略。不少质量较高数据集并不完全公开，对资源获取造成一定困难。因此，构建一个规范、标注内容丰富、细颗粒度的大规模情感语料库，以及具有较强的通用性、科学合理性、可迁移性的标注规范是非常必要的。

3 标注体系及标注规范

本文标注体系旨在对汽车产品评论短文本进行“三要素”细颗粒度的标注，共设置两类标签：（1）基础标签，“三要素”及对应的类别进行标注，最终形成评论“六元组”；（2）特殊标签，对隐含表达、建议类等文本、情感修饰词、主观情感词等进行单独标示。



图1：基于标注平台的示例文档

^③ <https://www.datafountain.cn/projects/2018CCF/index.html>

^④ <https://www.datafountain.cn/competitions/310/ranking?sch=1409>

^⑤ 第八届中文倾向性分析评测大纲（COAE2016）

^⑥ <http://www.cips-cl.org/CCL2018/index.html>

此外,基于 JAVAEE 的 MVC 框架并结合 JavaScript 及其他工具,开发了语料标注平台,支持在线提交和修改,并将数据以元组的形式存至后台数据库。标注示例如图 1 所示。

3.1 基础标签

基础标签标注主观文本中的评论对象、属性及极性“三要素”,不仅对文中的相关具体词语进行标注,同时标注该词语所属的类别标签,对每一要素以“词+标签”的形式进行标明,最终形成“六元组”形式,不仅对存在情感要素省略或隐含的文本补足情感要素信息,同时可以满足情感信息抽取、主题识别等多种应用需求。对象和属性的类别标签根据本体知识预先人工设定,并形成知识本体树;极性的标签主要采用三元分类:积极、消极和中性。对象和属性的词或短语,可以和标签位置的词相同,也可以是标签位置词的相关词或同义词。结合图 1 中的标注样例,基础标签的名称及说明如表 1 所示:

表1: 基础标签类型和说明

类别	标签名称	标签说明
对象	Target_pos (对象位置)	对象词
	Target_type (标签)	评论对象类型标签
属性	Attribute_pos (属性位置)	属性词
	Attribute_type (标签)	评论属性类型标签
极性	Polarity_pos (极性标签)	情感词
	Polarity_value (极值)	评论极性极值

评论对象是指评论面向的实体,即汽车的子系统或具体部件,如“后备箱”、“方向盘”、“发动机”、“座椅”等,具体分为对象词(Target_pos)和类别标签(Target_type)。标注对象的具体词语,既可以和对象类别标签相同,如“天窗”,也可以是类别标签的相关词、同义词或下位词。在类别标签的设定方面,考虑到汽车领域评论专业名词较多、表述相对单一等特点,为了更好地规范评论对象的标注以及为自动标注提供资源,参考姚天昉等^[14]的研究成果人工构建了汽车专业名词的知识本体库。该库采取树形结构,包括了各类标签下对应的汽车部件和具体表述。汽车专业名词知识本体树的内容示例如图 2 所示:

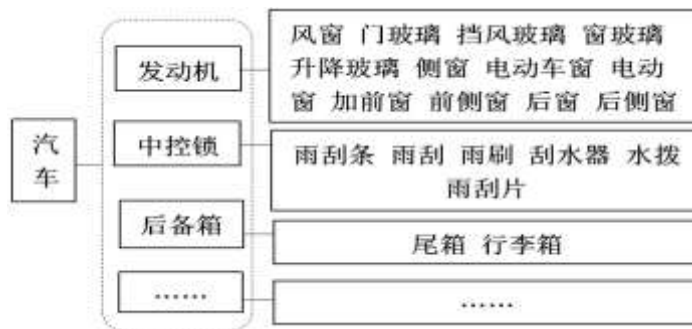


图2: 汽车专业名词知识本体树样例

最终,共设置 32 类对象标签,基本能涵盖汽车领域评论文本的实体。

值得一提的是,产品评论文本具有篇幅短小、省略主语等特点,这对评论对象的标注造成了困难。本体系主要有三种处理方法:一是默认评论对象为“汽车”这一整体且进行省略,如“(这辆车)外观特别霸气。”,对此赋予对象标签“汽车”即可。二是评论对象省略了汽车的某一具体部件,且该评论对象可通过先验知识的推断获得,如“没什么推背感!”中“推背感”主要针对的是汽车发动机的性能,因此赋予其对象标签“发动机”。三是由于评

论文本较短，评论对象出现在上下文中，且本句中无法通过推断进行判断，此时暂不予以对象的标注。

评论属性是指评论针对的具体方面、角度，属性标签同样也分为属性词（Attribute_pos）和类别（Attribute_type）。属性是对象的特征，如舒适性、空间、材质，能够从句子中直接体现出所关涉的属性的词，即为属性词；属性词有时与选择的标签相同，如图 1 中的“舒适性”；但更多情况不同，如句子“后背座椅靠背角度实在不敢恭维”的属性标签为“舒适性”，属性位置则为“靠背角度”。属性词省缺时同样可以赋予其标签。目前，知识本体库中的属性类别标签共有 12 个，分别为安全，材质，操控，动力，服务，价格，空间，实用性，舒适性，外观，油耗，质量（故障）。

极性主要标注短文本评论的情感倾向性，具体分为情感词（Polarity_pos）和极值（Polarity_value）。情感词实际上可以分为体现评论者态度的情感单元和情感词。情感单元以情感词为核心，包括了影响情感倾向或者强度的上下文的信息单元，是判断文本情感倾向的直接依据。隐含表达（即无明显情感词的表达）也是情感语料的常见表述，将在特殊标签部分进行重点说明。

极值是指句子的情感倾向，在标注体系中采取粗颗粒度的极值，分为 1、0、-1 三种，对应积极、中性和消极，对应关系如表 2 所示；值得说明的是，“0”对应中性情感，如“无功无过”、“一般”等；对于没有情感极性的句子，极性位置不予以标注。

表2：极值标签说明

标签名称	标签说明	情感单元/情感词范例
1	正向、积极的情感	喜欢；满意；比较软；漂亮；
-1	负向、消极的情感	生气；太贵；耗油；太小
0	中性情感	无功无过；规规矩矩；一般般

3.2 特殊标签

本标注体系设置了四类特殊标签，在基础标签“三要素”标注的基础上，分别对情感修饰词（程度副词和否定副词）、不含明显情感词的隐含表达、不能对产品特征进行描述的主观情感词及建议类文本等进行标识，一是为了对具有表达特色的内容进行细化标注，以适应更多的应用场景，如对隐含表达的标注；二是补足了基础标签无法充分标注的文本类型，如建议类文本。

情感修饰词标签主要标注对情感单元进行修饰的程度副词（如“有点”、“不太”）和否定副词（如“不”、“没有”）。否定副词造成极性反转，影响情感倾向性的判断，程度副词不仅能影响情感极性的强度，少数程度副词也能造成极性反转。对上述两类情感修饰词的标注将有利于情感计算等更加复杂的任务的实现。

隐含表达标签主要针对不含有明显情感词的短语，如“这个车跑长途也不累”的<Polarity_pos>为“跑长途不累”，同时赋予其隐含表达的标签。隐含表达尽管可以进行“三要素”的标注，但由于其标注内容灵活性较大，在情感词抽取等任务中识别效果并不理想，因此对这类句子赋予特殊标签。

主观情感词标签主要针对情感词和评价词的区分。就“喜欢年轻动感的颜色”而言，“年轻动感”是对外观这一属性的评价和描述，带有褒义，间接表达了对产品的肯定和认可，这类词往往需要借助先验知识和具体语境进行推断所表达的情感极性，本文称之为“客观评价词”；而“喜欢”作为心理动词单纯表达观点持有者的个人情感好恶，不涉及对对象及具体属性的概述，在词义上也具有绝对的积极色彩，本文称之为“主观情感词”。主观情感词与评价词的最大区别体现在能不能作为商品的描述词，本标注体系对此进行了区分，为不同需求的情感词抽取任务提供数据。

特殊文本标签标注语料库中含有特殊表达类型的句子:这些句子抽取出的六元组和短句的实际情感倾向相反。这类句子主要是建议类文本,以如下文本为例:

例 2: 要是倒车雷达倒车影像就好了。

其特殊性在于,能抽取的要素“<倒车雷达-实用性-有>”呈积极情感倾向,但是与整个短句观点持有人的主观情感相反,因此为这类句子赋予特殊标记标签。除建议类文本,类似还有比较句、反问句、反问句等。

3.3 数据来源与标注

语料选取方面,借助爬虫技术从“汽车之家”^⑦等产品论坛上抓取汽车产品评论语料,进行去噪后以句号、空格、叹号等标点作为断句标记进行分句,并根据哈工大 ltp 词性标记集^⑧进行词性自动标注,词性标签共 29 类;最终共获得 11810 句评论文本。

标注方面,由 2 名语言学专业硕士研究生进行标注,由 1 名计算语言学专业老师和一名计算机专业研究生进行检验。首先对标注人员进行沟通培训,标注人员随后选取同样的 200 条语料进行标注,利用 Cohen's kappa^[15]的 Kappa 值检测其标注一致性,如表 3 所示。其中 001-200 的数据为第一次标注的一致性统计,在讨论和研究后,二次标注并统计 Kappa 值,最终标注一致性达到不错的效果。最终,由该 2 名人员分别标注不同的语料,并进行统一检验:标注平台也开发了“疑难提交”功能,保证标注一致性和质量。

表3: 标注一致性结果检测

Kappa 值	对象	属性	极性
001-200	0.730	0.626	0.759
201-400	0.810	0.659	0.749

此外,标注平台支持对分词和词性标注的更改以及对六元组的数量的增加或删除,“三要素”对应词语可以借助分词结果,以词语在句中位置索引填入,简化标注操作。

4 语料库统计分析

在上述标注体系、标注人员及语料的获取基础上,本文完成了语料库的构建工作。本节将从语料库基本数据、基础标签和特殊标签统计三个方面给出相关数据,并介绍基于本语料库的自动标注实验。

4.1 语料库基本数据

经统计,语料库最终共计 9343 句有效句^⑨、103897 字,达到建库要求;平均每个有效句句长为 11.12 字,体现标注对象的短文本特征;具体如表 4 所示:

表4: 基础标签基本情况

类型	数量	平均长度(字)	平均数量/句
Target_pos	5271	2.64	1.49
Attribute_pos	4655	2.38	1.18
Polarity_pos	8408	3.51	3.16

语料库共获得 11073 组“三要素”标注信息,平均每一有效句标注 1.19 组“三要素”,其中,倾向为积极的 5672 组,消极的为 2870 组,中性的为 224 组。每一要素的相关具体数据如表 4 所示,平均每句句数最多的情感要素是极性,在文中表述使用的平均字数也最多。对象和属性的数量和字长的分布情况相差不大。

^⑦ <https://www.autohome.com.cn/beijing/>

^⑧ <http://ltp.ai/docs/appendix.html#id3>

^⑨ “有效句”是指标注了情感信息的句子

4.2 语料库基础标签统计

与之前的语料库相比，本语料库对“三要素”进行了细致而完整的标注，以往细颗粒度语料库往往除了极性之外，只标注对象或者属性。表 5 反映了本语料库情感要素中评论对象和属性标注的共现情况，表中的“广义共现”是指该短句同时具有<Target_type>及<Attribute_type>标签，而<Target_pos>及<Attribute_pos>可以为空的情况，即该句在语义上存在评论对象和评论属性两类信息，但省略了相关词语。从表格得知，这类“广义共现”在有效句中占比达到 77.54%，说明大部分评论是同时囊括了属性和对象两个要素的，因此对属性和对象分别进行标注非常必要。

表5：评论对象和属性标注共现情况

类型	数量（句）	比例 ^⑩
广义共现	7245	77.54%
严格共现	1413	15.12%

“严格共现”是指在<Target_type>及<Attribute_type>均不为空的前提下，<Target_pos>及<Attribute_pos>同样不为空，即类似于“方向盘手感很好”在表达层面没有对评论对象和属性进行省略的句子。由于评论文本短小简洁，往往习惯性地省略部分不影响理解的信息表述，这种情况的句子数量较少占比较小，如“这款车太贵”，尽管根据常识可知其属性为“价值”，但语言表达中很少将其表述为“这款车价格太贵”，因而造成了存在属性类别但<Attribute_pos>为空；再如“坐着太硬”，评论对象省略了“座椅”这一实体的表述。表 4 中属性位置数量和对象位置数量均较少也印证了这种省略现象的频繁。因此，能够根据上下文进行情感要素的归纳或补全是十分重要的，这也侧面反映了情感信息抽取任务的难点以及构建类别标签的必要性。

4.3 语料库特殊标签统计

在特殊标签的统计方面，表 6 对各类特殊标签的数量和占比对情感修饰词（程度和否定副词）在有效句中的使用情况进行了说明，可以看出，程度副词作为修饰的有效句比例较高，否定副词尽管数量较小，但其对于极性判断有着至关重要的影响，这也侧面证明本文对包含程度副词和否定副词的情感修饰词共同作为情感单元的标注于<Polarity_pos>，有利于情感要素的抽取、情感极性判断、情感计算等任务。

表6：特殊标签数量及占比表示

类型	情感修饰词		主客观情感词		特殊文本及隐含表达占比	
	程度副词	否定副词	主观情感词	客观评价词	特殊文本	隐含表达
数量（句）	2806	255	608	6675	416	427
比例	32.68%	2.73%	6.51%	71.44%	4.45%	4.57%

在主观的心理动词和客观评价词的统计方面，含有主观情感词的有效句共 608 句，客观评价词 6675 句，在有效句中高达 71.44%。因此，客观评价词依然是倾向性分析的研究重点。

特殊标签还标注了特殊文本和隐含表达。在特殊文本主要有建议类文本、比较句、反问句、反讽句四种类型，其中建议类和比较句文本最为常见。隐含表达是目前情感信息抽取和情感倾向性判别的难点，但是在本语料库中仅占 4.57%，这是因为通过“三要素”的信息抽取，不少看似为隐含表达的句子可以得到结构化的处理，且汽车领域的评论行业用语较多，用户们往往能用较为专业和精准的词汇表达个人情感态度，因此一定程度上削减了隐含表达

^⑩ 比例是指在有效句中的比例；下同。

出现的频次。总之，特殊文本和隐含表达尽管在语料库中占比较小，但是这类文本的情感倾向性判断具有一定的难度，本语料库则为相关研究提供了一定的语料基础和准备。

5 自动标注实验

为验证语料库细颗粒度的标注体系可计算性并检验标注质量，抽取了部分语料进行了“三要素”自动标注实验。实验采用 Bi-LSTM-CRF（双向长短时记忆网络-条件随机场）神经网络模型结构，并使用字标签的标注方法进行自动标注。Bi-LSTM-CRF 这一模型结构最早由百度研究院的 Huang^[16]等提出，其可以从标注语料中自动学习特征，CRF 可以对学习到的特征进行建模，给每一个字选择一个基于整句最优的字标注标签，最后用 viterbi（维特比）算法获得整个句子最优的标签序列。架构图如图 3 所示：

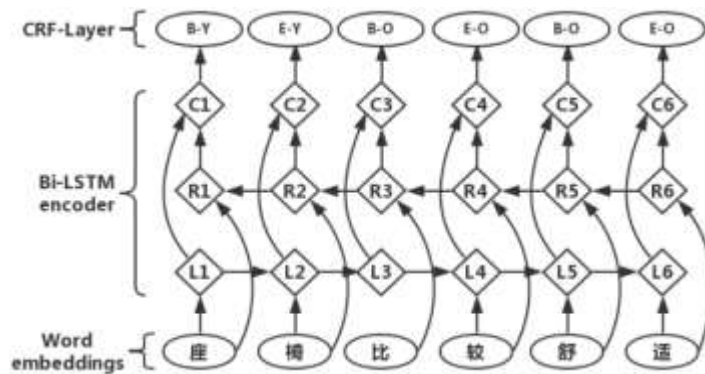


图 3：Bi-LSTM-CRF 架构图

将训练好的字向量作为 embedding 层加入 Bi-LSTM，L 代表当前字符上文，R 代表当前字符下文，C 代表当前时刻的上下文特征，将 Bi-LSTM 的输出输入到 CRF 层，通过 CRF 输出当前整个标签序列最优的标签序列。以属性词识别为例，每个标签由字在词中的位置（B、I、E、S）、词性标记、属性标签（Y）或非属性标签（N）组成，例如：“外 B-n-Y 观 E-n-Y”，B 表示该字位于词首、E 表示该字位于词尾，n 表示该字所属词性是名词，Y 表示该字所属词为属性词，通过这种字标注的方法，模型就把“外观”这个属性词识别出来了。

将语料按照 9：1 的比例分为训练语料和测试语料。对于自动标注结果的评价，采取 P（Precision）、R（Recall）和 F1（F1-Measure）作为评价指标。表 7 为最终的实验结果：

表 7：实验结果

类型	P	R	F1
对象	70.00%	62.31%	65.93%
属性	71.43%	67.21%	69.25%
极性	70.43%	62.33%	66.13%

实验结果表明，“三要素”细颗粒度标注体系能分别进行评论对象、属性和情感单元的自动标注并能取得不错的效果，对评论对象、属性和情感单元的标注正确率均能达到 70% 以上，而三项的 F1 值也接近 70%，与国内现有的评测结果相比^[12-13]，取得了不错的效果。这不仅说明本语料库的标注质量总体上有一定的保证，更说明了产品评论文本的细颗粒度标注方法具有较强的可计算性，也更有利于面向计算机的实际应用。

6 总结与未来工作

本文针对产品评论文本的情感分析任务,选取汽车评论文本,构建了对评论对象、属性和极性“三要素”进行充分标注的细颗粒度语料库,不仅标注了具体的词语,还将其所属的具体类别对应到知识本体树上,以更好服务于情感信息抽取、主题识别、情感倾向性分析等任务。针对含有情感修饰词、隐含表达、特殊文本、主观情感词等类型的文本,用六元组进行标注后赋予了特殊标签,以应用于更加精细的情感计算,为特殊类型的文本情感分析提供语料。最终,通过对标注信息的统计分析和基于语料库的自动标注实验,验证了本语料库标注体系不仅能更细致、全面地获取主观文本的情感要素,同时在实际应用中具有较强的可计算性,语料标注也有一定的质量保证。未来的工作,将进一步扩大语料库规模,并选取其他领域的产品评论文本进行标注,对该标注体系在不同产品领域的应用情况进行比较和分析;此外,利用现有语料和标注体系开发情感要素的自动标注器也是下一步的重要工作。

参考文献

- [1] 赵妍妍, 秦兵, 刘挺, 等. 文本情感分析[J]. 软件学报, 2010, (8): 1834-1848.
- [2] Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language[J]. Language resources and evaluation, 2005, 39 (2-3): 165-210.
- [3] Wiebe J, Wilson T, Cardie C, et al. Annotating Expressions of Opinions and Emotions in Language[J]. Language Resources and Evaluation, 2005, 39 (2-3): 164-210.
- [4] Hu M, Liu B. Mining opinion features in customer reviews[C]// Proceedings of the AAAI 2004. 2004:755-760.
- [5] Pang B, Lee, Vaithyanathan S, et al. Thumbs up? Sentiment Classification using Machine Learning Techniques[C]// Proceeding of EMNLP -20. 2002:79-86.
- [6] Li B, Zhou Liu, Feng S, et al. A Unified Graph Model for Sentence-based Opinion Retrieval[C] // Proceeding of ACL. 2010:1367-1375.
- [7] 戴敏, 朱珠, 李寿山, 等. 面向中文文本的情感信息抽取语料库的构建[J]. 中文信息学报, 2015, 29 (4): 67-73.
- [8] Cigdem Toprak, Niklas Jakob, Iryna Gurevych. Sentence and Expression Level Annotation of Opinions in User-Generated Discourse[C]// Proceedings of ACL 2010. 2010: 575-584.
- [9] Zhuang L, Jing F, Zhu X, et al. Movie review mining and summarization [C]// Proceedings of CIKM-2006. 2006:43-50.
- [10] 徐琳宏, 林鸿飞, 杨志豪, 等. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1):96-100.
- [11] 王中卿, 王荣洋. 情感倾向性分析技术报告[C]//第三届中文倾向性分析评测 (COAE2011). 2011.
- [12] 杨艳, 徐冰, 杨沐昀, 等. 一种基于联合深度学习模型的情感分类方法[J]. 山东大学学报, 2017, 52(9):19-25.
- [13] 谭松波, 王素格, 廖译文, 等. 第五届中文倾向性分析评测总体报告[C]//第五届中文倾向性分析评测 (COAE 2013).
- [14] 姚天昉, 聂青阳等. 一个用于汉语汽车评论的意见挖掘系统[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集. 2006: 260-281.
- [15] Toprack C., Jakob N., and Gurevych I. Sentence and Expression Level Annotation of Opinion in User-Generated Discourse[C]//Proceedings of ACL-2010. 2010:575-584.
- [16] Huang Z, Xu W, Yu K, et al. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.



曹紫琰 (1994—), 硕士生, 主要研究领域为计算语言学。
E-mail: caoziyannnu@foxmail.com



毛雪芬（1994—），硕士生，主要研究领域为计算语言学。
E-mail: 170102035@stu.njnu.edu.cn



程宁（1993—），硕士生，主要研究领域为计算语言学。
E-mail: chengninmo@foxmail.com

作者联系方式：曹紫琰 地址：江苏省南京市鼓楼区宁海路 122 号南京师范大学随园校区文学院
邮编：210097 电话：15601599238 电子邮箱：caoziyannnu@foxmail.com