基于语义分析的评价对象-情感词对抽取

江腾蛟 万常选 刘德喜 刘喜平 廖国琼

(江西财经大学信息管理学院 南昌 330013)

(江西财经大学数据与知识工程江西省高校重点实验室 南昌 330013)

摘 要 评价对象-情感词对是情感词及其所修饰评价对象的组合,评价对象-情感词对的识别是细粒度情感分析的一个原子任务和关键任务.现有的研究大多集中在商品评论上,随着金融大数据的涌现,金融评论的情感挖掘意义凸显.与商品评论不同,中文金融评论中评价对象数目繁多且构成形式复杂,虚指评价对象和隐式评价对象也更常见;情感词的词性更丰富,其在句中的句法成分也更灵活、语义更丰富.针对金融评论的这些特点,该文提出了基于浅层语义与语法分析相结合的评价对象-情感词对抽取方法.考虑到金融评论多动词情感词,设计了语义角色标注与依存句法分析相结合的评价对象-情感词对抽取规则,解决了评价对象构成的复杂性问题;基于语义和领域知识对虚指评价对象进行了判别和替换,以明确其实际的指向和含义;基于特殊情感词搭配表、上下文搭配表及频繁搭配表提出了隐式评价对象识别的新思路,能有效地识别出缺省和隐含评价对象.在大规模的中文金融评论上进行了详细的实验测试,实验结果表明了该方法的有效性.

关键词 情感分析;中文金融评论;评价对象-情感词对;语义角色标注;依存句法分析

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2017.00617

Extracting Target-Opinion Pairs Based on Semantic Analysis

JIANG Teng-Jiao WAN Chang-Xuan LIU De-Xi LIU Xi-Ping LIAO Guo-Qiong

(*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang* 330013)

(*Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang* 330013)

Abstract A Target-opinion pair is a combination of opinion word and the target it modified. The extraction of target-opinion pairs is an atomic and key task of fine-grained sentiment analysis. Most existing work on extracting target-opinion pairs considers product reviews. As more and more financial data accumulate on the Web, sentiment mining of financial reviews becomes an important task. In this work, we put focus on Chinese financial reviews. Compared with product reviews, Chinese financial reviews have some unique characteristics. First, the number of targets is very large in Chinese financial reviews and the structure of targets is usually more complex. Second, it is very common for a Chinese review to have ambiguous and implicit targets. What is more, the opinion words in Chinese financial reviews are more flexible in POS (Part-Of-Speech) and syntactic roles, and richer in semantics. In this paper, based on shallow semantic and syntactic parsing, we propose a new method for extracting target-opinion pairs from Chinese financial reviews. Considering that many opinion words in financial reviews are verbs, we design extracting rules of target-opinion pairs based on semantic role labeling and dependency parsing, which reserves

收稿日期:2015-05-25;在线出版日期:2016-01-24.本课题得到国家自然科学基金项目(61562032,61662027,61662032,61173146,61363039,61363010,61462037)、江西省自然科学基金重大项目(20152ACB20003)、江西省高等学校科技落地计划项目(KJLD12022,KJLD14035)资助.**江腾蛟**,女,1976 年生,博士研究生,讲师,主要研究方向为情感分析、Web 数据管理.E-mail:tj_jiang@163.com.**万常选**(通信作者),男,1962 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为 Web 数据管理、情感分析、数据挖掘、信息检索.E-mail:wanchangxuan@263.net.**刘德喜**,男,1975 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为 Web 数据管理、信息检索、自然语言处理.**刘喜平**,男,1981 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为信息检索、数据挖掘、Web 数据管理.**廖国琼**,男,1969 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为数据库、数据挖掘.

the structural complexity of targets. Our method distinguishes and replaces ambiguous targets based on semantics and domain knowledge, making the actual references and meanings of ambiguous targets explicit. To identify the implicit targets, we propose a novel approach based on the one-to-one correspondences of some special opinion words and targets, contextual co-occurrences of targets and frequent co-occurrences of targets, which is able to effectively identify the default and implied targets. We conduct comprehensive experiments on Chinese financial reviews, and experimental results show that the proposed methods are effective.

Keywords sentiment analysis; Chinese financial reviews; target-opinion pair; semantic role labeling; dependency parsing

1 引 言

细粒度的文本情感挖掘,是近年来数据挖掘和自然语言处理领域的热门研究.细粒度的文本情感挖掘又称为基于评价对象(特征)的情感挖掘,它需要挖掘出文本中每个评价对象的情感倾向性.评价对象-情感词对又称为评价搭配或情感评价单元,是用户评论中情感词及其评价的对象的搭配对.Bloom 等人[1]认为"在评论文本中,情感词总是针对具体的评价对象的,情感词及其评价对象的抽取是情感挖掘研究中的基础和重要的任务,这是因为评价对象-情感词组合所包含的信息比它们单独的任何一个都要多得多."

评价对象-情感词对的抽取,不仅可以解决情感抒发的对象问题,还可以解决评价对象对情感极性的影响问题.评价对象-情感词对的极性除了受情感词的语义极性影响,它还受到评价对象的影响,如情感词"下降"的语义极性为负,"利润下降"为负极性情感短语,但"成本下降"为正极性.即同一情感词修饰不同的评价对象时,其极性也可能不同.文献[2]将使得动态情感词的极性发生反转的评价对象称为奇异特征.因此评价对象-情感词对的识别有助于进一步判定该评价对象的情感倾向性.

现有的情感挖掘相关研究多集中在商品评论领域,文献[3]中指出"Web 金融评论具有实时性、全面性和真实性,基于 Web 金融评论构建和量化金融指标体系,用于上市企业财务预警模型是一项非常有意义的工作",同时文中也指出针对 Web 金融评论的情感挖掘是一件非常棘手的工作.

面向中文金融评论中的评价对象-情感词对抽取比商品评论的更加艰难.这是由于:

(1)评价对象数目繁多且构成形式更复杂

在商品评论中,评价对象一般指商品的特征或属性;在金融评论中,评价对象可以是国家政策,如"政策利好企业发展"中的"政策",财务报表中的一个子项,如"原材料上涨"中的"原材料";也可以是一个话题,如解读公司高层人事变更公告等.因此,金融评论中的评价对象包括财务指标、非财务指标及其子项,数目繁多.

商品评论中的评价对象一般为名词或名词短语,如手机评论中的"苹果"、"屏幕"和"按键布局"等;在金融评论中,评价对象除了名词或名词短语,如"原材料"、"宏观经济政策"等,还可能是从句,如"佣金率下降趋缓."一句中,情感词"缓"的评价对象为主语从句"佣金率下降".

(2)情感词在句中语法成分更灵活

在商品评论中,情感词多为形容词,现有的研究也多基于形容词情感词进行情感分析.在金融评论中,情感词的词性更为丰富,除了形容词,还有动词、名词,尤为突出的是动词情感词,如"下降"、"飙升"等.根据我们目前整理的金融域情感词典,动词情感词占情感词的比重为 34.3%,名词情感词占情感词的比重为 11.9%.

金融评论中情感词词性丰富,使得其在句中的位置和语法成分都更加灵活,相应的基于情感词的评价对象的抽取也就困难得多.

(3)虚指评价对象更常见

由于金融评论中评价对象的繁多及组成形式复杂,而中文表达要求言简意赅,因此,很多名词或名词短语虽然在句中充当主要成分,但本身却不能清楚表达语义,需结合上下文来理解,这类评价对象我们称之为虚指评价对象.如"1 季度投行业务净收入 1.72 亿元,同比下降 20%,贡献比降至 12.2%。"中

的“同比”和“贡献比”为虚指评价对象，若简单地将“同比”或“贡献比”列为评价对象，则很难理解其语义，不利于进一步情感分析中的评价对象分组.

(4) 隐式评价对象更频繁

由于中文的语言特点，在上下文提示下，句中经常会出现缺省主语或宾语，造成评价对象缺失；或根据特定词汇的暗示直接隐含评价对象. 为了简述，文中将缺省及隐含评价对象统称为隐式评价对象. 中文的语言表达特点决定了隐式评价对象更频繁.

同时，相比于商品评论，金融评论文档更正式、专业. 从长度上来看，金融评论文档及句子的长度均更长. 句子内的各个子句常从各个角度来描述同一主题评价对象；上下句间更注重语义相关性. 而商品评论往往整个评论只有一句，且该句内描述了商品各个属性的评价.

针对上述分析，本文的目标是在中文金融评论中根据情感词抽取其所修饰的评价对象，以及虚指评价对象的发现和替换、隐式评价对象的识别. 首先，根据情感词的词性及语法、语义分析，确定情感词与其修饰的评价对象间的语法路径及语义联系；其次，根据浅层语义分析，判定情感词为谓词及谓词的关联成分时其对应的评价对象；接着利用依存句法分析补充判定情感词在句中充当其他的语法成分时其对应的评价对象；然后，利用领域知识设计了虚指评价对象的发现及替换方法，利用领域知识和上下文语义知识设计了特殊情感词、上下文语义关联及频繁评价搭配 3 种方法识别隐式评价对象；最后，实验证明了本文方法的有效性.

本文的主要贡献包括：

(1) 从语法和语义的角度，分析了不同词性的情感词在句中可能充当的语法成分，接着进一步从语义上分析了情感词在句中充当不同语法成分时，情感词与其修饰的评价对象在句中的语法路径及语义角色的关系.

(2) 充分考虑评价对象构成的复杂性，设计了浅层语义角色分析下的评价对象-情感词对抽取规则，完备性起见，进一步设计了依存句法分析下的评价对象-情感词对抽取规则和评价对象的扩展规则.

(3) 根据评价对象的意义是否明确，结合领域知识提出了虚指评价对象的识别及替换方法.

(4) 根据特殊情感词与评价对象的一一对应关系，上下文语义的关联性及频繁评价对象-情感词搭配对，分 3 种情况有效地解决了隐式评价对象的识别问题.

(5) 以新浪财经的公司研究为数据源，通过详细的对比分析，验证了评价对象-情感词对抽取方法的有效性，同时也验证了虚指评价对象发现和替换、隐式评价对象识别的有效性.

本文第 2 节介绍相关工作；第 3 节介绍抽取评价对象-情感词对的语法和语义分析基础；第 4 节重点介绍基于语义角色标注和依存句法分析的评价对象-情感词对抽取规则，并同时设计了评价对象的扩展规则；第 5 节提出虚指评价对象、隐式评价对象的解决方案；第 6 节给出了评价对象-情感词对的抽取算法；第 7 节为实验及分析部分；最后部分是总结.

2　相关工作

现有的情感挖掘研究多集中在商品评论领域，在商品评论中，评价对象又称为特征(feature 或aspect)，情感词又称为极性词、观点词(opinion 或 attitude). 文献[1]首次提出了情感评价单元(appraisal expression)的概念，将情感评价单元定义为三元组〈*target*，*attitude*，*source*〉，其中 *target* 是评价对象，*attitude* 是情感词，*source* 是评价来源；文献[4]将评价对象-情感词对表示为二元组〈评价词语，评价对象〉. 评价对象-情感词对的抽取方法主要有以下 3 类：

(1) 基于最近距离的方法

文献[5]应用关联规则抽取频繁名词或名词短语作为特征，并把距离特征最近的形容词作为情感词. 文献[6]把动词或形容词作为情感词，然后把距离情感词 K 窗口内的名词或名词短语作为特征.

这类方法认为评价对象与情感词总是最近距离搭配的，存在的主要问题：① 没有考虑评价对象与情感词的长距离搭配；② 由于直接将名词短语作为特征，导致特征识别错误；③ 没有考虑特征与情感词间存在标号分隔；④ 无法解决隐式特征.

(2) 基于机器学习的方法

文献[7]通过隐马尔可夫模型来描述评价对象(商品特征)与情感词之间的语法依存关系，并结合上下文一致性同时提取评价对象和相应的情感词. 文献[8]使用条件随机场模型和规则相结合提取评价对象，根据依存关系和近邻法提取相应的情感词和评价对象. 文献[9]提出了商品导向的领域本体建模方法，将情感词与本体相结合，从中文评论文本中抽取特征-观点对，同时解决了与特征词一一对应的特殊情感词的隐含特征的识别问题. 文献[10]基于

条件随机场来识别商品特征，然后根据特征的最近邻和句法树来发现情感词. 文献[11]提出 ASUM 模型(Aspect and Sentiment Unification Model)，模型的主要任务是识别特征和极性，通过进一步的假设，即每个句子只有一个特征和相关联的极性，对 JST 模型进行了扩展. 作者还提出了SLDA(Sentence-LDA)模型，模型假设一个句子中词的生成都来自于一个特征，其主要任务就是发现特征；然后对 SLDA 进行扩展，将情感词和特征词合并从而实现对不同特征的情感进行建模，用来发现情感-特征对. 文献[12]通过统计情感词与评价对象的共现和两者间的依存模式，构建了一个情感图模型来发现评价对象和情感词. 文献[13]利用主题模型和词对齐方法来捕获评价对象间、情感词间及评价对象与情感词间的语义和情感关系，利用随机游走模型来估算候选词的置信度，通过这些方法的融合有效地抽取了评价对象和情感词. 文献[12-13]通过评价对象与情感词的语法和语义联系来发现评价对象和情感词，但都没有实现将评价对象与情感词成对抽取. 文献[14]利用词汇语义和上下文语义抽取商品特征，并通过实验证明基于语义的方法明显优于基于语法的方法.

有监督的机器学习方法将评价对象和情感词作为一个序列标注任务来识别，需要人工标注训练集. 在当前评论数据与日俱增的情况下，已标注的数据以越来越快的速度被淘汰，而新的数据还来不及标注就可能已经过时了，因此人工标注训练集的时效性和可迁移性差. 主题模型方法是一种无监督的机器学习方法，它在主题生成过程中较易发现海量文档下高频出现的评价对象，但却很难发现低频评价对象和局部文档中的高频评价对象.

(3)基于句法关联或规则的方法

文献[15]提出了一个无监督的信息抽取系统 OPINE，首先抽取名词和名词短语作为商品特征，接着基于依存句法制定了 10 个抽取规则来识别情感词. 文献[1]利用 Stanford Parser 手工构建了评价对象和情感词之间的搭配规则，以抽取评价对象-情感词对. 该方法能够很好地消除一些形容词的二义性，然而由于规则全部由人工构建，导致实验的召回率不高. 文献[16]在后续工作中使用置信度方法自动训练抽取规则，使得召回率有所上升，但准确率又略低于手工构建规则的方法.

文献[17]考虑到情感词的领域依赖性，提出一种半自动的搭配规则，为情感词和评价对象之间的关系定义了 8 个模板，由于模板过于简单，而且修饰关系仅仅停留在词表面，导致产生了大量的候选评价对象和候选情感词，需要人工筛选来完成评价对象-情感词对的抽取.

文献[18]在依存句法分析的基础上，首先使用限定词性的方法建立评价搭配候选集合，再使用最大熵模型的方法对候选评价搭配进行筛选，得到最终的评价搭配集合. 这种方法挖掘了评价对象和评价词的结构关系，相比使用最近距离进行匹配的方法有了很大改进，但是该方法限定了评价对象和评价词的词性，并且假设评价对象和评价词在一个单句中，对于评价对象和评价词分别在两个子句中的情况，无法做出正确处理.

文献[19]利用 SRL(语义角色标注)技术实现了从在线新闻文本中提取观点(情感词)、观点持有者以及评价对象. 文献[20]提出了一种“double propagation”方法，实现情感词和评价对象的同时识别与抽取. 文献[21]在对评论文档进行语言学和语义分析的基础上设计了相关规则，以实现评价对象-情感词对的抽取.

国内的学者也进行了相关研究. 文献[2]基于依存句法分析总结出“上行路径”和“下行路径”的匹配规则，进而总结出 SBV 极性传递的一些规则，用于识别评价对象-情感词对.

文献[4]提出了一种基于短语句法的句法路径，该句法路径是评价对象到评价词之间的句法结点序列，使用句法路径匹配的方法抽取评价对象-情感词对. 该文使用模式匹配的方法，首先通过短语句法分析获取评价对象与评价词的句法路径，再通过泛化、筛选等过程得到句法路径库，以此作为模式库采用匹配算法进行评价搭配的抽取.

文献[22]利用依存句法分析结果分别建立了名词、动词及形容词块规则，从而分别抽取出评价对象和评价词；在此基础上进一步利用词与词之间的搭配关系，设计评价对象与评价词的搭配算法.

文献[23]利用 SBV 极性传递法识别需抽取的评价对象和评价词，并引入 ATT 链算法以及互信息法确定评价对象的边界，进一步挖掘了评价对象与评价词的语法关系和评价对象的语义信息.

句法分析可以反映出句子各成分之间的语义修饰关系，它可以获得长距离的搭配信息，并与句子成分的物理位置无关. 因此，近几年对评价对象-情感词对的抽取研究都融入了句法分析，无论是基于机器学习的方法还是基于模板或规则的方法. 对于评价对象-情感词对的识别，基于模板的方法准确率

高，但现有的研究只分析了 SBV 主谓结构或 SBV 主谓结构和 ATT 定中结构，没有对 24 种依存关系中的情感词和评价对象间的语法关系进行一一分析；基于句法路径匹配的方法，由于中文表达灵活，路径复杂，再加上虚指评价对象及隐式评价对象的存在，都导致正确率和召回率不高.

上述研究较好地推动了数据处理和情感挖掘研究的进展，但同时我们也发现仍存在以下问题：

(1) 忽略了评价对象中的动词部分，导致评价对象组成不完备. 如“收入增长 35%，佣金率下降趋缓，份额继续上升。”，现有的研究中，抽取的评价对象-情感词对为：“收入-增长”、“佣金率-下降”、“佣金率-缓”和“份额-上升”. 这里的 “佣金率-缓”显然错误，应该是“佣金率下降-缓”，“佣金率-缓”为负极性，而“佣金率下降-缓”为正极性. 评价对象中的动词很多时候决定了评价对象的极性方向.

(2) 评价对象意义不明确，不利于进一步的分组. 如“1 季度投行业务净收入 1.72 亿元，同比下降 20%，贡献比降至 12.2%。”中，若简单地将情感词“下降”、“降”的评价对象分别抽取为“同比”、“贡献比”，则其含义不明确，应进一步抽取到“投行业务净收入”. 评价对象的意义不明确，不利于评价对象的进一步分组和情感计算.

(3) 隐式评价对象解决不完备，隐式评价对象包括隐含和缺省两种情况，现有的研究较少关注这部分，文献[9]研究了隐含问题，但没有探讨缺省问题.

基于最近距离的方法和基于机器学习的方法都是以分词的结果作为评价对象单位，或是以组合的名词短语作为评价对象单位，而无法根据句子的上下文语义来区别对待评价对象的组成. 这对于商品评论是合适的，因为商品评论中的评价对象为商品特征或属性，一般为名词或名词短语. 而由前面的分析可知，金融评论中，评价对象除了名词或名词短语，如财务指标和非财务指标等，还可能是包含谓语的主谓结构或从句结构. 经统计，在人工标注的 32529 个评价对象中，从句结构的评价对象占比达到 21.6%. 欲得到主谓结构或从句结构等形式的评价对象，须在分词后结合评价对象和情感词在句中的上下文具体考虑，这一点在最近名词和机器学习方法中是较难实现的.

考虑到评价对象组成的复杂性及金融评论文本中多动词情感词的特性，本文试图从中文语法及语义理解上来剖解这个难题，既不需要大量的人工标注，又能够找出情感词真正语义上的评价对象. 同时借助领域知识和上下文语义，解决虚指评价对象和隐式评价对象问题.

3　评价对象和情感词在句中的语法、语义分析

3.1　情感词的词性及语法分析

评价对象-情感词对由情感词和评价对象组成，情感词是指表现出某种情感极性的词；评价对象是这种情感的承载者. 文献[24]从现代汉语语法的角度分析了大多数的情感词语属于形容词、副词、名词和动词等. 由前文分析，金融评论中情感词的词性丰富，包括了情感词语的所有词性，即形容词、动词、副词和名词；评价对象为名词、名词短语或从句. 通过我们构建的 3947 个金融域情感词发现，动词情感词占情感词的比重为 34.3%，名词情感词占情感词的比重为 11.9%. 形容词情感词为最常见情感词，现有的商品评论情感挖掘文章中也以形容词情感词的研究居多.

从语法结构上分析，在汉语句子的基本结构中，主语和谓语是句子的必要成分，是表达一个完整意思的基本语言单位；当谓语的中心词为及物动词时后面带一个对象，即宾语；定语是名词前面的修饰和限制性的词语；状语一般用在动词或形容词前面的修饰语；补语一般用在动词或形容词后起到补充说明的作用.

评价对象为名词、名词短语或从句，其在句中充当主语、宾语、主语从句或宾语从句. 下面根据情感词的词性，分析其在句中的语法成分：

(1) 形容词情感词. 其在句法结构中可以充当定语、谓语、状语或补语，其中定语既可以是主语的定语也可以是宾语的定语.

(2) 动词情感词. 其在句中可以充当谓语和宾语. 在商品评论中动词作宾语比较少，因此也较少有研究者讨论，而在金融评论中，存在着动词情感词在句中充当宾语. 如“第三季度收入和利润均呈现快速增长”一句中，动词情感词“增长”在句中充当宾语.

(3) 副词情感词. 英文中，副词情感词往往通过形容词加副词后缀组成，副词情感词也表达着形容词情感词的语义，因此，英文中评价对象-情感词对的抽取，必须要考虑到副词情感词的评价搭配对抽取. 中文中的副词情感词在句中为状语或补语修饰形容词或动词，起到改变形容词或动词的情感程度

作用,因此只需要找出形容词情感词或动词情感词的评价对象即可.

(4)名词情感词.其在句中可充当主语或宾语,也可充当谓语或定语等.

说明:

(1)一些特定的形容词情感词,如大、小、有限、多、少、上、下、核心、合理、发达等作定语时仅表示限定作用,并不是情感词.

(2)在中文评论中,副词的评价对象可通过所修饰的形容词情感词或动词情感词来体现,副词往往起到说明所修饰的形容词或动词的程度作用,因此本文中不抽取副词情感词的评价对象.

(3)名词充当句中的主语或宾语,这时的名词情感词本身就是评价对象,它所具有的情感极性是评价对象自身所具有的,如"优势"、"涨幅"为正极性,"劣势"、"跌幅"为负极性,负极性的名词情感词的作用如同奇异评价对象词,因此充当主语或宾语的名词情感词不需抽取其评价对象;名词情感词充当谓语或定语时,作形容词用,如"最近股市很牛"中的"牛",这类名词情感词则需要抽取其评价对象.

3.2 情感词的语法及语义分析

一个句子除了满足词法和语法要求外,还需要清楚地表达出语义.一般在主动句中,主语是谓词的施事者,宾语是动词的受事者;而被动句则相反.

(1)情感词充当句中的谓语.情感词充当句子的谓语,可以为形容词或动词:

①形容词情感词作谓语.其修饰的评价对象一般为主语,语义角色标注中表现为施事者,如"公司收益良好."一句中,"良好"的评价对象为"公司收益".

②动词情感词作谓语.其修饰的评价对象相对较复杂,当动词情感词为情感类心理动词时,其修饰的评价对象为谓词的客体,即宾语[25],语义角色标注中表现为受事者,如"我们看好公司的发展前景。"一句中,"看好"为心理动词,其评价对象为"公司的发展前景";当动词情感词为非心理动词时,其修饰的评价对象为主语,语义角色标注中表现为施事者,如"股价快速上涨"一句中,"上涨"为非心理动词,其评价对象为"股价".

(2)情感词作宾语.情感词修饰的评价对象不再单为某个名词或名词短语,而为句子的主谓结构,谓语影响评价对象的极性.

(3)情感词作谓语的状语或补语.情感词修饰谓语时,谓语的修饰对象为主语,因此当情感词作谓语的修饰成分时,其评价对象为句子的主谓结构.

(4)情感词为宾语的状语或补语.情感词修饰宾语时,宾语的修饰对象为句子的主谓结构,因此当情感语为宾语的修饰成分时,其评价对象为句子的主谓宾结构.

(5)情感词为定语.情感词为定语时,其评价对象为定语所修饰限定的对象.

4 评价对象-情感词对的抽取

记评价对象-情感词对为〈评价对象,情感词〉.语义角色标注的目标是从句子中识别出与谓词相关的语义角色,通过语义角色标注可以方便地找出谓词所修饰的评价对象,但通过第3节的分析可知,情感词在句中并非只是充当谓语,当充当其他成分时,则需要在语义角色标注的基础上进行扩展.

4.1 基于语义角色标注抽取评价对象-情感词对

语义角色标注为目前浅层语义分析的主要实现形式,各标注的含义如表1所示[26].

表1 语义角色标注中的标注名及含义

标注	解释	标注	解释
A0	施事	A1	受事
A2	间接作用对象		
DIR	方向	DIS	篇章标记
EXT	范围	LOC	位置
MOD	一般修饰	NEG	否定
MNR	举止	PRD	第二谓词
PRP	提议	REC	反义
TMP	时间	ADV	副词修饰

哈尔滨工业大学开发的语言技术平台(Language Technology Platform,LTP)[27]提供了一系列的汉语语言处理模块,其中包括分词、词性标注、命名实体识别、依存句法分析和语义角色标注等.下面以一个简单的例子来说明LTP对一个句子的语义解析.

例1. "太阳能光伏发电目前市场需求稳定增长,长期发展前景看好。"的词性标注、依存句法及语义角色标注如图1所示.

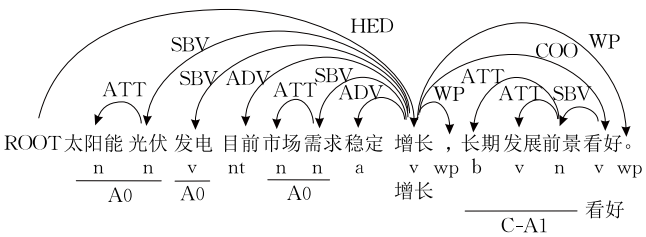


图1 例1的词性标注、依存句法分析及语义角色标注

为了简单起见,以下所有示例中的语义角色标注只标出与评价对象有关的语义角色,即A0、A1、

A2、C-A0 等."C-"表示子嵌套，一般多为主谓短语或动宾短语充当句子的某一成分，因此本文中将 C-A0 等同于 A0、C-A1 等同于 A1 处理；下文中提到的语义角色也都指的是 A0、A1 等与评价对象有关的语义角色.

在图 1 中，词之间的空格表示分词，词下面的一行为词性标注；词之间直接发生依存关系就构成了一个依存关系对，其中，一个词是核心词(也称支配词)，另一个词是修饰词(也称从属词)，依存关系对用一个有向依存弧表示，依存弧的方向总是从核心词(父亲节点)指向修饰词(儿子节点)，每个依存弧的上方都有一个标记，称为关系类型；词性下面的为语义角色标注. LTP 中标注了"增长"和"看好"两个谓词的语义角色标注.

根据第 3 节的分析，谓词的评价对象应满足下列 3 条规则：

规则 Pre1.　当谓词为非心理动词时，其评价对象的优先级 A0→A1→A2.

规则 Pre2.　当谓词为心理动词时，其评价对象的优先级 A1→A2→A0.

规则 Pre3.　当同时存在多个 A0、A1 或 A2 时，评价对象为多个 A0、A1 或 A2 的组合.

4.1.1　基于谓词的直接评价对象-情感词对的抽取

根据第 3 节的分析，情感词在句中充当的结构成分可以是谓语、宾语、谓语的状语或补语、宾语的状语或补语以及定语. 下面首先来讨论情感词为谓语时，其评价对象的识别.

规则 1.　当情感词为语义角色标注中的谓词时，则其评价对象-情感词对为〈谓词的评价对象，情感词〉.

在图 1 中，因为情感词"增长"为非心理动词，所以结合规则 Pre1 和规则 Pre3，其评价对象-情感词对为〈太阳能光伏发电市场需求，增长〉；而情感词"看好"为心理动词，因此结合规则 Pre2，其评价对象-情感词对为〈长期发展前景，看好〉.

语义角色标注只是标注句子中谓词的语义角色，而情感词在句中充当的成分除了谓语，还有宾语、谓语的状语或补语、宾语的状语或补语以及定语，接下来讨论情感词为非谓语的情况.

4.1.2　基于谓词的间接评价对象-情感词对的抽取

当情感词为宾语、谓语的状语或补语以及宾语的定语、状语或补语时，需要先找到其联系的谓语，再根据谓词找到其评价对象. 如何发现谓词与情感词之间的关系呢？这一过程可借助语法分析来实现.

依存句法通过分析语言单位内成分之间的依存关系揭示句子中各成分之间的语义修饰关系，即指出了句中词语之间在句法上的搭配关系，分析出一个句子的主、谓、宾、定、状、补结构. LTP 共定义了 24 种依存关系，如表 2 所示. 为了描述的一致性，文中的依存关系对统一记为：依存关系名(核心词，修饰词).

表 2　LTP 中依存关系名及含义

标记	解释	标记	解释
ATT	定中关系	DE	"的"字结构
QUN	数量关系	DI	"地"字结构
COO	并列关系	DEI	"得"字结构
APP	同位关系	BA	"把"字结构
LAD	前附加关系	BEI	"被"字结构
RAD	后附加关系	ADV	状中结构
VOB	动宾关系	MT	语态结构
POB	介宾关系	CMP	动补结构
SBV	主谓关系	IS	独立结构
SIM	比拟关系	CNJ	关联结构
HED	核心	IC	独立分句
VV	连动结构	DC	依存分句

根据我们的前期研究[28]，宾语、状语或补语可通过动宾结构 VOB、状中结构 ADV、动补结构 CMP 来发现. 下面分别讨论情感词为谓语的宾语、状语或补语及情感词为宾语的状语或补语两种情况.

(1) 情感词为谓语的宾语、状语或补语

当情感词为谓语的宾语、状语或补语时，情感词修饰的是整个主谓结构，即主谓结构为情感词的评论对象.

规则 2.　情感词为谓语的宾语、状语或补语时，需根据 VOB、ADV 或 CMP 找到其谓语，则该情感词的评价对象-情感词对为〈谓词的评价对象＋谓词，情感词〉.

说明：情感词为谓语的宾语、状语或补语时，情感词为 VOB、ADV 或 CMP 的修饰词，谓词为 VOB、ADV 或 CMP 的核心词.

例 2.　"股价上涨得很快。"的词性标注、依存句法及语义角色标注如图 2 所示.

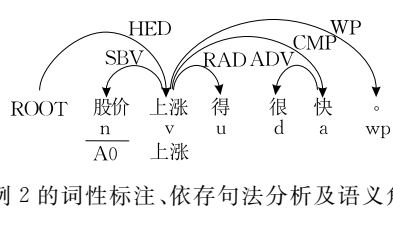


图 2　例 2 的词性标注、依存句法分析及语义角色标注

在图 2 中，情感词"快"通过 CMP(上涨，快)找到谓语"上涨"，谓词"上涨"的评价对象由规则 1 和规则 Pre-1 可知为"股价"，因此，情感词"快"的评价对象-情感词对为〈股价上涨，快〉.

(2)情感词为宾语的状语或补语

情感词为宾语的状语或补语时,虽然情感词的直接修饰对象是宾语,但为了保证其语义的完整性,其评价对象为整个句子的主谓宾结构.

规则 3. 情感词为宾语的状语或补语时,需根据 ADV 或 CMP 找到其宾语,再根据 VOB 找到谓语,则该情感词的评价对象-情感词对为〈谓词的 A0+谓词+宾语,情感词〉.

说明:情感词为宾语的状语或补语时,情感词为 ADV 或 CMP 的修饰词,宾语为 ADV 或 CMP 的核心词同时也是 VOB 的修饰词,谓语为 VOB 的核心词.

例 3. "硅原料供应问题得到有效解决。"的词性标注、依存句法及语义角色标注如图 3 所示.

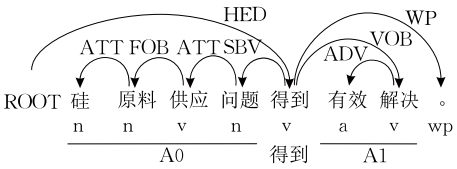


图 3 例 3 的词性标注、依存句法分析及语义角色标注

在图 3 中,由 ADV(解决,有效)可知情感词"有效"为宾语"解决"状语,由 VOB(得到,解决)可知"解决"为谓语"得到"的宾语,谓词"得到"的 A0 为"硅原料供应问题",因此,情感词"有效"的评价对象-情感词对为〈硅原料供应问题得到解决,有效〉.

规则 2 和规则 3 较好地解决了第 2 节中提到的评价对象为主谓结构或主谓宾结构问题.

4.2 基于依存句法分析补充抽取评价对象-情感词对及评价对象扩展

4.2.1 基于依存句法补充抽取评价对象-情感词对

语义角色标注只是给出了句中谓词的语义角色,定语由于与谓词未发生联系,因此在 4.1 节的讨论中并未涉及;又由于自然语言处理的复杂性,LTP 中并未对一些子句中的谓词进行语义角色标注.如例 4,LTP 对谓词情感词"大"给出了语义角色标注,但未对子句"新华保险上涨"中的"上涨"进行语义角色标注,这时可根据句法分析来确定"上涨"的评价对象.

(1)情感词为谓语

当情感词为主谓结构中的谓语时,其评价对象为主语,主谓结构在依存句法中用 SBV 表示.

规则 4. 若情感词为 SBV 的核心词,则该情感词的评价对象-情感词对为〈SBV 修饰词,情感词〉.

例 4. "新华保险上涨的弹性最大。"的词性标注、依存句法及语义角色标注如图 4 所示.

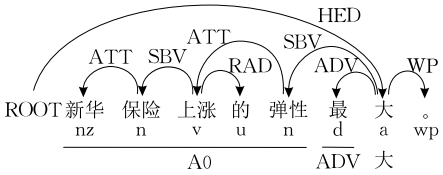


图 4 例 4 的词性标注、依存句法分析及语义角色标注

在图 4 中,情感词"上涨"的评价对象-情感词对为〈保险,上涨〉.

(2)情感词为定语

当情感词为主语或宾语的定语时,其评价对象为该定语所修饰的成分,即当情感词为定语时,其评价对象为定语的修饰对象,即 ATT 的核心词.

规则 5. 若情感词为 ATT 的修饰词,则该情感词的评价对象-情感词对为〈ATT 的核心词,情感词〉.

例 5. "1 季度延续去年年底的市场低迷状况。"的词性标注、依存句法及语义角色标注如图 5 所示.

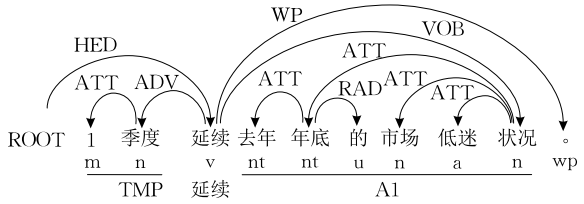


图 5 例 5 的词性标注、依存句法分析及语义角色标注

在图 5 中,情感词"低迷"为宾语的定语,其评价对象-情感词对为〈状况,低迷〉.

4.2.2 评价对象的扩展

规则 4 和规则 5,只抽取了核心的名词,这可能造成部分评价对象丢失或评价对象的语义不完整.如图 5 中,抽取情感词"低迷"的评价对象-情感词对为〈状况,低迷〉,由上下文语义分析可知评价对象采用"市场状况"比"状况"的语义更详实.

(1)评价对象有定语修饰

规则 Post. 评价对象有定语 ATT 修饰时,需将 ATT 的非情感词修饰词组合进评价对象.

在图 5 中,评价对象"状况"包含 3 个 ATT,分别为 ATT(状况,年底)、ATT(状况,市场)和 ATT(状况,低迷),其中"年底"又有 ATT(年底,去年),故扩展后的评价对象-情感词对为〈去年年底市场状况,低迷〉.

(2)评价对象为并列结构

规则 6. 评价对象为多个并列时,通过 COO

并列结构发现，分别抽取多个评价对象组成多个评价对象-情感词对，同时评价对象按规则 Post 扩展.

例 6.　“公司的重要优势是廉价的电和劳动力。”的词性标注、依存句法及语义角色标注如图 6 所示.

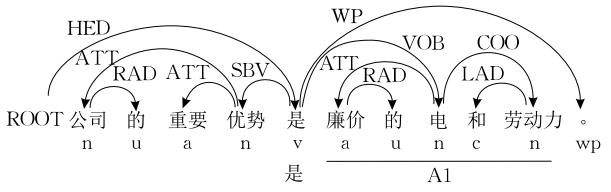


图 6　例 6 的词性标注、依存句法分析及语义角色标注

在图 6 中，情感词“廉价”是宾语“电”的定语，根据规则 5，识别出“廉价”的评价对象-情感词对为〈电，廉价〉，再根据 COO 发现另一评价对象“劳动力”，因此情感词“廉价”的评价对象-情感词对为〈电，廉价〉和〈劳动力，廉价〉.

4.3　情感词并列结构的评价对象-情感词对抽取

正如文献[27]中提到，情感词还可能出现在 COO 并列结构和 VV 连动结构中，即情感词通过 COO 或 VV 结构与句子的谓语、宾语、定语、状语或补语发生联系.

规则 7.　当情感词出现在 COO 并列结构或 VV 连动结构的修饰词位置时，该情感词的评价对象-情感词对为〈COO 并列结构或 VV 连动结构中的核心词的评价对象，情感词〉.

说明：①COO 并列结构或 VV 连动结构的核心词为情感词，则核心词的评价对象通过上述规则已找到；②COO 并列结构或 VV 连动结构中的核心词为非情感词，则需根据上述规则确定该非情感词核心词的评价对象；③当有多个 COO 或 VV 结构时，则通过追溯 COO 或 VV 的核心词，直到核心词为句中的谓语、宾语、定语、状语或补语，再通过该核心词发现评价对象.

例 7.　“同业资产规模大幅下降，较二季度减少 33.86%。”的词性标注、依存句法及语义角色标注如图 7 所示.

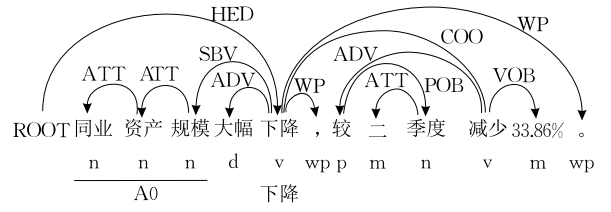


图 7　例 7 的词性标注、依存句法分析及语义角色标注

在图 7 中，由 COO(下降，减少)可知：“下降”与“减少”是并列结构；根据规则 1 发现情感词“下降”的评价对象-情感词对为〈同业资产规模，下降〉，所以根据规则 7 得到“减少”的评价对象-情感词对也为〈同业资产规模，减少〉.

如果通过上述规则仍未发现情感词的评价对象，则取情感词最近的名词或名词短语为其评价对象.

规则 8.　评价对象-情感词对〈情感词的最近名词，情感词〉，并按规则 Post 和规则 6 进行评价对象扩展.

若仍未发现情感词的评价对象，则该评价对象为缺省或隐含，按 5.2 节的隐式评价对象处理.

4.4　规则的执行顺序

根据情感词典识别句中的情感词后，根据情感词在句中充当的语法成分，选择相应的规则抽取其对应的评价对象. 由于情感词可能会同时满足两条或两条以上规则，如例 8 的图 8 中，情感词“增长”既满足规则 1 也满足规则 4，情感词“加大”既满足规则 1 又满足规则 7，因此需要制定规则的执行顺序. 规则的判断执行顺序如图 9 所示.

例 8.　“手续费收入稳定增长，拨备力度加大。”的词性标注、依存句法及语义角色标注如图 8 所示.

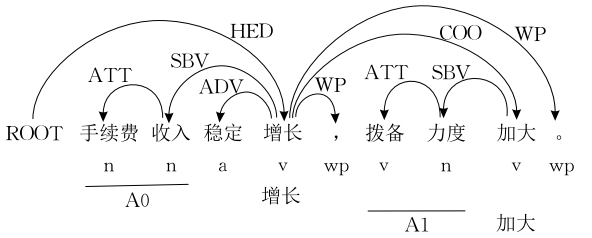


图 8　例 8 的词性标注、依存句法分析及语义角色标注

在图 9 中，规则 Pre1、规则 Pre2 和规则 Pre3 是规则 1、规则 2 或规则 3 的前置规则(事先要根据条件判断执行规则 Pre1、规则 Pre2 和规则 Pre3)，规

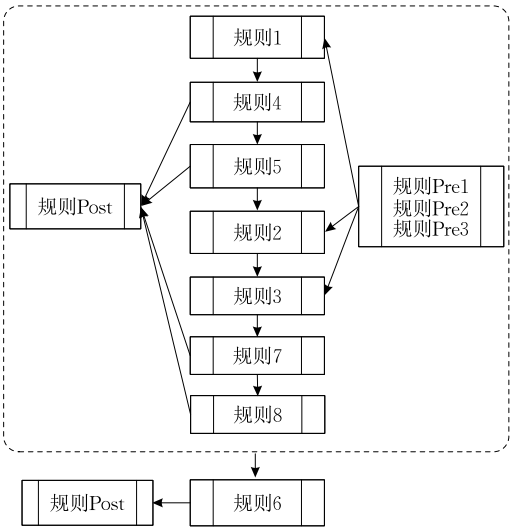


图 9　规则的执行顺序

则 Post 是规则 4、规则 5、规则 7 或规则 8 的后置规则(事后要执行规则 Post);规则 1～规则 5 及规则 7～规则 8 称为或选规则,它们的执行顺序为:1→4→5→2→3→7→8,如果一个情感词满足多条或选规则,则仅执行按上述顺序遇到的第一条或选规则;规则 6 称为必选规则,或选规则执行完毕之后,再执行必选规则.

规则的覆盖率如表 3 所示.

表 3 规则的覆盖率

规则	覆盖率/%
规则 1	54.9
规则 4	5.3
规则 5	13.8
规则 2	3.3
规则 3	2.8
规则 6	3.0
规则 7	4.5
规则 8	12.4

由表 3 可知,情感词在句中充当谓语的占比达到 60.2%(规则 1 和规则 4 之和),这与前面分析的 Web 金融评论中多动词情感词相吻合;规则的覆盖率与规则的执行顺序基本吻合;规则 1～规则 7 分别根据情感词在句中的句法成分抽取其对应的评价对象,这 7 个规则的覆盖率达到 87.6%,充分说明了规则的覆盖范围广.

规则 8 为最近名词规则,当通过上述 7 个规则未找到评价对象时采用规则 8,其出现的主要原因有:① 在上下文语境下,句子陈述时缺省了主语或宾语;② 评论者的句子表达错误或句法解析错误.

规则 6 是评价对象的并列情况,它也可以理解为评价对象的一个扩展,通过规则 1～规则 5 以及规则 7 和规则 8 找到情感词的评价对象后,均应再判断该评价对象是否有并列结构.

规则 1 和规则 4 的占比达到 60.2%,这也说明姚天昉等人[2]基于 SBV 结构抽取情感词的评价对象可以解决大部分情感搭配对的抽取问题.

5 评价对象的进一步处理

5.1 虚指评价对象的发现和替换

由于中文的表达习惯,中文评论中会出现很多代词、缩略词等虚指评价对象.如"2008 年度公司共实现主营业务收入 53 288.07 万元,同比增长 60.47%。"一句中,情感词"增长"修饰的评价对象为"同比",但显然"同比"语义指向不明确,不利于进一步的评价对象分组,由上下文语义理解可知该"同比"实际指向的评价对象为"主营业务收入".又如"前三季度资产管理业务收入增幅 111.06%,但由于绝对值相对较小,对公司整体业绩影响有限。"中情感词"小"的直接评价对象为"绝对值",同样需找到其实际指向的评价对象"资产管理业务收入".

虚指评价对象的出现是因为人们的表达习惯,在上下文语义提示下,为了表达的言简意赅,往往习惯于使用缩略或指代的形式.因此虚指评价对象所实际指向的评价对象就藏在它的上下文语义中.

(1) 虚指评价对象的发现

Web 金融评论涉及企业经济活动的各个层面,对企业的财务指标和非财务指标的各个指标项及其子项进行全方位的解读,因此金融评论中的评价对象应该反映企业财务指标或非财务指标项及其子项.记 Ψ 由新华 08 金融词典、非财务指标及其子集、财务指标及其子集及企业的产品名、原材料名等合成的数据集;为此,我们定义候选评价对象 t 在 Ψ 中出现的频率度量公式,如式(1)所示.

$$shamTarget(t)=\sum_{i=1}^{k} num(t_i \$ \Psi) \tag{1}$$

其中,t_i 是候选评价对象 t 分词后的一个子词;k 是 t 分词的个数;$num(t_i \$ \Psi)$ 是词 t_i 在 Ψ 中出现的次数.

如果候选评价对象 t 中的任一个分词都未出现在 Ψ 中,则 $shamTarget(t)=0$,称 t 为虚指评价对象.

(2) 虚指评价对象的替换

虚指评价对象一般指代的是最近提到的评价对象,因此可以用虚指评价对象前一名词短语(按评价对象扩展规则进行扩展)来替换虚指评价对象.同时需要注意,若前一名词短语仍为虚指评价对象,则仍需继续查找前一名词短语.重复上述过程,直到找到非虚指的评价对象.若直到句子的开始都未发现真正的评价对象,则按隐式评价对象处理.

5.2 隐式评价对象的识别

通过上述规则在抽取情感词的评价对象时,仍可能出现抽取不到评价对象的情况,均视为隐式评价对象.这主要是因为:① 通过规则未发现评价对象(通过规则 8 未找到名词);② 虽然找到评价对象,但评价对象为虚指评价对象,替换后也未发现真正的评价对象.

隐式评价对象的出现,主要有 3 种情况:

(1) 特殊情感词的强指示作用.情感词可以分为两类:特殊情感词和一般情感词.特殊情感词是指它修饰的评价对象是特定的,具有一一对应的搭配关系,如金融评论中的"增持"、"推荐"多用在评论的结尾,是指建议的"评级","增发"往往是指 "股票数

量”；而一般情感词是指它可以修饰多个评价对象. 正因为特殊情感词对评价对象的这种特指性，所以在评论中特殊情感词的评价对象经常缺省.

(2) 上下文语义的提示作用. 在人们的表达习惯中，当上下文语义明确时，常用代词(虚指评价对象)或直接缺省评价对象(隐式评价对象).

(3) 表达错误或句法解析错误造成的评价对象缺失.

针对上述隐式评价对象出现的 3 种情况，我们对应构建了 3 张表来解决隐式评价对象的识别. 第 1 张表为特殊情感词评价搭配表，用于记录特殊情感词的评价对象-特殊情感词对；第 2 张表为上下文语义关联评价搭配表，用于记录前一句中出现的各个评价对象-情感词对；第 3 张表为频繁评价搭配表，用于记录每个情感词在评论集中最频繁的评价对象-情感词对. 因为特殊情感词对评价对象的强指示性，人工指定特殊情感词的评价对象. 其他两张表均由机器统计生成.

隐式评价对象的识别步骤：

(1) 情感词是否在上下文语义关联情感搭配表中出现，若出现，则该隐式评价对象为上下文语义关联情感搭配表中的最近相同情感词的评价对象；

(2) 情感词是否为特殊情感词，若为特殊情感词则根据评价对象与特殊情感词的一一对应关系找到其评价对象；

(3) 根据频繁评价搭配表，给情感词匹配最频繁搭配的评价对象.

我们采用式(2)来度量情感词 o 的特殊程度，即用于区别特殊情感词与一般情感词.

$$special(o)=null(o)/count(o) \tag{2}$$

其中，情感词 o 出现的总次数记为 $count(o)$；情感词 o 没有找到评价对象的次数记为 $null(o)$. 实验表明：特殊情感词的 $special(o)$ 一般在 0.7 以上，而一般情感词的 $special(o)$ 一般小于 0.1. 实验中我们将阈值 α 设为 0.5.

6　评价对象-情感词对的抽取算法

综上所述，评价对象-情感词对的抽取包括 3 步，如算法 1 所示. 第 1 步，基于第 4 节的规则抽取情感词所修饰的评价对象，如算法 2 所示，其中抽取 COO 并列结构中的评价对象列表如算法 3 所示；第 2 步，虚指评价对象的发现与替换；第 3 步，隐式评价对象的识别.

算法 1.　抽取评价对象-情感词对.

输入：情感句列表 SL，每一个情感句中的情感词列表 O_s，特殊情感词评价搭配表 S，频繁评价搭配表 F

输出：评价对象-情感词对列表 $P=\{\langle t,w\rangle\}$

```
P=∅; //将〈评价对象,情感词〉列表 P 赋为空;
FOR (s∈SL) //SL 中的每一个情感句 s
  C=∅; //将上下文语义关联评价搭配表 C 赋为空
  FOR (o∈Os) //Os 中的每一个情感词 o
    t=ruleBasedExtraction(s,o); //调用算法 2
    IF (shamTarget(t)==0) //虚指评价对象判断
      t=targetReplacement(s,t); //虚指评价对象替换
    ENDIF
    IF (t==NULL) //隐式评价对象的发现
      IF (o∈C) //o 在上下文搭配表 C 中
        t=contextIdentify(C,o);
                          //从上下文搭配表 C 中识别
      ELSE IF (o∈special(o)) //o 是特殊情感词
        t=specialIdentify(S,o);
                          //从特殊搭配表 S 中识别
      ELSE
        t=frequencyIdentify(F,o);
                          //从频繁搭配表 F 中识别
      ENDIF
    ENDIF
    C=C∪〈t,o〉;
    P=P∪〈t,o〉;
  ENDFOR
ENDFOR
```

算法 2.　基于规则抽取情感词所修饰的评价对象.

输入：情感句 s，该情感句中的一个情感词 o

输出：情感词 o 的所有评价对象 t

```
IF (o 为谓词) //规则 1
  执行前置规则;
  t=谓词的评价对象;
ELSE IF (o 为 SBV 的核心词) //规则 4
  t=SBV 修饰词;
  执行后置规则对评价对象 t 进行扩展;
ELSE IF (o 为 ATT 的修饰词) //规则 5
  t=ATT 的核心词;
  执行后置规则对评价对象 t 进行扩展;
ELSE IF (o 为谓语的宾语、状语或补语) //规则 2
  执行前置规则;
  t=谓词的评价对象+谓词;
ELSE IF (o 为宾语的状语或补语) //规则 3
  执行前置规则;
  t=谓词的 A0+谓词+宾语;
ELSE IF (o 出现在 COO 或 VV 的修饰词位置) //规则 7
  t=COO 结构或 VV 结构中的核心词的评价对象;
  执行后置规则对评价对象 t 进行扩展;
```

ELSE //规则 8
t=情感词 o 的最近名词;
执行后置规则对评价对象 t 进行扩展;
ENDIF
IF (t 有 COO 并列的评价对象) //规则 6
$t=cooTargetsExtraction(s,o,t)$; //调用算法 3
ENDIF

算法 3. 抽取 COO 并列结构中的评价对象.
输入:情感句 s,s 中的情感词 o 和评价对象 t
输出:情感词 o 的 COO 并列评价对象列表 t
WHILE (t 有 COO 并列的评价对象) //规则 6
$temp$=COO 并列中的另一个评价对象;
执行后置规则对评价对象 $temp$ 进行扩展;
$t=t\cup temp$;
ENDWHILE

7 实验评测

依存句法分析和语义角色标注采用哈尔滨工业大学语言处理平台 LTP. 虚指评价对象替换的母文件 Ψ(见式(1))由人工标注、金融指标及企业信息 3 部分组成:(1)人工标注包括《新华 08 汉英金融词典》专家标注和数据集上的人工标注;(2)金融指标包括财务指标及其子集和非财务指标及其子集;(3)企业信息包括企业的股票代码、股票名词、产品名和原材料名等.

为了便于分析评价对象-情感词对抽取正确率在逐步语义挖掘下的攀升情况,将基于规则抽取评价对象-情感词对,记为 BSA;将在此基础上进行虚指评价对象替换,记为 XZSA;将再在此基础上利用领域知识和上下文语义解决隐式评价对象,记为 SSA.

7.1 实验数据集

实验数据采用新浪财经网上的公司研究信息,采用公司研究而不用股吧里的股评,主要是考虑公司的研究信息真实,可靠性更高,为下一步的金融域文本情感计算和上市公司财务预警提供可靠数据来源.

(1)数据集

根据上市企业在新浪财经的公司研究中的活跃程度,选取了新浪财经网上 2008 年 1 月至 2012 年 12 月间两个行业(制造业和金融业),每个行业 10 家上市企业的公司研究. 制造业企业涉及服装鞋类、生物制药、电子信息、传媒娱乐、环保、电子器件、化工和食品 8 个行业;金融业企业涉及证券和银行. 根据评论者对股票给予的评级,对每家上市企业选取了褒、贬义相对较强的 50 篇评论,共计 1000 篇评论文档,其中,包含情感词的句子有 16 208 句,共出现情感词 32 529 次(说明:准确地说应该是情感词与评价对象搭配对的次数,因为一个情感词在并列结构中会与多个评价对象搭配). 包含情感词的句子称为情感句.

实验数据集的具体信息如表 4、表 5 所示.

表 4 10 家制造业的基本情感数据信息

股票名称	股票代码	情感句数	情感词数
探路者	300005	830	1795
北陆药业	300016	711	1430
超图软件	300036	597	1092
互动娱乐	300043	696	1357
碧水源	300070	667	1279
数码视讯	300079	689	1254
长信科技	300088	745	1444
益佰制药	600594	706	1459
新安股份	600596	673	1447
青岛啤酒	600600	963	1952
合计		7277	14 509

表 5 10 家金融业的基本情感数据信息

股票名称	股票代码	情感句数	情感词数
宏源证券	000562	926	1684
国元证券	000728	783	1433
宁波银行	002142	840	1809
浦发银行	600000	969	2065
华夏银行	600015	940	2046
民生银行	600016	1010	2176
中信证券	600030	770	1454
交通银行	601328	859	1808
光大证券	601788	806	1491
建设银行	601939	1028	2054
合计		8931	18 020

(2)文本预处理

在实验中发现,LTP 只能处理中文标点符号,而爬虫抓取的 Web 金融信息里中英文标号不一,因此在预处理中把所有的英文标号统一成中文标号. 汉语句子中有时会用括号,括号内的内容表示进一步解释说明,不影响句子的情感和语义,因此预处理中我们去除了括号内的内容.

(3)标准数据集

为了选择出黄金标准集,本文选用 3 个人作为文本情感标注者,以少数服从多数决定正确答案;当出现无法决定的情形,则由 3 人讨论并经我们确认标注结果. 3 个人的组成为 2 位本科生和 1 位研究生,且都具有财经基础课程学习经历.

7.2 虚指评价对象与隐式评价对象的实验结果

(1)虚指评价对象

金融评论相对于一般的评论而言,更侧重数字的分析及与历史数据的对比,在评论中出现了大量的“同比”、“对比”、“环比”等虚指评价对象.

虚指评价对象的发现与替换,有助于进一步的评价对象分组,以便于细粒度的情感计算. 我们对两

个行业发现的虚指评价对象的总次数及正确替换的次数进行了统计，如表 6 所示.

表 6　虚指评价对象正确率统计

类别	发现总次数	正确替换次数	正确率/%
制造业	3424	2585	75.5
金融业	4600	3824	83.1
合计	8024	6409	79.9

由表 6 可知，共发现虚指评价对象 8024 次，其中正确替换虚指评价对象 6409 次，正确率达到 79.9%. 这充分说明了发现并替换虚指评价对象的重要性，同时也说明了本文方法的有效性.

(2) 隐式评价对象

根据式(2)，实验评论集中共发现 10 个特殊情感词，并根据领域知识人工给出了固定搭配的评价对象，如表 7 所示.

表 7　评价对象-特殊情感词对

评价对象	特殊情感词
股票评级	推荐
股票评级	增持
股票评级	减持
股票数量	增发
价值估值	低估
价值估值	高估
股票	分红
股票	转增
产量	减产
产量	增产

表 8 给出了一般情感词“增长”的搭配对排序. 表 9 给出了最频繁搭配表.

表 8　情感词“增长”的搭配对排序

评价对象-情感词对	出现次数
〈净利润，增长〉	248
〈营业收入，增长〉	245
〈收入，增长〉	138
〈业绩，增长〉	128
〈母公司净利润，增长〉	101
……	……

表 9　最频繁搭配对表

评价对象-情感词对	出现次数
〈净利润，增长〉	248
〈股票数量，增发〉	122
〈规模，扩张〉	48
〈毛利率，提升〉	40
〈成本收入，下降〉	38
……	……

评价对象-特殊情感词、上下文语义关联搭配及频繁搭配在评论集中出现的次数、正确识别的次数和正确率的统计结果如表 10 所示.

由表 10 可知，本文解决隐式评价对象的方法是有效的；同时特殊情感词搭配和上下文搭配的正确率要远高于频繁搭配，这正好与前文分析的两个原因相一致：① 特殊情感词对评价对象的指示能力强；② 上下文语义提示下，人们表述时往往采用缺省.

表 10　3 种搭配对解决隐式评价对象的正确率统计

搭配对	制造业			金融业			合计正确率/%
	出现次数	正确识别次数	正确率/%	出现次数	正确识别次数	正确率/%	
特殊情感词搭配	537	506	94.2	391	381	97.4	95.6
上下文搭配	9	7	77.8	11	7	63.6	70.0
频繁搭配	245	144	58.8	271	111	41.0	49.4
合计	791	657	83.1	673	499	74.1	79.0

(3) 虚指评价对象和隐含评价对象的作用分析

针对 BSA、XZSA 和 SSA 这 3 种情况，表 11 给出了制造业和金融业的正确率比较结果，图 10、图 11 分别给出了制造业、金融业中各 10 家上市企业的比较结果. 其中，图 10 和图 11 的横坐标为上市企业股票代码，纵坐标为正确率(单位：%).

表 11　BSA、XZSA 和 SSA 的正确率对比

类别	BSA/%	XZSA/%	SSA/%
制造业	59.5	77.3	81.8
金融业	59.7	80.9	83.7
合计	59.6	79.3	82.9

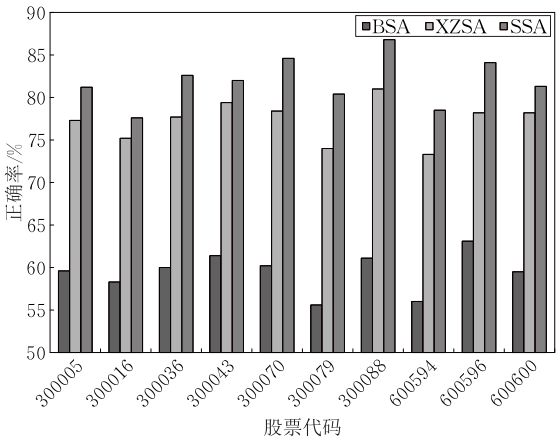


图 10　制造业 10 家企业的 3 种结果对比图

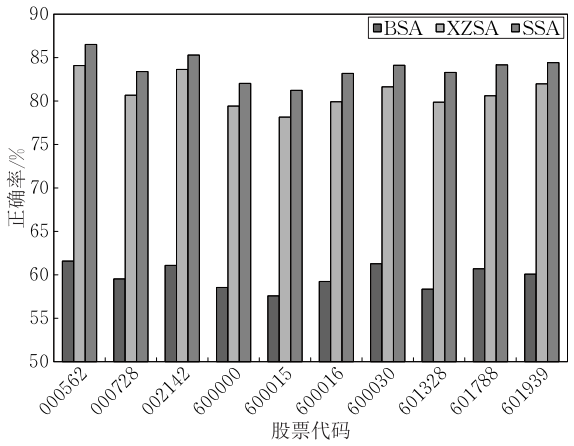


图 11 金融业 10 家企业的 3 种结果对比图

由表 11 和图 10、图 11 可以看出，金融业的正确率要略高于制造业，这是由于制造业 10 家企业涉及 8 个行业，而金融业 10 家企业只涉及到证券和银行两类，制造业的评价对象组成更为复杂.

(4) 句子长度对抽取正确率的影响

表 12 给出了句子长度(字数)与评价对象-情感词对(表中简称为搭配对)抽取错误率的关系.

表 12 句子长度与评价对象-情感词对抽取错误率的关系

句子长度 L	搭配对占比/%	抽取错误率/%
$0<L\leq 50$	62.13	13.20
$50<L\leq 60$	11.59	15.90
$60<L\leq 70$	8.83	17.10
$70<L\leq 80$	6.27	18.00
$80<L$	11.18	19.90

显然，金融评论中的长句子要比商品评论中多得多，评价对象-情感词对抽取的错误率与句子的长度成正比. 这是因为句子越长，句法成分越复杂，从而句法分析出错的概率也就越大.

7.3 与其他方法的实验对比

(1) 基线实验

本文选取了已有的两种评价对象-情感词对抽取方法作为基线实验.

Nearest 方法[5]. Hu 等人认为情感词修饰的是距其最近的评价对象. 最近名词方法作为一种简单实用的方法在中英文评论中均可方便使用.

EPM 方法[4]. Zhao 等人提出的基于句法路径的情感评价单元识别方法，它在英文语料集上表现出了较好的效果. 基于句法路径的方法是根据评价对象与情感词间的句法路径生成频繁路径库，再对频繁路径进行泛化形成最终的路径库. 然后根据情感词匹配的路径找到其评价对象.

机器学习方法和主题模型方法近几年在评价对象-情感词对抽取中也较为常用[7-8,10,12-13]，但这些方法在特征选择时也都加上了评价对象与情感词间的句法关系. 另外，这些方法基本应用在商品评论中，商品评论的评价对象较为单一为名词或名词短语，而金融评论中评价对象存在从句结构，直接移植这些方法肯定会造成评价对象的不完备.

(2) 3 种方法的正确率比较

将本文的 SSA 方法与最近名词方法、句法路径方法进行了实验对比.

由于金融评论的特殊性，为了公平起见，我们假设：① 对于最近名词方法和句法路径方法，也都进行了评价对象的扩展(规则 9)；② 虽然本文方法充分考虑了评价对象中动词的重要性，但考虑到两种对比方法没有对此进行研究，因此，只要两种对比方法中包含了评价对象中的主要名词，我们都认为是正确的；③ 传统的最近名词方法往往将评价对象与情感词限定在一定的窗口范围内，为了提高查全率本文将窗口范围扩大到整个句子.

3 种方法的正确率比较如表 13 所示. 显然，本文方法明显优于另外两种方法.

表 13 3 种方法的正确率比较

类别	Nearest/%	EPM/%	SSA/%
制造业	53.4	58.1	81.8
金融业	55.6	59.1	83.7
合计	54.6	58.6	82.9

这是因为：① 中文句法比英文要复杂得多，再加上金融评论中多长句使得路径更加复杂，这都不利于 EPM 方法；② 中文长句中情感词与评价对象距离较远的概率更大，因此也不利于 Nearest 中最近名词即为评价对象的方法.

在结果匹配中，通过与人工标注的标准答案进行匹配，结果分为完全匹配、部分匹配和完全不匹配. 在当前流行的商品评论情感挖掘评测中，如果评价对象包含了标准答案中的主要名词则认为是正确答案，因此本文中正确答案包括完全匹配答案和部分匹配答案. 为此，我们统计了完全匹配答案在正确答案中的占比情况，如表 14 所示.

表 14 3 种方法的完全匹配答案在正确答案中的占比

类别	Nearest/%	EPM/%	SSA/%
制造业	68.50	65.90	81.00
金融业	66.10	66.20	80.10
合计	67.20	66.10	80.50

由于 Nearest 和 EPM 方法的评价对象只考虑了名词或名词短语，而未考虑主谓结构、从句等复杂

结构，因此这两种方法的完全匹配答案在正确答案中的占比不是很高. 本文方法的完全匹配答案在正确答案中的占比要大大高于它们.

(3) 查全率

本文的目标是在中文金融评论中根据情感词正确识别其所修饰的评价对象，包括虚指评价对象和隐式评价对象的抽取. 因此，本文方法对于每一个情感词都一定会找到至少一个对应的评价对象，但仍有可能存在下列丢失评价对象-情感词对的情况：① 由于句法解析错误，导致并列结构的评价对象丢失；② 由于分词错误，导致情感词丢失；③ 由于情感词典的不完备，导致情感词的缺失. 对于另两种方法，除了上述可能以外，由于未进行隐式评价对象的处理，则还存在着隐式评价对象的丢失. 3 种方法对应的召回率如表 15 所示.

表 15　3 种方法的召回率

方法	召回率/%
Nearest	90.7
EPM	90.4
SSA	97.3

8　总结与展望

由于评价对象-情感词对的抽取一方面可以解决情感词与修饰评价对象的搭配问题，另一方面评价对象的奇异性又将影响着整个评价搭配对的情感极性，因此评价对象-情感词对的抽取是细粒度情感计算的原子问题和关键问题. 本文基于语义分析，充分抓住评价对象与情感词间的语义联系，利用领域知识和上下文语义，提出了基于语义角色标注和依存句法分析抽取评价对象-情感词对的规则，并同时提出了评价对象的扩展规则；针对中文金融评论，首次给出了虚指评价对象的发现与替换方法，系统地解决了隐式评价对象的识别问题. 本文的创新如下：

(1) 与商品评论不同，金融评论中情感词的词性丰富，尤其是动词情感词丰富. 针对这一特性，设计了语义角色标注与依存句法分析相结合的评价对象-情感词对抽取规则，根据情感词在句中的句法成分，抽取了不同构成形式的评价对象，保证了评价对象构成的复杂性. 浅层语义分析与依存句法分析相结合，充分考虑了 24 种依存关系中所有可能的评价对象与情感词间的语法路径，保证了抽取的完备性；而评价对象的扩展规则保证了评价对象-情感词对抽取的召回率.

(2) 虚指评价对象指的是评价对象的指代或缩略词，其含义不明确，不利于进一步的评价对象分组和细粒度情感计算. 文中基于语义和领域知识提出了虚指评价对象识别和替换方法，从而明确其实际的指向和含义.

(3) 基于特殊情感词搭配表、上下文搭配表及频繁搭配表提出了识别隐式评价对象的新思路，能有效识别出缺省和隐含评价对象.

进一步的研究工作主要有：① 基于评价对象的细粒度情感计算；② 不包含情感词的特殊情感句的评价对象识别及情感计算.

致　谢　本文的研究工作得到了美国伊利诺伊大学芝加哥分校刘兵教授的指导；利用了哈尔滨工业大学社会计算信息检索研究中心免费开放的 LTP 平台；中信证券南昌营业部的几位专家参与了《新华 08 汉英金融词典》的分词、词性及褒贬极性的标注工作；我校硕士研究生刘挺、吴双、陈煌烨、刘玉等以及 2013 级计算机 1 班的 10 余位本科生参与了实验测试数据集的整理、标注和实验程序的设计和分析. 在此一并表示感谢！最后，由衷地感谢论文评审专家和编辑对本文所提出的修改建议！

参　考　文　献

[1] Bloom K, Garg N, Argamon S. Extracting appraisal expressions //Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Poznan, Poland, 2007: 308-315

[2] Yao Tian-Fang, Nie Qin-Yang, Li Jian-Chao, et al. An opinion mining system for Chinese automobile review// Proceedings of the 25th Workshop of Chinese Information Processing Society of China. Beijing, China, 2006: 260-281(in Chinese)
(姚天昉，聂青阳，李建超等. 一个用于汉语汽车评论的意见挖掘系统//中国中文信息学会 25 周年学术会议. 北京，中国，2006：260-281)

[3] Wan C X, Jiang T J, Liu D X, Liao G Q. Sentimental analysis of Web financial reviews: Opportunities and challenges// Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval. Rome, Italy, 2014: 366-373

[4] Zhao Yan-Yan, Qin Bin, Che Wan-Xiang, Liu Ting. Appraisal expression recognition based on syntactic path. Journal of Software, 2011, 22(5): 887-898(in Chinese)
(赵妍妍，秦兵，车万翔，刘挺. 基于句法路径的情感评价单元识别. 软件学报，2011，22(5)：887-898)

[5] Hu M, Liu B. Mining and summarizing customer reviews//Proceedings of the 10th International Conference on Knowledge Disvovery & Data Mining. San Jose, USA, 2004: 168-177

[6] Kim S M, Hovy E. Extracting opinions, opinion holders and topics expressed in online news media text//Proceeding of the Workshop on Sentiment and Subjectivity in Text. Sydney, Australia, 2006: 1-8

[7] Lakkaraju H, Bhattacharyya C, Bhattacharya I, Merugu S. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments//Proceedings of the 2011 SIAM International Conference on Data Mining. Mesa, USA, 2011: 498-509

[8] Wang Zhong-Qing, Wang Rong-Yang, Pang Lei, et al. Technical report on Suda_SAM_OMS sentiment analysis system//The 3rd Chinese Opinion Analysis Evaluation. Jinan, China, 2011: 25-32(in Chinese)
(王中卿, 王荣洋, 庞磊等. Suda_SAM_OMS情感倾向性分析技术报告//第三届中文倾向性分析. 济南, 中国, 2011: 25-32)

[9] Yin P, Wang H, Guo K. Feature-opinion pair identification of product reviews in Chinese: A domain ontology modeling method. New Review of Hypermedia and Multimedia, 2013, 19(1): 3-24

[10] Zhang S, Jia W, Xia Y, et al. Extracting product features and sentiments from Chinese customer reviews//Proceedings of the 7th International Conference on Language Resources and Evaluation. Valletta, Malta, 2010: 1142-1145

[11] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York, USA, 2011: 815-824

[12] Zhao Q Y, Wang H, Lv P, Zhang C. A bootstrapping based refinement framework for mining opinion words and targets //Proceedings of the 23th ACM Conference on Information and Knowledge Management. Shanghai, China, 2014: 1995-1998

[13] Liu K, Xu L H, Zhao J. Extracting opinion targets and opinion words from online reviews with graph co-ranking//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 314-324

[14] Xu L H, Lai S W, Liu K, Zhao J. Product feature mining: Semantic clues versus syntactic constituents//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 336-346

[15] Popescu A M, Etzioni O. Extracting product features and opinions from reviews//Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada, 2005: 339-346

[16] Bloom K, Argamon S. Automated learning of appraisal extraction patterns. Language and Computers, 2009, 71(1): 249-260

[17] Kobayashi N, Inui K, Matsumoto Y, et al. Collecting evaluative expressions for opinion extraction//Proceedings of the International Joint Conference on Natural Language Processing. Hainan, China, 2005: 596-605

[18] Somprasertsri G, Lalitrojwong P. Mining feature-opinion in online customer reviews for opinion summarization. Journal of Universal Computer Science, 2010, 16(6): 938-955

[19] Kim S M, Hovy E. Extracting opinions, opinion holders, and topics expressed in online news media text//Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text. Sydney, Australia, 2006: 1-8

[20] Qiu G, Liu B, Bu J J, Chen C. Opinion word expansion and target extraction through double propagation. Computational Linguistics, 2011, 37(1): 9-27

[21] Kamal A, Abulaish M, Anwar T. Mining feature-opinion pairs and their reliability scores from Web opinion sources//Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. Craiova, Romania, 2012: 15-21

[22] Wang Su-Ge, Wu Su-Hong. Feature-opinion extraction in scenic spots reviews based on dependency relation. Journal of Chinese Information Processing, 2012, 26(3): 116-121(in Chinese)
(王素格, 吴苏红. 基于依存关系的旅游景点评论的特征-观点对抽取. 中文信息学报, 2012, 26(3): 116-121)

[23] Gu Zheng-Jia, Yao Tian-Fang. Extraction and discrimination of the evaluated object and its orientation. Journal of Chinese Information Processing, 2012, 26(4): 91-97(in Chinese)
(顾正甲, 姚天昉. 评价对象及其倾向性的抽取和判别. 中文信息学报, 2012, 26(4): 91-97)

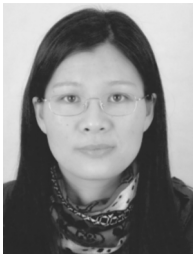
[24] Wang Zhi-Min, Zhu Xue-Feng, Yu Shi-Wen. Research on lexical emotional evaluation based on the grammatical knowledge-base of contemporary Chinese. International Journal of Computational Linguistics and Chinese Language Processing, 2005, 10(4): 581-592(in Chinese)
(王治敏, 朱学锋, 俞士汶. 基于现代汉语语法信息词典的词语情感评价研究. 中文计算语言学期刊, 2005, 10(4): 581-592)

[25] Lao Qing. The Semantic and Grammatical Peculiarity of the Psychology Verb in Modern Chinese [M. S. dissertation]. Shanghai Normal University, Shanghai, 2007(in Chinese)
(劳勍. 现代汉语心理动词语义、句法研究[硕士学位论文]. 上海师范大学, 上海, 2007)

[26] Wu Xiao-Feng, Zong Cheng-Qing. An approach to news paraphrase recognition based on SRL. Journal of Chinese Information Processing, 2010, 24(5): 3-9(in Chinese)
(吴晓锋, 宗成庆. 基于语义角色标注的新闻领域复述句识别方法. 中文信息学报, 2010, 24(5): 3-9)

[27] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: A Chinese language technology platform//Proceedings of the Coling 2010: Demonstrations. Beijing, China, 2010: 13-16

[28] Wan Chang-Xuan, Jiang Teng-Jiao, Zhong Min-Juan, Bian Hai-Rong. Sentiment computing of Web financial information based on the part-of-speech tagging and dependency parsing. Journal of Computer Research and Development, 2013, 50(12): 2554-2569(in Chinese)
(万常选，江腾蛟，钟敏娟，边海容. 基于词性标注和依存句法的 Web 金融信息情感计算. 计算机研究与发展，2013，50(12)：2554-2569)



JIANG Teng-Jiao, born in 1976, Ph. D. candidate, lecturer. Her current research interests include sentiment analysis and Web data management.

WAN Chang-Xuan, born in 1962, Ph. D., professor, Ph. D. supervisor. His current research interests include Web data management, sentiment analysis, data mining and information retrieval.

LIU De-Xi, born in 1975, Ph. D., professor. His current research interests include Web data management, information retrieval and natural language processing.

LIU Xi-Ping, born in 1981, Ph. D., associate professor. His current research interests include information retrieval, data mining and Web data management.

LIAO Guo-Qiong, born in 1969, Ph. D., professor. His current research interests include database and data mining.

Background

In recent years, opinion mining or sentiment analysis has been an active research area in data management, data mining and natural language processing. In this paper, we study a key problem in sentiment analysis, which is the extraction of target-opinion pairs. The main research focus is on extracting the corresponding targets of existing opinion words in Chinese financial reviews, including find and replacement of ambiguous targets and identification of implicit targets. On one hand, our work is helpful to find the corresponding target of opinion word; on the other hand, the singularity of target has influences on the sentimental tendency of the target-opinion pairs. The results of fine-grained sentiment analysis of the financial reviews can provide important indicators for enterprise financial early-warning.

Plenty of research efforts have been put on the problem. Most of the existing work focuses on product reviews, but they usually ignored the verbs in the targets and considered little about the default and implicit targets. In product reviews, opinion words are usually adjectives, and their targets are usually nouns or noun-phrases. Whereas, in financial reviews, opinion words have various POS (part of speech) and a large number of them are verbs. What is more, a financial review usually has more targets and more complex structure, and it is very common to have ambiguous targets and implicit targets.

In consideration of the characteristics of financial reviews, we capture the semantic relationships between opinion words and targets using domain knowledge and context information, and propose rules of extracting target-opinion pairs based on semantic roles labeling and dependency parsing. Rules of extending targets are also proposed. The paper also presents methods finding and replacing the ambiguous targets, and a systematic way of identifying implicit targets; to the best of our knowledge, our work is the first solution towards this problem.

The research is partially supported by the National Natural Science Foundation of China under Grant Nos. 61562032, 61662027, 61662032, 61173146, 61363039, 61363010 and 61462037, the Grand Natural Science Foundation of Jiangxi Province under Grant No. 20152ACB20003, and the Ground Program on High College Science & Technology Project of Jiangxi Province under Grant Nos. KJLD12022 and KJLD14035.

Our past work about sentiment analysis in Web financial reviews have been published in Proceeding of KDIR (Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 2014) and Journal of Computer Research and Development.