基于依存关系的中文情感要素抽取技术研究*

王倩[1,2] 何婷婷[1,2] 闻彬[1,2] 宋乐[1,2] 张茂元[1,2]

1(华中师范大学 计算机科学与技术系，湖北 武汉 430079)

2(国家语言资源监测与研究中心网络媒体分中心，湖北 武汉 430079)

E-mail wangqian1287@hotmail.com; tthe@mail.ccnu.edu.cn;

摘 要：近年来，中文倾向性分析在自然语言处理领域深受关注。针对情感要素抽取，本文提出了一种基于依存关系的抽取方法。该方法在已识别情感词语的基础上，利用分析器对包含情感词语的短句进行依存关系分析，抽取情感要素，并对其作倾向性判断。本文在第一届中文倾向性分析评测（COAE2008）比赛语料以及影评语料上进行训练和测试，情感要素抽取的 F 值最高值达到 0.7149，平均值达到 0.5673，充分验证了该方法的有效性。

关键词：依存关系，情感要素抽取，倾向性分析

Research on Dependency Tree-Based Chinese Sentimental Elements Extraction

Wang Qian[1,2] He Ting-ting[1,2] Wen Bin[1,2] Song Le[1,2] Zhang Mao-yuan[1,2]

(1Department of Computer Science, Huazhong Normal University, Wuhan, 430079;

2Monitor and Research Center for National Language Resource Network Multimedia Sub-branch Center, Wuhan, 430079)

E-mail wangqian1287@hotmail.com; tthe@mail.ccnu.edu.cn;

Abstract: Over the past few years, orientation analysis has been receiving a lot of attention in the field of natural language processing. Oriented to sentimental elements extraction, a dependency tree-based method has been proposed in this paper. On the basis of the identified sentimental words, the proposed method utilizes the parser to analyze the shot sentences which contain the sentimental words, and then extracts the sentimental elements and determines the orientation. To verify the effectiveness of this method, experimental tests have been carried out on the corpus of the First Chinese Opinion Analysis Evaluation (COAE2008) and film reviews. The evaluation results show that the highest F-measure of sentimental elements extraction is 0.7149 while the average value is 0.5673, which confirms the feasibility of the proposed method.

Key words: Dependency Tree; sentimental elements extraction; Orientation analysis.

1 引言

随着 Internet 的广泛应用和普及，越来越多的人通过网络来获得其他人对某一事件或者产品的讨论、发表自己的观点和评论。于是如何快速、自动地从海量信息中挖掘人们的观点信息，并对其进行分析就变得尤为重要。情感计算就是在这样的背景下应运而生的，利用情感计算技术可

*项目资助：国家自然科学基金(60773167)；国家十一五科技支撑计划课题“网络文化安全预警技术研究”(2006BAK11B03)；973 国家重点基础研究发展计划(2007CB310804)；教育部/国家外国专家局高等学校学科创新引智计划(B07042)。

以对网络信息进行有效的分析和挖掘，识别出观点信息并分析其观点的情感倾向性。目前，情感计算的研究工作大多是粗颗粒度的，侧重于对整篇文档进行褒贬极性分析，而对于情感要素抽取，即评价对象抽取方面的研究还很少。然而，评价对象的识别又是非常重要的，它直接影响到是否能够正确理解该评论。

本文提出了一种基于依存关系的情感要素抽取方法。该方法以情感词语为中心，通过挖掘句子中词语与词语之间所存在的关系，来抽取情感要素，即情感词语的评价对象。我们利用了第一届中文倾向性分析评测会议研讨会（COAE2008）任务三的语料进行训练和测试，并将该方法用在我们所下载的影评语料上做进一步的实验。实验结果在准确率、召回率和 F 值上都有一定的提高，这充分证明了该方法的有效性。

本文余下的内容做如下安排：第二部分：介绍了在这个领域目前的相关工作；第三部分：系统描述该方法，介绍其处理过程；第四部分：展示实验结果，并对结果进行分析；第五部分：总结，简述下一步工作。

2 相关工作

情感计算的研究工作最早可以追溯到 20 世纪 90 年代，当时的研究重点在于词汇的倾向性判断。在国外，情感计算的研究工作主要针对英文信息，而在国内，其研究的起步相对较晚，主要对中文信息进行处理，上海交大、复旦、大连理工、哈工大、中科院等研究机构针对情感计算的不同方面开展了研究，并且都取得了一定的成果。

目前在情感要素抽取方面，国内外主要有这样几种策略。2003 年，Yi等人根据名词短语的组成和位置特点,采用相似性测试(Likelihood test)方法来确定评价主题[1]。2004 年，Hu和Liu提出了一种根据主题和一些指示词的共显特征来识别常现(Frequent)和非常现(Infrequent)主题术语的方法[2],文中对于常现主题术语的识别采用了关联规则挖掘[3]的方法，而对于不包含常现主题术语的句子，若在其中找到一个或多个情感词语，就认为与这些情感词语最近的名词或名词短语为非常现的主题术语。2005 年，Popescu和Etzioni采用点互信息(PMI)的方法获取候选主题术语[4]。这个互信息值可以从整个互联网来获得。此外，在特定领域中，还可以利用领域本体来抽取主题术语。2007 年，Cheng等人研究了基于本体的主题抽取方法[5]。他们首先采用半自动的方法基于一些现存的本体资源构建有用的领域本体（汽车领域），然后将基于规则的命名实体识别技术和信息抽取引擎结合起来，识别被评价的汽车领域主题术语并给他们指派该应用领域的相关概念。同年，姚天昉等人也利用领域本体来抽取语句主题以及它的属性,然后在句法分析的基础上,识别主题和情感描述项之间的关系,从而最终决定语句中每个主题的极性[6]。

3 我们的方法

本文提出了一种新的情感要素抽取的方法。该方法首先采用一种改进了的基于HowNet 的情感词语识别方法，识别出文档中的情感词语，然后通过对情感词语上下文的选取，得到一个包含情感词语的，合适长度，语义相对完整的短句，再利用斯坦福大学设计的分析器，分析这个完整的短句，得到一系列依存关系，最后对这些依存关系进行分析，抽取出情感要素，其倾向性即为该情感词语的倾向性。

从以下例子可以清晰看出该方法的整体分析过程，例如：

说是“终于”，是 因为 从 最初 的 筹拍 就 开始 关注 了，据说，这 是 中国 的 第一 部 惊悚 片；据说，演员 阵容 强大；据说，开播 后，很 卖座，票房 成绩 很 好，据说 …… 昨天 抽空，终于 耳闻目睹 了 一番。【好】

对于这个句子，通过情感词语的识别，确定含有情感词语“好”，然后对该情感词语进行上下文抽取，抽取出包含该情感词语的短句：*票房 成绩 很 好*

接着利用分析器对其进行依存关系分析，得到结果：

[nmod(成绩-2, 票房-1), nsubj(好-4, 成绩-2), advmod(好-4, 很-3)]

根据依存关系的结果得出该情感词语“好”，修饰名词性主语“成绩”，故抽取出情感要素“成绩”，其倾向性和“好”一致。从最后的副词关系又可进一步得出，情感词语“好”的修饰词是程度副词“很”，故其情感强弱程度为强，这种副词关系的识别对以后的情感量化有很大的作用。

下面将逐一介绍该方法的各个阶段，情感词语的识别和情感词语的上下文选取这两个阶段，主要是在以前的工作基础上完成的，本文将不做重点介绍。

3.1 情感词语的识别

情感词语的识别是该方法中十分重要的一个阶段，情感要素的识别是在情感词语正确识别的基础上进行的。本文采用了一种基于HowNet的改进后的情感词语识别方法[7]，该方法利用了HowNet中的“良”、“莠”情感义原进行一种极性相似度计算，识别出词汇的情感倾向性。该方法在第一届中文倾向性分析评测（COAE2008）比赛中的中文情感词语褒贬分析任务中取得了很好的效果，能够有效地识别情感词语、区分情感词语的极性。

3.2 情感词语上下文抽取

为了保证后续阶段能够更加准确的进行依存关系分析,在期间加入了情感词语上下文抽取阶段。对于一个结构复杂的句子，依存分析的结果必将也很复杂，不仅会产生很多无用的依存关系，甚至还有可能产生错误的依存关系，这将对后续的分析过程产生很大的干扰。

在这里主要是根据标点符号和情感词语来进行子句划分，抽取上下文的。对于句号、感叹号等明显需要划分的标点符号，则直接将其划分成两个句子；而对于逗号、分号等具有不确定性的标点符号，则根据该句所含有的情感词语个数来决定是否进一步做子句划分，当在句中出现了两个及以上的情感词语时，就根据其间的标点符号做子句划分。经过划分处理之后，将得到一些包含情感词语的短句。若该短句经依存关系分析得不到情感要素，则合并前一个子句，对一个稍长的句子进行分析。

3.3 依存关系分析

依存关系的分析是本方法的核心。我们利用了斯坦福大学设计的分析器，对已经经过多重处理的包含情感词语的短句进行依存分析，通过对依存关系分析结果研究，识别出我们认为有用的关系，从而确定情感词语所修饰的对象即情感要素，其倾向性多数情况下与该情感词语一致。

那么，识别关键依存关系是十分重要的，直接影响到最后的抽取结果。本文利用第一届中文倾向性分析评测（COAE2008）任务三的参考答案，将其中手机领域的文本作为训练语料，抽取关键依存关系。实验得到了 44 种常用依存关系，对这些关系进一步分析，抽取关键依存关系。其抽取过程如下：

步骤 1：预处理，对训练语料进行分词，分句，词性标注等。

步骤 2：利用会后参考答案，找到包含情感要素的句子。

步骤 3：针对每一个包含情感要素的句子，识别出其中含有的情感词语。

步骤 4：利用斯坦福大学设计的分析器，对同时包含情感词语和情感要素的句子做依存分析，得到一系列依存关系。

步骤 5：分别从情感要素和情感词语开始，遍历这些依存关系，找到他们之间的一条依存关系路径。

步骤 6：对所有得到依存关系路径进行频率统计，并选择频率较高的依存关系路径。

在实验过程中，我们对选择频率阈值的选取做了多次的尝试，对选择频率较高的前十种、二

十种、三十种、四十种、五十种依存关系路径分别做了实验，实验结果显示，随着所选路径种类的增多，情感要素抽取的准确率逐渐降低，而召回率逐渐升高。

最后，本文选用了频率较高的四十种依存关系路径作为最终结果，并将其分为两大类别。一类是简单关系，可以直观的得到情感要素。例如：*“完美的音乐表现”[rcmod(音乐表现-3, 完美-1)，cpm(完美-1, 的-2)]；外形佳：nsubj(佳-2, 外形-1)*；另一类是具有传递性的关系，需要通过对这些关系之间的传递性进行分析才能得到情感要素，例如：*“自带软件还总无缘无故出各种各样的毛病”: [nsubj(出-5, 自带软件-1)，advmod(出-5, 还-2)，advmod(出-5, 总-3)，advmod(出-5, 无故-4)，rcmod(毛病-8, 各种各样-6)，cpm(各种各样-6, 的-7)，dobj(出-5, 毛病-8)]*.这里通过 *nsubj(出-5, 自带软件-1)，dobj(出-5, 毛病-8)* 这两个关系递推找出，“自带软件”是“出”的主语，而“出”又是“毛病”的谓语动词，利用这种传递关系，可以识别出情感词语“毛病”所修饰的要素是“自带软件”。

下面分别对这两类依存关系系列表说明，表 1 为部分简单依存关系及其描述，表 2 为部分传递性依存关系及其描述。（其中 S 为情感词语）

表 1　部分单一性依存关系及其描述

序号	依存关系对	说　明	实　例
1	nsubj(S, A)	含形容词性情感词语的主谓结构	nsubj(准确-5, 电量显示-2)
2	rcmod(A, S)	含形容词性情感词语的偏正短语结构	rcmod(外形设计-4, 精妙-1)
3	amod (A, S)	形容词短语结构	amod(音乐播放效果-3, 最佳-2)
4	nmod(S, A)	含名词性情感词语名词短语结构中	nmod(毛病-2, 光盘自带软件-1)
5	comod(A, S)	仅含一个情感词语的并列短语结构	comod(灵敏度-1, 高-2)
6	assmod(A, S)	含名词性情感词语的偏正短语结构	assmod(摄像功能-6, 时尚-4)
7	dobj(A, S)	含副词性情感词语的动宾结构	dobj(按键的使用-6, 生硬-8)
8	ccomp(A, S)	含使动性情感词语的主谓结构	ccomp(屏幕效果方面-13, 令人满意-16)

表 2　部分传递性依存关系及其描述

序号	依存关系对	说　明	实　例
1	nsubj(S1, A) comod(S1, S2)	情感词语 S1 修饰情感要素 A，S1 和 S2 又为并列结构，故同时修饰 A	nsubj(清晰-2, 通话-1) comod(清晰-2, 明亮-3)
2	nmod (A, B) nsubj(S, B)	情感词语 S 修饰情感要素 B, B 和 A 组成名词结构，故 S 同时修饰 A 和 B	nmod(使用-2, 操作系统-1) nsubj(烦琐-5, 使用-2)
3	nsubj(A, B) dobj (A, S)	B 为 A 的主语，A 不为情感词语，S 为 A 的宾语，故 S 修饰 B	nsubj(出-5, 自带软件-1) dobj(出-5, 毛病-8)
4	nsubj(是, B) attr(是, S)	“是”结构。递推关系为“B 是 S”故 S 修饰 B	nsubj(是-3, 屏幕的显示面积-1) attr(是-3, 令人满意-21)
5	ccomp(A, S1) comod(S1, S2)	情感词语 S1 修饰情感要素 A, S1 和 S2 为并列结构，故 S1、S2 同时修饰 A	ccomp(拍照的各项功能-11, 齐全-12) comod(齐全-12, 实用-13)
6	rcmod(A, S1) comod(S1, S2)	情感词语 S1 修饰情感要素 A, S1 和 S2 为并列结构，故 S1、S2 同时修饰 A	rcmod(外形设计-4, 精妙-1) comod(精妙-1, 独特-2)
7	conj(A, B) nsubj(S, A)	A 和 B 为并列结构，S 修饰 A，故 S 同时修饰 A 和 B	conj(收听效果-7, 通话质量-5) nsubj(不错-9, 收听效果-7)

除此之外还有两类重要依存关系。一个是 neg(A,B),表示否定关系，若其含有情感词语，则会将其情感倾向进行反向，例如“neg(准确-5, 不-4)”。另一个是 advmod(A,B)，表示副词修饰关系，若其含有情感词语和程度副词，则会对其情感强度进行修改，例如“advmod(烦琐-5, 非常-4)”。

4 实验和分析

4.1 语料及测试方法介绍

本文选用了第一届中文倾向性分析评测研讨会（COAE2008）任务三的语料进行训练和测试，共包含 478 篇主观性文本，包括笔记本、手机、数码相机和汽车四个领域，约 3000 个句子，其中选用了手机这个领域的文本 123 篇作为训练语料，以抽取重要的依存关系。除此之外，还从网上下载了影评文章 100 篇，约 400 个句子，做了进一步的测试。

本文采用了传统的准确率(Precision= 识别出的正确情感要素数/识别出的情感要素数)、召回率(Recall=识别出的正确情感要素数/正确情感要素数)、以及 F 值（F =（2*precision*recall）/(precision+recall)）作为评测标准。由于 COAE2008 研讨会所给答案的情感要素有可能是短语，简单的分词并不能得到其结果，故我们在利用他提供的语料做实验时，首先将可能组成情感要素的词语合并成短语，再做进一步的实验，以保证评测结果的准确性。对于影评文本，我们则对语料进行手工标注，标注出可能出现的情感要素及其倾向性，以评测本文所提出的方法。

4.2 实验结果及分析

本文对四个领域的文本做了测试实验，其评测结果如表 3。

表 3　情感要素抽取结果

语料	准确率（P）	召回率（R）	F 值（F）
笔记本领域	0.6357	0.4757	0.7149
数码相机领域	0.6248	0.4593	0.5294
汽车领域	0.6272	0.4479	0.5226
影评文本	0.5983	0.4326	0.5021

从实验结果可以看出，该方法在各领域的效果较为平均。将实验结果和 COAE2008 研讨会结果相比，抽取效果在准确率和召回率上都有一定的提高，总体来说，均高于平均水平，其中在笔记本、数码相机、汽车领域文本中的结果相对较好，然而也不是很令人满意。分析主要原因是在于情感要素的抽取涉及到了多方面的技术。除了倾向性分析的方法和技术即情感词语的识别以外，还包含了中文分词、短语识别、未登陆词识别、句法分析等技术。在前三个领域的文本处理中，本文首先将情感要素短语进行了合并，这对情感要素抽取的准确率提高起到了很大的作用。除此之外，分析器的依存分析效果严重依赖于句子结构，情感词语的上下文抽取将直接影响分析结果，如“且 整体外观 透 着 一 种 无与伦比 的 尊贵【尊贵】”对于这个短句，依存分析结果为：[nsubj(透-3, 且-1), advmod(透-3, 整体外观-2), asp(透-3, 着-4), numod(种-6, 一-5), clf(尊贵-9, 种-6), rcmod(尊贵-9, 无与伦比-7), cpm(无与伦比-7, 的-8), dobj(透-3, 尊贵-9)] ，无法准确的抽取出情感要素，而对于“整体外观 透 着 一 种 无与伦比 的 尊贵【尊贵】” 这个短句，依存分析结果为：[nsubj(透-2, 整体外观-1), asp(透-2, 着-3), numod(种-5, 一-4), clf(尊贵-8, 种-5), rcmod(尊贵-8, 无与伦比-6), cpm(无与伦比-6, 的-7), dobj(透-2, 尊贵-8)] ，则可以很容易的通过递推关系找到情感要素“整体外观”。对于中文文本，提取出含有连接词的短句是十分常见的，这也严重影响了准确率。

5 结论

本文提出了一种基于依存关系的情感要素抽取方法，这是一种通用的方法，并不专门针对某一个特定领域。本文使用了多个领域的语料进行试验，发现依存关系可以很好的帮助我们识别情感要素。

然而，利用该方法进行情感要素抽取时，发现对于关键依存关系的挖掘是十分重要的，今后应该进一步实验，以更加准确的挖掘关键依存关系。此外，依存关系分析器的准确率还不够高，仅仅依靠分析器进行抽取可能并不能达到一个很好的抽取效果，可以结合词性标注、短语结构、偏移位置等信息来综合考虑，以达到一个更加好的抽取效果。

参 考 文 献

[1] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques[A]. In: Proceedings of the 3rd IEEE International Conference on Data Mining(ICDM-2003)[C].Melbourne,Florida:2003,427-434.

[2] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews[A].In Proceedings of Nineteeth National Conference on Artificial Intellgience(AAAI-2004)[C].San Jose,USA:2004.

[3] HU M,LIU B. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining.New York: ACM Press,2004:168-177.

[4] A.-M.Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews [A]. In: Proceeding of HLT-EMNLP-05,the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing[C].Vancouver,Canada:2005,339-346.

[5] X. Cheng. Automatic Topic Term Detection and Sentiment Classification for Opinion Mining [D].Master Thesis. Saarbr cken, Germany: The University of Saarland,2007.

[6] 娄德成，姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用，26（11），2006，2622-2925.

[7] 何婷婷，闻彬，宋乐，王倩，罗乐. 词语情感倾向性识别及观点抽取研究. 第一届中文倾向性分析评测研讨会. 2008, 89-93.