许洪波　　孙乐　　姚天昉　　主编
Xu Hongbo　Sun Le　Yao Tianfang　(**Eds.**)

第三届中文倾向性分析评测(COAE2011)

The Third Chinese Opinion Analysis Evaluation

中国科学院计算技术研究所
2011-10-20

第三届中文倾向性分析评测委员会

评测主办机构

中国中文信息学会信息检索专业委员会

评测网址

http://www.ir-china.org.cn/Information.html

评测组织单位

中国科学院计算技术研究所
中国科学院软件研究所
福州大学

评测资源提供单位

福州大学
中国科学院计算技术研究所

评测指导专家

程学旗、刘挺、马少平

评测委员会

主席：

许洪波	中国科学院计算技术研究所
孙乐	中国科学院软件研究所
姚天昉	上海交通大学

委员：

廖祥文	福州大学
黄萱菁	复旦大学
赵军	中科院自动化研究所
林鸿飞	大连理工大学
朱靖波	东北大学
王厚峰	北京大学
吕学强	北京信息科技大学
张 敏	清华大学
秦 兵	哈尔滨工业大学
徐睿峰	哈尔滨工业大学
关毅	哈尔滨工业大学

谭松波　　中科院计算技术研究所
唐慧丰　　解放军外国语学院
王小捷　　北京邮电大学
徐蔚然　　北京邮电大学
徐冰　　哈尔滨工业大学
王素格　　山西大学
陈竹敏　　山东大学
万小军　　北京大学

前　言

众所周知，文本倾向性（观点和情感等）分析近几年持续成为自然语言处理领域研究的热点问题之一。TREC 评测、NTCIR 评测以及前两届中文倾向性分析评测推动和加速了倾向性分析研究的发展。在 SIGIR、ACL、WWW、CIKM、WSDM 等著名国际会议上，针对这一问题的研究成果层出不穷。随着研究的深入展开，也出现了一些新的研究关注点，如 Aspect-Based Opinion Mining，Context-sensitive Opinion Mining 等。在国内，对于文本倾向性分析的研究正处于快速发展中。如何结合中文处理的特点，进一步推动中文情感分析的发展是目前亟待解决的问题。

第一、二届中文倾向性分析评测共吸引了来自日本国立德岛大学、香港城市大学、哈尔滨工业大学、东北大学、大连理工大学、山西大学、北京大学、北京邮电大学、南京大学、上海交通大学、复旦大学、华中师范大学、福州大学、中科院软件所、中科院自动化所、中科院计算所、富士通研究开发中心、北京拓尔思信息技术股份有限公司、上海语天信息技术有限公司等近 20 家国内外一线科研单位的 30 多个科研团队参加，在清华大学（第一届）和上海交通大学（第二届）举行的 workshop 也得到了国内外同行的热情支持，可以说，两次评测均取得了圆满成功。

为了持续推动中文倾向性分析技术的发展和应用，中文信息学会信息检索专业委员会在成功组织前两届中文倾向性分析评测的基础上，以在山东大学举行的第七届全国信息检索学术会议（CCIR2011）为依托，继续组织第三届中文倾向性分析评测（The Third Chinese Opinion Analysis Evaluation-COAE2011）。本届评测将继续致力于探索中文倾向性分析的新技术、新方法，推动中文倾向性分析理论和技术的研究及应用，在此基础上建立、完善中文倾向性分析研究的基础资源库和评测标准。

中文信息学会信息检索专业委员会

第三届中文倾向性分析评测委员会

2011 年 10 月

目 录

第三届中文倾向性分析评测（COAE2011）总结报告*

许洪波[1]，孙乐[2]，姚天昉[3]，廖祥文[4]
[1]中国科学院计算技术研究所，北京，100190
[2]中国科学院软件研究所，北京，100190
[3]上海交通大学，上海，200240
[4]福州大学，福州，350108
Email: hbxu@ict.ac.cn, sunle@iscas.ac.cn, tf_yao@yahoo.com.cn, liaoxw@fzu.edu.cn

摘　要：本文详细描述了第三届中文倾向性分析评测（COAE2011）的总体情况。在前两届评测（COAE2008、COAE2009）的基础上，本次评测把领域知识和上下文语境（Context）对倾向性的影响融入到相关任务中，同时沿袭了词—句子—篇章的三级评测体系，以便考察不同粒度的倾向性分析效果。具体设置了 4 个评测任务：领域观点词的抽取与极性判别、中文观点句抽取、评价搭配抽取和观点检索。本次评测共有 21 个队伍报名，18 个队伍成功提交结果。本文对评测的任务、数据、指标以及评测结果等情况进行了详细介绍。

关键词：中文倾向性分析评测，领域，上下文语境，评价搭配，观点检索

Overview of the Third Chinese Opinion Analysis Evaluation (COAE2011)

Hongbo Xu[1], Le Sun[2], Tianfang Yao[3], Xiangwen Liao[4]
[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
[2]Institute of Software , Chinese Academy of Sciences, Beijing,, 100190
[3] Shanghai Jiaotong University, Shanghai, 200240
[4]Fuzhou University, Fuzhou, 350108
Email: hbxu@ict.ac.cn, sunle@iscas.ac.cn, tf_yao@yahoo.com.cn, liaoxw@fzu.edu.cn

Abstract: This paper proposes the overview of the third Chinese Opinion Analysis Evaluation (COAE 2011). Based on COAE2008 and COAE2009, this year we try to measure how the domain and context affect opinion analysis. We also follow the setup of word-sentence-full text, in order to investigate the impact of different Granularity. Accordingly, we conduct four tasks in COAE 2011, as follows: Domain Opinion Words Identification, Opinion Sentences Identification; Opinion Combination Extraction and Polarity Analysis, Opinion Retrieval. In this evaluation, 18 participants out of 21 applicants submit results successfully. This paper shows the details of task design, corpus information, evaluation metrics and evaluation results.

Keywords: Chinese Opinion Analysis Evaluation, Domain, Context, Opinion Combination, Opinion Retrieval

1　引言

文本倾向性（观点和情感等）分析近几年持续成为自然语言处理领域研究的热点问题

*本文工作受到国家自然科学基金重点课题（60933005）、国家自然科学基金课题（61173064、90920010）的资助

之一。TREC 评测、NTCIR 评测以及前两届中文倾向性分析评测推动和加速了倾向性分析研究的发展。在 SIGIR、ACL、WWW、CIKM、WSDM 等著名国际会议上，针对这一问题的研究成果层出不穷。随着研究的深入展开，也出现了一些新的研究关注点，如 Aspect-Based Opinion Mining，Context-sensitive Opinion Mining 等。在国内，对于文本倾向性分析的研究正处于快速发展中。如何结合中文处理的特点，进一步推动中文情感分析的发展是目前亟待解决的问题。

第一、二届中文倾向性分析评测共吸引了来自日本国立德岛大学、香港城市大学、哈尔滨工业大学、东北大学、大连理工大学、山西大学、北京大学、北京邮电大学、南京大学、上海交通大学、复旦大学、华中师范大学、福州大学、中科院软件所、中科院自动化所、中科院计算所、富士通研究开发中心、北京拓尔思信息技术股份有限公司、上海语天信息技术有限公司等近 20 家国内外一线科研单位的 30 多个科研团队参加，在清华大学（第一届）和上海交通大学（第二届）举行的 workshop 也得到了国内外同行的热情支持，可以说，两次评测均取得了圆满成功。

为了持续推动中文倾向性分析技术的发展和应用，中文信息学会信息检索专业委员会在成功组织前两届中文倾向性分析评测的基础上，以在山东大学举行的第七届全国信息检索学术会议（CCIR2011）为依托，继续组织第三届中文倾向性分析评测（The Third Chinese Opinion Analysis Evaluation-COAE2011）。本届评测继续致力于探索中文倾向性分析的新技术、新方法，推动中文倾向性分析理论和技术的研究及应用，在此基础上建立、完善中文倾向性分析研究的基础资源库和评测标准。

本文对评测的任务、数据、指标以及评测结果等情况进行了详细介绍。具体安排如下：第二节介绍 COAE2011 评测的任务设置，包括每个子任务的详细描述。第三节介绍评测数据集情况。第四节介绍评测方法和评测指标。第五节介绍各评测任务的评测结果。第六节对评测过程的一些问题进行了说明和分析探讨。最后是结论和展望。

2 任务描述

随着文本倾向性分析研究的深入，国内外的研究发现：（1）领域知识对文本倾向性分析有重要的影响，纷纷开展诸如领域观点词表构建、跨领域倾向文本分类等研究。（2）上下文语境（Context）对倾向性判别至关重要，吸引了国内外研究学者的广泛关注。因此，在前两届评测的基础上，本次评测把领域知识和上下文语境（Context）对倾向性的影响融入到相关任务中，同时沿袭了词—句子—篇章的三级评测体系，以便考察不同粒度的倾向性分析效果：首先从给定的三个领域数据集中抽取观点词并判断极性；然后从三个领域数据集随机抽取一定比例的数据经过自动分句构成测试集，从中抽取观点句并进行观点极性判别；接下来，从得到的观点句中抽取评价搭配<观点句，评价对象,评价短语，极性>；最后，把三个领域数据集作为一个大数据集，结合领域知识和上下文语境对给定的查询对象进行观点检索。为了适应上述评测要求，本次评测使用了全新的评测数据集。具体评测任务设置如表 1 所示。

表1 COAE2011评测任务设置

Tab.1 The Tasks in COAE 2011

任务号	类型	任务名称	任务说明
任务 1	要素级	领域观点词的抽取与极性判别	考虑领域对倾向性的影响，识别给定的三个领域的观点词，并判断极性。
任务 2	句子级	中文观点句抽取	从三个领域数据集随机抽取一定比例的数据经过自动分句构成测试集，从中抽取所有观点句并判别观点极性。
任务 3	要素级	评价搭配抽取	考虑上下文语境对词语倾向性的影响，抽取评价对象、评价短语并判别评价极性。
任务 4	篇章级	观点检索	面向特定对象的中文文本观点检索

下面对每个任务的要求分别进行详细介绍。

2.1 任务 1：领域观点词抽取与极性判别

不同领域之间所使用的观点词存在明显的差异，同一观点词在不同领域也有不同的倾向性表达。例如：

[Doc1] 该机采用了高端笔记本流行的金属机身，“城市流光”纹理，外形边角处理十分圆滑。

[Doc2] 沙溢扮演的郭洋港在大学时就充满梦想，一心想着毕业后能大展鸿图，性格却决定他无法像袁浩东那样变得圆滑，只能选择一再躲避现实。

Doc1 和 Doc2 中都出现了观点词“圆滑”，但是由于出现在不同的领域，同一个观点词却有明显不同的倾向性。

考虑到领域对观点表达的影响，本届评测设置了领域观点词抽取与极性判别任务。给定三个领域的数据集 digital、entertainment、finance，要求参赛者分别从三个数据集中抽取观点词，并判断该观点词的极性（即褒义、贬义）。特别需要注意的是：观点词是表达对外评价的褒贬观点词，不是表达人物自身情绪的倾向词，也不包括评价短语（可根据其英文翻译是否为一个单词来简单判断，但某些成语例外），故例句 1 中的观点词是“圆滑”而不是“十分圆滑”。每个领域提交 2000 个观点词，要求生成三个结果文件：D_Lexicon, E_lexicon, F_Lexicon，并按置信度降序排列。

提交结果格式：

id run-tag sentiment-word docid context-string polarity confidence-score

说明：

id：结果序号

run-tag：参加单位标识

sentiment-word：观点词

docid：来源文本 id

context-string：以观点词为中心的前后各 20 字节组成的文本片段

polarity：观点词极性，-1 代表贬义，1 代表褒义

confidence-score：观点词判别的置信度（用于排序）

注意：不同字段之间以 tab 键分隔，下同。

例如，对 Doc1 和 Doc2 的领域观点词识别结果如下：

Doc1

1 abcd圆滑 Doc1 理，外形边角处理十分圆滑。 1 1.0

Doc2

1 abcd圆滑 Doc2 无法像袁浩东那样变得圆滑，只能选择一再躲避现实 -1 1.0

2.2 任务 2：中文观点句抽取

从 digital、entertainment、finance 三个领域数据集中分别随机抽取了 2000 篇文章，经过自动分句后构成任务 2 和任务 3 的三个测试集 digital_sentence、entertainment_sentence、finance_sentence，即测试集直接以句子集合的形式发布。本任务要求参赛者从每个领域的测试集中自动识别出所有观点句及其表达观点的总体极性（褒义、贬义或混合观点），要求不仅能准确识别观点句，而且要尽可能找出所有的观点句。输出三个结果文件：D_sentence.txt, E_sentence.txt, F_sentence.txt,结果按照置信度降序排列，同时要求给出观点句所在的文章编号。与前两届评测任务的区别是本任务中所有返回结果都会参与评分。

提交结果格式：

id run-tag docid sentence-string polarity confidence-score

说明：

id：结果序号

run-tag：参加单位标识

docid：观点句所在文档的 id

sentence-string：观点句

polarity：观点句总体极性，-1 代表贬义，1 代表褒义，0 代表既有褒义又有贬义

confidence-score：观点句判别的置信度

例如：

[Doc3] 数字键之上的导航键是金属制，冰冰冷冷很特别，按键手感则是有深度且偏软，按起来很爽快。

[Doc4] 诺基亚 5320 在设计上采用了比较传统的直板造型，108×46×15 的机身尺寸和变焦的圆润处理看起来中规中矩，外观上并没有什么太大的突破。

[Doc5] LG GRS25DDH 冰箱，这款冰箱外形非常精美，整体采用珍珠白色 VCM 面板，镶嵌有韩式绣球花纹，非常具有艺术感，这种精巧的设计也比较受女性用户的青睐。

其识别结果如下：

Doc3

1 abcdDoc3 数字键之上的导航键是金属制，冰冰冷冷很特别，按键手感则是有深度且偏软，按起来很爽快。 1 0.8

Doc4

1 abcdDoc4　　诺基亚 5320 在设计上采用了比较传统的直板造型，108×46×15 的机身尺寸和变焦的圆润处理看起来中规中矩，外观上并没有什么太大的突破。　-1　0.8

Doc5

1 abcdDoc5　　LG GRS25DDH 冰箱，这款冰箱外形非常精美，整体采用珍珠白色 VCM 面板，镶嵌有韩式绣球花纹，非常具有艺术感，这种精巧的设计也比较受女性用户的青睐。　1　1.0

2.3 任务 3：评价搭配抽取

本任务关注上下文语境对观点识别和倾向性判断的影响。采用跟任务 2 同样的测试集（经过自动分句的句子集合），要求找出任务 2 抽取的每个观点句中观点所针对的评价对象、评价短语，并对评价的倾向性做出判别。任务输出结果是由七元组<id, 单位标识，文档 id, 观点句，评价对象,评价短语，极性>构成的“评价搭配”，其中评价对象是指评论针对的对象或对象的属性，越具体越好；评价短语是指修饰成分和评价词语组合而成的评价单元，比如“十分喜欢”、“不讨厌”等；修饰成分是指加强、减弱或置反观点的语言成分，可以是程度副词、否定词等。需要注意的是，一个句子可能包括多个评价搭配，故其中包含的七元组可能是一个或者多个。本任务要求找出所有观点句中的所有评价搭配七元组，按照三个领域分别输出结果文件：

D_Sentiment_pair.txt, E_Sentiment_pair.txt, F_Sentiment_pair.txt。

提交结果格式：

id run-tag docid sentence-string sentiment-object sentiment-string polarity

说明：

id：结果序号

run-tag：参加单位标识

docid：观点句所在文档的 id

sentence-string：观点句

sentiment-object：句子中观点的评价对象

sentiment-string：评价短语

polarity：对该对象评价的观点极性，-1 代表贬义，1 代表褒义

例如：

[Doc6] 凯越的油耗非常高。

[Doc7] 海尔乐家家 K3-D209 台式电脑，外观时尚简洁,影音效果好，无论是玩游戏、还是看电影，效果都还不错。

[Doc8] 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。

[Doc9] 国外 MP3 很容易让人萌发购物欲望，不过它的缺点在于功能相对单一，而且价格普遍偏高；而国产产品的缺点则在于音质有待提升，设计和做工也有较大的进步空间。

其识别结果如下：

Doc6

1 abcdDoc6 凯越的油耗非常高。 油耗 非常高 -1

Doc7

1 abcdDoc7 海尔乐家家 K3-D209 台式电脑，外观时尚简洁,影音效果好，无论是玩游戏、还是看电影，效果都还不错。 外观 时尚简洁 1

2 abcdDoc7 海尔乐家家 K3-D209 台式电脑，外观时尚简洁,影音效果好，无论是玩游戏、还是看电影，效果都还不错。 影音效果 好 1

3 abcdDoc7 海尔乐家家 K3-D209 台式电脑，外观时尚简洁,影音效果好，无论是玩游戏、还是看电影，效果都还不错。 效果 都还不错 1

Doc8

1 abcdDoc8 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。 运行速度 顺畅 1

2 abcdDoc8 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。 扩展方面 优秀 1

3 abcdDoc8 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。 机身 轻巧的 1

4 abcdDoc8 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。 价格 实惠的 1

5 abcdDoc8 诺基亚 5230 手机搭载塞班系统，运行速度顺畅，扩展方面优秀，搭配轻巧的机身以及实惠的价格，是一款千元机型的不错之选，感兴趣的朋友不妨多关注一下。 诺基亚 5230 手机 不错之选 1

Doc9

1 abcdDoc9 国外 MP3 很容易让人萌发购物欲望，不过它的缺点在于功能相对单一，而且价格普遍偏高；而国产产品的缺点则在于音质有待提升，设计和做工也有较大的进步空间。 功能 相对单一 -1

2 abcdDoc9 国外 MP3 很容易让人萌发购物欲望，不过它的缺点在于功能相对单一，而且价格普遍偏高；而国产产品的缺点则在于音质有待提升，设计和做工也有较大的进步空间。 价格 普遍偏高 -1

3 abcdDoc9 国外 MP3 很容易让人萌发购物欲望，不过它的缺点在于功能相对单一，而且价格普遍偏高；而国产产品的缺点则在于音质有待提升，设计和做工也有较大的进步空间。 音质 有待提升 -1

4 abcdDoc9 国外 MP3 很容易让人萌发购物欲望，不过它的缺点在于功能相对单一，而且价格普遍偏高；而国产产品的缺点则在于音质有待提升，设计和做工也有较大的进步空间。 设计和做工 也有较大的进步空间 -1

2.4 任务 4：观点检索

在篇章级上继续前两届评测的观点检索任务。给定 20 个观点检索对象（topic/target），针对每个对象，把 digital、entertainment、finance 看成一个大数据集，要求找出包含评价该对象的倾向性观点的文章。给定对象可能是人物、商品、组织机构或者概念、事物、事件等。该任务是信息检索和观点识别的组合任务。输出针对给定对象的评论性文章并按观点相关度降序排列。

提交结果格式：

tid run-tag docid polarity float-score

说明：

tid：检索对象编号

run-tag：参加单位标识

docid：文本 id

polarity：文本表达观点的总体极性，-1 代表贬义，1 代表褒义，0 代表既有褒义又有贬义

float-score：观点相关程度

例如，对于检索对象“诺基亚 5230”，在 Doc1～Doc9 中只有 Doc8 是包含观点的相关文档，其他文档均无关，则正确的检索结果如下：

1 abcdDoc8 1 0.9

3 评测数据集

本届评测所有任务均采用全新的评测数据集，分别来自电子产品（digital）、影视娱乐（entertainment）、金融证券（finance）三个领域，数据规模比前两届评测有所扩大。评测数据集总体情况如表 2 所示：

表 2 评测数据集总体情况

Tab.2 Corpus Description

语料领域	语料说明
电子产品（digital）	全部文章都采集自互联网，经过内容抽取得到简体中文 txt 文本，每个领域约 15，000 篇，总计 44245 篇文章；包括真实用户的评论、博客、微博等，主观文本和客观文本混合；文章长度不一，按长度可以分为三个级别：篇章级、段落级、句子级。
影视娱乐（entertainment）	
金融证券（finance）	

评测时利用上述数据构造了两个数据集：COAE2011_Corpus_All_Text 和 COAE2011_Corpus_Sample_Sentence，两个数据集均包括 digital（电子产品）、entertainment（影视娱乐）、finance（金融证券）三个领域。其中 COAE2011_Corpus_All_Text 是本届评测数据的全集，包含所有未作分句处理的文本，总计 44245 篇。

COAE2011_Corpus_Sample_Sentence 是 COAE2011_Corpus_All_Text 数据集的子集，分别从每个领域随机抽取了 2000 篇文本组成，其中的文本均采用中科院计算所提供的自动分句工具进行了断句处理，每篇文本均组织成一句话占一行的格式，每行以\r\n 结束。

评测中任务 1 和任务 4 使用 COAE2011_Corpus_All_Text 数据集，任务 1 需要分领域处理，任务 4 的检索主题面向所有领域，检索结果不区分领域。任务 2 和任务 3 使用 COAE2011_Corpus_Sample_Sentence 数据集，针对文本中已经分好的句子进行处理，不必重新断句，结果以句子集合的形式提交，同时要分领域处理。

评测数据集与评测任务的对应关系说明如表 3 所示：

表 3 评测数据与评测任务的对应关系

Tab.3 Corpus for Tasks

对应任务	任务 1	任务 4	任务 2	任务 3
评测结果要求	分领域	不分领域	分领域	分领域
数据特点	未断句的文本集		已断句的文本集	
数据集 / **领域**	COAE2011_Corpus_All_Text		COAE2011_Corpus_Sample_Sentence	
Digital（电子产品）	14799		2000	
Entertainment（影视娱乐）	14904		2000	
Finance（金融证券）	14542		2000	
合计	**44245**		**6000**	

本届评测的任务 4 共设置了 20 个主题（topic），每个主题的观点相关文档从几十到数百不等，主题列表如表 4 所示。

表 4 任务 4 的评测主题

Tab.4 Topics for Task 4

编号	**主题（Topic）**
T301	新浪
T302	人民币升值
T303	微博
T304	Android
T305	诺基亚 5800
T306	创业板

T307	明基 BR1001
T308	索尼爱立信
T309	广发证券
T310	IPO
T311	房产税
T312	90 后
T313	万科
T314	存款准备金率
T315	非诚勿扰
T316	哈利波特
T317	蜗居
T318	裸婚
T319	iphone
T320	ipad

4 评测方式和评测指标

本届评测的任务 1 和任务 4 采用跟 TREC 评测相同的人工评判方法（即 pooling 方式）进行评测。对于每个任务，按顺序截取各单位提交结果的前 K 个组成评测池（任务 1 中 K=500，任务 4 中 K=100），组织专家评判小组对评测池结果进行人工评判，由其中所有正确结果组成标准答案；再根据标准答案，对各个单位提交的所有结果从头到尾进行自动评价，计算相应的评价指标得分，如准确率、召回率、F1、MAP、Raccuracy 等。

任务 2 和任务 3 已经标注了全部答案，采用自动评价方法对所有提交结果进行评测，主要指标为 F1，其他指标包括准确率、召回率等。特别的，考虑到任务 2 中在缺少上下文的情况下对于句子总体极性的判断有一定的不确定性，故该任务评测中同时考察观点句识别和观点句极性判断的能力。

任务 1、2、3 分别考察三个领域的识别情况，同时提供了宏平均和微平均的综合结果。任务 3 分别考察评价对象、评价短语、极性的识别情况。任务 1 的考察对象是观点词，置信度排序仅作为参考，所以主要评测指标选择 Precision 和 P@1000；任务 2 和任务 3 的结果排序相对可信，而且有完整的标注答案，所以首选的评测指标为宏平均 F1。

任务 4 的评测包括两方面：首先考察识别出主题相关文章并进行正确排序的能力，称为主题相关性检索子任务；然后考察识别出主题相关的有观点文章并能正确判断观点褒贬极性的能力（也就是说，只有跟主题相关且极性类型判别正确的结果才是正确结果），称为观点文档检索与极性识别子任务。由于 MAP 是比较严格的排序指标，因此作为任务 4 的首选评测指标。

每个参加单位每项任务允许提交 1-3 组结果，最少 1 组，最多 3 组，多组结果时之间按优先级排序，评测时优先选择第一组结果。不同单位、同一单位的不同任务均通过唯一标识 run-tag 进行区别，confidence-score 和 float-score 用于实现答案的降序排列。

5 评测结果

第三届中文倾向性分析评测时隔一年之后继续举行，依旧得到了国内同行的热情支持，共有来自中科院软件所、哈尔滨工业大学计算机学院语言技术研究中心网络智能研究室、北京大学信息科学技术学院计算语言学研究所、苏州大学自然语言处理实验室、大连理工大学信息检索实验室、苏州大学 Suda_SAM_OMS、北京邮电大学信息与通信工程学院 PRIS、北京邮电大学智能科学技术研究中心、南京理工大学经管学院信息管理系、重庆邮电大学中文信息处理研究所、清华大学、山东师范大学信息科学与工程学院、湖南工业大学计算机与通信学院、哈尔滨工业大学计算机科学与技术学院机器智能与翻译实验室、哈尔滨工业大学社会计算与信息检索研究中心、哈尔滨工业大学深圳研究生院智能计算研究中心、杭州电子科技大学计算机学院、山西大学计算机与信息技术学院等 18 支队伍成功提交结果，另有三个单位报名但未能及时提交结果。其中任务 3 共有 20 个单位报名，但最终成功提交结果的单位只有 12 个，可见难度较大。18 个单位参加任务的具体情况统计如表 5 所示。

表 5 各单位参加任务情况

Tab.5 The Information of each Participants

提交顺序	**报名团队**	**任务 1**	**任务 2**	**任务 3**	**任务 4**
1	中科院软件所	√	√		
2	哈尔滨工业大学计算机学院语言技术研究中心网络智能研究室	√	√		
3	北京大学信息科学技术学院计算语言学研究所	√	√	√	√
4	苏州大学自然语言处理实验室	√	√		
5	大连理工大学信息检索实验室	√	√	√	√
6	苏州大学 Suda_SAM_OMS	√	√	√	√
7	北京邮电大学信息与通信工程学院 PRIS	√	√	√	√
8	北京邮电大学智能科学技术研究中心	√	√	√	
9	南京理工大学经管学院信息管理系		√		
10	重庆邮电大学中文信息处理研究所	√	√	√	√
11	清华大学		√	√	
12	山东师范大学信息科学与工程学院		√	√	
13	湖南工业大学计算机与通信学院	√	√		

14	哈尔滨工业大学计算机科学与技术学院机器智能与翻译实验室	√	√	√	
15	哈尔滨工业大学社会计算与信息检索研究中心	√		√	
16	哈尔滨工业大学深圳研究生院智能计算研究中心	√	√		
17	杭州电子科技大学计算机学院		√	√	
18	山西大学计算机与信息技术学院	√	√	√	
任务提交总数		14	17	12	5

5.1 任务 1 评测结果

任务 1 共有 14 支队伍提交了 15 个 run，最终评测结果如下。

表 6 电子产品领域观点词识别的评测结果

Tab.6 The Evaluation Results for "**Digital**" Opinion Words Identification

电子产品领域				
参评系统标识	**Precision@1000**	**Precision**	**Recall**	**F1**
ISCAS-Opinion	0.539	0.259	0.0775	0.1193
HITWI	0.577	0.3927	0.0956	0.1538
PKUICL	0.63	0.3815	0.1141	0.1757
SoochowUniversity-1	0.526	0.309	0.0924	0.1423
DUTIR	0.593	0.364	0.1089	0.1676
Suda_SAM_OMS	0.516	0.3208	0.0779	0.1254
pris_t1	0.518	0.2765	0.0827	0.1273
buptailab	0.594	0.338	0.1011	0.1557
CQUPT	0.507	0.252	0.0754	0.1161
Hut	0.509	0.263	0.0787	0.1211
MI&Tlab	0.502	0.254	0.076	0.117
HIT-SCIR	0.567	0.3155	0.0944	0.1453
HITSZ	0.572	0.3425	0.1025	0.1577
SXU-31	0.635	0.34	0.1017	0.1566
SXU-32	0.674	0.34	0.1017	0.1566
Average	**0.563933333**	**0.3165667**	**0.09204**	**0.1425**
Best	**0.674**	**0.3927**	**0.1141**	**0.1757**

表 7 影视娱乐领域观点词识别的评测结果

Tab.7 The Evaluation Results for "**Entertainment**" Opinion Words Identification

影视娱乐领域				
参评系统标识	Precision@1000	Precision	Recall	F1
ISCAS-Opinion	0.548	0.2765	0.0838	0.1286
HITWI	0.596	0.4528	0.1009	0.1651
PKUICL	0.614	0.3605	0.1093	0.1677
SoochowUniversity-1	0.578	0.303	0.0918	0.141
DUTIR	0.601	0.388	0.1176	0.1805
Suda_SAM_OMS	0.512	0.3487	0.0793	0.1292
pris_t1	0.542	0.3113	0.0864	0.1352
buptailab	0.59	0.377	0.1143	0.1754
CQUPT	0.508	0.256	0.0776	0.1191
hut	0.501	0.2575	0.0781	0.1198
MI&Tlab	0.503	0.255	0.0773	0.1186
HIT-SCIR	0.55	0.316	0.0958	0.147
HITSZ	0.622	0.394	0.1194	0.1833
SXU-31	0.606	0.353	0.107	0.1642
SXU-32	0.624	0.353	0.107	0.1642
Average	**0.566333333**	**0.3334867**	**0.096373**	**0.14926**
Best	**0.624**	**0.4528**	**0.1194**	**0.1833**

表 8 金融证券领域观点词识别的评测结果

Tab.8 The Evaluation Results for "**Finance**" Opinion Words Identification

金融证券领域				
参评系统标识	Precision@1000	Precision	Recall	F1
ISCAS-Opinion	0.592	0.2775	0.0845	0.1295
HITWI	0.641	0.6125	0.0936	0.1624
PKUICL	0.62	0.378	0.1151	0.1764
SoochowUniversity-1	0.524	0.277	0.0843	0.1293
DUTIR	0.632	0.373	0.1135	0.1741
Suda_SAM_OMS	0.5245	0.5175	0.0787	0.1366
pris_t1	0.5612	0.5394	0.0791	0.138
buptailab	0.63	0.39	0.1187	0.182

CQUPT	0.504	0.252	0.0767	0.1176
hut	0.521	0.2635	0.0802	0.123
MI&Tlab	0.502	0.261	0.0794	0.1218
HIT-SCIR	0.604	0.4305	0.0948	0.1554
HITSZ	0.638	0.3905	0.1189	0.1822
SXU-31	0.587	0.361	0.1099	0.1685
SXU-32	0.672	0.361	0.1099	0.1685
Average	**0.583513333**	**0.37896**	**0.09582**	**0.15102**
Best	**0.672**	**0.6125**	**0.1189**	**0.1822**

表 9 三个领域观点词识别的宏平均和微平均结果（根据 Precision 排序）

Tab.9 The **Macro- and Micro**-Evaluation Results for Three Domains of Opinion Words Identification

参评系统标识	宏平均 Macro				微平均 Micro			
	P@1000	Precision	Recall	F1	P@1000	Precision	Recall	F1
HITWI	0.6047	0.486	0.0967	0.1613	0.6047	0.4681	0.0967	0.1603
Suda_SAM_OMS	0.5175	0.3957	0.0786	0.1312	0.5175	0.3786	0.0786	0.1302
HITSZ	0.6107	0.3757	0.1136	0.1744	0.6107	0.3757	0.1135	0.1744
pris_t1	0.5404	0.3757	0.0827	0.1356	0.5404	0.3426	0.0828	0.1333
DUTIR	0.6087	0.375	0.1133	0.1741	0.6087	0.375	0.1133	0.1741
PKUICL	0.6213	0.3733	0.1128	0.1733	0.6213	0.3733	0.1128	0.1733
buptailab	0.6047	0.3683	0.1114	0.171	0.6047	0.3683	0.1113	0.171
HIT-SCIR	0.5737	0.354	0.095	0.1498	0.5737	0.3462	0.095	0.1491
SXU-32	0.6567	0.3513	0.1062	0.1631	0.6567	0.3513	0.1062	0.1631
SXU-31	0.6093	0.3513	0.1062	0.1631	0.6093	0.3513	0.1062	0.1631
SoochowUniversity-1	0.5427	0.2963	0.0895	0.1375	0.5427	0.2963	0.0896	0.1375
ISCAS-Opinion	0.5597	0.271	0.0819	0.1258	0.5597	0.271	0.0819	0.1258
hut	0.5103	0.2613	0.079	0.1213	0.5103	0.2613	0.079	0.1213
MI&Tlab	0.5023	0.2567	0.0776	0.1191	0.5023	0.2567	0.0776	0.1191
CQUPT	0.5063	0.2533	0.0766	0.1176	0.5063	0.2533	0.0766	0.1176
Average	0.5689	0.3328	0.0946	0.1469	0.5689	0.3286	0.0946	0.1466
Best	0.6567	0.3957	0.1136	0.1744	0.6567	0.3786	0.1135	0.1744

表 10 三个领域观点词识别的宏平均和微平均结果（根据 P@1000 排序）

Tab.10 The **Macro- and Micro**-Evaluation Results for Three Domains of Opinion Words Identification

参评系统标识	宏平均				微平均			
	P@1000	Precision	Recall	F1	P@1000	Precision	Recall	F1
SXU-32	0.6567	0.3513	0.1062	0.1631	0.6567	0.3513	0.1062	0.1631
PKUICL	0.6213	0.3733	0.1128	0.1733	0.6213	0.3733	0.1128	0.1733
HITSZ	0.6107	0.3757	0.1136	0.1744	0.6107	0.3757	0.1135	0.1744
SXU-31	0.6093	0.3513	0.1062	0.1631	0.6093	0.3513	0.1062	0.1631
DUTIR	0.6087	0.375	0.1133	0.1741	0.6087	0.375	0.1133	0.1741
HITWI	0.6047	0.486	0.0967	0.1613	0.6047	0.4681	0.0967	0.1603
buptailab	0.6047	0.3683	0.1114	0.171	0.6047	0.3683	0.1113	0.171
HIT-SCIR	0.5737	0.354	0.095	0.1498	0.5737	0.3462	0.095	0.1491
ISCAS-Opinion	0.5597	0.271	0.0819	0.1258	0.5597	0.271	0.0819	0.1258
SoochowUniversity-1	0.5427	0.2963	0.0895	0.1375	0.5427	0.2963	0.0896	0.1375
pris_t1	0.5404	0.3757	0.0827	0.1356	0.5404	0.3426	0.0828	0.1333
Suda_SAM_OMS	0.5175	0.3957	0.0786	0.1312	0.5175	0.3786	0.0786	0.1302
hut	0.5103	0.2613	0.079	0.1213	0.5103	0.2613	0.079	0.1213
CQUPT	0.5063	0.2533	0.0766	0.1176	0.5063	0.2533	0.0766	0.1176
MI&Tlab	0.5023	0.2567	0.0776	0.1191	0.5023	0.2567	0.0776	0.1191
Average	0.5713	0.3430	0.0947	0.1479	0.5713	0.3379	0.0947	0.1475
Best	**0.6567**	**0.486**	**0.1136**	**0.1744**	**0.6567**	**0.4681**	**0.1135**	**0.1744**

5.2 任务 2 评测结果

任务 2 共有 17 支队伍提交了 20 个 run，最终评测结果如下。

表 11 电子产品领域观点句识别及极性判别的评测结果

Tab.11 The Evaluation Results for "**Digital**" Opinion Sentences Classification

电子产品领域				
参评系统标识	Precision	Recall	F1	P@1000
ISCAS-Opinion_D	0.462408	0.554792	0.504404	0.72
hitwi_D	0.49268	0.391297	0.436174	0.608
SoochowUniversity-1_D*	0.056905	0.119274	0.07705	0.103
DUTIR_D	0.439797	0.518675	0.47599	0.622
Suda_SAM_OMS——RUN1_D	0.600413	0.563073	0.581144	0.707
Suda_SAM_OMS——RUN2_D	0.599925	0.560078	0.579317	0.712
Suda_SAM_OMS——RUN3_D	0.600413	0.563073	0.581144	0.707

PRIS_COAE_t2_1_D	0.440073	0.468992	0.454072	0.552
PRIS_COAE_t2_2_D	0.365987	0.42389	0.392816	0.391
buptailab_D	0.348272	0.340909	0.344551	0.565
8-NJUST_D	0.135667	0.120155	0.127441	0.146
CQUPT_D*	0.416937	0.453665	0.434526	0.58
THUHUANG_D	0.483157	0.611522	0.539813	0.527
SDNU_D	0.360115	0.663319	0.466803	0.532
hut_D*	0.352539	0.5821	0.439128	0.502
MI&Tlab_D	0.0441	0.050564	0.047111	0.073
HITSZ_D	0.729751	0.660324	0.693304	0.8
hdu_D	0.330416	0.798097	0.467348	0.652
SXU-3_D	0.437077	0.602713	0.506702	0.588
PKUICL_D	0.404299	0.490486	0.443242	0.571
Average	0.4050466	0.4768499	0.429604	0.5329
Best	0.729751	0.798097	0.693304	0.8

表 12 影视娱乐领域观点句识别及极性判别的评测结果

Tab.12 The Evaluation Results for "**Entertainment**" Opinion Sentences Classification

影视娱乐领域				
参评系统标识	**Precision**	**Recall**	**F1**	**P@1000**
ISCAS-Opinion_E	0.213675	0.366151	0.269865	0.264
hitwi_E	0.303493	0.2262	0.259207	0.278
SoochowUniversity-1_E*	0.054861	0.203417	0.086415	0.076
DUTIR_E	0.186184	0.460537	0.265167	0.272
Suda_SAM_OMS——RUN1_E	0.303249	0.341741	0.321347	0.332
Suda_SAM_OMS——RUN2_E	0.332685	0.278275	0.303057	0.334
Suda_SAM_OMS——RUN3_E	0.303249	0.341741	0.323147	0.332
PRIS_COAE_t2_1_E	0.188156	0.364524	0.248199	0.204
PRIS_COAE_t2_2_E	0.163753	0.312449	0.214885	0.167
buptailab_E	0.115678	0.320586	0.170011	0.223
8-NJUST_E	0.068022	0.149715	0.093543	0.071
CQUPT_E*	0.156368	0.302685	0.206208	0.212
THUHUANG_E	0.213079	0.471928	0.293597	0.233
SDNU_E	0.157414	0.505289	0.240046	0.223
hut_E*	0.172962	0.540277	0.262036	0.205
MI&Tlab_E	0.017224	0.041497	0.024344	0.011

HITSZ_E	0.537607	0.511798	0.524385	0.531
hdu_E	0.118904	0.724166	0.204269	0.199
SXU-3_E	0.178439	0.50773	0.264071	0.191
PKUICL_E	0.161808	0.401953	0.230733	0.23
Average	0.1973405	0.368633	0.2402266	0.2294
Best	0.537607	0.724166	0.524385	0.531

表 13 金融证券领域观点句识别及极性判别的评测结果

Tab.13 The Evaluation Results for "**Finance**" Opinion Sentences Classification

金融证券领域				
参评系统标识	**Precision**	**Recall**	**F1**	**P@1000**
ISCAS-Opinion_F	0.127632	0.386847	0.191939	0.149
hitwi_F	0.204598	0.172147	0.186975	0.089
SoochowUniversity-1_F*	0.010023	0.075435	0.017695	0.015
DUTIR_F	0.104384	0.386847	0.164406	0.131
Suda_SAM_OMS——RUN1_F	0.20801	0.311412	0.249419	0.161
Suda_SAM_OMS——RUN2_F	0.220141	0.181818	0.199153	0.094
Suda_SAM_OMS——RUN3_F	0.220141	0.181818	0.199153	0.094
PRIS_COAE_t2_1_F	0.121248	0.398453	0.185921	0.12
PRIS_COAE_t2_2_F	0.091438	0.334623	0.143628	0.054
buptailab_F	0.09571	0.336557	0.149036	0.13
8-NJUST_F	0.040834	0.090909	0.056355	0.047
CQUPT_F*	0.076193	0.528046	0.133171	0.131
THUHUANG_F	0.108069	0.435203	0.173144	0.112
SDNU_F	0.076735	0.485493	0.132524	0.128
hut_F*	0.070588	0.510638	0.124031	0.084
MI&Tlab_F	0.019139	0.015474	0.017112	0.008
HITSZ_F	0.336722	0.512573	0.406442	0.265
hdu_F	0.054286	0.647969	0.100179	0.085
SXU-3_F	0.092314	0.439072	0.152554	0.108
PKUICL_F	0.122947	0.535783	0.2	0.16
Average	0.1200576	0.3483559	0.1591419	0.10825
Best	0.336722	0.647969	0.406442	0.265

表 14 三个领域观点句识别的宏平均和微平均结果(根据宏平均 F1 排序)

Tab.14 The **Macro- and Micro-**Evaluation Results for Three Domains of Opinion Sentences Identification

参评系统标识	宏平均				微平均			
	Precision	Recall	F1	P@1000	Precision	Recall	F1	P@1000
HITSZ	0.5759	0.6055	0.5834	0.5600	0.7031	0.6719	0.6872	0.5600
Suda_SAM_OMS——RUN1	0.4278	0.4706	0.4444	0.4430	0.5715	0.5759	0.5737	0.4430
SXU-3	0.3300	0.7444	0.4346	0.3853	0.4224	0.7844	0.5491	0.3853
Suda_SAM_OMS——RUN3	0.4340	0.4190	0.4251	0.4163	0.5881	0.5651	0.5763	0.4163
THUHUANG	0.3328	0.6480	0.4205	0.3500	0.4295	0.6937	0.5306	0.3500
Suda_SAM_OMS——RUN2	0.4468	0.3942	0.4185	0.4210	0.6052	0.5505	0.5766	0.4210
buptailab	0.3453	0.6319	0.4138	0.4113	0.4245	0.6165	0.5028	0.4113
ISCAS-Opinion	0.3198	0.5384	0.3890	0.4457	0.4101	0.5792	0.4802	0.4457
PKUICL	0.3030	0.6383	0.3868	0.3960	0.3776	0.6203	0.4695	0.3960
hitwi	0.4326	0.3423	0.3820	0.3777	0.5499	0.4341	0.4852	0.3777
DUTIR	0.3022	0.5742	0.3766	0.4117	0.3903	0.6126	0.4768	0.4117
PRIS_COAE_t2_2	0.3044	0.5498	0.3735	0.3360	0.3866	0.5631	0.4584	0.3360
PRIS_COAE_t2_1	0.3062	0.5183	0.3665	0.3483	0.3909	0.5334	0.4511	0.3483
hut	0.2541	0.7332	0.3549	0.3370	0.3140	0.7171	0.4367	0.3370
SDNU	0.2407	0.6954	0.3427	0.3310	0.3101	0.7382	0.4367	0.3310
hdu	0.2080	0.9535	0.3216	0.3480	0.2548	0.9394	0.4008	0.3480
CQUPT	0.2500	0.5102	0.3005	0.3390	0.3010	0.4923	0.3736	0.3390
MI&Tlab	0.2394	0.3166	0.2645	0.2573	0.3653	0.4867	0.4174	0.2573
8-NJUST	0.0999	0.1532	0.1153	0.1073	0.1209	0.1447	0.1317	0.1073
SoochowUniversity-1	0.0490	0.1661	0.0733	0.0703	0.0567	0.1555	0.0831	0.0703
Average	**0.3101**	**0.5302**	**0.3594**	**0.3546**	**0.3986**	**0.5737**	**0.4549**	**0.3546**
Best	**0.5759**	**0.9535**	**0.5834**	**0.5600**	**0.7031**	**0.9394**	**0.6872**	**0.5600**

表 15 三个领域的观点句识别及极性判别的宏平均和微平均结果(根据宏平均 F1 排序)

Tab.15 The **Macro- and Micro-**Evaluation Results for Three Domains of Opinion Sentences Classification

参评系统标识	宏平均				微平均			
	Precision	Recall	F1	P@1000	Precision	Recall	F1	P@1000
HITSZ	0.5347	0.7234	0.5414	0.5320	0.4945	0.6544	0.6254	0.6396
Suda_SAM_OMS——RUN1	0.3706	0.4054	0.3840	0.4000	0.3652	0.5048	0.5089	0.5068
Suda_SAM_OMS——RUN3	0.3746	0.3622	0.3678	0.3777	0.3549	0.5200	0.4999	0.5097
Suda_SAM_OMS——RUN2	0.3843	0.3401	0.3605	0.3800	0.3401	0.5326	0.4846	0.5075
THUHUANG	0.2681	0.5062	0.3355	0.2907	0.2776	0.3567	0.5761	0.4406
ISCAS-Opinion	0.2679	0.4359	0.3221	0.3777	0.3113	0.3624	0.5119	0.4244
SXU-3	0.2359	0.5165	0.3078	0.2957	0.2675	0.3099	0.5756	0.4029
DUTIR	0.2435	0.4554	0.3019	0.3417	0.2791	0.3185	0.4999	0.3891
PRIS_COAE_t2_1	0.2498	0.4107	0.2961	0.2920	0.2616	0.3274	0.4468	0.3779
hitwi	0.3336	0.2632	0.2941	0.3250	0.2632	0.4417	0.3487	0.3897
PKUICL	0.2297	0.4761	0.2913	0.3203	0.2752	0.2916	0.4790	0.3625
SDNU	0.1981	0.5514	0.2798	0.2943	0.2734	0.2624	0.6248	0.3696
hut	0.1987	0.5443	0.2751	0.2637	0.2318	0.2556	0.5837	0.3555
CQUPT	0.2165	0.4281	0.2580	0.3077	0.2559	0.2653	0.4338	0.3292
hdu	0.1679	0.7234	0.2573	0.3120	0.2623	0.2103	0.7754	0.3309
PRIS_COAE_t2_2	0.2071	0.3570	0.2504	0.2040	0.1947	0.2741	0.3992	0.3250
buptailab	0.1866	0.3327	0.2212	0.3060	0.2447	0.2322	0.3372	0.2750
8-NJUST	0.0815	0.1203	0.0924	0.0880	0.0772	0.1028	0.1230	0.1120
SoochowUniversity-1	0.0406	0.1327	0.0604	0.0647	0.0534	0.0475	0.1302	0.0696
MI&Tlab	0.0268	0.0358	0.0295	0.0307	0.0254	0.0350	0.0466	0.0400
Average	**0.2408**	**0.4060**	**0.2763**	**0.2902**	**0.2554**	**0.3153**	**0.4505**	**0.3579**
Best	**0.5347**	**0.7234**	**0.5414**	**0.5320**	**0.4945**	**0.6544**	**0.7754**	**0.6396**

5.3 任务 3 评测结果

任务 3 共有 12 支队伍提交了 14 个 run，最终评测结果如下。

表 16 评价对象抽取结果（根据宏平均 F1 排序）

Tab.16 Evaluation Results of Opinion Target Extraction

参评系统标识	宏平均			微平均		
	Precision	Recall	F1	**Precision**	**Recall**	**F1**
Suda_SAM_oms	0.104141978	0.081763	0.091606	0.140579845	0.149070779	0.1447009
MI&Tlab	0.122980583	0.051811	0.072907	0.120942075	0.131474891	0.1259887

thuhuang(run2)	0.080730769	0.061293	0.069682	0.121899273	0.096184263	0.1075257
thuhuang(run3)	0.086527502	0.058184	0.06958	0.135611401	0.087485172	0.1063574
thuhuang(run1)	0.087982554	0.057156	0.069296	0.137322293	0.086892052	0.1064358
HIT-SCIR	0.084461117	0.052887	0.065045	0.122505725	0.074041123	0.0922982
buptailab	0.133932716	0.03756	0.058668	0.159847036	0.04132068	0.0656665
DUTIR	0.06639795	0.045311	0.053864	0.104914079	0.068801898	0.0831045
CQUPT	0.052153434	0.054176	0.053146	0.065183537	0.086892052	0.0744884
PRIS_COAE_t3	0.050105951	0.053378	0.05169	0.066396527	0.113384737	0.0837501
hdu	0.064799429	0.029555	0.040595	0.115865283	0.054072756	0.0737346
SXU-3	0.075496326	0.020733	0.032532	0.0876444	0.034499802	0.0495106
SDNU	0.017597867	0.018828	0.018192	0.021880387	0.03262159	0.0261926
PKUICL	0.012676275	0.014414	0.013489	0.016713454	0.088869118	0.0281355
Average	**0.074284604**	**0.045504**	**0.054307**	**0.101236094**	**0.081829351**	**0.0834207**
Best	**0.133932716**	**0.081763**	**0.091606**	**0.159847036**	**0.149070779**	**0.1447009**

表 17 评价短语抽取结果

Tab.17 Evaluation Results of Opinion Phase Extraction

参评系统标识	宏平均			微平均		
	Precision	Recall	F1	**Precision**	**Recall**	**F1**
Suda_SAM_oms	0.07695003	0.100699	0.087237	0.099561853	0.105575326	0.1024804
DUTIR	0.07733472	0.055652	0.064726	0.117425384	0.077006722	0.0930149
PRIS_COAE_t3	0.045453706	0.095188	0.061527	0.061765557	0.105476473	0.0779088
thuhuang(run3)	0.07379251	0.049768	0.059445	0.106650322	0.068801898	0.0836438
thuhuang(run2)	0.063780425	0.05395	0.058455	0.094086695	0.07423883	0.0829926
thuhuang(run1)	0.069691025	0.045589	0.05512	0.106077175	0.067121392	0.0822183
CQUPT	0.042140968	0.073171	0.053481	0.054653319	0.072854883	0.062455
SXU-3	0.085672109	0.031919	0.04651	0.100954294	0.039739027	0.0570294
HIT-SCIR	0.053941268	0.034862	0.042352	0.086849853	0.052491103	0.0654344
MI&Tlab	0.071534321	0.029599	0.041873	0.051468582	0.055950969	0.0536163
Buptailab	0.05405951	0.016032	0.02473	0.066156788	0.017101621	0.0271778
Hdu	0.038452969	0.017791	0.024327	0.062486761	0.029161724	0.0397655
PKUICL	0.011887376	0.080848	0.020727	0.015542211	0.08264136	0.0261638
SDNU	0.010677485	0.023185	0.014621	0.012531494	0.018683274	0.0150012
Average	**0.055383459**	**0.05059**	**0.046795**	**0.074015021**	**0.061917472**	**0.0620644**
Best	**0.085672109**	**0.100699**	**0.087237**	**0.117425384**	**0.105575326**	**0.1024804**

表 18 评价对象和评价短语抽取以及极性判别结果

Tab.18 Evaluation Results of both Opinion Combination Extraction and Polarity Analysis

参评系统标识	宏平均			微平均		
	Precision	Recall	F1	Precision	Recall	F1
Suda_SAM_oms	0.032301645	0.037671	0.03478	0.044187564	0.046856465	0.0454829
DUTIR	0.038508614	0.026116	0.031124	0.021817714	0.018677942	0.0201261
thuhuang(run2)	0.033214595	0.027482	0.030078	0.053620646	0.042309213	0.047298
thuhuang(run3)	0.037530482	0.025044	0.030042	0.061599755	0.039739027	0.0483115
thuhuang(run1)	0.035265037	0.022829	0.027716	0.060459303	0.038256228	0.0468608
HIT-SCIR	0.026954522	0.016649	0.020584	0.045632974	0.027580071	0.0343808
CQUPT	0.016075468	0.027016	0.020157	0.021505376	0.028667457	0.0245752
PRIS_COAE_t3	0.015169519	0.028815	0.019876	0.022575977	0.038552788	0.0284765
SXU-3	0.039638941	0.010363	0.016431	0.047212456	0.018584421	0.0266704
buptailab	0.031490539	0.008952	0.013941	0.035181644	0.009094504	0.0144529
MI&Tlab	0.02340895	0.008164	0.012106	0.012457943	0.013542902	0.0129778
PKUICL	0.005360044	0.034717	0.009286	0.008087155	0.043001186	0.013614
hdu	0.013044018	0.005528	0.007765	0.025206524	0.011763543	0.016041
SDNU	0.002700839	0.006293	0.00378	0.002718472	0.004052985	0.0032542
Average	**0.025047372**	**0.020403**	**0.019833**	**0.033018822**	**0.027191338**	**0.027323**
Best	**0.039638941**	**0.037671**	**0.03478**	**0.061599755**	**0.046856465**	**0.0483115**

表 19 三个领域的抽取结果对比

Tab.19 Comparison between Results of Three Domains

内容	项目	领域	Precision	Recall	F1
评价对象抽取	average	电子产品	0.145336	0.083806	0.0965769
	best	电子产品	0.22267	0.160298	0.1864046
	average	影视娱乐	0.050095	0.032552	0.0398785
	best	影视娱乐	0.202899	0.053385	0.0774681
	average	金融证券	0.027811	0.02381	0.0282651
	best	金融证券	0.090226	0.044547	0.0449612
评价短语抽取	average	电子产品	0.0972	0.076239	0.0772441
	best	电子产品	0.218182	0.112751	0.1256875
	average	影视娱乐	0.033837	0.027344	0.0351168
	best	影视娱乐	0.130435	0.095703	0.0639548
	average	金融证券	0.0253	0.041475	0.0279531

	best	金融证券	0.043818	0.121352	0.0490609
评价对象和评价短语抽取以及极性判别	average	电子产品	0.047714	0.038277	0.0371933
	best	电子产品	0.108485	0.051457	0.0626223
	average	影视娱乐	0.014107	0.010742	0.0117333
	best	影视娱乐	0.043478	0.029297	0.0198394
	average	金融证券	0.009498	0.014593	0.0093718
	best	金融证券	0.022556	0.044547	0.0170543

5.4 任务 4 评测结果

任务 4 共有 5 支队伍提交了 9 个 run。

评测采用Pooling方式。经过统计，各家单位提交的结果中文档数量基本都在100以上，因此取每个提交结果的前 100 个组成标注集合，经过裁判员人工评判后作为标准答案，然后对各提交结果进行评测打分。分别对主题相关性检索和观点文档检索与极性识别两个子任务进行了评测。其中主题相关性检索子任务的评测结果如表 20 所示，评测指标为MAP_REL。

表 20 主题相关性检索的评测结果

Tab.20 The Evaluation Results for Topic Relevance Retrieval

参评系统标识	未插值 MAP_REL	只对返回的相关文档计算 MAP_REL
PRIS_COAE_t4_2	0.148384	0.940712
PRIS_COAE_t4_1	0.13599	0.919228
Suda_SAM_OMS	0.133931	0.878175
DUTIR1	0.129834	0.957106
PKUICL3	0.0836135	0.982308
PKUICL1	0.0790097	0.988516
DUTIR2	0.078913	0.823916
CQUPT	0.0666762	0.50059
PKUICL2	0.0618901	0.999816
Average	**0.102027**	**0.887819**
Best	**0.148384**	**0.999816**

观点文档检索与极性识别子任务的评测结果如表 21 所示，评测指标为 MAP_SENTI和宏平均的准确率、召回率、F1、Raccuracy100以及P@10，主要指标为未插值的MAP_SENTI。

表 21 观点文档检索与极性识别的评测结果

Tab.21 The Evaluation Results for Opinion Retrieval and Classification

参评系统标识	未插值 MAP_SENTI	只对返回的相关文档 MAP_SENTI	MACRO_P	MACRO_R	MACRO_F1	Raccu-racy	P@10
PKUICL3	0.0724	0.6574	0.5621	0.0305	0.0571	0.1150	0.9150
PRIS_COAE_t4_2	0.0698	0.4306	0.3524	0.0406	0.0708	0.1431	0.9200
PKUICL1	0.0686	0.6451	0.5262	0.0283	0.0531	0.1092	0.9250
PRIS_COAE_t4_1	0.0646	0.4043	0.2874	0.0382	0.0635	0.1304	0.9100
DUTIR1	0.0593	0.5043	0.3123	0.0339	0.0589	0.1242	0.9150
Suda_SAM_OMS	0.0581	0.4548	0.1838	0.0388	0.0538	0.1689	0.8600
DUTIR2	0.0462	0.4621	0.2575	0.0287	0.0493	0.1010	0.7800
PKUICL2	0.0439	0.6024	0.5359	0.0212	0.0399	0.0711	0.8100
CQUPT	0.0244	0.2009	0.0881	0.0277	0.0336	0.1535	0.5000
Average	**0.0564**	**0.4846**	**0.3451**	**0.0320**	**0.0533**	**0.1240**	**0.8372**
Best	**0.0724**	**0.6574**	**0.5621**	**0.0406**	**0.0708**	**0.1689**	**0.9250**

6 评测分析和讨论

本届评测采用了全新的评测数据，跟 2009 年的第二届评测相比，区别主要有以下几点：

- 第二届评测不区分领域，本届评测的任务 1、任务 2 和任务 3 区分领域，即需要在三个领域内分别识别观点词和观点句。
- 第二届评测将倾向性（包括词、句子、篇章）区分为观点和情感，其中观点界定为对其他对象的评价，情感界定为内心自我情感（比如“心情”），两类倾向性均列入考察范围；本届评测的倾向性则只限于观点（即对其他对象的评价），而不包括情感（内心自我情感）。
- 第二届评测只标注了部分答案，所有任务的评测均采用 pooling 方式，属于不完全评测；本届评测的任务 2 和 3 标注了全部答案，属于完全评测，任务 1 和任务 4 仍然采用 pooling 方式评测。

必须指出，任何评测都难以做到尽善尽美，特别是对于标注标准的量化和标注尺度的争议，贯穿了评测的整个过程，评测指标的选择也是一个难题。下面将一些评测中遇到的问题列出来供大家讨论和参考。

关于观点依赖性，是指倾向词具有领域依赖性、上下文语境依赖性或评价对象依赖性，也就是说，倾向词的极性不仅与领域相关，而且与倾向词所在的篇章、句子语境以及评价的对象相关，可能存在同一个倾向词在同一领域内、不同文章中、或者不同语境中表现出不同的褒贬极性。同一个词，由于领域变化、上下文语境变化或者评价对象变化导致的极性发生改变，都视为不同的倾向词。但每个倾向词每种极性只计算一次，领域、上下文或评价对象不同而极性相同时则认为是重复的结果。

关于观点句重复问题，评测时是以观点句为单位来精确去重的，完全相同的句子在评测中只计算一次，这样，其中的评价搭配自然也就实现了一定的去重；但是并不直接以评价搭配为单位来进行完全去重，不同的句子中可以抽出相同的评价搭配，均视为正确结果。此外，由于数据噪声导致部分只相差几个字节的句子被视为不同的句子，允许同时提交，其实这些句子应该作为重复句去掉的。

关于金融证券数据的倾向性，大家争议较多。标注人员的原则是，在表达一个事实或现象时，标注为无倾向性，比如“昨日股市涨了 200 点但今天跌了 300 点”、“某股票领涨、某股票跌穿”等，但也有老师认为其中包含一定的倾向性，这里面的细微区别很难分得清。同时，为了在一定程度上消除模糊和歧义，采用了 2 个评判员同时进行人工标注的方法，如遇到分歧再增加第三个评判员进行投票。这种依赖于人的主观性评判标准很难量化表达，而且肯定存在一定的偏差，但是至少可以保证所有参评单位基于一个公平的标注结果进行对比。这里列举两个有一定代表性的标注例子：

例句 1：韩国 3 月消费者信心指数为 2009 年 7 月以来最低，韩国央行周四公布的数据显示，韩国 3 月消费者信心指数连续第二个月下跌，至八个月低位，令对经济复苏放缓的担忧加重。

该句子判为有极性，标注为：-1（贬义）

例句 2：至收盘，沪指报 2587.81 点，跌 0.27%，成交 789.8 亿元；深成指报 9991.40 点，跌 0.12%，成交 629.4 亿元。

该句子判为无极性。

关于评价对象（Opinion Object）和评价短语（Opinion Phrase）的评测粒度问题，标注的时候 Opinion Object 和 Opinion Phrase 均取最大字串，评测时采用了严格尺度，只有跟标注答案完全一致的结果才被判为正确，如“LED 显示效果非常清晰”的正确结果为“LED 显示结果　非常清晰　+”，其他结果均判为错误，比如“LED　清晰　+ ”、“LED　非常清晰　+”、“LED 显示　清晰　+”等等。这样的尺度虽然减轻了标注和评测的工作量，但由于过于严格也漏掉了一些近似的正确结果。

7　结论与展望

本文对 COAE2011 评测的总体情况进行了详细的介绍和说明，描述了任务设置、评测数据集、评测方法和评测结果，并对评测中遇到的一些问题进行了分析讨论。

从评测结果来看，发现一个有意思的现象，在观点词识别任务中，金融证券领域的准确率跟电子产品和影视娱乐领域差不多；而在观点句识别和评价搭配抽取任务中，情况则不同，金融证券领域的准确率、F1 值均显著低于电子产品和影视娱乐领域。这应该意味着在句子和篇章的大粒度上，金融证券领域的情感倾向不如电子产品和影视娱乐领域强，电子产品领域的倾向性最强，影视娱乐领域次之。

评价搭配的抽取仍然是一个很困难的问题，这一方面是由于评测的尺度比较严格，另一方面也确实反映出自然语言表达的灵活性、复杂性和不确定性。

观点相关性检索任务中，大家普遍重视了检出结果的正确性，而对结果的全面性有所忽视，导致准确率较高、召回率很低。与此同时，大家也更重视首页命中率（P@10），这

一点跟传统搜索引擎追求的目标是一致的。

同样的，标注标准的量化、标注尺度的争议、裁判员的主观偏差等仍然不可避免，需要我们不断完善和提高。

8 致谢

本次评测的成功举行离不开大家的热情支持和无私帮助。

感谢中文信息学会各位老师和信息检索专委会程学旗研究员、马少平教授、刘挺教授、孙乐研究员、姚天昉教授和刘悦副研究员的指导和协调；感谢山东大学计算机科学与技术学院的马军教授、陈竹敏老师和 CCIR2011 组委会对于评测的大力支持，你们的悉心安排帮助评测免去了繁琐的组织程序的困扰；感谢第三届中文倾向性评测委员会各位老师特别是林鸿飞、王厚峰、秦兵、徐睿峰、关毅、王小捷、徐蔚然、王素格、徐冰等诸位老师的热情响应和积极参与；感谢清华大学黄民烈老师给评测提出的宝贵意见和建议；感谢徐睿峰、王仲卿、吴云芳、王素格、李素建、周延泉等老师在评测中的讨论和意见。

特别感谢福州大学廖祥文老师和他带领的 20 多人的评测团队，以及中科院计算所网络数据科学与技术中心的朱亮、李赫元、林祥辉等同学在大纲制定、语料标注、结果评测和问题答疑过程中的艰苦工作和无私贡献，他们是本次评测能够顺利完成的重要保证。

最后感谢所有支持本次评测的朋友。

参 考 文 献

[1] Ellen M. Voorhees, Overview of TREC 2010，In Proceedings of tThe Nineteenth Text REtrieval Conference Proceedings (TREC 2010). NIST, 2010.

[2] Iadh Ounis, Craig Macdonald, Ian Soboroff. Overview of the TREC2010 Blog Track. In Proceedings of tThe Nineteenth Text REtrieval Conference Proceedings (TREC 2010). NIST, 2010.

[3] Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen and Noriko Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis, In Proceedings of the Eighth NTCIR Workshop, 2010

[4] 许洪波，姚天昉，黄萱菁，唐慧丰等，第二届中文倾向性分析评测技术报告，第二届中文倾向性分析评测会议（COAE2009），上海，2009。

[5] 赵军、许洪波、黄萱菁等，中文倾向性分析评测技术报告，第一届中文倾向性分析评测会议（COAE2008），北京，2008。

Suda_SAM_OMS 情感倾向性分析技术报告*

王中卿，王荣洋，庞磊，代大明，居胜峰，苏艳，李寿山
苏州大学 自然语言处理实验室，苏州，215006
E-mail: wangzq870305@gmail.com

摘　要：文本情感倾向性分析已经成为自然语言处理中的一个热点问题。本文介绍了 Suda_SAM_OMS 系统在 COAE-2011 评测中针对每一项任务使用的方法。具体的任务包括：领域观点词的抽取与极性判别、中文观点句抽取、评价搭配抽取以及面向特定对象的中文文本观点检索。结果表明，本系统在 COAE-2011 语料上测试这四项任务取得了较好的成绩。特别是在评价搭配抽取任务中，本系统得到了非常好的抽取效果。

关键词：情感分析；观点抽取；观点检索

Technical Report on Suda_SAM_OMS Sentiment Analysis System

Wang Zhongqing, Wang Rongyang, Pang Lei, Dai Daming, Ju Shengfeng, Su Yan, Li Shoushan
Natural Language Processing Lab, Soochow University, Suzhou 215006
E-mail: wangzq870305@gmail.com

Abstract: Recently, sentiment analysis of text has become a hotspot in the research area of natural language processing. In this technical report, we introduce the system named Suda_SAM_OMS, which is designed for joining COAE-2011. This report includes the methods towards the four tasks in sentiment analysis: domain-dependent word extraction and polarity identified, Chinese opinion sentence extraction and polarity, opinion relation extraction and Chinese opinion retrieval. Our system can achieve good performance across the four tasks. Especially, our system can achieve a very high performance on the task of opinion relation extraction.

Keywords: Sentiment Analysis; Opinion Extraction; Opinion Retrieval

1 简介

COAE2011 中文倾向性分析评测共包含 4 个任务。我们的系统 Suda_SAM_OMS（Suda Sentiment Analysis teaM Opinion Mining System）参加了全部任务并提交了结果。这些任务包括（1）要素级的领域观点词的抽取与极性判别，该任务需要考虑领域对观点词倾向性的影响，识别给定的领域的观点词，并判断极性；（2）句子级的中文观点句抽取，主要目的是分析给定数据集中的观点句，并判别极性；（3）要素级的评价搭配抽取，主要需要考虑上下文语境对词语倾向性的影响，抽取评价对象、评价短语并判别评价极性；（4）最后一个任务是篇章级的面向特定对象的中文文本观点检索。

Suda_SAM_OMS 针对不同粒度的情感分析任务分别设计了系统。首先，对评测语料进行预处理，包括分词和词性标注，针对不同任务，我们分别设计了不同的模型：任务 1 为领域观点词的抽取与极性判别，我们从网络上收集了多份情感词典并进行人工校对，获得了一个大规模的情感词典，以此为基础我们利用上下文关系在语料中选择领域相关的观点

*本文承国家自然科学基金 61003155 及模式识别国家重点实验室开放课题基金项目资助

词，并标注极性。任务 2 为中文观点句抽取，我们通过结合情感词典信息识别观点句并判别极性。任务 3 为评价搭配抽取，由于没有充足的训练语料，我们通过条件随机场模型和规则相结合的方法识别评价对象，在识别出评价对象后找出距离评价对象最近的情感词作为它的评价词语，并判定它的极性。任务 4 为中文文本观点检索，我们借助 Lemur 检索系统进行 ad-hoc 检索，并结合情感词典对检索文本进行重排序。

2 系统描述

针对不同粒度的中文情感倾向性分析任务，本系统分别使用了基于机器学习的方法和基于规则的方法完成。

2.1 领域观点词的抽取与极性判别

为了抽取领域观点词以及判别观点词的极性，我们首先网络上收集了多份情感词典并进行人工校对，获得了一个大规模的没有歧义的领域无关的情感词集，更近一步的，我们上下文关系在语料中选择领域相关的观点词。具体流程如下：

为了收集领域无关情感词集，我们收集了 Hownet 情感词集[1]，NTU 情感词集[2]等多部情感词典，并进行人工筛选，将一些有歧义的词就从字典中剔除。

在获得领域无关情感词集之后，我们主要通过一些上下文关系在语料中选择领域相关的观点词。具体的，我们考虑如果是动词或形容词，则这个词为情感词的可能性较高。其次，共现的情感词数量，如果共现的情感词较多，则这个词是情感的的可能性也较高。对于出现在情感词典中的词，则是情感词的可能性也相对较高。

对于前面三个因素进行综合考虑就得到一个打分，根据该打分进行排序。

为了获得每个观点词的极性，我们考虑如果词出现在情感词典中，则可以参考词典中的极性，给予一定的权值。其次，根据情感词共现的词的情感分类是哪一类的，如果一个情感词基本和正极情感词共现，则这个词为正极的可能性比较高，所以以同现词的极性来判断词的极性是一个因素。综合以上两种因素，按各自的得分，最终获得极性。

2.2 观点句抽取

2.2.1 观点句的抽取

由于情感文本单元表现格式比较自由，区分主、客观文本单元的特征并不明显，在很多情况下，情感文本的主客观识别比主观文本的情感分类更有难度。所以，我们在中文情感词褒贬性判断的基础上，使用情感词典中的词语对中文文本进行主客观的分类。我们对任务 1 中获得的情感词典进行进一步的分析，将每个情感词分为极性强的情感词和极性弱的情感词两类。本文假设，只要包含一个强情感词的文本或者包含两个以上弱情感词的文本就是主观文本。

2.2.2 观点句的极性判别

在文本主客观分析时，根据文本中包含的情感词个数及其极性对文本进行打分。文本的主观分数的计算方法可以分为两种，一是采用判别函数：根据句子中情感词的个数判别句子的主客观性。另一种方法则是采用第三节中计算出的情感词分数，将文本中所含情感词

[1] http://www.keenage.com
[2] http://nlg18.csie.ntu.edu.tw:8080/opinion/

的分数之和作为句子的主观分数。本文采用第一种方法，计算文本极性分数的公式如下所示：

$$Score(T) = \alpha Max(V(word_1), V(Word_2)...)$$

其中 $Score(T)$ 为文本T的情感极性分数，α 判别函数系数（当 $len(T)/N < 20$ 时，$\alpha = 1$；当 $20 \le len(T)/N < 40$ 时，$\alpha = 0.9$；当 $len(T)/N \ge 40$ 时，$\alpha = 0.8$；$len(T)$ 为文本长度，N为文本中情感词个数），$V(word_i)$ 为第i个情感词极性强度权值。

在实验过程中，由于情感词前面出现否定词会使极性发生逆转，所以我们还需要从抽取的句子中将情感词前面含有否定形式的句子极性归到相反的类别中去。例如：“姜文的电影果然没有让我们失望。”在第一步处理中由于“失望”是负面情感词，它前面出现了“没有”，所以要将这句话重新归入到正面评论里面。经过上面的处理，我们完成了对文本否定形式的处理。

在主观文本褒贬分类中，还有一类是含有混合极性的文本，由于本届中任务2和往届不同，本届任务2是基于句子级别的，所以混合极性的文本相对较少，我们仅通过情感词典，对同时含有正、负面情感词的文本进行抽取，以获取含有混合极性的文本。

2.3 评价搭配抽取

评价对象（Opinion Target）是指某段评论中所讨论的主题，具体表现为评论文本中评价词语所修饰的对象。评价对象抽取是情感信息抽取任务重要的研究课题之一。而且这项研究的开展有助于为上层情感分析任务提供服务。评价对象抽取任务可以被建模成信息抽取。由于条件随机场序列标注模型能较好地捕捉上下文信息，它已经被成功应用于多个任务中，在评价对象抽取中也得到了很好的应用。

本节基于条件随机场模型和规则相结合的方法识别评价对象，CRF（Conditional Random Field）是一个序列标注模型，近年来，随着 CRFs 在中文分词、词性标注、命名实体识别等自然语言处理任务取得的进展，特别是在情感分析领域主客观分类、观点持有者识别等任务上的成功应用。评价对象抽取与命名实体识别有着相似的地方，它们都是针对某类特定的名词进行识别。

任务要求不仅要提取出评价对象，还需要识别出针对评价对象的情感词语以及情感极性。我们首先识别出评价对象，然后利用情感词典找出距离评价对象最近的情感词作为它的评价词语，并利用一些规则判断它的极性。

2.3.1 基于机器学习识别评价对象

我们采用基于CRFs的评价对象抽取系统，参考文献[2]的工作，加入的特征如下表所示：

表 1 基于 CRFs 识别评价对象中特征的描述

Tab.1 Feature description of target extraction with CRFs

词法特征	词 tk	该特征表示当前单词的字符串特征
	词性 pos	该特征表示当前单词的词性标记特征
依存关系特征	依存关系	布尔型特征，当前词与情感词是否有直接的依存关系，有直接的依存关系记为 1，否则为 0

2.3.2 情感词评价对象关系的识别

由于评价对象和情感词存在着某种联系，因此我们采用基于规则的方法进一步识别情感词和评价对象的关系，步骤如下：

1) 利用任务1构建的情感词典识别出情感词。
2) 把一句话中的主语，宾语，以及和情感词存在依存关系的名词列为候选评价对象，这样做的原因是一句话中的评价对象大部分属于以上三种情况。
3) 对识别出来的每个情感词，寻找最近的候选评价对象，并把它们作为一个<情感词，评价对象>对。
4) 对CRF识别出来的每个评价对象，寻找最近的情感词，并把它们作为一个<情感词，评价对象>对。
5) 最后和CRF的识别结果融合，得到最终的<情感词，评价对象>对。

识别出<情感词，评价对象>对之后，我们再扩充情感词，比如“我非常喜欢这款电脑”，识别出的情感词是“喜欢”，我们把“喜欢”前面的修饰词“非常”添加上，并判断修饰词是不是否定词，比如“不”等，如果不是，则判断它的极性为情感词在词典中的极性，如果前面的修饰词是否定词，则判断它的极性为情感词在词典中的极性取反。

2.4 文本观点检索

任务 4 针对给定对象（topic/target），检索出包含评价该对象的倾向性观点的文章。给定对象可能是人物、商品、组织机构或者概念、事物、事件等。该任务由信息检索与观点识别两部分任务组成，输出针对给定对象的评论性文章并按照观点相关度降序排列。本次评测包含了 20 个主题对象，我们使用 Lemur[3]检索系统进行 ad-hoc 检索，并利用情感词典来判断文本的情感倾向，从而对检索结果排序。

Lemur 是卡耐基梅隆大学（CMU）以及马萨诸塞大学（UMass）联合推出的支持信息检索以及语音模型的工具箱。Lemur 中提供了基本的检索算法，像 TFIDF 和 Okapi。Lemur 检索系统具有强大的结构化查询语言，能够支持构造复杂的查询，我们使用主题词来构造查询。例如，对于 t2 的查询构造如下：

<DOC 2>
人民币 升值
</DOC>

Lemur 检索系统会根据文本与查询的相关度返回文本与对象主题间的相关性评分 relScore。同时我们通过情感词典计算来判别文本的情感极性倾向。由于需要检索出与对象相关且具有观点

[3] http://www.lemurproject.org/

倾向性的文章，所以采用如下模型：

$$P(Q,S/D)=P(S/Q,D)\cdot P(Q/D)$$

并将上式改写成对数形式：

$$\log P(Q,S/D)=(1-\lambda)\log P(S/Q,D)+\lambda\log P(Q/D)$$

$P(Q,S/D)$ 表示由检索到的某个文档 D 与查询对象 Q 相关并是主观性文档的概率，$P(S/Q,D)$ 表示在文档 D 是主观性文档的概率。$P(Q/D)$ 表示某个文档 D 生成该查询对象 Q 的概率。这里 λ 表示对于主题检索以及主观性判别的权重，当 λ 越大，表示最后结果的排序偏向于主题检索结果的排序。在我们提交的结果中，设定 $\lambda=0.7$。通过这个模型，我们有效地把文档的主题检索打分和主观性打分融合在一起。

为了更好的体现出主题检索以及主观性判别的权重，对 $P(S/Q,D)$、$P(Q/D)$ 归一化。归一化过程如下：

$$P(Q,D_i)=\frac{relScore(Q,D_i)}{\sum_{i=1}^{DocumentCount(Q)}relScore(Q,D_i)}$$

$$P(S/Q,D_i)=\frac{P(S/Q,D_i)}{\sum_{i=1}^{DocumentCount(Q)}P(S/Q,D_i)(Q,D_i)}$$

由于对于一个给定的对象主题，主题只有一个，所以想通过查询扩展，得到更多与主题相关的查询词，从而检索到更多更全面的文档。这里采用 PMI-IR 技术在基于 Wikipedia 的基础上对每个主题进行扩展。我们利用 yahoo 搜索引擎来计算每个查询词与其扩展候选词之间 PMI 值。计算方法如下：

$$PMI(Word,Word_{\exp})=\frac{Count(Word,Word_{\exp})}{Count(Word)\cdot Count(Count_{\exp})}$$

这里 $Count(\bullet)$ 表示搜索引擎返回的结果数目；Word 表示原始查询词，$Word_{\exp}$ 表示查询扩展的候选词。通过计算，我们选择 PMI 最高的至多 20 个扩展候选词作为我们的查询扩展关键词。

对于文档主观性的判别，我们采用基于情感词典的方法对文档主观性判别。即 $P(S/Q,D)$：

$$P(S/Q,D)=\sum_{s\in S_{lexcion}}\frac{TermCount(s,Q,D)}{TermCount(Q,D)}$$

$TermCount(s,Q,D)$ 表示情感词典 $S_{lexicon}$ 中的情感词 s 在返回文档 D 中主题与查询相关的句子中出现的次数。$TermCount(Q,D)$ 表示返回文档 D 的长度。

对于观点倾向性的判别，我们使用情感词典的方法判别文档的倾向性。计算方法如下：

$$Score(P/Q,D)=\frac{TermCount(P,Q,D)/positiveLexiconSize}{TermCount(N,Q,D)/NegativeLexiconSize}$$

这里 $Score(S/Q,D)$ 表示在返回文档中观点的倾向性得分，$TermCount(P,Q,D)$ 表示褒义的情感词在文档中出现的数目，$TermCount(N,Q,D)$ 表示表示贬义的情感词在文档中出现的数目。当 $TermCount(P,Q,D)$、$TermCount(N,Q,D)$ 其中存在 0 时，可以轻易知道文档的倾向性，但是

当都不为 0 时，我们设定一个阀值 $\lambda_p = 1.5$ 来判断文档的倾向性，则有

$$Polarity(D) = \begin{cases} Positive & Score(P/Q,D) \geq \lambda_P \\ Neutral & 1/\lambda_p \prec Score(P/Q,D) \prec \lambda_P \\ Negative & Score(P/Q,D) \leq 1/\lambda_P \end{cases}$$

对于情感词典，我们选取我们在任务 1 中构建的情感词典。

3 实验结果和分析

3.1 任务 1 的评测结果及分析

表 2 是我们的系统在任务 1 上的评测结果，从结果上看，我们的结果不是很理想的，主要是由于我们为了获得观点词而选择的上下文关系过于简单，并不能找到一些潜在的观点词，同时对于如何对获得观点词的极性的比较简单。

表 2 任务 1 评测结果

Tab.2 Result of Task 1

类别	Precision@1000	Precision	Recall	F1	Raccuracy
电子产品	0.516	0.321	0.078	0.125	0.078
影视娱乐	0.512	0.349	0.079	0.129	0.079
金融证券	0.525	0.518	0.079	0.137	0.079
Median	0.571	0.343	0.095	0.148	0.095
Best	0.674	0.613	0.119	0.183	0.119

3.2 任务 2 的评测结果及分析

表 3 为我们的系统在任务 2 上的评测结果，在本任务上我们的结果属于中等水平，由于单纯考虑情感词对句子的影响，无法获得最优的观点句，所以在下一步的工作中我们可以考虑加入机器学习的方法。

表 3 任务 2 评测结果

Tab.3 Result of Task 2

	Precision	Recall	F1	P@1000	Raccuracy
电子产品	0.600	0.560	0.579	0.712	0.560
影视娱乐	0.333	0.278	0.303	0.334	0.278
金融证券	0.220	0.182	0.199	0.094	0.182
Median	0.241	0.398	0.276	0.290	0.255
Best	0.730	0.798	0.693	0.800	0.660

3.3 任务 3 的评测结果及分析

表 4 到表 6 是我们的系统在任务 3 上的评测结果，由于结合的机器学习的方法和规则

的方法所以本系统在该任务上表现出较好的性能。

表 4 评价对象识别结果

Tab.4 Result of Taget Extraction

	P@1000	Precision	Recall	F1	Raccuracy
电子产品	0.244	0.223	0.160	0.186	0.160
影视娱乐	0.061	0.056	0.053	0.054	0.053
金融证券	0.028	0.034	0.032	0.033	0.032
All	0.111	0.104	0.082	0.092	0.082
Median	0.066	0.074	0.046	0.054	0.046
Best	0.111	0.134	0.082	0.092	0.082

表 5 评价短语识别结果

Tab.5 Result of Opinion Extraction

	P@1000	Precision	Recall	F1	Raccuracy
电子产品	0.158	0.150	0.108	0.126	0.108
影视娱乐	0.058	0.048	0.096	0.064	0.056
金融证券	0.031	0.033	0.098	0.049	0.035
All	0.082	0.077	0.101	0.087	0.066
Median	0.051	0.055	0.051	0.047	0.035
Best	0.118	0.086	0.101	0.087	0.066

表 6 评价对象，评价短语以及评价短语极性识别结果

Tab.6 Result of Taget Extraction, Opinion Extraion and Polarity Idefine of Opinion

	P@1000	Precision	Recall	F1	Raccuracy
电子产品	0.081	0.071	0.051	0.060	0.051
影视娱乐	0.021	0.015	0.029	0.020	0.015
金融证券	0.009	0.011	0.032	0.016	0.011
All	0.037	0.032	0.038	0.035	0.026
Median	0.027	0.025	0.020	0.020	0.015
Best	0.072	0.040	0.038	0.035	0.026

3.4 任务 4 的评测结果及分析

表 7 为任务 4 的评测结果，从结果来看，由于利用情感信息对检索出得文档结果进行

重排序的方法还是过于简单，所以本系统在该任务上只达到了中等水平。

表 7 任务 4 评测结果

Tab.7 Result of Task 4

	MAP	Precision	Recall	F1	Raccuracy	P@5	P@10
Suda_SAM_OMS	0.455	0.184	0.039	0.054	0.169	0.880	0.860
Median	0.455	0.184	0.039	0.054	0.169	0.880	0.860
Best	0.657	0.562	0.041	0.071	0.169	0.980	0.925

4 总结与工作展望

本报告介绍了我们组（Suda_SAM_OMS）参加的 COAE-2011 的具体方法，我们主要实现了四个方面的任务：（1）要素级的领域观点词的抽取与极性判别；（2）句子级的中文观点句抽取；（3）要素级的评价搭配抽取；（4）篇章级的面向特定对象的中文文本观点检索从评测结果来看，我们的系统在某些任务上能够去的较好的结果，但在一些任务上同最好的结果仍然有一定的差距，说明我们的方法在结构和细节上还需要改进。

参 考 文 献

[1] Jakob N. and Gurevych I. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields. In *Proceedings of EMNLP-2010*.

[2] Li B., Zhou L., Feng S. and Wong K. A Unified Graph Model for Sentence-based Opinion Retrieval. In *Proceedings of ACL-2010*.

[3] Lu B. Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts. In *Proceeding of the NAACL HLT-2010 Student Research Workshop*.

[4] Pang B., Lee L., and Vaithyanathan S. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*-2002.

[5] Pang B. and Lee L.. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts. In *Proceedings of ACL-2004*.

[6] Turney P. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of reviews. In *Proceedings of ACL*-2002.

[7] Zagibalov T. and Carroll J. Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Test. In *Proceedings of COLING-2008*.

[8] 王荣洋,鞠久朋,李寿山,周国栋.基于 CRFs 的评价对象抽取特征研究. In *Proceeding of CCL-2011(Chinese)*.

[9] 黄萱菁,许洪波,赵军. 第一届中文倾向性分析评测论文集. 2008.

DUTIR COAE2011 评测报告*

杨亮，王昊，李雪妮，任巨伟，林鸿飞
大连理工大学计算机科学与技术学院，大连，116024
E-mail: yangliang@mail.dlut.edu.cn

摘　要：中文情感分析的发展离不开相关评测的推动作用。第三届中文倾向性分析评测（COAE2011）包括了领域观点词的抽取与极性判别、中文观点句抽取、评价搭配抽取以及观点检索等四个子任务。本文介绍了大连理工大学信息检索研究室在 COAE2011 中的所用的各种方法及资源。评测结果表明了，大连理工大学信息检索实验室的大连理工大学文本倾向性分析知识库对于中文情感分析工作具所起到的辅助作用，但方法创新性仍需要进一步加强，以获得更加理想的成绩。

关键词：中文情感分析；COAE2011；文本倾向性分析知识库

DUTIR at COAE2011

Yang Liang, Wang Hao, Li Xueni, Ren Juwei, Lin Hongfei
Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024
E-mail: yangliang@mail.dlut.edu.cn

Abstract: Chinese Sentiment Analysis Evaluation is indispensable to the development of Chinese sentiment analysis. There are four subtasks in COAE2011, including opinion word and opinion sentence identification, comment object extraction, and opinion retrieval. In this paper, the methods and resources of DUTIR at COAE2011 were detailed introduced. The evaluation result proved the effectiveness of the sentiment ontology by DUTIR. On the other hand, the method innovation still needs to be improved in order to get a better result.

Keywords: Chinese sentiment analysis; COAE2009; sentiment ontology

1　引言

大连理工大学信息检索研究室（DUTIR）参加了第三届中文倾向性分析评测（COAE2011）的全部四个子任务。本文介绍了 DUTIR 在四个子任务中所用方法及评测结果的分析。

2　领域观点词抽取与极性判别

本文使用大连理工大学信息检索实验室的“DUTIR 文本倾向性分析知识库”中的情感词汇本体库部分[1]，该本体库目前已经包含词汇两万余条，收录的基本知识来源于词典、语义网络、网络用语等。情感该本体库共包含 7 大类，20 个小类情感，包括快乐、安心、尊敬、赞扬、相信、喜爱、愤怒、悲伤、失望、疚、思、慌、恐惧、羞、烦闷、憎恶、贬责、妒忌、怀疑、惊奇。每个词语通过一个三元组来描述：

Lexicon = (B, R, E)

*基金项目：国家自然科学基金资助项目（编号：60673039，60973068）、国家 863 高科技计划资助项目（编号：2006AA01Z151）、教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金资助课题（编号：20090041110002）。作者简介：杨亮，王昊，李雪妮，任巨伟，研究方向为情感计算；林鸿飞，男，博士，教授，博士生导师，研究方向为搜索引擎、文本挖掘和自然语言处理，hflin@dlut.edu.cn。

其中 B 表示词汇的基本信息，主要包括编号，词条，对应英文，词性，录入者和版本信息；R 代表词汇之间的同义关系，即表示该词汇与哪些词汇有同义的关系；E 代表词汇的情感信息，包括情感类别、情感强度、情感极性，是情感词汇描述框架中重要的一部分。

2.1 观点词汇的抽取

观点词汇的抽取过程如图 1 所示。

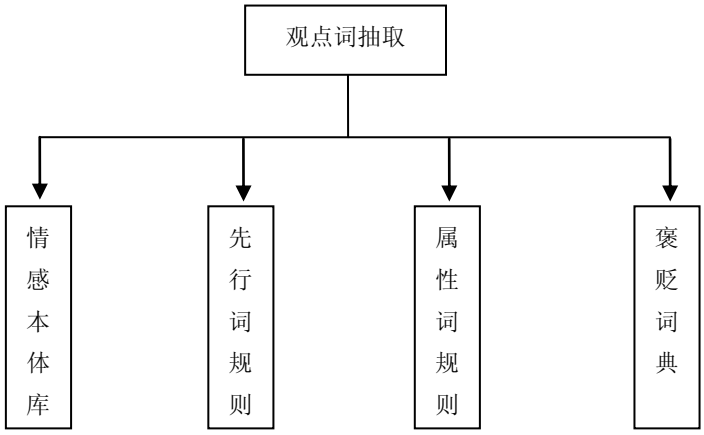

图 1　观点词抽取流程

Fig.1 The Flow of the extraction of opinion words

（1）情感词汇本体库：首先我们将语料集里在情感词汇本体库中的词汇识别出来，过滤，保留情感强度较强的词汇及含有褒贬极性的词汇。就词性而言，由于观点词多集中于形容词、名词和动词之中，而相应副词、连词、介词、代词及量词为观点词的可能性较小[2]。所以，在置信度计算时给予形容词、名词和动词较高的置信度权重。

（2）先行词规则：通过先行词与观点词的搭配规律来定位观点词[2]。这里的先行词主要包括助词、程度副词和否定词，通过与先行词的共现来识别观点词。

（3）属性词规则：在网络中通过爬虫获取相关领域（数码、娱乐、财经）的属性词。之后在该属性词的上下文片段中获取形容词及名词作为观点词。

（4）褒贬词典：利用本组往年参加 COAE2009 任务中对于语料分析获取的褒贬词典，加入到观点词的判别中去。

2.2 不同领域观点词的极性判别

判别流程如图 2 所示。

（1）对于出现在褒贬词典中的词，直接根据褒贬词典赋予褒贬极性。

（2）对于出现在本体中的具有单一褒贬极性的词语，直接根据情感本体赋予极性，并根据情感本体库中情感强度的大小进行置信度设定，情感强度较强的给予较高的置信度。

（3）没有出现在本体库中的词，或出现在本体库中但在不同领域具有多个极性的词，则根据 N 元（本次评测 N=5）共现信息进行二次极性判断。根据与其共现的前后 N 个词语

的极性预测当前观点词的极性。

（4）考虑到部分副词在极性判别中具有指导意义，利用本组往年参加 COAE2009 中对于语料分析获取的类似副词（如太、过于、异常），加入到领域观点词的极性判别中去。

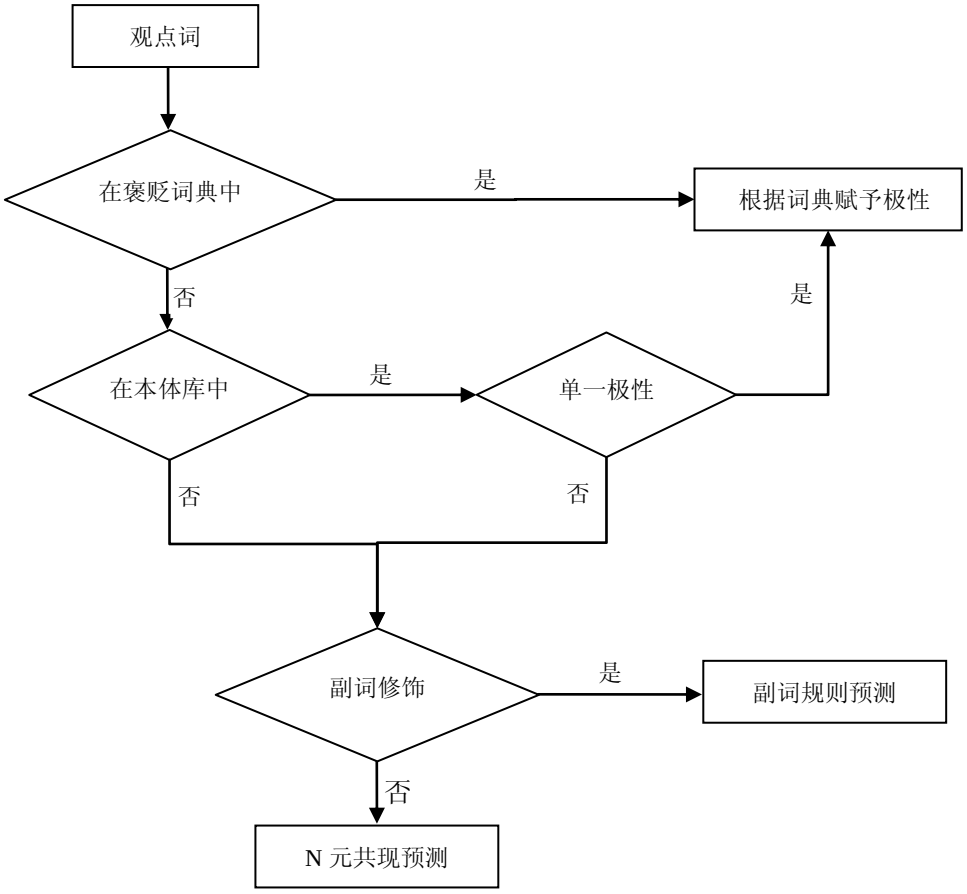

图 2 观点词极性判断流程

Fig.2 The flow of the polarity judgement of opinion words

2.3 结果分析

表 1 任务一实验结果

Tab.1 The results of task 1

标识	类别	Precision@1000	Precision	Recall	F1	Raccuracy
DUTIR	电子产品	0.593	0.364	0.1089	0.1676	0.1089
DUTIR	影视娱乐	0.601	0.388	0.1176	0.1805	0.1176
DUTIR	金融证券	0.632	0.373	0.1135	0.1741	0.1135
Median		0.57126	0.343	0.0947	0.1476	0.0947
Best		0.674	0.6125	0.1194	0.1833	0.1194

本次结果基本达到了预期目的，但实验中还有缺憾，比如观点词中的单字词语没有得到识别，比如“高”、“低”，“贵”等词语，同时情感本体中一些具有情感但不具有观点极性的词以及一些动词和名词的加入也为实验带来了噪音，对任务一的准确率产生影响。

3 中文观点句抽取

由于本任务要求参赛者从每个领域的测试集中自动识别出所有观点句及其表达观点的总体极性（褒义、贬义或混合观点），并没有要求对观点进行综合分析，因此本文在沿用 DUTIR COAE2009 的基础上稍作修改[3]。

3.1 使用资源

DUTIR 文本倾向性分析知识库（主要包括 DUTIR 情感词汇本体，DUTIR 语料库，DUT 常识库，DUTIR 搭配词典）

3.2 观点句识别流程

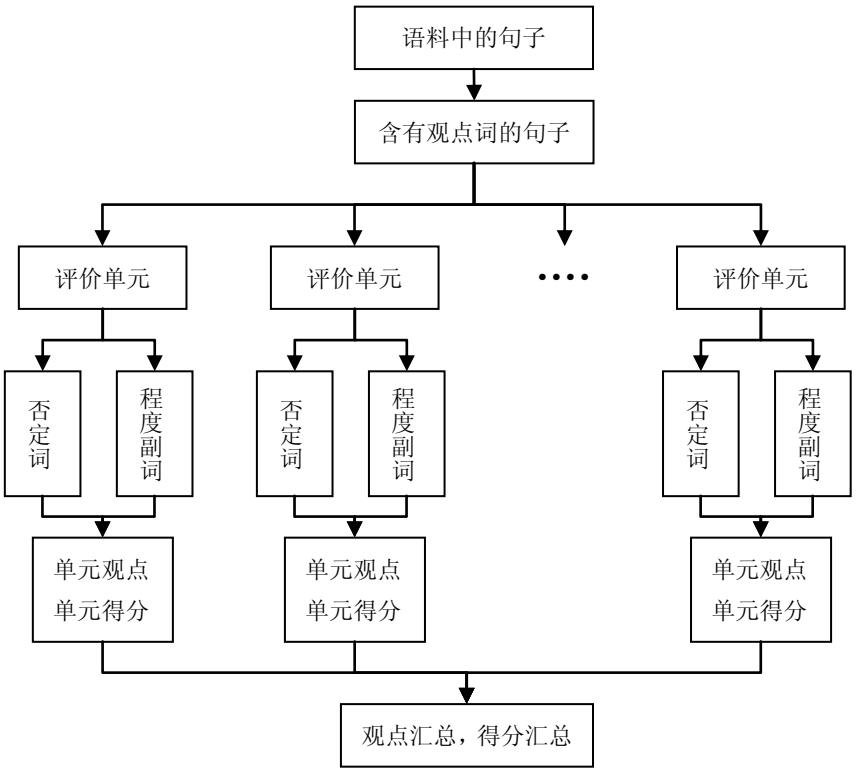

图 3 观点句识别流程图

Fig.3 Opinion sentence detection flow chart

3.3 观点句的识别

由图 3 所示，首先根据 DUTIR 情感词汇本体找到句子中含有的观点词，如果句子中不含有观点词，则该句子没有观点。对于含有观点的句子具体处理过程如下：

（1）以观点词为中心，设置窗口，把窗口内出现的副词与否定词与观点词作为一个评价单元。

（2）根据否定副词和程度副词，计算该单元的极性与强度

（3）将所有评价单元的极性汇总，若是出现混合极性，则为 0，褒义为 1，贬义为-1

（4）将所有评价单元的强度汇总，并考虑长度信息按照下面的公式计算强度得分

$$S = (1/L)\sum \alpha * \beta * s(w) \quad (1)$$

其中，α、β 为否定副词和程度副词的放大倍数，$s(w)$ 为该评价词的强度。

4 评价搭配抽取

本任务要求从语料中抽取评价对象与评价短语，并且要求抽取的结果越具体越好。因此，本文将任务分解为抽取核心搭配与扩展核心搭配两个步骤。

4.1 抽取核心搭配

4.1.1 抽取流程

在本任务中,我们依靠搭配词典，情感词汇本体结合 CRF++工具包进行抽取。抽取流程如图 4 所示。

本文根据“,”“;”tab 等，将目标句子分成子句。对每一个子句，首先采取搭配词典进行匹配，如果匹配成功，则返回，如果失败，则采用出现在以观点词为中心的窗口中的名词与观点词组成搭配，如果匹配成功，则返回。最后，对于没有找到搭配的子句，采用 CRF 的方法进行自动识别。

4.1.2 CRF++抽取

评价词和评价对象的抽取可以转换为序列标注的问题，而对于序列标注问题，常用的算法是利用条件随机域（Confidence Random Field，简称 CRF）[4]。本任务中沿用了 09 中组块（Chunk）分析的思想[3,5]。

本文采用了 CRF++工具包，将规则标注出的结果作为训练集的正例，人工标出等量的不含有搭配的语料作为负例，组成训练集。对语料训练后得到模型。

4.2 核心搭配的扩展

本文根据形容词，副词，助词，否定词等信息对已经抽取的核心搭配进行扩展。同时利用窗口长度限制扩展的总长度。

扩展规则也会对之前抽取的搭配进行检查。由于抽取核心搭配时是以评价词为中心向两边寻找评价对象，因此会带来噪音。本文制定了一些规则过滤这些噪音，比如评价对象

与评价词中间出现“和”或者“并”等表连接或者表转移的词，就将该核心搭配从结果集中删除。

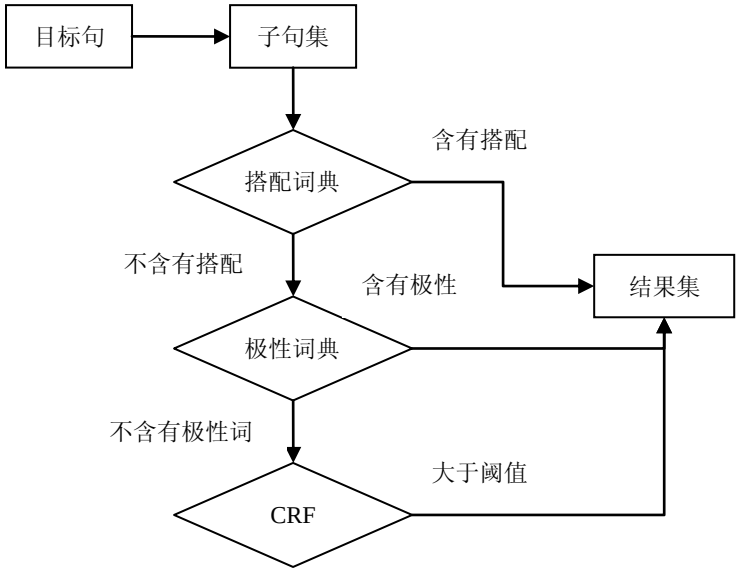

图 4 抽取流程图

Fig.4 Extraction flow chart

4.3 结果分析

表 2 任务三结果

Tab.2 The results of task 3

评价对象、短语、极性都正确						
run-tag	domain	P@1000	Precision	Recall	F1	Raccuracy
DUTIR	D	0.104	0.088581315	0.048429815	0.062622309	0.048429815
DUTIR	E	0.018	0.013513514	0.013020833	0.013262599	0.013020833
DUTIR	F	0.012	0.013431013	0.016897081	0.014965986	0.015360983
宏平均		0.044666667	0.038508614	0.02611591	0.03112402	0.025603877
微平均		0.024888889	0.021817714	0.018677942	0.020126109	0.017995231

评价对象、短语、极性都正确				
	宏平均		微平均	
	Precision@1000	F1	Precision@1000	F1
average	0.027190476	0.019833191	0.025777778	0.027323013
max	0.071666667	0.034780117	0.071666667	0.048311501

在三组结果中，D 领域的结果最好。在标注训练集的样本中，采样的结果是 D 领域中

含有的搭配最多。所以，根据 D 领域的结果最好，以及我们预计 D 领域正确答案最多这两个条件，可以得出的结论此次的评测结果是微平均值大于宏平均值。但是评测结果显示，微平均值远远小于宏平均值，因此，本文得到结论，在参赛时，E 与 F 领域的正确答案据有高比例，而且我们对抽样样本标注在这两个领域标注出现了比较大的偏差。

如图 4 所示，由于本文采用的方法是在构建 CRF 训练集的时，负例通过人工挑选，因此，在标注出现偏差的时候 CRF 值影响了最后的结果。

而宏平均值不错的表现得益于 D 领域的表现，这主要是因为我们的搭配词典比较完善，同时训练集的选取也比较合理。

5 观点检索

5.1 任务简述

本届评测将继续前两届评测的观点检索任务。给定 20 个观点检索对象（topic/target），针对每个对象，把电子、娱乐、财经三个领域的数据集看成一个大数据集，要求找出其中包含评价该对象的倾向性观点的文章。给定对象可能是人物、商品、组织机构或者概念、事物、事件等。该任务是信息检索和观点识别的组合任务，要求输出针对给定对象的评论性文章并按观点相关度降序排列。

5.2 方法简介

此次的观点检索任务是前两届的延续，借鉴于 COAE 2009[3]和 Blog TREC2007[6]中所采用的方法，在 COAE 2011 的中文观点检索任务中也采用了相似流程，如图 5 所示。开始阶段，利用 Lucene 对原始语料集建立索引，进行主题相关性检索；其次，在主题检索的基础上，利用建立的词典文件进行观点相关性检索，并最终得到针对给定对象的评论性文章；最后，则利用 SVM 分类器对观点检索的结果进行分类。具体做法如下。

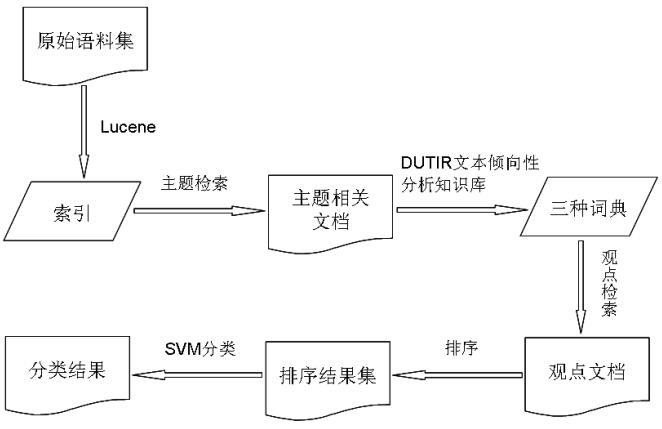

图 5 观点检索处理流程

Fig.5 The flow chart of opinion retrival

在主题检索阶段，本文先对给定的 20 主题词进行适当的扩展，例如，将“iphone”扩

展为“苹果手机”等。之后再利用所有的主题词，对全部语料进行主题检索，并将扩展词的检索结果与原主题词的检索结果进行合并。

在观点检索阶段，主要是通过建立词典，并结合查询窗口的方法来实现观点文档的检索。这里我们一共建立了三种词典——倾向性词典、指示词词典、产品属性词词典。其中倾向性词典来源于 DUTIR 文本倾向性分析知识库中的情感词汇本体，产品属性词词典与任务一种使用的一致，指示词词典则是观察语料所得。为了确保了所捕获的观点是针对特定 topic 的观点，这里使用窗口的方法先选定 topic 的上下文信息，再分别结合三种词典进行检索。

在对所得的观点文档进行排序的阶段，综合考虑了主题检索和观点检索两部分的得分情况。Lucene 在主题检索完成时，会根据相关性高低自动对每篇文档给出一个分值，我们对其进行最大值归一化，并将结果作为主题相关性得分，记为 $Score_{rel}$，在观点检索阶段，先统计每篇文档在指定的窗口范围内，特定 topic 和观点部分同时出现的次数，得出最大值，再利用该最大值对全部次数进行归一化，最后将归一化所得的数值作为该 topic 的观点相关性得分，记为 $Score_{opn}$，由此我们得到以下的排序公式：

$$Score = Score_{rel} + Score_{opn} \quad (2)$$

利用公式（2）对每个 topic 的返回文档进行排序，得到最终的结果文档集。

在观点分类阶段，从结果文档集中选取排序靠前的部分文档作为训练集，并采用 SVM 分类器将结果集分为褒义、贬义、既褒既贬三类。这里按照信息增益的方法选取部分特征，信息增益是指在在保留某个特征和去掉某个特征之后，信息量变化的大小。其公式如下：

$$IG(Y \mid X) = H(Y) - H(Y \mid X) \quad (3)$$

信息增益越大，说明这个特征越重要。此外，我们还将倾向性词典中的词，以及程度副词、否定词等作为特征考虑进去，并利用训练的 SVM 完成分类任务。

5.3 结果分析

本次实验，我们提交了两组结果，run_1 是综合考虑了三种词典的结果，run_2 是只考虑了倾向性词典的结果。

表 3 任务四结果

Tab 3. The results of task 4

	MAP			
	未插值 AP.MAP_REL	未插值 AP.MAP_senti	只对返回的相关文档进行计算 AP.MAP_REL	只对返回的相关文档进行计算 AP.MAP_senti
DUTIR1	0.129834	0.059291	0.957106	0.504276
DUTIR2	0.078913	0.04617	0.823916	0.462087
MEDIAN	0.104374	0.05273	0.890511	0.483182
BEST	0.148384	0.072356	0.999816	0.657434

从 DUTIR1 的实验结果中看出，在只考虑主题相关性时，MAP 的值较好，但是加入情感观点信息后，MAP 值与最好结果的差距明显加大，这说明本文使用的观点分析的方法有

所不足。如上所述，任务四的主要方法是词典结合窗口，所以窗口的大小对 topic 所在的上下文观点分析有很大影响，窗口小，容易遗失有观点的部分，窗口大，则容易引入噪音。同时，词典的方法在潜在语义分析方面存在缺陷，使得类似情况下的观点分析受到影响。

参 考 文 献

[1] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185

[2] 陈建美，林鸿飞，杨志豪. 基于语法的情感词汇自动获取[J].智能系统学报, 2009, 4(2):100-106

[3] 潘凤鸣，王宇轩，常富洋等. DUTIR COAE2009 评测报告[C]. 第二届中文倾向性分析评测

[4] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[A]. In: Proc. 18th International Conf. on Machine Learning[C]. San Francisco: 2001, 282-289.

[5] 宋锐,洪莉,林鸿飞. 基于 ChunkCRF 的观点持有者识别及其在观点摘要中的应用[J].小型微型计算机系统，2009，7:1462-1466

[6] S. Rui, T. Qin, D. Shi, H. Lin, Z. Yang. DUTIR at TREC 2007 Blog rack. [EB/OL]. Proceedings of Text Retrieval Conference 2007, 2007.

PRIS_COAE COAE2011 评测报告

李岩，张佳玥，林宇航，宋洋，徐饶，周燎明，谢乾龙，肖晶，李晓宁，徐蔚然，郭军

北京邮电大学，北京，100876

E-mail: buptly@yahoo.com.cn

摘　要：PRIS_COAE 在第三届中文倾向性分析评测（COAE2011）中参加了全部四项任务。本文分别对每项任务所用的方法进行简要说明。任务一主要通过条件随机场模型（CRFs）结合词激活力模型（WAF）进行观点词的抽取和极性判断；任务二分为基于句法和 CRFs 的观点句抽取、基于情感词个数的观点句抽取；任务三通过句法分析和规则的方法抽取评价搭配；任务四主要利用检索模型进行篇章级的观点检索。文章最后列出四项任务的评测结果。

关键词：句法分析；CRFs；褒义词；贬义词

PRIS_COAE at COAE2011 Track

Yan Li, Jiayue Zhang, Yuhang Lin, Yang Song, Rao Xu, Liaoming Zhou, Qianlong Xie, Jing Xiao, Xiaoning Li, Weiran Xu, Jun Guo

Beijing University of Posts and Telecommunications, Beijing 100876

E-mail: buptly@yahoo.com.cn

Abstract: PRIS_COAE participated in all the four tasks in the Third Chinese Opinion Analysis Evaluation (COAE2011). It is briefly introduced that the method used in each task. In order to extract opinion words and analyze their polarities, the Conditional Random Fields model (CRFs) combining with the Word Activation Force model (WAF) is applied for task 1. Task 2 used two methods respectively: syntactic analysis and CRFs based opinion sentence extraction, sentiment word based opinion sentence extraction. Task 3 extracts object-opinion pairs by syntactic analysis and some rules. The retrieval model is mainly applied in task 4 for opinion document retrieval. The evaluation results for each task are listed in the last section of this paper.

Keywords: syntactic analysis; CRFs; positive word; negative word

1　介绍

COAE2011（The Third Chinese Opinion Analysis Evaluation）是中文信息学会信息检索专业委员会举办的第三届中文倾向性分析评测。

近几年来，文本倾向性（观点、情感等）分析持续成为自然语言处理领域研究的热点问题之一。TREC评测、NTCIR评测以及前两届中文倾向性分析评测推动和加速了倾向性分析研究的发展。在SIGIR、ACL、WWW、CIKM、WSDM等著名国际会议上，针对这一问题的研究成果层出不穷。随着研究的深入展开，也出现了一些新的研究关注点，如Aspect-Based Opinion Mining，Context-sensitive Opinion Mining等。在国内，对于文本倾向性分析的研究正处于快速发展中。如何结合中文处理的特点，进一步推动中文情感分析的发展是目前亟待解决的问题。

与前两届中文倾向性分析评测有所不同，COAE2011拟把领域知识和上下文语境（Context）对倾向性的影响融入到相关任务中。数据集按领域分为三类：电子产品类、影

视娱乐类、财经类。COAE2011根据抽取粒度的不同分为四个任务：首先从给定的三个领域数据集中抽取观点词并判断极性；然后从三个领域数据集随机抽取一定比例的数据经过自动分句构成测试集，从中抽取观点句并进行观点极性判别；接下来，从得到的观点句中抽取评价搭配<观点句，评价对象,评价短语，极性>；最后，把三个领域数据集作为一个大数据集，结合领域知识和上下文语境对给定的查询对象进行观点检索。

我们参加了全部四项任务，本文的第2至5部分分别介绍我们四项任务所采用的方法，第6部分是评测结果及分析。

2 任务一：领域观点词的抽取与极性判别

任务一旨在提取文档的观点词并进行极性判别。由于不同领域文本的用词存在差异，且同一词在不同领域也有不同的倾向表达，本任务提供电子、财经、娱乐三个领域的数据集，要求分别提取观点词，并判别其褒贬极性。该任务的处理流程如图 1 所示。

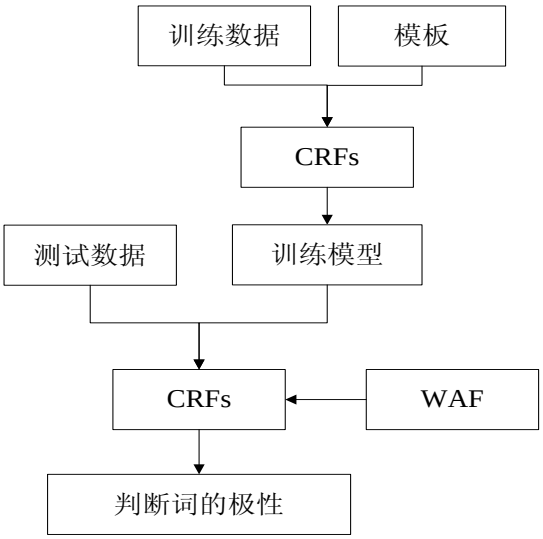

图 1 任务 1 系统框图

Fig.1 The framework of task 1

2.1 条件随机场

条件随机场(Conditional Random Fields, CRFs)，是一种判别式图模型，因为其强大的表达能力和出色的性能，得到了广泛的应用。从最通用角度来看，CRFs 本质上是给定了观察值集合的马尔可夫随机场。

任务一采用 CRFs 训练词情感倾向模型，并得到形容词的极性。使用的是 CRF++工具包。

2.2 训练

训练数据语料采用 2009 年 COAE 评测提供的数据集。使用哈工大分词工具包 HIT-LTP 进行分词和词性判别后，提取语料中所有的形容词以及该词所在的句子，构成原始标注集。由于提取出的所有词与句子数量过多，采用随机抽样，最终得到 1000 个倾向词和 2 万条例

句构成待标注语料集。采用人工标注的方法，正倾向标为 1，负倾向标为-1，中性标为 0。为了增加观点词（标注为 1 或-1）在训练语料中的比例，将标注集中标签全部为 0 的句子删除。最终得到 CRFs 的训练集并形成训练模型。

2.3 测试

测试语料采用 COAE2011 提供的三个领域的数据集，对于不同领域的文本分别使用哈工大分词工具包 HIT-LTP 进行分词，并对分词结果进行词性判别和句法结构分析，得到每个词的词性以及词在句子中的句法结构。然后依据这些信息整理成 CRFs 测试数据的格式，形成测试语料。

用 2.2 中得到的 CRFs 训练模型对测试集进行标注。每个词会以一定的置信度被标注为 0，1 或-1。1 表示正倾向观点词，-1 表示负倾向观点词，0 表示中性词或其他词。

对标注结果进行处理，分别提取词倾向标注为 1 和-1 的词，对于不同句子中标注为相同倾向的同一个词，选择其倾向判别置信度最高的一个进行保留；对于不同句子中标注为不同倾向的同一个词，则全部保留。

2.4WAF 模型

在使用 CRF 进行词性判别的时候会出现所标注的词个数很少的情况（召回率低）。于是我们考虑使用词激活力模型（Word Activation Force，WAF）的方法提取语料中的观点词库，作为 CRFs 词性判别的参考。

WAF 通过计算词之间的共现频率与其在语料集中的结构信息来挖掘它们之间的亲密度与相似性关系。其计算步骤如下：

2.4.1 共现矩阵

对于一个词（称之为中心词）设定一个词窗，在这个词的左边窗口内出现的词，我们认为它们是中心词的入链，每出现一个新词，入链加一，每出现一个已经出现过的词，这个词对于中心词的入链程度加一。同理，对于词的右边窗口内出现的词，我们称之为出链，我们同样会统计出现的词和词的出现次数。最后，我们就会得到一个共现矩阵。

之所以我们会把中心词左边与右边的词分开统计，是因为我们认为在语法上，两边的词的功能不一样，作用也不一样，在句子中扮演的角色也不一样。

2.4.2WAF 矩阵

WAF 矩阵是由词之间的激活力 (wafs)组成的，对于词 A 和词 B，设 f_a 是词 A 出现的频率，f_b 是词 B 的出现频率，f_{ab} 是词 B 在词 A 右窗口内出现的频率，则词 A 激活词 B 的 WAF 值为：

$$waf_{ab} = (f_{ab}/f_a)（f_{ab}/f_b）/（d_{ab}）^2 \quad (1)$$

其中 d_{ab} 是词 A 与词 B 的平均距离。

显然 waf_{ab} 语 waf_{ba} 是不相等的，故 WAF 矩阵是一个非对称矩阵，或称为有向图。

2.4.3A 值矩阵

相比于 WAF 值，A 值体现的是在相同环境中出现的词的关系，A 值由下面的公式获得：

$$A_{ij}^{waf} = [\frac{1}{|K_{ij}|}\sum_{k\in K_{ij}} OR(waf_{ki}, waf_{kj}) \cdot \frac{1}{|L_{ij}|}\sum_{j\in L_{ij}} OR(waf_{ij}, waf_{ji})]^{1/2} \quad (2)$$

其中 K_{ij} 是同时出现在词 i 和词 j 入链的词的集合，L_{ij} 则是同时出现在词 i 和词 j 的出链的词的集合，$OR(x,y) = \min(x,y)/\max(x,y)$，A 值可以比较好的体现两个词出现在相同语境的概率。

通过以上方法，以部分人工标记的词为种子，通过 A 值矩阵扩展其亲密度较高的词构成完整的观点词词库。

3 任务二：中文观点句抽取

该任务旨在从三个领域的数据集中自动识别观点句及其表达观点的总体极性（褒义、贬义或混合观点）。我们采用了两种方法进行中文观点句的抽取：基于句法和 CRFs 的观点句抽取；基于情感词个数的观点句抽取。下面分别介绍。

3.1 基于句法和 CRFs 的观点句抽取

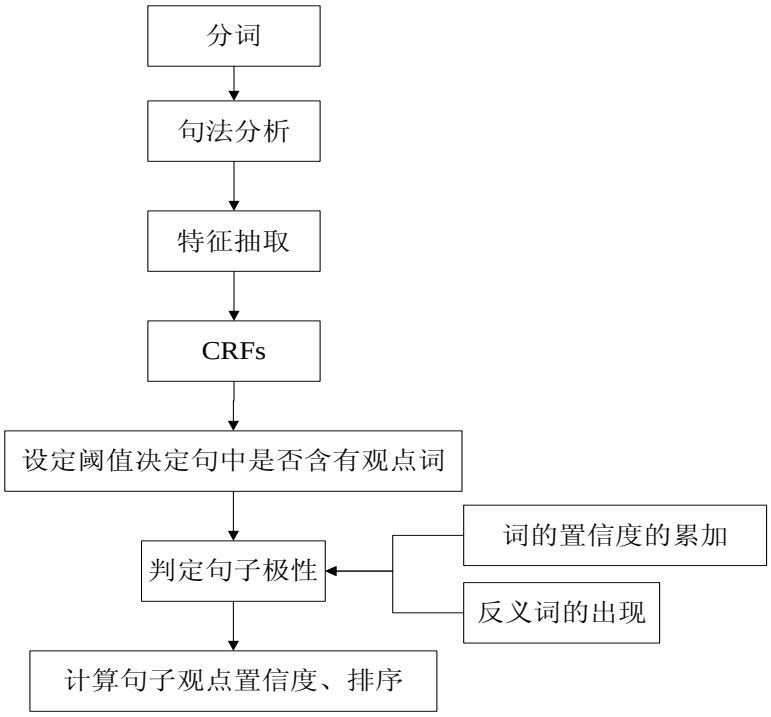

图 2 基于句法和 CRFs 的观点句抽取框图

Fig.2 The framework of syntax and CRFs based opinion sentence extraction

该方法的系统框图如图 2 所示。首先，与任务 1 的预处理相同，先用哈工大分词工具包 HIT-LTP 对测试集进行分析，得到每一个词（分词结果）在句子中的句法结构及其是否为实体等特征，然后依据这些信息整理成条件随机场模型（CRFs）测试数据的格式。再用任务 1 得到的 CRFs 训练模型对测试数据进行标注。每个分词会以一定的置信度被标注为 0，1 或-1。1 表示褒义观点词，-1 表示贬义观点词，0 表示其他词。

从标注结果中提取出每一句进行分析。对每句标为褒义和贬义的初始置信度都设为 0。

为了防止否定词、转折词影响范围过长，我们按照标点符号将句子 S 分为 n 个子句 S_i(i=1,2,…n)。我们定义如下标记：

- $label_t$:S 中每个词 t 的标签，其值可取 0、1、-1，分别表示词 t 为无观点词、褒义词、贬义词；
- $score_t$:CRFs 对词 t 标签的打分。
- $sign_i$:子句 S_i 中是否含有转折词或否定词。如果不包含，则值为 1;否则为-1。

$$ConfScore_S = |PolScore_S| = \left|\sum_{S_i}\sum_{t\in S_i} sign_i \cdot label_t \cdot score_t\right| \quad (3)$$

$$Pol_S = sgn(PolScore_S) \quad (4)$$

公式(3)中 $ConfScore_S$ 是句子 S 的最终置信度，公式(4)中 Pol_S 是句子 S 的极性。首先查看 S_i 中是否有词被标注为“1”或“-1”，如果有，再检查 S_i 中是否包含转折词或反义词。例如词 t 被标为 1（褒义词），CRFs 的打分为 0.3。如果在词 t 所在子句中，在词 t 前出现了否定词或者在词 t 后出现了转折词，则句子的贬义置信度就加 0.3。

3.2 基于情感词个数的观点句抽取

该方法通过计算句子中情感词的个数来判断句子的褒贬倾向性。如果句子中不含情感词，则判别为无观点；如果句子中褒义词的个数大于贬义词的个数，则句子判别为正倾向性；如果褒义词的个数小于贬义词的个数，则句子判别为负倾向性；如果褒义词贬义词个数相等，则判别为混合观点。

另外，我们对以下两种语言现象做了特殊处理：

(1) 句子中出现否定词：如果一个分句中，情感词前面存在否定词，则该情感词的倾向性反转。

(2) 对“不是……而是……”句型做了处理。例如：“*诺基亚手机受欢迎不是因为它的外表而是因为它的实用。*”句子中的“实用”一词，并不因为前面的否定词“不是”而反转倾向性。

最终句子S的置信度为计算公式为：

$$ConfScore_S = SenNum/SenNumMax \quad (5)$$

其中SenNum是该句中情感词的个数，SenNumMax为所有句子中情感词个数的最大值。

4 任务三：评价搭配抽取

本任务关注上下文语境对观点识别和倾向性判断的影响，要求找出任务2抽取的每个观点句中观点所针对的评价对象、评价短语，并对评价的倾向性做出判别。在这一部分中，针对任务二挑选出的观点句，基于中文句子的结构信息和语法特点，应用词性标注、句法分析方法，借助知网倾向性词典，对观点句中的评价词和评价对象进行抽取。

4.1 中文句子结构信息

目前，自然语言处理技术已在文本分析中发挥着巨大的作用。我们使用哈工大句法分析器对观点句进行语义分析来获取其上下文信息，包括短语分析和句法分析。例如：“凯越的油耗非常高”的句法分析结果如下：

凯　　越　　SBV
越　　的　　DE
的　　油耗　　ATT
油耗　　高　　SBV
非常　　高　　ADV
高　　NONE　　PUN

其中：SBV 表示主谓关系，ATT 表示定中关系，ADV 表示状中关系，DE 表示“的”字结构。通过上述分析，可以获得句子语义的层次结构及其上下文依存修饰关系。

4.2 抽取步骤

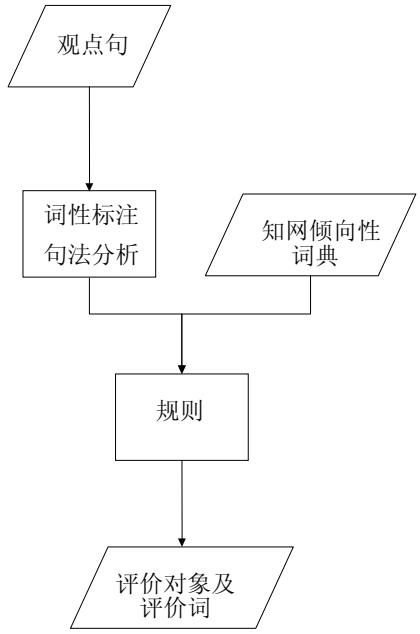

图 3 任务 3 系统框图

Fig.3 The framework of task 2

任务三的系统框图如图 3 所示。其中，观点句来自于任务二挑选出来的具有倾向性的评价句子。具体的抽取过程描述如下：

从任务二中获得挑选后的观点句，这些评价句都具有倾向性观点，用于进行评价对象和评价词对的抽取。使用的形容词词典是中文知网提供的正负形容词词典，已经带有倾向性判断，可以直接得出评价词的倾向性。

用哈工大的句法分析工具包对观点句进行句法分析，使用如下规则进行抽取：

(1)先根据评价词词表定位形容词，然后根据句法分析合并修饰它的副词，合并后的结果作为最后的评价词。

(2)在每一个分句中，如果存在 SBV 关系，抽取出存在 SBV 关系的词对（评价对象，评价词）；如果不存在 SBV 关系，则查找 DE 关系，向后查找跟当前形容词有 DE 关系的名词，若没有 DE 关系，直接向前查找名词。

(3)对于找到的评价对象，合并其父节点，合并的结果作为最后的评价对象。

5 任务四：观点检索

该任务给定 50 个观点检索对象（topic/target），针对每个对象，把三个数据集看成一个大数据集，要求找出包含评价该对象的倾向性观点的文章。该任务是信息检索和观点识别的组合任务。该任务中明确提到：在给定查询实体的条件下，找出包含评价该对象的倾向性观点的文章。给定对象可能是人物、商品、组织机构或者概念、事物、事件等。综合多个因素，我们在该任务中引入了 Indri 索引，它是一种基于概率语言模型的检索工具。

在对中文文档建立索引之前，需要按前面任务的流程进行分词预处理等工作，得到按空格分割的词序列。之后，对各个文档加入对应的文档编号，文档所属的类别，使用 indri 工具包建立文档索引。索引建成后，使用任务中提供的 20 个实体作为查询对象，按关键字匹配的规则返回所有包含对应的实体文档集与文档编号集合，用于下一步处理。

得到每个 topic 的相关文档后，我们发现 topic“新浪”对应的文档往往是新浪网站上的一些文章而并非对“新浪”本身的评价，所以我们对其做了特殊处理：

(1) 去掉其中字数少于 20 的文档。

(2) 利用现有的明星姓名词表过滤文档，含有明星姓名的文档去掉。

(3) 我们采用两种方法处理：一种是采用任务三的方法对这些文档抽取评价对象，如果这些评价对象中包含“新浪”，则该文档作为相关文档，否则不作为相关文档。

得到最终的相关文档后，我们采用基于情感词个数的方法来判断文档的情感倾向性，其中情感词由情感词表确定。对于不含情感词的文档，认为他们不包含观点。如果文档中褒义词的个数大于贬义词的个数，则文档判别为正倾向性；如果褒义词的个数小于贬义词的个数，则文档判别为负倾向性；若相等，则判别为混合观点（类似于任务二中用情感词个数来判断句子极性的方法）。

表 1 任务一评测结果

Tab.1 Evaluation result of task 1

标识	领域	Precision@1000	Precision	Recall	F1	Raccuracy
pris_t1	电子产品	0.518	0.2765	0.0827	0.1273	0.0827
pris_t1	影视娱乐	0.542	0.3113	0.0864	0.1352	0.0864
pris_t1	金融证券	0.5612	0.5394	0.0791	0.138	0.0791
Median		0.57126	0.343004	0.094744	0.147593	0.094744444
Best		0.674	0.6125	0.1194	0.1833	0.1194
宏平均						
pris_t1		0.5404	0.3757	0.0827	0.1356	0.0827
Median		0.571266667	0.342993333	0.09474	0.14788	0.09474
Best		0.6567	0.486	0.1136	0.1744	0.1136

微平均						
pris_t1		0.5404	0.3426	0.0828	0.1333	0.0828
Median		0.571266667	0.337933	0.09474	0.147547	0.09474
Best		0.6567	0.4681	0.1135	0.1744	0.1135

6 评测结果

6.1 任务一

任务一的评测结果见表 1。

6.2 任务二

我们对任务二提交了两个结果，“PRIS_COAE_t2_1”和“PRIS_COAE_t2_2”分别对应基于句法和 CRFs 的观点句抽取、基于情感词个数的观点句抽取，评测结果见表 2。

表 2 任务二评测结果

Tab.2 Evaluation result of task 2

标识	Precision	Recall	F1	P@1000	Raccuracy
PRIS_COAE_t2_1_D	0.440073	0.468992	0.454072	0.552	0.446265
PRIS_COAE_t2_1_E	0.188156	0.364524	0.248199	0.204	0.205045
PRIS_COAE_t2_1_F	0.121248	0.398453	0.185921	0.12	0.133462
PRIS_COAE_t2_2_D	0.365987	0.42389	0.392816	0.391	0.357646
PRIS_COAE_t2_2_E	0.163753	0.312449	0.214885	0.167	0.147274
PRIS_COAE_t2_2_F	0.091438	0.334623	0.143628	0.054	0.079304
Median	0.240815	0.397946	0.276324	0.290183	0.255445
Best	0.729751	0.798097	0.693304	0.8	0.660324
宏平均					
PRIS_COAE_t2_1	0.249826	0.410656	0.296064	0.292	0.261591
PRIS_COAE_t2_2	0.207059	0.356987	0.250443	0.204	0.194741
Median	0.240815	0.406039	0.276324	0.290183	0.255445
Best	0.534693	0.723411	0.541377	0.532	0.494511
微平均					
PRIS_COAE_t2_1	0.327377	0.44678	0.37787	0.292	0.384533
PRIS_COAE_t2_2	0.274073	0.399219	0.325015	0.204	0.303422

Median	0.315252	0.450532	0.357871	0.290183	0.36317
Best	0.654448	0.775397	0.639614	0.532	0.611425

6.3 任务三

表 3 任务三评测结果

Tab.3 Evaluation result of task 3

评价对象正确							
标识		领域	P@1000	Precision	Recall	F1	Raccuracy
PRIS_COAE_t3		D	0.074	0.086096	0.088914	0.087483	0.088914
PRIS_COAE_t3		E	0.043	0.039579	0.048177	0.043457	0.048177
PRIS_COAE_t3		F	0.02	0.024642	0.023041	0.023815	0.023041
宏平均			0.045667	0.050106	0.053378	0.05169	0.053378
微平均			0.045667	0.066397	0.113385	0.08375	0.07849
宏平均	Median		0.065524	0.074285	0.045504	0.054307	0.045504
	Best		0.111	0.133933	0.081763	0.091606	0.081763
微平均	Median		0.065524	0.101236	0.081829	0.083421	0.069819
	Best		0.111	0.159847	0.149071	0.144701	0.135726
评价短语正确							
标识		领域	P@1000	Precision	Recall	F1	Raccuracy
PRIS_COAE_t3		D	0.075	0.083031	0.112751	0.095635	0.087275
PRIS_COAE_t3		E	0.017	0.026303	0.068359	0.037988	0.024089
PRIS_COAE_t3		F	0.021	0.027027	0.104455	0.042943	0.024578
宏平均			0.037667	0.045454	0.095188	0.061527	0.045314
微平均			0.037667	0.061766	0.105476	0.077909	0.073646
宏平均	Median		0.051429	0.055383	0.05059	0.046795	0.035003
	Best		0.118333	0.085672	0.100699	0.087237	0.066468
微平均	Median		0.051429	0.074015	0.061917	0.062064	0.052096
	Best		0.118333	0.117425	0.105575	0.10248	0.095492
评价对象、短语、极性都正确							
标识		领域	P@1000	Precision	Recall	F1	Raccuracy
PRIS_COAE_t3		D	0.034	0.032135	0.043637	0.037013	0.033043
PRIS_COAE_t3		E	0.004	0.007014	0.018229	0.01013	0.00651
PRIS_COAE_t3		F	0.006	0.006359	0.024578	0.010104	0.004608
宏平均			0.014667	0.01517	0.028815	0.019876	0.014721
微平均			0.014667	0.022576	0.038553	0.028477	0.027185
宏平均	Median		0.02719	0.025047	0.020403	0.019833	0.015183

	Best		0.071667	0.039639	0.037671	0.03478	0.025728
微平均	Median		0.025778	0.033019	0.027191	0.027323	0.02309
	Best		0.071667	0.0616	0.046856	0.048312	0.043298

6.4 任务四

我们对任务四提交了两个结果，“PRIS_COAE_t4_1”和“PRIS_COAE_t4_2”分别对应第 5 部分介绍的对“新浪”topic 做特殊处理和不做任何特殊处理，评测结果见表 4。

表 4 任务四评测结果

Tab.4 Evaluation result of task 4

标识	MAP				宏平均			微平均			P@N	
	未插值		相关文档		Precision	Recall	F1	Precision	Recall	F1	P@5	P@10
	REL	senti	REL	senti								
PRIS_COAE_t4_1	0.136	0.065	0.919	0.404	0.287	0.038	0.063	0.202	0.140	0.165	0.93	0.91
PRIS_COAE_t4_2	0.148	0.070	0.941	0.431	0.352	0.041	0.071	0.268	0.149	0.192	0.95	0.92
MEDIAN	0.142	0.067	0.930	0.417	0.320	0.039	0.067	0.235	0.145	0.179	0.94	0.915
BEST	0.148	0.072	0.9998	0.657	0.562	0.041	0.071	0.575	0.149	0.192	0.98	0.925

参 考 文 献

[1] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[A]. Proceedings of the 18th International Conf on machine Learning[C]. 2001. p.282-p.289.

[2] 洪铭材，张阔，唐杰. 基于条件随机场(CRFs)的中文词性标注方法[J]，计算机科学，2006：p.148-p.155.

[3] Guo J, Guo H. & Wang Z. An Activation Force-based Affinity Measure for Analyzing Complex Networks[J]. Sci. Rep, 2011,1:113; DOI:10.1038/srep00113.

[4] 许细清，林世平. Web 文档评价对象抽取研究[J]. 计算机工程，2011 年，第 37 卷第 6 期：p.30-p.34.

词语搭配情感倾向的自动判别方法*

王菲, 吴云芳, 徐艺峰, 王宇昕, 李素建
北京大学计算语言学研究所 北京 100871
计算语言学教育部重点实验室
E-mail: sxjzwangfe@163.com

摘要：文章针对 COAE 中的所有评测任务对情感词、搭配及句子的极性判别做了详细的讨论。文章构建了高质量的核心词典，针对度量形容词计算了被修饰词的正/负向情感期待；引入京东种子语料进行领域观点词的抽取；引入语义相似度进行修饰词和被修饰词的同/近义词扩展；采用 LDA 模型进行观点检索结果的生成。实验证明，质量良好的核心词典影响巨大，度量形容词的情感极性处理得当，同/近义词扩展有效，话题模型对观点检索作用明显。

关键词：核心词典，度量形容词，同/近义词扩展，LDA

Predicting the Semantic Orientation of Word Collocations

Fei Wang, Yunfang Wu, YifengXu, Yuxin Wang, Sujian Li
Institute of Computational Linguistics, Peking University, Beijing 100871
Key laboratory of Computational Linguistics, Ministry of Education
E-mail: sxjzwangfe@163.com

Abstract: This paper gives an in-depth discussion of the sentiment orientation computation of sentiment-bearing words，opinionated patterns and sentences. This paper carried out the construction of a high quality core dictionary, the sentiment-expectation of target words modified by measure adjectives, the field sentiment extraction boosted by Jingdong seed texts, the synonyms expansion facilitated by word semantic similarity computation et. al, and the LDA aided indexing progress. The experimental results proved the role of the core dictionary, the effectiveness of computation towards target words modified by measure adjectives, the importance of seed text,the usefulness of synonyms expansion and the significance of topic model complementation in document indexing.

Keywords:core dictionary, measure adjectives, Jindong seeds, synonyms expansion, LDA

1. 引言

互联网提供的购物、信息、社交等平台为商家、机构和个人提供了海量的文本数据。这些数据中，少量属于结构/半结构化数据，诸如京东商城的产品评论包含优点、缺点、使用心得;而绝大多数数据非结构化的数据，诸如论坛，博客，微博上蕴涵的海量关于某话题的信息。这些文本中蕴涵了大量的观点/情感信息，具有非常大的潜在价值。通常情况下，情感被分成诸如 Positive/Negative 两类，或[1-n]多个级别。为此，情感分析通常看成是分类问题。

*该工作由自然科学基金(基金号:60703063,60875042 和 90920011)和社会科学基金(基金号:No: 10CYY023)支持。

在以往的研究中，研究着从不同的粒度和侧面对情感分析任务进行了探索，包括文档级别的情感分类（Dave,2003），词汇级别的情感分类（Hatzivassiloglou,1997），句子级别的情感分类（Yu,2003），主观表达的抽取（Pang,2004），评价源的抽取（Choi,2005），等等。从研究方法上，有带监督的情感分类(Gamon,2004)，半监督的情感分类(Sajib,2009)和无监督的情感分类（Turney,2002）。随着情感分类技术的发展和应用需求的增加，考虑到标注成本和训练学习的开销，夸领域的情感分类和领域词的极性消歧也成为研究的热点（Danushka,2011. Wu,2010）。除情感挖掘本身的研究之外，情感挖掘也服务与观点检索，即检索出的主观文本赋予相应的观点。

本文主要介绍了在 COAE 评测中使用的方法。我们首先进行了任务 3,即抽取搭配中的修饰成分和被修饰成分，并判断修饰成分在修饰被修饰成分时所表现的极性;随后进行了任务 1,抽取领域词汇和任务 2,观点句子的抽取和观点判别。任务 4 独立于其它三个任务单独进行，分信息检索和极性判别两步。在下文中，我们首先介绍了模式的抽取方法;其次介绍了本文中主要使用的特殊单子词极性判别、核心辞典构建、同/近义词扩展等技术;随后给出了各个任务的实现方式,最后列出了实验结果和结论。

2. 搭配的抽取

在搭配抽取的过程中，考虑到句法分析器的复杂性和准确率，我们采用基于规则的方法猜测短语中修饰成分何被修饰成分。为此，我们采用中科院分词工具 ICTCLAS 对语料进行分词和词性标注，并采用如下模式猜测搭配中的修饰成分何被修饰成分。

通常情况下，构成评价搭配的基本情况大致如下：

名词短语+形容词短语　　例如：外观整洁　私生活混乱

形容词短语+助词的+名词短语　例如：整洁的外观 混乱的私生活

动词短语+形容词短语　　例如：反应迅速　运行流畅

动词短语+助词得+副词短语　　例如：跑得快 消耗得少

形容词短语+助词地+动词短语　例如：流畅地运行 迅速地反应

其中，名词短语主要由如下扩展方式：

名词短语->名词　例如：做工优良

名词短语->名词短语+名词短语 例如：长宽高比例匀称

名词短语->名词短语+连词+名词短语　　例如：做工和设计精良

名词短语->名词短语+助词+名词短语　　例如：产品的造型独特

形容词短语主要由如下扩展方式：

形容词短语->形容词

形容词短语->副词短语+形容词短语 例如：绯闻很多

形容词短语->形容词短语+了+量词短语　例如：差了一些

动词短语主要有如下扩展形式：

动词短语->动词

动词短语->动词短语+名词短语 例如：处理信息非常迅速

为此，我们根据上述规则，制定了如下模式：

表 1. 模式的抽取

Table1. Extraction Patterns

模式	修饰成分	被修饰成分	备注
nouns —> (vn|ns|nr|nt|nz)+ nouns —> nouns (的|之|和|与) nouns adjs —> ((d)? (a|an)+)+ 注：本行中，用“+”表示一次或多次出现。“？”表示0次或1次出现。“*”表示0次或多次出现。			下文中，用 end 表示分句末尾，用 begin 表示分句起始位置。用“+”表示衔接关系。“？”和“*”解释同左。
nouns+adjs+end	nouns	adjs	如：做工/n 精致/a
nouns+v+adjs+end	nouns+v	adjs	如：色彩/n 表现/v 完美/a
adjs + 的 +nouns	nouns	adjs	如：丰富/a 的/u 场景/n
adjs + 的 + v|vn +nouns	nouns	adjs	如： 最高/a 的/u 写入/v 速度/n
adjs+地+v+nouns?	v+nouns?	adjs	如：流畅/a 地/u 运行/v
v+得+adjs	v	adjs	如： 跑/v 得/u 快/a
begin|nouns+v+nouns+adjs	nouns? +v+nouns	adjs	如：处理/v 信息/n 迅速/a
begin+v+ adj+end	v	adjs	如：反应/v 迅速/a
nouns+z+adjs*+end	nouns	z+adjs*	如：做工/n 优良/z 如：功能/n 依旧/z 单一/a
nouns+d+v+end	Nouns	d+v	如：游戏/n 有点/d 卡/v

对于上表，我们仍有如下需要解释：由于分词中词性标注的原因，出现了状态词 z。语料中有大量的修饰关系诸如“星光熠熠”、“做工优良”中的“熠熠”、“优良”被标注为状态词 z。为此，我们添加了模式“nouns+z+adjs*+end”。出于对语料的观察，我们补充了“adjs + 的 + v|vn +nouns”和“nouns+d+v+end”。而在实现的过程中，我们发现，“动词短语->动词短语+名词短语”模式会引入较大的噪音，因此，我们将动词短语简化为动词。在上述模式中，有些出现了 begin 和 end 约束，这是为了提高模式抽取所得到的搭配的准确率，当然，在提高准确率的同时也牺牲了部分召回率。

3. 情感词的识别

本文以词的极性为基础，计算搭配的机型和句子的极性。词的极型包括修饰成分的极性，被修饰成分的极性期待。同时，对数码领域进行了特殊的领域扩展。为了扩大带极性的词和种子搭配的影响范围，我们引入了同/近义词的扩展。

3.1. 度量形容词的情感倾向判别

在观点挖掘的过程中，单子词是十分特殊而且十分难以捕捉的一部分，其中，大、小、多、少、高、低、厚、薄、深、浅、重、轻又是其中非常具有代表性的一部分。这十二个单子词频繁地出现在文本中，且其情感倾向绝大多数依赖于它所修饰的目标词。为此，我们引用了（Wu，2010）提出的方法，对这些词进行了特殊的处理。

我们将{大，多，高，厚，深，重}称为类积极形容词，将{小，少，低，薄，浅，轻}称为类消极形容词。此外，将与类积极形容词搭配表示积极情感倾向的名词称为正向期待词，如工资（高）、福利（多）;将与类消极形容词搭配表示积极情感倾向的名词称为负向期待词，如税（少），负担（轻）。同时，类消极形容词修饰正向期待词时往往具有负向的情感倾向，如工资低，福利少;同样的，类消极形容词修饰正向期待词时也通常表现出负向的情感倾向，如税多，负担重。为此，当某名词与这十二个单字词搭配时，该词组的情感倾向的计算转化成为该名词是正向期待词或者负向期待词的计算。

$$C(a)=\begin{cases}1 & a\text{ 是类积极形容词}\\ -1 & a\text{ 是类消极形容词}\end{cases}$$

$$C(n)=\begin{cases}1 & n\text{ 是正向期待词}\\ -1 & n\text{ 是负向期待词}\end{cases}$$

$$SO(n+a)=C(n)*C(a)$$

上述公式中，C(a)是形容词 a 类积极/消极的特性，C(n)是目标词 n 的正/负向期待特性。当 a 修饰 n 时，搭配显示的极性 SO 由 C(a)和 C(n)共同决定。

在自然语言中，有一些模式具有非常好的情感倾向暗示作用，诸如，我们常用“有点儿”暗示一种负向情感，通常用于衔接正向期待词与类消极情感词表示抱怨，例如“工资有点儿低”。（Wu，2010）采用了四个模式用于捕捉名词的正/负向期待。

表 2. 期待模式

Table 2. Expectation Patterns

正向期待模式	负向期待模式
<n>有点<类消极形容词> eg：工资有点低	<n>有点<类积极形容词> eg：价格有点高
<n>有点儿<类消极形容词> eg：职称有点儿低	<n>有点儿<类积极形容词> eg：负担有点儿重
<n><类消极形容词>，怎么办? eg：收入低，怎么办?	<n><类积极形容词>，怎么办? eg：血压高，怎么办?
嫌<n><类消极形容词> eg：嫌钱少	嫌<n><类积极形容词> eg：嫌噪音大

名词的正/负向期待计算方法如下：

$$PT\_SO(n) = \sum_{b \in Na} \sum_{i=1}^{4} PositivePatternHit_i(n,b) - \sum_{a \in Pa} \sum_{i=1}^{4} NegativePatternHit_i(n,a)$$

$$n \text{ 是} \begin{cases} \text{正向期待如果 } PT\_SO(n) > 0 \\ \text{负向期待如果 } PT\_SO(n) < 0 \\ \text{无法预测如果 } PT\_SO(n) = 0 \end{cases}$$

其中，n 待计算期待极性的名词，Na 代表类消极形容词集合，Pa 代表类积极形容词集合，PositivePatternHit 和 NegativePatternHit 分别代表正向期待模式和负向期待模式在百度搜索引擎中返回的 Hit 值。本文直接引用了其计算结果。

3.2. 核心褒贬义词典的构建

为了获得一个良的核心褒贬义词集，我们参考了《褒贬义辞典》,《小学生褒贬义辞典》与“Hownet 情感辞典”。我们认为 Hownet 的单子词中生辟或者情感倾向模糊的情况较多，故而将 Hownet 中的单子词剔除，取其余部分与褒贬义辞典与小学生褒贬义辞典取并集。在取并集的过程中，如若某词既出现在褒义词集中，又出现在贬义词集中，则将其从核心词集重剔除。至此，褒义词集合共得到 5223 个词，贬义词集合得到 3760 个词。这些词构成了核心褒贬义辞典。由于小学生褒贬义词典和褒贬义词典中不存在单子词，因此，最终的核心词典中不包含单子词。不包含单子词固然有其好处，但是我们也丢掉了诸如“好”、“佳”等词的极性信息，为此，后续的同、近义词扩展也是很有作用的。

3.3. 领域观点词的获取

由于褒贬义词集的获取主要源自于常规辞典，故其收录的词多采用其最常用最生活化的意向，如“轻薄”。褒贬义辞典，小学生褒贬义辞典，Hownet 都收录“轻薄”为贬义词，取其意向“言行不庄重、不敦厚”，而忽略其于电子产品上的意向“轻巧纤薄”。考虑到电子产品的领域性比较强，而娱乐领域中用词的意向与日常常用意向相似，财经领域用词相对专业，我们特为数码领域构建领域词辞典。电子领域中的领域词有一个比较大的特点就是，当它背离常用意向时，它往往是两个词的缩写，注重的是每个字的意思，而不是连起来词的意思，诸如上面提到的“轻薄”——轻巧纤薄，还有“圆滑”——圆润光滑等。领域辞典中我们采用 double-anchor 的方式确认词的极性，即某词的极性不仅取决于词本身，也取决于它所修饰/搭配的目标名词词组，这是为了适应某些搭配多样的词再修饰不同目标时表现出不同的极性，例如“噪声大”表示贬义，而“照明面积大”表示褒义。

为了捕捉数码领域极性词，我们引入京东种子语料。京东网提供了一种结构化的评论，有优点，缺点文本框供买家评论产品的优缺点。此外，考虑到产品介绍中往往是针对产品的优点而写的，我们也将产品介绍纳入优点的范畴。我们随机从手机、相机、电脑、电脑

配件中随机抽取了 10 个产品以获得其评论及介绍，将其设为种子语料。采用前文介绍的搭配的抽取方法，我们获得了京东种子语料中，某修饰成分修饰被修饰成分时表现出的极性。将这些搭配记录下来，用以指导某修饰词修饰被修饰词时所表现出的极性。我们将出现在优点/产品介绍中的搭配记为正向种子，将出现在缺点中的搭配记录为负向种子。可以看到，在产品介绍中，某些领域极性词诸如“圆滑”是可以被捕捉到正向极性的。需要意识到的是 double-anchor 在保守的保证了种子扩展的正确性外，也有覆盖率不足的问题。这个缺点一部分可以用同、近义词扩展来解决，另一部分却由于 double-anchor 而无法解决，例如:“圆滑”在京东种子中只修饰了“设计”和“外观”二词，为此使用此方法只能解决“圆滑”修饰“设计”和“外观”或其近义词时的领域极性，而无法覆盖测试语料中的所有情况，诸如“圆滑的曲线”、“圆滑的触感”等。

3.4. 评价对象和情感词语的自动扩充

考虑到核心褒贬义辞典覆盖能力有限，我们期望采用同义词扩展的方式获得更多的褒贬义词。同样的，为了使得京东种子搭配中被修饰成分获得扩展，而不仅仅局限于种子语料中出现的被修饰词，也同样需要近义词对被修饰成分进行扩展。

我们使用了（石静，2011）相似词的获取方法。我们采用 1991~2004 年共 14 年的新华社新闻语料和搜狗互联网语料中 130000 个常用词所出现的句子，选取窗口大小为 2 或 3 的上下文词语特征、以上下文与目标词之间的互信息作为权值构建特征向量，通过计算向量之间的 cosine 夹角作为词语相似度，采用平均分数的集成方法，获得每个词的语义相似词。需要注意的是，由于此处采用上下文作为特征，采用此方法计算出的是分布相似词，而非严格意义上的同义词和近义词，为此，它并不适用于同义情感词的计算（情感词多为形容词，而形容词的分布相似词可能混杂了大量的同义词、反义词等）。但是对于目标成分而言，由于目标成分多为名词，名词的分布相似词多是该名词的近义词，因此，我们仅使用此方法计算名词的近义词，用以扩充京东搭配中的被修饰成分。

对于修饰成分（多为形容词）而言，我们则采用哈工大《同义词词林扩展版》，期以获得严格意义上的同义词。我们认为，出现在同义词词林中同一行的两个词互为同义词。

4. 技术实现

在评测实现过程中，我们按照 task3,task1,task2 的顺序完成前三个任务。第四个任务则独立于前三个任务单独实现。

4.1. Task3 搭配中修饰词的极性与置信度计算

基于上述陈述，我们采用如下方法计算修饰词的极性与执行度。

a) 应用模式抽取修饰成分和被修饰成分。获取修饰成分中的形容词成分（当没有形容词的时候，依据其抽取自模式 nouns+z+adjs*+end 或 nouns+d+v+end，取其中

的状态词或动词代替）与被修饰成分中的中心词（猜测被修饰成分中的最后一个词为该成分的中心词）。当搭配中存在词“不”、“没有”时，搭配的极性取反。

b) 若搭配关系抽取自模式 nouns+adjs|v+了+一些|一点|一点儿+end，则搭配极性判断为 Negative，置信度为 1。

c) 若当前领域为数码领域，且修饰形容词（或其同义词）与被修饰中心词（或其近义词）在京东种子搭配中命中：

   1) 若修饰形容词为类积极/类消极形容词，且被修饰成分中的中心词的正/负向期待的分数 PT_SO(n)，如若该分数的绝对值大于某一阈值 threshold，则以 PT_SO(n)的符号的正/负判断极性 Positive/Negative，置信度为 min(1,|PT_SO(n)|/normalize-factor)。

   2) 若当前搭配中的修饰形容词词与被修中心词再京东正向/负向种子中命中，则判断其极性为 Positive/Negative。置信度为 0.9。

   3) 若当前搭配中，修饰形容词的近义词与被修饰中心词在京东正向/负向种子中命中，或者修饰形容词与被修饰中心词再京东正向/负向种子中命中，则判断其极性为 Positive/Negative。置信度为 0.8。

   4) 若当前搭配中，修饰形容词的近义词与被修饰中心词的近义词在京东正向/负向种子中命中，则判断其极性为 Positive/Negative。置信度为 0.7。

d) 若当前搭配中，修饰形容词为类积极/类消极形容词，且被修饰成分中的中心词的正/负向期待的分数 PT_SO(n)。则以 PT_SO(n)的符号的正/负判断极性 Positive/Negative，置信度为 min(1,|PT_SO(n)|/normalize-factor)。

e) 若当前搭配中，修饰形容词在核心褒贬义辞典中，则以该形容词的极性为该搭配的极性，置信度为 0.6。

f) 若当前搭配中，修饰形容词的近义词在褒贬义辞典重出现，则以其近义词的极性为该搭配的极性，置信度为 0.4。若修饰形容词为类积极/类消极形容词，但是被修饰中心词未能在正/负向倾向词的结果中命中，则置信度折半。

在上述陈述中，有如下几点需要说明：

a) normalize-factor 在实验中取值为 1000。

b) 京东正向种子和负向种子在用于上述过程之前需要过滤。如若互为同义的形容词在修饰同一个中心词时表现出不同的极性，则将这两个搭配从种子中删除。此外，由于在优点/产品介绍中常常出现“比普通的显示器”，“相对于一般的手机而言”等比较成分，导致“普通”和“一般”与其修饰成分大量出现在正向种子中，为此，我们剔除了包含“普通”和“一般”作为修饰成分的种子。但是，我们也需要承认，在采用京东种子的过程中，我们冒了一定的风险：如“对称的设计”/“不对称的设计”，它们并没有孰好孰坏。但是，如果“对称的设计”出现在正向京东语料中，“不对称的设计”就可能会被判断为负向。

c) 否定的判断仅局限在搭配中，且仅局限于词“不”和“没有”。

4.2. Task1 其它极性词的抽取和置信度计算

由于模式的匹配能力有限，在模式之外仍然存在部分极性词尚未被发觉出来。为了尽可能提高极性词的召回率，我们在使用模式之外，也使用极性词直接在文本中进行匹配。直接匹配的过程用以补充 task3 的极性词结果。为了使结果更加可靠，我们将极性词限制在了褒贬义核心辞典中。方法如下：

如果极性词出现在句中，且其未被 task3 中的模式捕捉到：

1) 若该极性词的词性为形容词（包括名形词），则该词在上下文中的极性被指定为该词在褒贬义核心辞典中的极性。若其为双字词，置信度置为 0.5。若其为 3 字或 3 字以上的词，置信度设为 0.65。如“得人心”、“盲目”、“贪财”、“稀松”、“肤浅”、“扫兴”等等。
2) 若该极性词为名词，且该词在褒贬义辞典中的极性为 Negative。若其为双字词，置信度置为 0.5。若其为三字或三字以上词，置信度设为 0.65。如“吸血鬼”,“篓子”、“过失”、“弊端”、“地狱”、“土匪”等。之所以没有取极性为 Positive 的词，是因为其中包含大量出现在褒义词典中的词，诸如“钢铁”、“黄金”、“专家”、“业绩”等，而这些词在测试语料中仅用于形容“钢铁板块”、“黄金板块”、“专家意见”等等，而“业绩”可能没有明显的倾向。
3) 若该极性词为副词，则该词在上下文中的极性被指定为该词在褒贬义核心辞典中的极性。若其为双字词，置信度置为 0.5。若其为 3 三字或三字以上词，置信度设为 0.65。如“一味”、“擅自”、“反而”、“贸然”、“真心”等。
4) 若该极性词为动词，且该词在褒贬义辞典中的极性为 Negative。若其为双字词，置信度置为 0.5。若其为三字或三字以上词，置信度设为 0.65。如“蠢蠢欲动”、“张冠李戴”、“斗殴”、“蹂躏”等。同样的，之所以删除 Positive 的词，是因为包含诸如“建设”、“推动”、“带动”、“强化”等极性不明确的词语。
5) 若该极性词为非时间词的其它词性，则该词在上下文中的极性被指定为该词在褒贬义核心辞典中的极性。若其为双字词，置信度置为 0.45。若其为三字或三字以上词，置信度设为 0.65。如 “乌烟瘴气”、“闲言碎语”、“大刀阔斧”、“大无畏”、“非法”、“悠悠”、“全新”、“恶性”等等。这些词中包含了区别词、状态词、处所词（“幕后”）、成语、代词（“各色”、“鄙人”）、连词（“不是”）等。由于此项中此类混杂，故比其他各项略略降低了权重，但是多字词的权重仍然与其它保持一致。

上述结果与 task3 的结果汇总并按执行度排序得到了 task1 的结果。上述过程中，需要加以阐述的是：

1. 在词性为名词和形容词时仅选择了极性为 Negative 的词。
2. 之所以提升了三字或三字以上词的权重是基于如下考虑：三字词如“得人心”，“绊脚石”，“假猩猩”，三字以上的词如“威风凛凛”，“貌和神离”，“捕风捉影”，“令人发指”等较少出现歧义或多义的现象。通常情况下，多字词作为一种稳定的语言搭配，其含义或者背后的故事通常多余双字词或单子词。多字词较双字/单字词具有更好的夸领域通用性，多字词的极性相对较少地依赖于其修饰的对象。

但是，虽然三字或多字词具有更好的夸领域的特性，仍然存在一些意外。例如：“保

护伞”，“为子女的教育撑起保护伞”不等同与“恶势力的保护伞”；“小辫子”，“扎起了小辫子”也不等同于“抓住了某某的小辫子”。前者情况下，“保护伞”被××收录基于何种考虑尚且不清，而后者较容易出现，即通常的褒贬义词典倾向于收录其带有情感倾向的意向，忽略而其中性意向。

4.3. Task2 句子极性和置信度的计算

句子极性定义为句子中所有搭配（task3）或者搭配外的极性词（task1）的总和。与task3 中否定略微不同的是，否定的判断不仅局限在模式内部，也出现在模式/极性词的上下文中。

4.4. Task4 观点检索

在本任务中，我们首先检索得到包含该对象的文章，按照文章与话题的相关度对检索到的文章排序，然后从这些文章中找出实为评价该对象的文章，并对它的总体极性打分。

4.4.1. 信息检索及相关度排序

我们在信息检索时采用最简单的方式，直接查找包含检索对象的文章。但是，这样查找得到的结果中，有许多文章只是提及该检索对象，而不是对该检索对象评价。为此，我们想要借助从文章中查找与检索对象相关度较高的词汇，来辅助判断文章是否是对检索对象的提及，是否是对检索对象的评论，并据此给出相关度分数。

为了获得与检索对象相关度较高的词汇，我们采用 LDA 话题模型，从话题结果文件中获取与检索对象出现在同主题下的相关词汇。我们选取 GibbsLDA++作为 LDA 模型平台，将所有话题的检索文件汇总成单一数据集合，话题数选为 20 个。

我们根据下列公式为每篇文章打分：

$$\text{score} = (\text{Ns} + \alpha\text{Nr})\text{LogL}$$

其中，Ns 表示检索词在文章中出现的次数，Nr 表示相关词汇在文章中出现次数的总和，α表示相关词汇的权，L 表示文章的长度。

根据检索词与相关词汇的关系，我们为不同的检索词对应的相关词汇给予了不同的权重。对于诸如创业板（T306）、IPO（T310）等话题而言，其相关词较好地反映了文章与检索词的相关程度，为此，我们为α赋予了一个较高的正权。对于诸如存款准备金率(T314)、裸婚(T318)等，相关词汇的出现有限地反映了文章与检索词的相关程度，为此，我们为α赋予了一个较低的正权。对于诸如新浪（T301）、微博（T303）等话题，相关词汇的出现反而妨碍了文章与检索词相关程度的计算，此时，我们为α赋予了一个负值。

得到每篇文章的分数后，选取分数大于某阈值的文章作为新的话题下的检索文章集合输出到下一步计算中。该阈值设为原话题下索引文件的平均得分。我们对新的话题下检索文件的集合按得分数进行排序并对分数进行归一处理。

4.4.2. 检索文档的情感判断

对于检索文档的情感判断，我们简单地使用 Hownet 情感词表，对文档中出现的正/负向情感词进行简要的统计。如若正情感词和负情感词的词频都为 0，则视为无情感倾向，将其从检索文档集合中剔除。此外，若正情感词频为 0 或负情感词频为 0，则置文档的情感倾向为负或正向。若正情感词频和负情感词频都小于 3 且都不为 0，则置文档倾向为混合倾向。此外，若正情感词频不小于负情感词频的 M 倍或负向情感词频不小于正向情感词频的 M 倍，置文档极性问正或负（试验中设 M 为 2）。其余情况置文档情感倾向为混合倾向。之所以设定如上规则是因为我们将容忍 Hownet 中可能出现的差错。

5. 实验结果与分析

依据评测的结果，我们对实验结果及我们采用的方法做简要的分析。

5.1. Task3

很遗憾的是，我们在 task3 的实验中使用了错误的数据，错误地返回了 COAE2011_Corpus _All_Text 的结果,导致结果极差。由于错误的数据,导致精确率极低。由于数据的错误使用，再此不予汇报。

虽然结果数据不具有参考意义，但是观察实验数据，我们仍然能够发现使用模式抽取的缺点。例如，对于答案中的如“D00846.txt 惠普 4 是近期上市的一款热门机型，超薄的机身外观设计和出色的配置，让该机在众多机型之中脱颖而出。 惠普 4 超薄的 1 惠普 4 出色的 1”，我们只能获得“超薄的—机身”、“出色的—配置”这样的搭配关系，而无法将修饰目标延伸到“惠普 4”。又如，“D00846.txt 在外观方面惠普 41022 采用了高端笔记本流行的金属机身，应用高端系列金属蚀刻技术，雕刻出“城市流光”纹理，外形边角处理圆润，内部键盘周围以及触控板也采用圆滑设计，面除键盘外全面采用金属材质，高贵细节值得品触。 高贵细节 值得 1”，采用基于模式抽取的方法将抽取出“高贵—细节”这样的修饰—被修饰关系。对于答案，他对于人而言可能很容易理解，对于机器而言，为什么“高贵细节”不是修饰“惠普 41022”，而是被“值得”修饰呢？而对于我们的方法而言，比较容易保持一致性，即我们认为我们抽取出的修饰—被修饰关系也是有一定道理的。人工观察我们结果，从我们的初衷来看，我们的修饰与被修饰关系捕捉的还是比较理想的。

5.2. Task2

Task2 的结果，正如我们所料，是比较差的，略超平均值。这是由于我们采用的方法简单，而且我们对否定的情感逆转也计算过于简略。此外，在结果明细中，数码领域计算得到的精确率（44%），召回率（49%），F 值（44%）都远大于娱乐领域（分别为 16%，40%，

23%）和财经领域（12%，54%，20%），此外，我们的召回率颇高，但精确率不足。从后文可知，我们 task1 的结果颇好，也即在极性计算过程中，短语的极性计算应该是颇为准确的，只是简单的极性搭配权重相加忽略了否定、中心等问题，导致结果一般。

表 3. Task2 实验结果

Table3. Task2 Results

	宏平均					微平均				
	Precision	Recall	F1	P@1000	Raccuracy	Precision	Recall	F1	P@1000	Raccuracy
PKUICL	0.2297	0.4761	0.2913	0.3203	0.2752	0.2916	0.4790	0.3625	0.3203	0.4124
Median	0.2408	0.4060	0.2763	0.2902	0.2554	0.3153	0.4505	0.3579	0.2902	0.3632
Best	0.5347	0.7234	0.5414	0.532	0.4945	0.6544	0.7754	0.6396	0.532	0.6114

5.3. Task1

Task1 的结果颇好，可以从下表中看出。我们的结果逼近了最优的结果，远远的超过了平均结果。

表 4. Task1 实验结果

Table4. Task1 Results

	宏平均					微平均				
	Precision	Recall	F1	P@1000	raccuracy	Precision	Recall	F1	P@1000	raccuracy
PKUICL	0.6213	0.3733	0.1128	0.1733	0.1128	0.6213	0.3733	0.1128	0.1733	0.1128
Median	0.5713	0.3430	0.0947	0.1479	0.0948	0.5713	0.3379	0.0947	0.1475	0.0947
Best	0.6567	0.486	0.1136	0.1744	0.1136	0.6567	0.4681	0.1135	0.1744	0.1135

回顾我们在 Task1 中的计算方法，我们借鉴了 Task3 的输出，同时也是用了核心词典命中的方式。虽然 Task3 的评测失效，但是我们仍然可以在 Task1 中看到 Task3 中方法成功的影子。此外，精确率得益于良的核心词典，这证明了一个良的核心词典的正确性，也证明了我们词性过滤 Positive 极性词的正确性。我们猜测，这是因为一般词典对 Positive 的偏移，（如词典中词的个数：褒义词集合共得到 5223 个词，贬义词集合得到 3760 个词）。一般情况下，正向情感词可能有多个意向，词典往往因为其中一个意向而将其收录(如“*钢铁*制造业”、“*钢铁*一般的意志”)，而负向情感词较少出现这样的情况。

5.4. Task4

实验四的结果也是很不错的。我们可以看到，我们提交的第一组和第三组结果都超过了平均值，且第三组结果和第一组结果的最优值也逼近了最优结果。这个结果说明，我们采用话题模型辅助 Task4 的实现是有效的。

表 5. Task4 实验结果

Table5. Task4 Results

标识	MAP				宏平均				微平均			P@N	
	未插值 AP.MAP _REL	未插值 AP.MAP _senti	返回的文档 AP.MAP _REL	返回的文档 AP.MAP _senti	Macro _P	Macro _R	Macro _F1	Macro _Raccuracy	MICRO_P	MICRO_R	MICRO_F1	P@5	P@10
PKUICL1	0.079	0.069	0.989	0.645	0.526	0.028	0.053	0.109	0.575	0.092	0.159	0.980	0.925
PKUICL2	0.062	0.044	1.000	0.602	0.536	0.021	0.040	0.071	0.560	0.068	0.122	0.970	0.810
PKUICL3	0.084	0.072	0.982	0.657	0.562	0.030	0.057	0.115	0.557	0.099	0.167	0.970	0.915
MEDIAN	0.075	0.062	0.990	0.635	0.541	0.027	0.050	0.098	0.564	0.086	0.150	0.973	0.883
BEST	0.148	0.072	1.000	0.657	0.562	0.041	0.071	0.169	0.575	0.149	0.192	0.980	0.925

6. 结论与展望

在本次的评测中，我们在构建高质量核心词典外，针对度量形容词计算了被修饰词的正/负向期待分数；引入京东种子语料进行领域词的抽取；引入语义相似度等进行修饰词和被修饰词的同/近义词扩展；采用 LDA 模型进行辅助检索结果的生成。实验证明，质量良好的核心词典影响巨大，文章对度量形容词处理得当，同/近义词扩展有效，话题模型对检索辅助作用明显。同时我们也意识到，我们对 Task3 跨度较远的被修饰目标和修饰短语采用基于模式的方法抽取是不完善的，我们对 Task2 的处理也过于简略。我们将在今后的学习中，对上述两点给予更多的关注。

参考文献

[1]. Bo Pang &Lillian Lee. *A Sentimental Education: Sentiment Analysis Using SubjectivitySummarization Based on Minimum Cuts.* In Proceeding of ACL,2004.

[2]. Choi.Y.,Cardie,C.,Rilogg,E.,&Patwardhan,S. *Identifying sources of opinions with conditional random fields and extraction patterns*. In Proceeding of ENLP,2005.

[3]. DanushkaBollegala, David Weir, John Carroll. *Using multiple sources to construct a sentiment sensitive thesaurus for cross domain sentiment classification*. In Proceeding of ACL,2011.

[4]. Dave,K,Lawrence,S. &Pennock,D.M. *Mining the peanut gallery:opinion extraction and semantic classification of product reviews*. In Proceeding of the 12th international WWW conference, 2003.

[5]. Gamon,M. *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis*. In Proceedings of COLING,2004.

[6]. Hatzivassiloglou,V. &McKeown,K.R. *Predicting the semantic orientation of adjectives*.In Proceeding of the 8th conference on European chapter of the association for computational

linguistics. 1997.

[7]. SajibDasagupta& Vincent Ng. *Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification*.In Proceedings of ACL,2009.

[8]. Turney,P.D. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. In Proceedings of ACL,2002.

[9]. Yu,Hong&VasileiosHatzivassiloglou. *Towards answering opinion and identifying the polarity of opinion sentences*. In Proceedings of EMNLP, 2003.

[10]. YunfangWu,Miaomiao Wen. *Disambiguating dynamic sentiment ambiguous adjectives*. In Proceedings of COLING, 2010.

[11]. 石静 ，邱立坤，王菲，吴云芳. 相似词获取的集成方法. CNCCL, 2011.

规则与统计相结合的观点极性分类与观点抽取

史兴*，房磊*，何蔼[$]，杨文婕[#]，黄民烈*，朱小燕*

[*]清华大学计算机系，北京 100084

[$]北京航空航天大学，[#]北京邮电大学

E-mail: aihuang@tsinghua.edu.cn

摘　要：文本倾向性分析是自然语言处理中的一个热点问题，正确分析文本的倾向性在引导用户决策等方面具有较高的应用和商业价值。本文综合规则和基于 Deep Learning 的跨领域情感分类方法，实现句子级别的情感分类。针对文本中的属性词-观点词对，融合语法词性规则，互信息等统计信息以及机器学习的方法，对属性词-观点词对进行抽取和极性分类。

关键词：跨领域情感分析；观点分类；观点倾向

Opinion Polarity Classification and Opinion Extraction by Combining Patterns and Statistics

Xing Shi*, Lei Fang*, Ai He[$], Wenjie Yang[#], Minlie Huang*

Dept. Of Computer Science and Technology, Tsinghua University, Beijing 100084

E-mail: aihuang@tsinghua.edu.cn

Abstract: Recently, opinion and review mining come to be a hot topic in natural language processing. Accurate sentiment analysis plays an important role in guiding decision making and business intelligence. In this paper, we integrate deep learning based cross-domain sentiment classification with rule-based methods to classify the opinion polarity at sentence level. Meanwhile, we employ grammar patterns and mutual information to extract opinion-aspect pairs, and predict polarity of pairs using various information. We present preliminary analysis to explain the advantages and disadvantages of our approach.

Keywords: cross domain sentiment classification, opinion mining, deep learning, opinion polarity

1　引言

随着网络的迅猛发展，万维网成为大量信息的载体。论坛，旅游，购物，电子商务等网站的蓬勃发展提供了海量用户数据，包含很多评论及倾向性的文本，对此，通用搜索引擎缺乏对评论数据及文本倾向性的分类。对这些信息进行情感分析及统计，是正确引导其他用户决策的关键，具有较高的应用和商业价值。

针对句子级别的情感分类，本文融合跨领域和基于规则的方法，提出句子级别的情感分类模型。其中，规则的方法主要基于词典，虽然有比较大的局限性，但是区分观点句和非观点句等方面的优势较为明显。综合两种方法，给出句子级别的情感倾向，实验结果表明，该分类模型具有一定的应用前景。

文本中属性词与观点词对的抽取是实现文档摘要、属性分析等研究的基础工作，准确的判断属性词与观点词是正确抽取的前提。本文综合使用了语法词性规则，互信息等统计学方法以及机器学习的方法对评论中的属性词-观点词对进行了抽取，同时利用了互联网评论中用户的评注信息及上下文转折词对属性词-观点词对进行了极性判断。实验结果表明，本文提出的属性词-观点词对的抽取方法具有较高的准确性，在综合考虑属性词-观点词对和文本倾向方面具有较高的性能。

2 句子级别的情感分类

清华大学中文评论挖掘系统 cReviewMiner [1]抓取了近 1000 万数码产品，酒店和餐馆的点评数据，并且提供标注过的部分餐馆点评数据，本文通过已经标注的这部分数据，采用跨领域的情感分类方法和规则的方法进行句子级别的情感分类。

2.1 跨领域情感分类

跨领域的情感分类需要模型具有较高的领域适应性，即对于不同领域，训练数据和测试数据从不同的分布中采样。假设有两个数据集合，源领域 S 和目标领域 T，其中在源领域 S 上有标注的训练数据。学习的过程需要将领域 S 学习到的知识更好的迁移到领域 T。在这里，本文采用 Deep Learning[2] 的方法进行分类。Deep Learning 算法能够学习到原始输入和目标之间的一些中间概念，简单来说，中间概念就是对原始输入特征的一种重新表示。例如，中间概念可以间接地表征了产品的质量，价格及售后服务等，这当中的一些概念在很多领域上具有通用性。文[3]认为，Deep Learning 学习得到的中间概念能够更好的实现领域的迁移，相同的词或者词组可能在各个领域上被使用，来刻画这些中间概念。同时，Deep Learning 能够采用无监督学习的方法学习出这些中间概念，挖掘出数据之间的隐藏信息。2006 年，Hinton 等人提出了 Deep Belief Networks（DBN）模型[4]。DBN 不同于简单的多层神经网络，DBN 采用无监督学习的方法（有约束的波尔兹曼机[2]）对网络中的每一层进行训练。后续的研究者提出了自动编码机[2]，去噪自动编码机[5]等算法模型，奠定了 Deep Learning 理论的应用基础。

2.1.1 Deep Learning 结构

Deep Learning 结构上和多层神经网路类似。在 2006 年之前，深层结构相比于层次较少的神经网络并没有取得较好的结果。同时，多层神经网络的底层在优化方面也遇到一些问题。Deep Learning 相比于传统的多层神经网络，其最高两层的联合概率分布构成有约束的波尔兹曼机。对于有约束的波尔兹曼机，同一层之间的节点没有边相连，只有不同层之间的边相连，使得联合概率的分布较为方便的分解。 多层次级联，实现了由底层特征向高层特征的转变，从而学习出更能表示数据特征的高层次概念。

Deep Learning 采用具有稀疏的分布表示的特征，而自动编码机算法能够无监督地学习出这些特征。自动编码机算法的基本想法是对于给定的输入 x 进行特征的编码，通过非线性变换，得到编码后的结果 c(x)，然后反过来通过译码得到 x’， 标准就是译码得到的 x’ 与 x 之间的误差最小。优化过程采用梯度算法调节变换矩阵和偏置的参数。自动编码机算法存在的一个问题是，如果没有其他限制标准，当编码得到输出的维度大于输入的维度时，学习得到的编码函数可能只是简单的恒等函数，部分位只是简单的复制输入，多出来的编码的位上的值大多是 0。随机梯度下降方法能够避免这种情况的出现，得到较好的表示（从分类结果的角度）。后来，研究人员提出采用去噪的自动编码机，去噪的自动编码机不仅避免上述恒等函数的情况，同时还能学习出特征之间的统计信息和结构信息。将去噪自动编码机以栈的方式堆叠起来，就得到栈式去噪编码机，通过对每一层采用随机梯度下降优化每一层的参数，从而实现了从底层特征到高层特征的步步转化。

2.1.2 Deep Learning 应用于跨领域分类

Deep Learning 能够学习出具有分布式特征的概念表示，如果学习出来的概念能够较好

的体现数据变化的因素，且具有一定的通用性，那么这些概念就可以用来对不同领域的数据进行分类。因此，文[3]认为，Deep Learning 可以应用在跨领域的情感分类上。步骤分为两步：第一步，采用栈式去噪自动编码机，实现底层的文本特征到高层概念的转换，训练的数据为各个领域的情感句子，这部分为无监督学习；第二步，选取标注数据，用学习得到的栈式去噪编码机进行编码，获取高层次的概念，这些概念较为通用，然后采用支持向量机训练，训练得到的分类器就具有领域一般性。对于新的测试数据，经过第一步编码之后，用第二步得到的分类器进行分类就可以得到分类的结果。

2.2 基于规则的情感分类方法

本文在词典基础上，同时采用基于规则的句子情感分类。对于一些特定分类，运用规则进行分类能有很好的效果。本文先构建词典，并与用词典对句子进行标注，最后再运用规则进行极性判断。

2.2.1 词典结构

一般来讲，影响句子极性的词分为 4 类：褒义词、贬义词、否定词和转折词。首先，从给定的语料库中总结出跨领域的褒义词词典和贬义词词典。其次，由于取反词汇和转折词汇的固定和少量的特点，也可以很容易地总结出取反词词典和转折词词典。

2.2.2 规则假设

一个句子由若干具有褒义、贬义或中性的意群构成。一般认为一个意群对外只显示一个极性。具体来说，句子中的每个逗号、分号看作是意群的分隔符。另外，为了让一个一群分句只体现出一种极性，句子中出现的转折词也可作为意群的分隔符。需要注意的问题是句子中的转折词可能是引导一个分句的，也可能是就在一个分句中。例如，句子“这个手机用起来真的不错，但是就是价格太贵。”和句子“这个手机屏幕大但是耗电。”，第一个句子就是转折词“但是”引导一个分句，而第二个句子就是转折词“但是”就在分句中，而不引导分句。同时考虑词性对词汇的影响。能够体现出褒义贬义的词不可能是介词，连词或者专有名词；而取反词、转折词也只可能是介词，连词、动词或者副词。

2.2.3 规则内容

判断意群对外显示的极性，需要看句中出现的褒、贬义词和取反词的位置关系。

对于一个取反词，看它前面一个取反词是否有效且两词距离在适当距离范围之内，若是则为“双重否定”效果，两个取反词的取反效果消失；否则，本词取反效果对后文有效。但若这个取反词是分句中最后一个取反词，看它前面一个词是否为褒义或者贬义词且两次距离在适当范围之内且两次中间的词汇的词性是一些虚词，则认为这个是否定后置的情况，前面一词褒贬义被反转。

对一个褒义词或者贬义词，看他前面是否含有取反词，且两次距离在适当范围之内，若是取反有效，褒贬义词极性反转，否则取反无效。

本文定义两种规则：一个称之为弱规则，主要判断非观点句和混合观点句，其分类为褒贬义的句子作为中间结果临时保存；另一个为强规则，具体为只出现一种极性意群的句子中，如果这种极性意群个数超过 2 个或者在某个此极性意群分句中出现该极性的词数超过两个，则认为整个句子体现该极性。强分类器与跨领域的分类器融合，处理跨领域分类器与弱规则不一致的情况，通过三种分类器的投票结果决定句子的总体极性。

2.3 总体的分类流程

综合跨领域分类和规则的分类方法，对句子基本的总体流程如图 1，首先采用弱规则

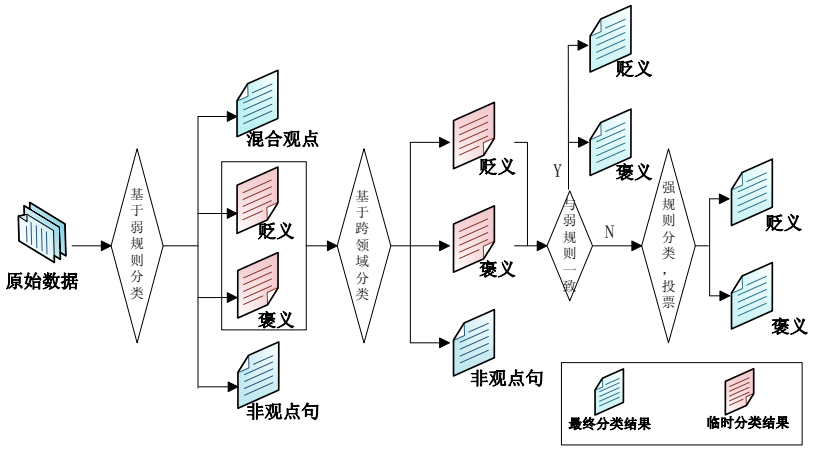

图 1 句子级别情感分类流程

对原始数据进行分类，用规则确定非观点句和混合观点句；然后对观点句（褒、贬义）采用跨领域分类器分类，确定非观点句，对于观点句，与弱规则一致则确定其观点倾向，不一致的句子用强分类器分类，通过中间结果的投票确定最终句子的分类结果。

3 评论中属性词与观点词对的抽取及极性标注

3.1 属性词-观点词对的抽取

我们对语料中出现的观点词和属性词进行了观察分析，得出了属性词和观点词具有的如下的特征：第一，属性词和观点词在句子内部有一点的语法关系上的连结。比如，在“苹果手机具有很不错的外观”一句中，“不错”是观点词词，“外观”是观点词的评价主体，也即属性词。这一对属性词和观点词在语法上属于修饰关系；第二，属性词和观点词具有一定的搭配关系。比如，在“屏幕很大，性价比很高”中，“屏幕”和“大”搭配，“性价比”和“高”搭配，而很少出现“屏幕很高”这样的说法。同时，这种搭配关系是同人们的语言习惯以及语料所属的领域相关。例如，在手机领域，可能会出现这样的语料：“诺基亚很呆板，苹果却很时尚”，而在餐饮领域，同样的“苹果”一次，可能会出现在“他们家的苹果很美味”这样的语句中。第三，属性词和观点词都具有各自的词性特点。一般地，观点词多为形容词或形容词短语等这类具有修饰关系的词组，而属性词则多为名词或者名词词组。但我们的实验并不仅仅限于上述两种词性的考虑。基于上面三点分析，我们将属性词和观点对抽取系统的流程设计如图 2 所示：一，对原始的语料利用一些现有的自然语言处理工具进行预处理，包括词性标注和语法关系的提取；二，利用语法关系和词性限制，对属性词-观点词对进行预抽取，并生成相应的领域相关的属性词-观点词搭配表；三，根据属性词-观点词搭配表以及其他额外的词典资源，通过查表，查字典的方式进行第二遍的属性词-观点词对的抽取，结合前一步生成的属性词-观点词表，生成一个如下方式表示的三元组的表：<属性词，观点词，句子>；四，计算评测语料、三元组表的相关统计信息，为

下一步的过滤做准备；五，对是否满足搭配关系，以上一步计算出来的多项特征为基础，分别采用多项特征综合评分和 SVM 的方法，对第三步中产生的三元组进行过滤；六，以互信息等度量条件对三元组进行进一步的过滤，等到最终的三元组的集合，完成属性词-观点词对的抽取工作。

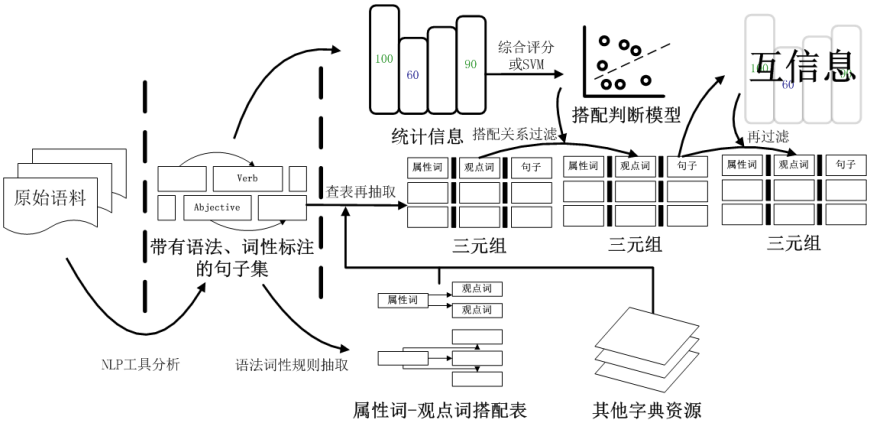

图 2 属性词-观点词对抽取主要流程

3.1.1 基于语法词性关系的属性词-观点词对抽取

对于语法关系的使用，我们选择利用依存文法（由法国语言学家 L.Tesniere 在其著作《结构句法基础》（1959 年）中提出）。依存文法通过分析一个句子内部语言单位成分的依存关系来揭示句子的句法结构。具体到属性词和观点词之间存在的依存关系，主要可以分为如下两类：第一，属性词和观点词存在直接的依存关系。比如，在“性价比很高”这句话中，“性价比”和“高”直接存在直接的主谓关系（在 Stanford Parser 中判定为 nsubj 关系）。第二，属性词和观点词存在间接的依存关系。即二者共同与第三个语言单位存在依存关系。比如，在“11 寸的体型确实是同类产品中最小巧的”这句话中，“体型”和“小巧”属于属性-词观点词对。二者并没有直接的依存关系，而是共同与“是”存在依存关系：“体型”与“是”构成了主谓关系（在 Stanford Parser 中判定为 nsubj 关系），“是”与“小巧”构成了表语关系（在 Stanford Parser 中判定为 attr 关系）。在二个语言单位间没有直接依存关系，或者直接依存关系不属于属性词和观点词间的依存关系，可以利用间接依存关系。在我们的实验中，我们采用了 Stanford Parser 的中文自然语言处理工具来分析语料的依存文法。

由于现有的中文语言处理工具所得到的句子依存关系中，会出现不可避免的错误。所以仅仅依靠依存关系来提取属性词-观点词对可能同时带来很大的噪音。为此，我们在利用依存关系的同时，增加了对词语词性的限制。在先前的其他学者的工作中，往往将观点词限定为形容词或者形容词词语，而属性词则限定为名词或名词词组。我们并没有采用这样的限制，而是通过分析原始语料，对观点词和属性词的词性选择范围进行了更多的扩展。对于词性的标注，我们综合了其他中文分词工具和 Stanford Parser 中文自然语言处理工具的词性标注结果。这样，一个属性词-观点词对的抽取规则（见表 1）就可以用如下的三元组来表示<依存关系，属性词词性，观点词词性>。在对每个句子进行观点词属性词抽取之

后，我们对观点词和属性词进行合并，便得到了一个领域相关的属性词-观点词搭配表。

表 1 属性词-观点词对的抽取规则

依存关系	属性词词性	观点词词性	示例
rcmod	l/n/v/vn	a/b/z	具有/v 良好/a 的/u 抗/v 盐水/n 腐蚀性/n rcmod(腐蚀性，良好)
nsubj	j/l/n/nr/nx/v/vn	a/b/z	市场/n 还是/c 相当/d 平淡/a nsubj（市场，平淡）
amod	n/vn	a/z	可/v 播放/v 高清/a 视频/n 文件/n amod（高清，视频）
assmod	j/l/n/vn/v	a/b/z	加上/v 精致/a 的/u 不锈钢/n 刀头/n assmod(精致，刀头）
nn	j/l/n/vn/v	a/b/z	连/n 上/f 电视/n 欣赏/v 更/d 清晰/a 的/u 动人/a 画质/n 。/w nn(动人，画质）
vmod	j/l/n/vn/v	a/b/z	微微/d 扬起/v 的/u 飘逸/a 长发/n vmod（飘逸，长发）
Nsubj +attr	j/l/n/vn/v	a/b/z	这个/r 价格/n 对于/p 20x/nx 刻录机/n 来讲/u 是/v 相当/d 便宜/a 的/u nsubj（价格，是） attr（是，便宜）

3.1.2 基于候选词典的属性词观点词对再抽取

仅仅利用语法词性关系来进行属性词-观点词的提取，由于依存语法判断的准确性较低，可能会导致如下问题：一，部分句子由于句子太长或者其他原因，使得依存语法分析失败；二，部分句子的依存语法分析出错，使得一些属性词-观点词对之间的关系并不满足表一种所示的规则，同时，满足了表一中所示规则的词语对可能并非是属性词-观点词对。上述两个问题均降低了召回率，第二个问题降低了提取的准确率。为了解决召回率的问题，我们变采用了本节所介绍的观点词再提取技术。而准确率的提升，则留到后面的搭配关系的认定和利用互信息过滤来解决。

我们的属性词观点词再抽取的方法如下：首先，利用上一步生成的属性词和观点词搭配表，生成一个观点词表和一个属性词表；然后，使用外部的一些已有的字典对属性词和观点词进行扩展。在我们的实验中，我们使用了外部的正负面评价词各约 3000 个，来对我们的观点词表进行了扩充；接着，利用查字典的方式在每个句子中去匹配属性词和观点词表中的词，作为候选属性词-观点词对。在这一步中，我们默认了如下的规则，即属性词只可能与其最近的观点词进行搭配，从而减少了同一句中不同的属性词和观点词所可能出现的组合情况。这样，我们生成了一个<属性词，观点词，句子>的三元组的集合。我们采用这样的方法可能提高 recall 的原因是因为纯粹的匹配查找方法突破了原句的语法限制，使得那些不符合语法抽取规则的属性词-观点词对同样能提取出来。但是由于相同的原因，此方法带来了很多的噪音。

3.1.3 三元组的过滤

在上一步过程中，得到的三元组表存在有很大的噪音。这样的噪音可能有三个原因：一是三元组中的属性词和观点词不满足搭配关系，比如，在三元组<“性价比”,“粗”,“这个手指一般粗的播放器性价比还不错”>中，“性价比”和“粗”的搭配便不恰当。二是是提取出来的属性词跟所在的领域没有关系，或者说关系不大。比如，在“大卖场里面的手机质量都很不错”一句中，“卖场”和“大”是一对搭配合理的属性词-观点词对，但是由于卖场并不算电子领域中的一个主要的属性。三是提取出来的观点词对属性词没有明显的评价关系，并不能表明属性的好或者差。比如，“有限公司”这一词组中，“有限”和“公司”是一对合理的搭配，但是“有限”一词仅仅是对公司类型的一个描述，并没有对该公司做出评价。为此，我们将从搭配关系，领域相关性以及观点词是否具有倾向性三个方面来对三元组进行过滤。

3.1.3.1 搭配关系的评价

对于不满足搭配关系的情况，可能会有如下两种：一是提取出来的属性词和观点词本身不满足搭配关系，比如“性价比”和“粗”；另一种情况是提取出来的属性词和观点词本身满足搭配关系，但是在句子中并没有搭配关系。比如，在“质量很高而且性价比也不错”一句中，“性价比”和“高”是一组良好的搭配，但是在句子中他们并没有修饰关系，“高”描述的是“质量”而非“性价比”。对于前一种情况，我们将采用统计的方法对两个词的搭配关系进行考虑；对于后者，我们则根据句子中的具体情况来进行判断。

孙茂松在《汉语搭配定量分析初探》中给出的关于搭配的性质：搭配是重复出现的；搭配是任意的；搭配通常是具有一定结构的；搭配是领域相关的。根绝上述性质，我们对在依据句法词性规则提取出来的属性词-观点词搭配表上，进行了统计计算，分别计算了如下的统计量（见表 2）：一个属性词-观点词对出现的次数 NOccur；同一对属性词-观点词之

表 2 判断搭配关系所用特征量

变量名称	特征意义
Noccur	一个属性词-观点词对在由语法词性提取出来的属性词-观点词表中出现的次数
rel[]	$\mathrm{NOccur}=\sum_{i=1}^{7}\mathrm{rel}[i]$, 同一对属性词-观点词之间的不同依存关系的频次
loc[]	$\mathrm{NOccur}=\sum_{i=-5}^{5}\mathrm{loc}[i]$，观点词相对于属性词的位置的频次
iMax	$\mathrm{iMax}=\{i\|\mathrm{loc}[i]\geq\mathrm{loc}[j],i\neq j\}$，loc[]数组中最大值的下标
relation	属性词和观点词之间是否存在直接的依存关系,没有则等于 null
foDistance	属性词和观点词之间的距离
frDistance	$\lvert\mathrm{foDistance}-\mathrm{iMax}\rvert$,观点词相对于属性词的位置与 iMax 的差值
flag[]	观点词与属性词之间是否存在标点符号，其间是否间隔介词，其间是否间隔动词

间的不同依存关系的频次 rel[i]（i=1,2,3……7），分别对应 7 中依存关系的出现次数，理论上，$\mathrm{NOccur}=\sum_{i=1}^{7}\mathrm{rel}[i]$；观点词相对于属性词的位置的频次 loc[i]（i=-5,-4,-3,-2,-1,1,2,3,4,5），理论上，$\mathrm{NOccur}=\sum_{i=-5}^{5}\mathrm{loc}[i]$。一般的，两个词如果满足搭配关系，那么他们之间的相对距离的分布图会出现尖峰，尖峰的位置就是观点词最可能出现的位置，loc[]数组中最大值的下标 iMax。上述特征用于将用于判断属性词和观点词再没有上下文的情况下是否搭配一

致。对于每个句子中具体的搭配关系的判断，我们计算了如下特征量：属性词和观点词之间是否存在直接的依存关系 relation；属性词与观点词的距离 foDistance；观点词相对于属性词的位置与 iMax 的差值 frDistance，其值越小，说明观点词距最可能出现的单词的距离越近，其越可能成为一个合适的搭配；三个标志位 flag[]：观点词与属性词之间是否存在标点符号，其间是否间隔介词，其间是否间隔动词。

现在，对于一个三元组，我们有 8 个特征值可以用来判断属性词和观点词是否满足搭配关系。在实验中，我们分别采用了 SVM 的方法和特征值加权评分这两种方法来判断某一三元组是否满足搭配的要求。在使用 SVM 的方法前，我们首先人工标注了正确搭配和错误搭配的三元组各 300 个用来训练 SVM 分类器。而在特征值加权评分的方法中，我们使用了公式 1 进行加权评分：

$$\mathrm{Score} = (\mathrm{OccurScore} + \mathrm{RelationScore}) * \mathrm{FrDistanceFactor} * \mathrm{FlagFactor} \quad (1)$$

其中，各个量的计算如下：OccurScore 和 RelationScore 分别为 Noccur 和 relation 的分段函数，而 FrDistanceFactor 和 FlagFactor 则分别是惩罚因子，同 frDistance 和 flag 有关。对最后的 Score 取一定的阈值即可对搭配关系进行判定。

3.1.3.2 基于领域互互信息的过滤

由于三元组表中存在着一写于领域无关的属性词，比如，“公司”，“感觉”等。对于这类的噪音，我们将利用互信息来进行过滤。互信息（公式 2）可以很好地表示一个词 w 与领域 D 的相关度。

$$\mathrm{M}(\mathrm{w},\mathrm{D}) = \sum_{\mathrm{i}=0}^{\mathrm{n}} \log \frac{\mathrm{p}(\mathrm{w},\mathrm{D}(\mathrm{i}))}{\mathrm{p}(\mathrm{w}) * \mathrm{p}(\mathrm{D}(\mathrm{i}))} \quad (2)$$

其中，p(w)为词 w 在所有语料中出现的概率，p(D(i))为第 i 个领域出现的概率，等于第 i 个领域中的词的个数占所有语料中词的个数的比例，p(w, D(i))为词 w 在第 i 个领域中出现的次数处理所有语料中的单词数。按照上述公式计算得出互信息，对于那些互信息小于一定阈值的属性词 w，我们可以认为他同该领域 D 的关系较小，因此可以将包含 w 的三元组过滤掉。

3.1.4 观点词和属性词的扩展

三元组中的属性词和和观点词都是单个的单词（分词结果中的一个词），我们需要在原句中扩展出修饰观点词的副词，以及属性词前面的修饰词。对于观点词和属性词的修饰词，主要有以下这几个特征：第一，修饰词的词性多种多样，观点词的修饰词多为副词，而属性词的修饰词可能为名词，形容词，专有名词，英语缩写等等；第二，修饰词可能有多个，可能前面连续连个词均为修饰词。比如在词组“诺基亚 5233 手机”中，属性词“手机”前面就有“5233”，“诺基亚”两个词来修饰；第三，多个修饰词连续出现；第四，修饰词和被修饰词在较大量的语料中重复出现。

根据上面四个特征，我们的修饰词扩展系统设计如下：首先，在原始语料中计算相邻两个词的互信息。相邻两个词 w1,w2 的互信息计算公式如公式 3 所示。

$$\mathrm{M}(\mathrm{w1},\mathrm{w2}) = \sum_{\mathrm{i}=0}^{\mathrm{n}} \log \frac{\mathrm{p}(\mathrm{w1},\mathrm{w2})}{\mathrm{p}(\mathrm{w1}) * \mathrm{p}(\mathrm{w2})} \quad (3)$$

其中，p(w1,w2)为相邻两个词相邻出现的概率，p(w1)和 p(w2)则分别是 w1,w2 两个词出现的概率。然后，对于单词串[w1,w2,w3…wi]（其中 wi 是被修饰词），我们从最 w(i-1)

来逐个向前判断某个单词是否属于修饰词组中的单词。如果 M(w(j-1), wj)大于一定的阈值 a，同时 w(j-1), wj 满足修饰词的词性要求（属性词：a/b/z/n; 观点词：d），wj 已经属于修饰词组，那么 w(j-1)属于修饰词组。

3.1.5 属性词-观点词对的极性判断

对属性词和观点词极性，我们认为，一些观点词有确定性倾向，即无论他们与哪个属性词搭配，他们都表示固定的倾向，比如“不错”，总是表示正面倾向；另外，一些观点词的倾向性并非确定性的，对不同的属性词会有不同的倾向性判断。比如“高”，当同“性价比”搭配在一起时，表示正面倾向；而同“价格”搭配在一起时，则表示负面倾向。对于确定性的观点词的极性判断，我们将采用查字典的方法进行极性标注。而对于那些非确定性的观点词的极性判断，我们主要采用以下两种方法进行极性标注：一，我们将利用一些点评类网站，电子商务类网站的评论资源，进行极性的判断。这类评论资源中，有不少网站要求对产品的优点和缺点分别进行评论。这样，我们就相当于拥有了一个规模极大的人工标注的语料。其中,笔记本类评论约 4 万条,相机类评论约 3 万条,手机类评论约 14 万条,餐馆宾馆类评论约 120 万条。在这些优缺点分类的语料中，我们分别利用同样的方法抽样出属性词-观点词搭配表，同时记录下每一对属性词-观点词在优点分类的评论中出现的次数 NAdv 和在缺点评论中出现的次数 NDisadv。按照公式 4 算出一个倾向性的评分 polarity：

$$\text{polarity} = \begin{cases} \left(\frac{\text{NAdv}}{\text{NAdv} + \text{NDisadv}} - 0.5\right) * 2, & \text{NAdv} \geq \text{NDisadv} \\ \left(\frac{\text{NDisadv}}{\text{NAdv} + \text{NDisadv}} - 0.5\right) * 2, & \text{NAdv} < NDisadv \end{cases} \tag{4}$$

polarity 的取值范围为[-1,1]。二，利用上下文和句中的转折关系来进行极性判断。首先，用查字典的方法，对那些确定性的属性词-观点词对进行初始的极性赋值，如果为正面倾向，则 polarity=1，否则 polarity=-1。再利用上述的网站评论资源进行极性赋值。这时，仍会有一些属性词没有得到极性赋值，令其 polarity=0。然后，我们以一个句子 S 为单位进行分析，假设 S 是有 n 个小句（由逗号分开）组成，令 psub(i)(i=1 至 n)表示第 i 个小句的示性标注，初始值全部预设为 1。然后从 1 开始逐个对 n 个小句进行极性判断，如果第 i 个小句的第一个词为一个转折词，那么 psub(i)=-psub(i-1)。同时 S 中找到了 m 个观点词-属性词对，令 pfo(j)(j=1 至 m)表示第 j 个属性词-观点词对的示性标注。如果第 j 个属性词-观点词对出现在第 i 个小句中，那么 pfo(j)=psub(i)；如果第 i 个小句内部（并非第一个单词）出现了转折词，不妨设在转折词左边的属性词-观点词对为第 j 个，右边的为第 j+1 个，那么 pfo(j)=psub(i)，pfo(j+1)=-pfo(j)。这样，我们可以通公式 5 对一个属性词-观点词对的极性 polarity 进行计算：

$$\text{polarity(j)} = \frac{\sum_{i=1;i!=j}^{m} \text{pfo (i)} * \text{polarity (i)} * \text{weight (i,j)}}{\sum_{i=1;i!=j}^{m} \text{weight (i,j)}} \tag{5}$$

其中，weight(i, j)为第 i 个和第 j 个属性词-观点词对的权值，主要和其间的距离成反比。

在计算完一遍后，对同一个属性词-观点词对在不同句子中的 polarity 求平均值，作为新的 polarity。迭代计算后，便得到了所有属性词-观点词对的 polarity。然后根据一定的阈值，过滤掉那些没有明显倾向性的属性词-观点词对。

4 实验

4.1 句子级别的情感分类

为了提高结果的准确率，本文同时使用规则和跨领域情感分类的方法进行分类。跨领域的方法，我们选取清华大学中文评论挖掘系统[1]提供的大众点评网的数据作为标注数据，共有 1574 条褒义句，430 条贬义句，1379 条非观点句，数码、娱乐和金融领域各随机选择 2000 句作为未标注数据，训练采用栈式去噪编码机，提取出高层次的概念。特征采用词袋特征，过滤掉文档频率小于 3 的词，得到的特征总数为 5890。训练结束后，用得到已标注数据的编码训练支持向量机，得到跨领域分类器。

根据标注的结果和本文提出的基于规则和跨领域的方法，分析句子级别的情感分类，结果如表 3 所示。对照标注结果，本文提出的方法在识别混合观点类型的句子性能比较低。混合观点的句子在[1]提供的餐馆点评的数据中出现较少，训练跨领域分类时未考虑这部分句子，所以通过规则来判断。规则判断的结果基于词表和转折词等信息，这些信息还不能完全实现句子混合观点的分类。另外，本文提出的方法在识别非观点句方面有待提高。本文方法认为是褒义和贬义的句子中有近一半的句子被标注为非观点句，如何提高非观点句的识别，也是今后的工作方向。

表 3 句子级别情感分类结果

领域分类		Precision	recall	F1
Digital	褒义	0.53	0.72	0.61
	混合观点	0.11	0.26	0.16
	贬义	0.33	0.19	0.23
Entertainment	褒义	0.25	0.57	0.34
	混合观点	0.07	0.38	0.12
	贬义	0.17	0.17	0.17
Finance	褒义	0.12	0.58	0.19
	混合观点	0.03	0.26	0.04
	贬义	0.16	0.20	0.17

4.2 评论中属性词与观点词对的抽取及极性标注

本文在实验中运用不同的方法，得到 3 组结果，见表 4,5,6 所示。其中：run3 代表使用机器学习 SVM 的方法判断搭配关系的结果；run2 代表使用特征综合评分的方法判断搭配关系的结果；run1 代表在 run3 的结果的基础上再使用句子级判断的结果过滤后生成的结果。

表 4 评价对象正确

run-tag	P@1000	Precision	Recall	F1	Raccuracy
宏平均(run1)	0.071667	0.087983	0.057156	0.069296	0.057156
宏平均(run2)	0.068	0.080731	0.061293	0.069682	0.061293
宏平均(run3)	0.068	0.086528	0.058184	0.06958	0.058184

微平均(run1)	0.071667	0.137322	0.086892	0.106436	0.086892
微平均(run2)	0.068	0.121899	0.096184	0.107526	0.094998
微平均(run3)	0.068	0.135611	0.087485	0.106357	0.087485

表 5 评价短语正确

run-tag	P@1000	Precision	Recall	F1	Raccuracy
宏平均(run1)	0.064667	0.069691	0.045589	0.05512	0.045589
宏平均(run2)	0.062	0.06378	0.05395	0.058455	0.048318
宏平均(run3)	0.061667	0.073793	0.049768	0.059445	0.049768
微平均(run1)	0.064667	0.106077	0.067121	0.082218	0.067121
微平均(run2)	0.062	0.094087	0.074239	0.082993	0.073151
微平均(run3)	0.061667	0.10665	0.068802	0.083644	0.068802

表 6 评价对象、短语、极性都正确

run-tag	P@1000	Precision	Recall	F1	Raccuracy
宏平均(run1）	0.035667	0.035265	0.022829	0.027716	0.022829
宏平均(run2)	0.034333	0.033215	0.027482	0.030078	0.024922
宏平均(run3)	0.035667	0.03753	0.025044	0.030042	0.025044
微平均(run1)	0.035667	0.060459	0.038256	0.046861	0.038256
微平均(run2)	0.034333	0.053621	0.042309	0.047298	0.041815
微平均(run3)	0.035667	0.0616	0.039739	0.048312	0.039739

通过对标准答案的分析，发现出现错误的原因主要有以下几个方面：一、在观点词属性词扩展的部分上，扩展地比较粗糙，导致很多结果不同；二、由于 Stanford Parser 的限制，本文在属性词，观点词抽取的过程中限制了属性词和相应的观点词必须出现在同一小句中，这样导致一部分跨小句出现的属性词-观点词对没有被提取出来；三、由于规则的涵盖面不够广导致的错误。四、标准答案标注中存在一些非领域相关的词，比如在电子产品领域，出现了“这”,“大腿”一类的属性词；五、标准答案中出现了一个属性词-观点词对整体作为一个观点词去修饰某个属性词的情况，本文并没有考虑到。比如，在“手机运行稳定”一句中，属性词为“手机”，观点词则为“运行稳定”，这其实也是一个属性词-观点词对。以上几种情况，将会是今后进一步改进的角度。

参 考 文 献

[1] http://www.qanswers.net:1880/cReviewMiner/,清华大学中文评论挖掘系统

[2] Yoshua Bengio. Learning Deep Architectures for AI[J]. Foundations and Trends in Machine Learning，2009，Vol.2, No.1, pp1-127

[3] Xavier Glorot ，Antoine Bordes ，Yoshua Bengio. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach[A]. Proceedings of the Twenty-eighth International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[4] G.E. Hinton and R.R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks[J],

Science, 28 July 2006, Vol. 313. no. 5786, pp. 504 – 507

[5] P. Vincent, H. Larochelle Y. Bengio and P.A. Manzagol, Extracting and Composing Robust Features with Denoising Autoencoders[A], Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML‘08), pages 1096 - 1103, ACM, 2008.

[6] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, et al. Discriminative Reordering with Chinese Grammatical Relations Features[D]. Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation. 2009

[7] 孙茂松，黄昌宁，方捷. 汉语搭配定量分析初探[D]. 中国语文 1997 年第 1 期(总第 256 期). 1997

[8] Li Zhuang, Feng Jing, Xiao-Yan Zhu. Movie Review Mining and Summarization[D]. CIKM’06. 2006

[9] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, et al. Cross-Domain Sentiment Classification via Spectral Feature Alignment[D]. WWW’2010. 2010

[10] 李智超. 面向互联网评论的情感资源构建及应用研究. 清华大学. 2011

基于多知识源融合和多分类器表决的中文观点分析*

徐睿峰 王亚伟 徐军 张玥 郑海清 桂林 叶璐
哈尔滨工业大学深圳研究生院网络环境智能计算重点实验室，深圳，518055
E-mail: xuruifeng@hitsz.edu.cn

摘　要：本文介绍了哈尔滨工业大学深圳研究生院参加 2011 年第三届中文倾向性分析评测中的任务 1(领域观点词的抽取与极性判别)和任务 2（中文观点句抽取和极性判别）的系统实现。在任务 1 中，使用了高质量的情感词语知识库作为核心，利用核心观点词的形态变化、连词和并列结构、同义词扩展、邻接词语共现强度等特征对观点词集进行扩展，从文本中识别出新的观点词。在观点句抽取和极性判别任务中，首先采用了基于图的句子排序算法，对不同来源的观点句标注语料进行优选用于训练。通过对多个分类器判别结果进行集成，实现较高精度的观点句识别和极性判别。本系统在评测中取得的较好成绩显示了高质量的情感词语知识库和有效的分类器集成方法对观点分析的重要性。

关键词：中文观点分析 情感词语知识库 多分类器表决 句子排序算法

Chinese Opinion Analysis based on Multi Knowledge Integration and Multi Classifier Voting

Xu Ruifeng, Wang Yawei, Xu Jun, Zhang Yue, Zheng Haiqing, Gui Lin, Ye Lu
Key Lab of Network Oriented Intelligent Computing, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen 518055
E-mail: xuruifeng@hits.edu.cn

Abstract: This paper presents the works of Harbin Institute of Technology, Shenzhen Graduate School (HITSZ) for Task 1 (Domain opinion word extraction and polarity determination) and Task 2 (Chinese opinion sentence identification and polarity determination) in COAE 2011. In Task1, a high quality opinion word knowledge base is firstly applied as the core words. By utilizing features of morphological variant, conjunctions words and parallel structures, synonymy, and neighboring words, the core opinion words are expanded to recognize new opinion words in the text. In Task 2, a graph-based sentence ranking algorithm is firstly applied to rank the annotated opinion examples from several corpora and identify good examples for classifier training. By integrating the output from multi opinion sentence classifier, a good performance on opinion sentence identification and polarity determination is achieved. The encouraging results of this system shows the importance of high quality opinion word knowledge base and effective integration method for multi classifier in Chinese opinion analysis.

Keywords: Chinese Opinion Analysis, Opinion Word Knowledge Base, Multi Classifier Voting, Sentence Ranking Algorithm

1　简介

随着互联网技术的高速发展，网络上的信息出现了爆炸性的增长。这种增长使得人们可以有机会从更多信息源了解、获知他人的信息、想法、态度和意见。而这些主观性的情感和评价，无法通过传统的基于关键词和自动索引的信息检索技术来获取。因此，在最近

*本研究由哈尔滨工业大学创新科技基金资助 HIT.NSFIR.2010124.

的十年里，基于文本的情感分析技术在世界范围引起了广泛关注。情感分析的主要任务就是挖掘和分析文本中的主观情感和评价[姚天昉 2008]。早期的情感分析研究主要集中在词语级别，包括识别新的情感词汇以及确定其极性。在此基础上，大量的文档级和句子级的粗粒度情感分析技术得到了充分研究。根据情感分类使用的知识来源和采用的分类策略，大多数的粗粒度情感分类技术可以分为基于词典和语言知识的、基于监督学习和基于非监督学习、半监督学习的三大类方法。基于监督学习的情感分类技术研究重点首先在于根据情感分析任务的需要和分类器特点，设计、选择和优化分类特征和选取适当的分类器。目前主要使用的分类器包括 Naïve Bayesian[Pang et al., 2002]、支持向量机(Support Vector Machine, SVM)、K-nearest 分类器[Wiebe et al., 2002]、最大熵(Maximum-Entropy, ME)以及条件随机场(Conditional Random Fields, CRFs)等[Pang et al., 2002; Dave et al., 2003; Matsumoto et al., 2005; 樊娜 2009]。在基于半监督学习方法中，基于 Bootstrapping [Riloff and Wiebe, 2003], Self-training [Wiebe and Riloff, 2001]，以及基于上下文情感转移趋势的学习方法[Xu et al., 2007]得到了较多研究。随着情感分析技术的不断深入，基于抽取的细粒度情感分析技术吸引了越来越多的注意。识别的细粒度情感表达信息包括情感表达子句、情感/意见的发出者、情感/意见的对象、情感的陈述、情感极性等等[Min and Park 2009]。

随着文本观点分析研究的深入，国内外的研究发现：（1）领域知识对文本倾向性分析有重要的影响，（2）上下文语境（Context）对倾向性判别至关重要[许洪波等 2009]。因此，在 COAE2011 评测中，强调提出了对来自电子、金融和娱乐三个领域的文本进行观点分析并进行比较[许洪波等 2011]。哈工大深圳研究生院（HITSZ）在本次评测中参加了任务 1(领域观点词的抽取与极性判别)和任务 2（中文观点句抽取和极性判别）。在任务 1 的实现中，使用了高质量的情感词语知识库作为核心，利用核心观点词的形态变化、连词和并列结构、同义词扩展、邻接词语共现强度、语句情感密度等多种特征对观点词集进行扩展，从文本中产生新的观点词候选。而后利用情感句判别结果和情感密度特征，对文本中出现的候选实例进行分析，从中筛选出新的观点词。在任务 2 观点句抽取和极性判别任务中，首先采用了基于图的句子排序算法，对不同来源的观点句标注语料进行优选用于训练。通过对多个分类器判别结果进行集成，实现较高精度的观点句识别和极性判别。本系统在所参加的两个任务中均取得了较好成绩。这一结果显示了高质量的情感词语知识库和有效的分类器集成方法对观点分析的重要性。

2 领域观点词的抽取和极性判别

2.1 情感词语知识库

在本课题的研究工作中，课题组构建了一个高质量的情感词语知识库。首先，课题组综合以下情感词语资源，建立一个大规模的情感词语词汇表：

1. 褒义词词典
2. 贬义词词典
3. HowNet 评价词典
4. HowNet 情感词典
5. 台湾大学情感词典
6. NTCIR 情感词典
7. CUHK 情感词典

这个情感词词汇表共包含正面情感词 10435 个，负面情感词 13513 个。而后，课题组人工

对上述词表进行了进一步标注。在人工去除了原始词表中错误词语后，将情感词语分为极性确定核心观点词(对一个评价对象进行主观评价的观点词且极性在不同上下文中稳定)、极性不确定核心观点词(对一个评价对象进行主观评价的观点词且极性在不同上下文中稳定)、情感词(表明评价人的主观情感和情绪的词语)、情感语境词(本身不带有倾向性，但其出现经常对下文的观点倾向或者主观情感具有一定的影响或者预示作用)四大类。通过使用多个情感标注语料库的实例，对上述分类结果进行统计验证，获得的分类后情感词语情况如下表所示。

表 1 情感词语分类标注结果及样例

情感词类别	极性	词汇个数	样例
极性确定核心观点词	正面	1287	优异、聪慧、完美
	负面	1317	恶劣、拙劣、肤浅
极性不确定核心观点词	不确定	928	圆滑、粗、滑
情感词	正面	1287	高兴、喜悦、钟爱
	负面	1317	深恨、惋惜、黯然
情感语境词	正面	5467	屹立、盛况、观摩
	负面	7236	冒牌货、窘况、糟蹋

最后，利用搭配抽取技术和人工验证，为极性不确定词建立了典型搭配表及相应的情感极性，最终建立了一个情感词语知识库。

2.2 领域极性词的抽取和极性判别

在任务 1 的实现中，首先利用情感词语知识库中的观点词对分词后的测试语料进行匹配，在电子、金融和语料三个领域各 2000 个文档的匹配情况见表 2 (分词系统以 ICTCLAS 为主、辅以 HKPolyU Segmnetor)。

表 2 核心观点词匹配情况

领域	极性	匹配数	样例
电子	正面	997	雅观、出类拔萃、神速
	负面	488	粗糙、粗制滥造
金融	正面	648	顺利、精明
	负面	311	偏激、混乱
娱乐	正面	1009	深刻、勤奋
	负面	510	碌碌无为、粗鲁

从核心观点词匹配结果可以看出，电子和娱乐领域，核心观点词匹配较高，金融领域则相对较低，而所有领域的匹配词汇数都少于 2000 词，因此有必要对领域观点词进行一定量的扩展。在极性词扩展研究中，采用了如下方法：

1. 利用词典中的长词对未分词文本进行直接匹配，以消除由于长词分词错误导致未匹配情况。如“有的放矢、众望所归”等。此类候选词往往具有非常高的正确率。
2. 从分词结果中抽取连词，使用 Pattern “评价词 A 连词/cc 评价词 B”，利用一端已匹配观点词去发现另一端候选观点词。典型连词包括 “和、与、且、又”等
3. 利用与已知观点词紧邻且高频共现，同时要求同词性。
4. 使用程度副词为线索，使用 Pattern “程度副词 评价词（限名词和形容词）”从测试

文本中发现候选观点词。

5. 利用已知观点词的各种词形变化，对测试文本进行匹配发现候选观点词。包括
   a) 增加重叠成分，例如由已知情感词“漂亮”，产生“漂漂亮亮”
   b) 减少多字观点词中一个字到两个字
   c) 两两交换四字观点词的次序
   d) 拼接两字情感词和已知情感词词尾
6. 基于同义词词林和 HowNet，对核心观点词进行同义词/近义词扩展，并对测试文本进行匹配以发现候选观点词
7. 基于情感词预示作用，利用情感词为线索，分析与之相邻的观点评价对，以获得候选观点词。
8. 基于评价词左右邻接词信息熵，利用与已知核心观点词相同的左右邻接词为特征，发现新的候选观点词。

表 3 给出了利用上述八种方法，在三个领域中的发现候选词的情况。

表 3 观点词扩展匹配情况

扩展方法	领域	极性	扩展观点词候选
1. 长词匹配	电子	正面	71
		负面	87
	金融	正面	37
		负面	50
	娱乐	正面	81
		负面	88
2. 连词扩展	电子	正面	21
		负面	17
	金融	正面	9
		负面	16
	娱乐	正面	27
		负面	14
3. 邻接高频词	电子	正面	27
		负面	11
	金融	正面	16
		负面	14
	娱乐	正面	25
		负面	14
4. 程度副词	电子	正面	37
		负面	26
	金融	正面	14
		负面	8
	娱乐	正面	20
		负面	13
5. 词形变换	电子	正面	62

		负面	37
	金融	正面	31
		负面	19
	娱乐	正面	68
		负面	29
6. 同义词扩展	电子	正面	312
		负面	75
	金融	正面	176
		负面	67
	娱乐	正面	278
		负面	130
7. 情感词预示	电子	正面	57
		负面	16
	金融	正面	14
		负面	5
	娱乐	正面	96
		负面	37
8. 左右邻接词	电子	正面	21
		负面	14
	金融	正面	8
		负面	6
	娱乐	正面	22
		负面	13

需要指出的是，表 3 中扩展候选观点词均为相对核心观点词而言的。不同方法扩展出的新候选可能出现重复。

下一步，利用候选观点词句子中已知核心观点词倾向性、情感词倾向性、语境词倾向性及相对于候选观点词距离，采用下列公式粗略估算其极性。

$$plo(w)=\sum_{i}\lambda_1\cdot plo(w_{\text{核心观点词-}i})/d_{\text{核心观点词-}i}+\sum_{\text{j}}\lambda_2 plo(w_{\text{情感词-j}})/d_{\text{情感词-j}}+\sum_{k}\lambda_3\cdot plo(w_{\text{语境词-k}})/d_{\text{语境词-k}}$$

公式中，$\lambda_1,\lambda_2,\lambda_3$分别为对应于核心观点词、情感词和语境词对候选词极性的影响权重且$\lambda_1+\lambda_2+\lambda_3=1\ \lambda_1\prec\lambda_2\prec\lambda_3$；$d$ 则为当前核心观点词/情感词/语境词到候选词词语的距离，距离越大意味着其影响越小；*Plo()*函数取值则为对应词语的极性，*Plo()*=1 意味着该词语为正面，*Plo()*=-1 则意味着该词语为负面。

通过结合情感句极性判别结果和词性粗估算结果，最终判定句子中新观点词的极性。

最后，在生成任务 1 输出结果，利用情感句分析结果以及由句子中核心观点词倾向性、情感词倾向性、语境词倾向性共同组成的情感密度，挑选出情感最密集、可信度最高的句子作为情感词实例输出。

3 观点句的抽取和极性判别

3.1 观点句抽取和极性判别训练语料

为了保障观点词抽取和极性判别分类器的性能，课题组使用和建立了如下训练语料。

1. 电子类产品评价细粒度语料库，该语料是在[Xu et al. 2008]工作的基础上，进一步细化和补充形成的。共包含数码相机和手机的产品评价 1,700 篇，约 9,000 句的细粒度标注信息。
2. 电子类产品评价粗粒度语料。分别来自北京大学万小军和建立[Xu et al.2008]语料过程中获得的文档级粗粒度情感语料约 20,000 篇
3. 课题组成员徐军在 2010 年建立的金融和股票投资领域句子级情感标注语料
4. 课题组建立的小规模娱乐新闻相关领域的句子级标注语料库。

这些情感标注语料库将用于分类器训练。

3.2 基于图的 Sentence Rank 算法

由于人工标注训练语料集合，具有一定主观性。特别是使用短文档级别的粗粒度标注结果，利用句子与短文档极性应相同的假设，直接获得句子标注结果的过程中，可能带来一定数量的噪声文本，从而影响分类器的性能。这里，本文提出一种基于图的句子排序算法 SentenceRank，计算训练语料中句子间的相关性，去除掉潜在噪音数据，从而保障了分类器的性能。

在 SentenceRank 算法中，每一个句子被看成是图的一个顶点，句子间的关系看成是边，每类训练语料集被用来构建一个多重图，算法 1 给出具体的算法描述，句子采用向量空间模型来表示。

算法 1　SentenceRank 算法

1） 对每一类训练语料集，构建多重图 G=<V,E>。其中 V 是顶点集合，训练集中每个句子看成是图中的一个顶点，E 是边集合，任意两个句子间存在相同的特征就增加一条边，两个顶点间的边数为公共特征数。

2） 删除图 G 中的孤立顶点，剩余的顶点数为 N，将 G 用矩阵 A 表示，其中

$$a_{ij} = k \qquad (1)$$

$k(k \geq 0)$为顶点i和顶点j之间的边数,$1 \leq i, j \leq N$

3） 计算概率转移矩阵 C，

$$c_{ij} = \frac{a_{ij}}{d_i} \qquad (2)$$

其中d_i表示顶点i的度数，$d_i = \sum_{j=1}^{N} a_{ij}$。

4） 由于矩阵 C 满足$\sum_{j=1}^{N} c_{ij}$=1,所以我们可以称 C 是一个马尔科夫链的随机矩阵。

5） 初始化顶点的初始概率分布$P_0(P_0(1)，P_0(2),\cdots,P_0(N))^T$，并且$\sum_{i=1}^{N} P_0 = 1$，这样就可以采用幂迭代方法计算出特征值为 1 的住特征向量，步骤如下：

在第 K 步转移之后顶点 j 的概率为

$$P_k(j) = \sum_{i=1}^{N} \alpha C_{ij} P_{k-1}(i) + (1-\alpha)\frac{1}{N} \qquad (3)$$

以矩阵的形式表示为$P_k = (1-\alpha)\frac{e}{N} + \alpha C^T P_{k-1}$,其中$\mathbf{e}$为全 1 的列向量，$\alpha$为衰减系数，取值范围为 0 到 1 之间，我们取为 0.9。

6）当$\|P_k - P_{k-1}\| < \varepsilon$（这里$\varepsilon$为 0.7）时停止迭代，特征向量的分量值作为对应顶点的重要程度值，然后对其进行排序，从训练集合中去除掉重要程度值比较低的句子。

通过去除重要性较低，同时有较大错误可能的样例，可以有效的改善分类器训练结果。

3.3 基于线性分类器的观点句分类器

此分类器将观点句分类分成观点句识别和倾向性判定两个子任务，并分别提取不同特征集合，最后通过一个简单的线性分类器进行判别，线性分类器的参数通过训练样本训练获得，取在训练集合上性能指标的综合值最高的取值作为最终的线性分类器系数。

特征集合

1）观点句识别

考虑到一个观点句通常包含：(1)观点持有者；(2)观点目标；(3)观点词；(4)观点主张词；(5)观点指示词等五个组件或者其中的部分组件。为了有效区分观点句和非观点句，本文从句子中抽取组件对应的词语特征作为识别线索，抽取的特征主要有如下四类：

(1) 观点主张词，能够指示观点表达或者言语事件的动词。典型的如“斥责”, “称赞”和“指出”等。

(2) 观点指示词，主要指的是相关的连词、副词和副词短语。包括：①转折连词，如“但是”、“尽管”等，其后句子/子句的情感倾向与前面相反。②表承接的连词，如“并且”、“而且”等，其后句子/子句的情感倾向与前面一致。③能够直接指示观点句情感倾向的副词或副词短语，如“不幸的是”。

(3) 观点词，具有明显的语义倾向的词语，表达正面、负面或中性的语义，并在观点表达中其决定作用。

(4) 观点持有者和目标，句子中包含的名实体和代词，他们都是观点持有者和目标的候选词。表 4 出了观点句识别任务的详细的特征列表。

表 4 观点句识别特征集（分类器 1）

特征编号	特征描述	样例
1	包含观点指示词的数目	遗憾的是、好在等 79 个
2	包含主张词的数目	强调、扬言、发表等 130 个
3	特殊句式	疑问句和感叹句
4	包含连词的数目	连词但是、然而，若是等
5	包含代词的个数	他，我们等
6	包含评价词的数目	
7	包含领域情感词的数目	

倾向性判定

考虑到文本的情感倾向由其所使用的词语及词语搭配来体现，某些词语的出现与否能够指示文本情感倾向。因此，这里通过线性分类器来结合文本中包含的情感词、副词、否定词语搭配，以及句子间的转折关系来计算句子的情感倾向值，从而来判断其倾向性。在对句子的倾向性识别任务中，提取了的特征如表 5 所示。

表 5 观点句极性判别特征集（分类器 1）

特征编号	特征描述
1	特殊句式，如疑问句和感叹句
2	包含评价词的数目
3	包含领域情感词的数目
4	否定词修饰的评价词和领域情感词的数目
5	连词修饰的评价词和领域情感词的数目

针对观点句识别和情感倾向性识别任务特点，使用简单线性判别函数$g(\mathbf{x}) = \mathbf{a}^t\mathbf{x} + b$来结合前文提出相关特征，其中，$\mathbf{x}$为句子的特征向量表示，$\mathbf{a}$为权向量，b 为阈值。$\mathbf{a}$ 和 b 的值通过训练语料训练获得。分类器在训练时，不断迭代调整 $\mathbf{a}$ 和 b 取值，最终的分类器是在训练语料集上的宏 F1 值最高者。对倾向性判定任务，由于其是 3 类分类，$g(\mathbf{x})$> +score 正面， 给 g(x)< -score 负面， 否则 mixed。算法的过程下所示：

begin initialize $\mathbf{a}$，b
  do i← i + 1
    调整a_i的大小
      do z← (z+1) mod n
        根据$\mathbf{ax^z}$+b 的大小设置句子 z 的标签L_z
until z=N-1
根据宏F1值选择最优的a_i
  until i 等于$\mathbf{a}$的长度
end

3.4 基于多特征、逐步求精的观点句分类器

此分类器为课题组在 COAE2009 任务中设计实现的观点句分类器的改进[徐睿峰 2009]。这个分类器融合了多种特征，采用逐步求精的策略。利用粗分类获得句子倾向性产生上下文和篇章特征，而后将上下文特征融入进来，对句子倾向性进行细分类。其主要流程如图 1 所示。

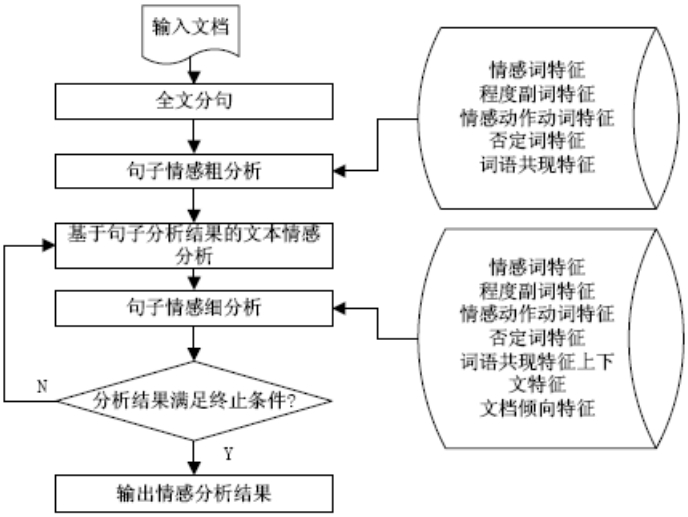

图 1 观点句分类器流程（分类器 2）

在粗分析中我们主要使用句内特征，包括：

（1）情感词特征。包括情感词类别，强度和隶属度。

（2）程度副词特征。即情感词与程度副词依存的比例

（3）情感动词特征。即情感词与情感动作动词共现的比例

（4）否定词特征。

（5）词语共现特征。

而在粗分析结果上，我们加入

(6) 句子级别情感状态转换概率这种上下文特征

(7) 全文的情感倾向特征

到细分析分类器。完成每次细分析之后，更新句子和文本的情感分析结果，当结果基本稳定以后就可以输出情感分析结果

类似的，可以建立极性判别分类器，在此不再重复。

3.5 多分类器分类结果的表决

这里我们采用了相对简单的表决策略。由于分类器 2 具有高精确率、相对低得召回率的特点，在观点句/非观点句分类器表决中，输出观点句基本为两分类器结果的并集，除非分类器 2 有非常高的可信度确定一个句子不是观点句。而在极性判别任务中，采用了对两个分类器加权求和的方法，并赋予分类器 2 较高的权重。

4 系统性能及结论

本系统在 Task1 取得的评测结果如表 6 所示。

表 6 HITSZ 任务 1 性能

标识	类别	Precision@1000	Precision	Recall	F1	Raccuracy
HITSZ	电子产品	0.572	0.3425	0.1025	0.1577	0.1025

HITSZ	影视娱乐	0.622	0.394	0.1194	0.1833	0.1194
HITSZ	金融证券	0.638	0.3905	0.1189	0.1822	0.1189
Median		0.57126	0.3430	0.0947	0.147593	0.0947
Best		0.674	0.6125	0.1194	0.1833	0.1194

系统的宏平均 Precision@1000 为 0.6107，居于评测平均水平的 0.57 到最好水平的 0.6567 之间，宏平均 Recall 为 0.1136 为评测的最好水平，最终的 F1 为评测的最好水平。

本系统在 Task2 取得的评测结果如表 7 所示。

表 7 HITSZ 任务 2 性能

标识	类别	Precision@1000	Precision	Recall	F1	Raccuracy
HITSZ	电子产品	0.8	0.729751	0.660324	0.693304	0.660324
HITSZ	影视娱乐	0.531	0.537607	0.511798	0.524385	0.511798
HITSZ	金融证券	0.265	0.336722	0.512573	0.406442	0.311412
Median		0.290183	0.240815	0.397946	0.276324	0.255445
Best		0.8	0.729751	0.798097	0.693304	0.660324

系统的宏平均 Precision 为 0.534，居于评测最好水平，Recall 为 0.7234，居于评测最好水平，最终的 F1 为 0.541，为评测最好水平。

对比各领域文本的分析结果，可以看到电子产品类结果的分析正确率远远优于金融证券类文本。一方面这个性能和文本本身的特点有关，电子产品类文档相对简单，情感密集且连续性好，系统实现的分类器 2 非常适宜处理此类文本，性能较好。而金融类文本观点稀疏，导致了分类器 2 的性能下降较多。此外，这个结果也充分显示了高质量训练数据对分类器的影响。电子领域的高质量、大规模细粒度标注的训练数据明显提高了分类器对此类文本的分析正确率。而金融类文本，由于用于训练的标注数据中将股票的高涨、下跌等视为评价句而评测答案中将此类作为事实句，也就意味着训练样本中具有较多的错误，最终导致此领域分类性能的明显下降。

本系统的实现和性能分析显示高质量的情感词语知识库、情感标注语料库和有效的分类器集成方法可以有效地改善观点分析系统的性能。

参 考 文 献

[1] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of WWW, pp. 519–528, 2003.

[2] S. Matsumoto, H. Takamura, and M. Okumura, “Sentiment classification using word sub-sequences and dependency sub-trees,” in Proceedings of PAKDD’05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2005

[3] H. J. Min and J.C. Park, “Toward finer-grained sentiment identification in product reviews through linguistic and ontological analyses”, in Proceedings of the ACL-IJCNLP 2009 Conference, pp. 169–172,Singapore, 2009

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in Proceedings of the Conference on Empirical Methods in

Natural Language Processing (EMNLP), pp. 79–86, 2002.

[5] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury, “Recognizing and organizing opinions expressed in the world press,” in Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 2003.

[6] R.F. Xu, K. F. Wong and Y. Q. Xia, "Opinmine – Opinion Analysis System by CUHK for NTCIR-6 Pilot Task," In Proceedings of NTCIR-6, pp. 50–35, Toyko, 2007.

[7] Ruifeng Xu, Yunqing Xia, Kam-Fai Wong and Wenjie Li, Opinion Annotation in On-Chinese Product Reviews, in Proceedings of Language Resources and Evaluation Conference 2008 (LREC 2008)

[8] 樊娜,蔡皖东,赵煜,李慧贤，“中文文本情感主题句分析与提取研究”，计算机应用, vol. 29, no. 4, pp. 1171-1176, 2009

[9] 许洪波等，"中文倾向性分析评测技术报告 2009"，第二届中文倾向性分析评测会议(COAE)，上海, 2009.

[10] 许洪波等，"第三届中文倾向性分析评测大纲"，第三届中文倾向性分析评测会议(COAE)，济南, 2011.

[11] 徐睿峰,揭春雨，“基于多种特征的情感/倾向分析技术”，第二届中文倾向性分析评测会议(COAE)，上海, 2009.

基于多策略的中文文本倾向分析技术

赵立东[1]，王素格[1,2]，王瑞波[3]，张鹏[1],吕云云[1]，薛宾[4]，李亚红[1]，张彩琴[4]

[1]山西大学计算机与信息技术学院，山西 太原，030006

[2]山西大学计算智能与中文信息处理教育部重点实验室，山西 太原，030006

[3]山西大学计算中心, 山西 太原，030006

[4]山西大学数学科学学院，山西 太原，030006

E-mail: wsg@sxu.edu.cn

摘 要：近年来，文本观点分析已成为自然语言处理领域的一个热点问题。在 COAE2011 设立的 4 个任务中，任务 1 要求从文本中抽取观点词，任务 2 从文本中抽取观点句并分析其极性，任务 3 要求在任务 2 的基础上抽取评价对象搭配及其情感倾向。本文介绍了课题组参加 COAE2011 任务 1－任务 3 所采用的具体方法与技术。任务 1，系统采用词汇资源与统计方法相结合抽取各领域的观点词。任务 2，系统采用观点词汇、机器学习与规则相结合的方法抽取文本中的观点句。任务 3，采用模板匹配与近邻法抽取了评价搭配，并利用所包含的观点词汇对其极性进行判别。评测结果表明，本文提出的方法和技术对任务 1 的各项指标、任务 2 的召回率以及任务 3 的精确率都取得了比较理想效果。

关键词：文本倾向分析，观点词、观点句抽取，评价搭配抽取，倾向性判别

Chinese Text Orientation Analysis Technique Based on Multi Strategy

Zhao Li-dong[1],Wang Su-ge[1,2], Wang Rui-bo[3], Zhang Peng[1], Lv Yun-yun[1], Xue Bing[4], Li Ya-hong[1], Zhang Cai-qin[4]

[1] School of Computer & Information Technology, Shanxi University, Taiyuan 030006

[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education

[3]Computer Centre, Shanxi University, Taiyuan 030006

[4]School of Mathematics Science, Shanxi University, Taiyuan 030006

E-mail: wsg@sxu.edu.cn

Abstract: Text opinion analysis has received a lot of attention from the natural language processing in recent years. Among the 4 tasks of COAE2011 the task 1 demands to extract opinion words from texts, and the task 2 demands to extract opinion sentences, the task 3 demands to extract evaluated collocations and they are demanded to identify their polarity. This paper introduces the methods and techniques which we adopted for the tasks 1- tasks 3 of COAE2011. For Task 1, proposed method integrates Words resources and statistical methods for extracting opinion words in multi domains. For Task 2, integrate method is proposed with opinion words, machine learning and rules for extracting opinion sentences. For Task 3, we extract the evaluated collocations by using templets method and near neighbour method. The evaluation results given by COAE2011 indicate that the proposed methods and technolgies obtained quite ideal effects in evaluated measures for the task1, recalls for the

---

作者简介：赵立东(1986－)，男，山西朔州人，硕士生，主要研究方向为信息检索。

基金项目：国家自然科学基金资助项目(60875040，60970014，61175067)；教育部高等学校博士点基金(200801080006)；山西省自然科学基金资助项目(2010011021-1)；山西省科技攻关项目(No. 20110321027-02)

tasks 2 and pricision for the task 3.
Key words: text analysis, opinion sentence extraction, orientation identifying, Fisher’s discriminant ratio, evaluated object

1 引言

近几年，随着 Web2.0 的迅猛发展，网络上出现越来越多的主观性文本，中文文本倾向性分析（观点和情感等）就是对主观性文本进行分析、处理、归纳和推理的过程[1]，它持续成为自然语言处理领域的研究热点问题之一。对于该问题的研究具有重要的现实意义，它可在市场预测分析、社会舆情分析、有害信息过滤、智能导购等诸多领域中有着广阔的应用空间和发展前景。

2011年第三届中文倾向性分析评测（COAE2011）在前两届中文倾向性分析评测的基础上，将继续致力于探索中文倾向性分析的新技术、新方法，推动中文倾向性分析理论的技术研究及应用。本届评测共设置了4个任务，从要素级、句子级和篇章级三种类型进行评测。与前几届评测最大的不同在于本次评测所给定的数据集分为电子产品、影视娱乐和金融证券三个领域，任务1－任务3分别对这三个领域进行倾向性挖掘。我们课题组参加了任务1－任务3。在这三个任务中，任务1是领域观点词抽取与极性判别；任务2是中文观点句抽取；任务3是评价搭配抽取。下面详细阐述了完成每个任务的技术与方法，并对评测结果进行了分析，最后，对未来的研究工作进行了展望。

2 领域观点词构建与极性判别

该任务要求从给定的三个领域数据集中分别抽取 2000 个观点词，并判断观点词的极性（即褒义、贬义）。大纲中特别强调指出，观点词是表达对外评价的褒贬观点词，不是表达人物自身情绪的倾向词，也不包括评价短词。因此，本系统采用词汇资源与统计方法相结合抽取各领域的观点词，其系统流程图如图 1。

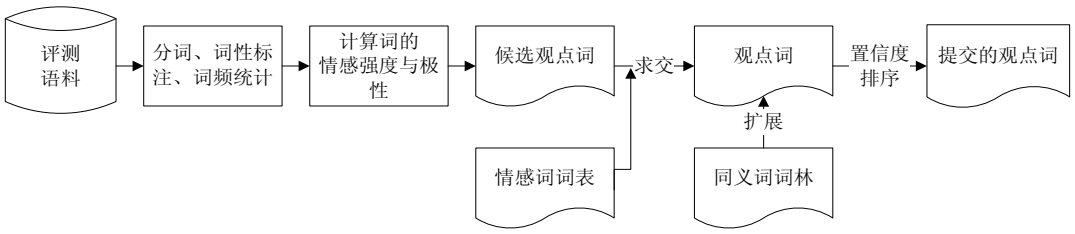

图 1 领域观点词构建流程图

（1）词汇资源

系统中用到的词汇资源包括 Hownet 评价词词表、自建的汉语情感词词表以及哈工大提供的同义词词林。

（2）词的情感强度计算方法

点互信息（PointWise Mutual Information,PMI）是信息论中度量两个随机变量间统计依

赖性的一种测度。利用 PMI 可以度量人们在使用某两个词的统计依赖性[2]。

一个词与褒义（贬义）基准词集的关联强度其计算公式如下：

$$S(word)=\sum_{word_1\in Words} PMI(word,word_1) \tag{1}$$

这里的$Words$为褒义或贬义基准词集。

一个词与褒义基准词集的关联强度越大，则该词倾向于褒义的程度越大，反之，它与贬义基准词集的关联强度越大，则其倾向于贬义的程度就越大。词的情感倾向强度$SO_PMI(word)$刻画了一个词更倾向于褒义还是贬义的程度。因此，一个词 word 情感强度$SO_PMI(word)$是由该词的褒义关联强度与贬义关联强度是差决定的。其计算公式如下：

$$SO\_PMI(word)=\sum_{pword\in Pwords} PMI(word,pword)-\sum_{nword\in Nwords} PMI(word,nword) \tag{1}$$

$$SO\_PMI(word)\approx \log_2\left(\frac{\prod_{pword\in Pwords} hits(word,pword)\ \prod_{nword\in Nwords} hits(nword)}{\prod_{pword\in Pwords} hits(pword)\ \prod_{nword\in Nwords} hits(word,nword)}\right) \tag{2}$$

（3）观点词极性判别

主要利用情感词表及其同义词扩展的极性进行判别，其中同义词扩展的极性是由其在情感词表中的同义词的极性决定的。

（4）置信度排序方法

根据图 1 领域观点词构建流程图，可获取三个领域各 2000 观点词。为了得到最可信的观点词，我们采用了两种置信度排序方法。

① 基于词汇词性置信度排序方法（WPC）：根据文献[3-4], 采用按形容词、动词和名词三种词性降序排序，相同词性的按 2.2 中计算出的词的情感倾向强度降序排列。

② 基于资源的置信度排序方法（RC）：该方法对选取的 2000 个观点词，按照 Hownet 正负评价词、汉语情感词词表、同义词词林中扩展词依次优先级设定为 3、2、1。同一优先级中按词的情感倾向强度降序排列。

3 中文观点句抽取与极性判别

此任务要求从每个领域的测试集中自动识别出所有观点句及其表达的观点的总体极性（褒义、贬义或混合观点），要求不仅能准确识别观点句，而且要尽可能找出所有的观点句。本系统采用规则与统计相结合的方法（RSM）。整个系统流程图如图 2。

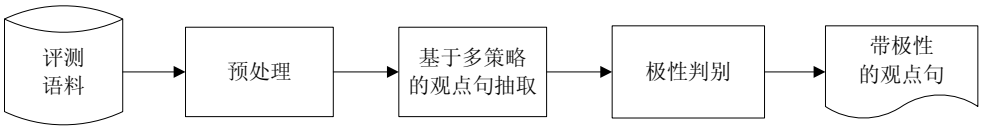

图 2 观点句抽取与极性判别流程图

3.1 预处理

为了提高系统抽取观点句的精度，对句子进行预处理。将明显不是观点句的句子采用如下规则将其去除。

Rule 1 如果句子中包含 WordSet1 中的词，则将该句去除。

WordSet1=“咨询电话，订购热线，联系方式，配送地址，QQ，…..”。

Rule 2 如果句子中包含“？”，则将该句去除。

Rule 3 如果句子中包含 WordSet2 中的词，则将该句去除。

WordSet2=“通知，公告，宣布，快讯，…..”。

Rule 4 如果句子为假设性和情感意愿的句子，则将该句去除。

3.2 观点句抽取

为了抽取观点句，系统首先采用以下方法抽取候选观点句集。

（1）基于情感词的候选观点句的抽取

对于文本中已切分的句子，利用匹配技术，将包含情感词汇的句子进行抽取，构成候选观点句子集 COSSet1。

（2）基于机器学习方法的候选观点句抽取

利用文献[5]的文本情感分类方法，建立观点句和非观点句分类器。

Step 1. 将 COAE2009 的评测观点数据作为训练数据集，训练观点句分类器；

Step 2. 采用 Step1 训练的分类器，对待抽取观点句集 SSet 进行分类，将其分成观点句和非观点句，得到的观点句记为候选观点句集 COSSet2；

Step 3. COSSet3= COSSet1∩ COSSet2；

Step 4. 利用 Step3 中的 COSSet3 作为新的训练数据，重新训练观点句分类器；

Step 5. 利用 Step4 的新的分类器，重新对（COSSet1∪ COSSet2）- COSSet3 的句子集进行观点句与非观点句分类，得到候选观点句 COSSet4。

（3）基于规则的候选观点句抽取

为了提高观点句的召回率，系统对于句子集 SSet- COSSet3- COSSet4 的句子，采用基于规则的方法再次抽取。其抽取规则如下：

Rule 5：如果句子中包含“相比”、“不如”、“不输给”等词，则认为是观点句。（对比句）

Rule 6：如果在句首部分包含“点评”、“小结”、“看点”、“评价”、“整体印象”等词，则认为是观点句。（点评类句子）

Rule 7：如果领域是“金融证券”且句子中包含“上扬”、“走高”、“走低”、“走强”、

“高开”、“低开”、“强势”、“弱势”等词，则认为是观点句。（金融证券）

通过（1）—（3）可得到候选观点句，最终的观点句为 COSSet3∪COSSet4∪COSSet5。

3.3 观点句极性判别

为了判断第 3.2 节中得到的观点句的极性，本系统采用基于情感词与句子类型相结合的极性判别方法。

设 $\text{Sets}=\cup_{word\in Sentence\wedge word\in PNWord}\{word\}$

这里 PNW=PWord∪NWOrd，PWord 表示褒义词集，NWord 表示贬义词集。采用如下规则进行观点句极性判别。

Rule 8：如果 $Sets\neq\varnothing$ 则采用如下子规则判别其极性。

Rule 8.1 如果 $Sets\subseteq PWord$ ，则句子 sentence 倾向为 1；

Rule 8.2 如果 $Sets\subseteq NWord$ ，则句子 sentence 倾向为-1；

Rule 8.3 如果 $Sets\not\subset NWord$ and $Sets\not\subset PWord$ ，则句子 sentence 倾向为 0；

Rule 9：如果句子是转折句，则只判别转折部分的极性为全句极性。

Rule 10：如果某观点词前面是否定词，则先对该观点词的极性逆转。

4 评价搭配抽取

该任务要求找出任务 2 抽取的每个观点句中观点所针对的评价对象、评价短词，并对评价的倾向做出判别。本系统采用模板匹配与规则相结合的方法（TRM），对此任务进行了尝试性的工作。系统流程图如图 3

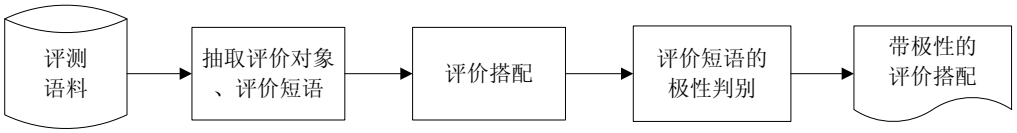

图 3 评价搭配抽取流程图

4.1 评价对象与评价短语的抽取

本系统利用模板对句子中的核心词汇进行扩展，得到评价对象和评价短语。

评价对象的核心词汇认为具有领域实体性质的名词。

评价短语的核心词汇则为评价词表中的形容词、动词。

评价对象和评价短语抽取：利用模板技术将核心词汇扩展成评价对象和评价短语。其模板如下：

模板 1：“nn”、“n、n”、“n 和 n”、“n”。句子中出现这些模板，将其扩展成评价对象，作为评价对象，记为 obj。

模板 2：d+a、d+v、v+v、a+u+v。句子中出现这些模板，将其扩展成评价短语，作为评价短语。

4.2 评价搭配

评价搭配是指评价对象与评价短语搭配。本系统采用两种策略融合的方法构成评价搭配。

策略 1：利用模板对评价对象和评价短语直接进行组合构成搭配。

模板 3：a+u+obj、d+a+u+obj、n+u+obj，把 obj 作为评价对象，a+u、d+a、n 作为评价短语，组成评价搭配。

模板 4：obj+a、obj+d+a，将 obj 作为评价对象，a、d+a 作为评价短语，组成评价搭配。

模板 5：obj+v、obj+d+v，将 obj 作为评价对象，v、d+v 作为评价短语，组成评价搭配。

模板 6：i+u+obj，将 obj 作为评价对象，i+u 作为评价短语，组成评价搭配。

策略 2：利用 4.1 节中的模板 1 与模板 2 中获得评价对象和评价短语采用近邻法直接进行搭配。即在一句话中，如果评价对象与评价短词相邻，则将其认为评价搭配。

本系统将策略 1 和策略 2 获得评价搭配合并，并去重后，得到所有的评价搭配。

4.3 评价短语的极性判别

利用 3.3 中的情感词表，查找评价短语中是否包含了情感词，以此来判断评价搭配的极性。如果包含的词是褒义词，则认为此评价短语为褒义。反之，则为贬义。

5 实验结果与分析

根据系统对任务 1-任务 3 的测试，得到如下测试结果。

（1）观点词的抽取

根据 2.1 节置信度的不同，三个领域的观点词提交了两组结果，见表 1。

表 1　三个领域观点词的抽取结果与比赛各项平均和最优成绩

方法	类别	Precision@1000	Precision	Recall	F1	Raccuracy
WPC	电子产品	0.63500	0.340000	0.101700	0.156600	0.101700
	影视娱乐	0.60600	0.353000	0.107000	0.164200	0.107000
	金融证券	0.58700	0.361000	0.109900	0.168500	0.109900
RC	电子产品	0.67400	0.340000	0.101700	0.156600	0.101700
	影视娱乐	0.62400	0.353000	0.107000	0.164200	0.107000
	金融证券	0.67200	0.361000	0.109900	0.168500	0.109900
Median		0.57126	0.343004	0.094744	0.147593	0.094744
Best		0.67400	0.612500	0.119400	0.183300	0.119400

由表 1 和表 2 可以看出各领域的 Precision@1000 和平均的 Precision@1000 都比较高，其中宏平均和微平均均达到了最大值 0.656700。其他各项指标均超过了平均值。说明本系统对于任务 1 是有效的。

（2）观点句的抽取

利用第 3 节观点句抽取流程，得到观点句的各项指标见表 3 和表 4。

由表 3 可以看出，本系统得到的电子产品类的各项指标均取得了比较好的结果，而金融

证券类的精确率比较差，说明对系统该类观点句的界定有待提高。但三个领域召回率普遍优于中间值，说明本系统侧重召回率的策略取得了预期的效果。

表 2　领域观点词的抽取结果与比赛各项平均和最优成绩的宏平均和微平均

评价指标	方法	Precision@1000	Precision	Recall	F1	Raccuracy
宏平均	WPC	0.609300	0.351300	0.106200	0.163100	0.106200
	RC	0.656700	0.351300	0.106200	0.163100	0.106200
	Median	0.571267	0.342993	0.094740	0.147880	0.094740
	Best	0.656700	0.486000	0.113600	0.174400	0.113600
微平均	WPC	0.609300	0.351300	0.106200	0.163100	0.106200
	RC	0.656700	0.351300	0.106200	0.163100	0.106200
	Median	0.571267	0.337933	0.094740	0.147547	0.094740
	Best	0.656700	0.468100	0.113500	0.174400	0.113500

表 3　三个领域观点句抽取结果与比赛各项平均和最优成绩

方法	领域	Precision	Recall	F1	P@1000	Raccuracy
RSM	电子产品	0.437077	0.602713	0.506702	0.588000	0.484848
	影视娱乐	0.178439	0.507730	0.264071	0.191000	0.203417
	金融证券	0.092314	0.439072	0.152554	0.108000	0.114120
Median		0.240815	0.397946	0.276324	0.290183	0.255445
Best		0.729751	0.798097	0.693304	0.800000	0.660324

表 4 观点句抽取与比赛各项平均和最优成绩的宏平均和微平均

评价指标	方法	Precision	Recall	F1	P@1000	Raccuracy
宏平均	RSM	0.235943	0.516505	0.307776	0.295667	0.267462
	Median	0.240815	0.406039	0.276324	0.290183	0.255445
	Best	0.534693	0.723411	0.541377	0.532000	0.494511
微平均	RSM	0.309947	0.575586	0.402924	0.295667	0.412422
	Median	0.315252	0.450532	0.357871	0.290183	0.363170
	Best	0.654448	0.775397	0.639614	0.532000	0.611425

（3）评价搭配抽取及极性判别

为了获取评价搭配，需要首先获取评价对象、评价短语，在此基础上，得到评价搭配。各目标得到的结果见表 5-表 7。

由表 5-表 7 可以看出，本系统获得评价对象、评价短语以及评价对象、评价短语、极性一致正确的 Precision@1000 以及宏平均的 Precision 均达到了评测小组中的最好成绩 0.071667 和 0.039639，说明采用模板与规则的方法在精确率上取得效果是明显的，但由于模板和规则方法的局限性，获得召回率是比较低，均低于中值。另外电子产品的各项指标均优于其他两个领域，这与我们小组最近在产品类领域的研究相对比较深入有密切的关系，而其他两个领域从未涉足过。

表 5 评价对象抽取及与比赛各项平均和最优成绩

领域与指标	方法	P@1000	Precision	Recall	F1	Raccuracy
电子产品	TRM	0.280000	0.190909	0.039728	0.065769	0.039728
影视娱乐		0.018000	0.025788	0.011719	0.016115	0.011719
金融证券		0.009000	0.009792	0.010753	0.010250	0.010753
宏平均	TRM	0.102333	0.075496	0.020733	0.032532	0.020733
	average	0.065524	0.074285	0.045504	0.054307	0.045504
	max	0.111000	0.133933	0.081763	0.091606	0.081763
微平均	TRM	0.102333	0.087644	0.034500	0.049511	0.033610
	average	0.065524	0.101236	0.081829	0.083421	0.069819
	max	0.111000	0.159847	0.149071	0.144701	0.135726

表 6 评价短语判别结果及与比赛各项平均和最优成绩

		Precision@1000	Precision	Recall	F1	Raccuracy
电子产品	TRM	0.322000	0.218182	0.045403	0.075164	0.045403
影视娱乐		0.016000	0.022923	0.010417	0.014324	0.010417
金融证券		0.017000	0.015912	0.039939	0.022757	0.016897
宏平均	TRM	0.118333	0.085672	0.031919	0.046510	0.024239
	average	0.051429	0.055383	0.050590	0.046795	0.035003
	max	0.118333	0.085672	0.100699	0.087237	0.066468
微平均	TRM	0.118333	0.100954	0.039739	0.057029	0.038256
	average	0.051429	0.074015	0.061917	0.062064	0.052096
	max	0.118333	0.117425	0.105575	0.10248	0.095492

表 7 评价对象、评价短语、极性均正确的判别结果与比赛各项平均和最优成绩

领域与指标	方法	P@1000	Precision	Recall	F1	Raccuracy
电子产品	TRM	0.205000	0.108484	0.022575	0.037373	0.022575
影视娱乐		0.007000	0.008596	0.003906	0.005372	0.003906
金融证券		0.003000	0.001836	0.004608	0.002626	0.004608
宏平均	TRM	0.071667	0.039639	0.010363	0.016431	0.010363
	average	0.027190	0.025047	0.020403	0.019833	0.015183
	max	0.071667	0.039639	0.037671	0.034780	0.025728
微平均	TRM	0.071667	0.047212	0.018584	0.026670	0.018584
	average	0.025778	0.033019	0.027191	0.027323	0.023090
	max	0.071667	0.061600	0.046856	0.048312	0.043298

6 总结与展望

本文主要介绍了参加COAE2011评测系统的基本情况。该系统采用多策略技术和方法

从三个领域的大量文本中抽取评价词，并对三个领域的各2000篇文本中抽取所有的观点句、评价对象、评价短语以及评价搭配及其情感倾向。在评价词抽取、观点句召回率、评价对象、评价短语以及评价短语以及评价搭配及其情感倾向的精确率方面取得了比较理想的结果，但在观点句精确率、评价对象、评价短语以及评价短语以及评价搭配及其情感倾向的召回率方面还需深入研究。

感谢：董振东先生提供的知网中的评价词汇和情感词汇，中科院计算所提供的分词与词性标注软件，哈工大社会计算与信息检索研究中心提供的《同义词词林扩展版》，笔者在此深表谢意。

参 考 文 献

[1] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报. 2010,21(8):1834−1848

[2] 王素格,李德玉,魏英杰,宋晓雷. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报. 2009,23(5): 68-74

[3] Peter D. Turney, Michael L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association[J]. ACM Transaction on information systems. 2003,21(4):315-346

[4] V. Hatzivassilolou Kathleen, R. Mckeown. Predicting the semantic orientation of adjectives[C]. In Proceeding of the 35th Annual meeting of the association for computational linguistics and the 8th conference of the European Chapter of the ACL. 1997:174-181

[5] A Feature Selection Method Based on Improved Fisher's Discriminant Ratio for Text Sentiment Classification[J]. Expert Systems with Application. 2011,38(7):8696-8702

基于最大熵模型和最小割模型的中文词与句褒贬极性分析*

董喜双，邹启波，关毅，高翔，闫铭
哈尔滨工业大学，哈尔滨，150001
dongxishuang@gmail.com，zouqibo2009@163.com，guanyi@hit.edu.cn，hngaoxiang@gmail.com，
mingitouch@gmail.com

摘　要：本文运用最大熵模型和最小割模型预测中文词和句子的褒贬极性。词级情感分析首先构建领域情感词典，然后根据领域情感词典提取候选词，并使用最大熵模型预测候选词的极性，最后采用最小割模型优化极性结果。句级情感分析首先根据领域情感词典识别观点句，将观点句切分成短句并基于规则提取特征，应用最大熵模型预测短句的极性，最后根据短句的极性预测长句的极性。

关键词：情感极性，情感分析，最大熵，最小割；

Positive and Negative Polarity Analysis on Chinese Words and Sentences Based on Maximum Entropy Model and Min-Cut Model

Dong Xi-Shuang, Zou Qi-Bo, Guan Yi, Gao Xiang, Yan Ming
Harbin Institute of Technology， Harbin，150001
dongxishuang@gmail.com，zouqibo2009@163.com，guanyi@hit.edu.cn，hngaoxiang@gmail.com，
mingitouch@gmail.com

Abstract: In this paper, Maximum Entropy Model and Min-Cut Model were adopted to predict the positive and negative polarities of Chinese words and sentences. First, we built a domain sentiment lexicon. Then, the candidate sentiment words were recognized by the lexicon, and the polarity was predicted by Maximum Entropy model. Finally, we used Min-Cut model to optimize the polarity results. On the side of sentiment analysis on sentences, opinion sentences were obtained by the lexicon. These opinion sentences were split up into short sentences, and sentiment features were extracted by rules. Then the polarity of short sentences was predicted by Maximum Entropy model. Finally, polarities of opinion sentences were predicted according to polarities of short sentences.

Keywords: Sentiment Polarity, Sentiment Analysis, Maximum Entropy, Minimum Cut;

1　引言

情感分析的基本任务指对给定文本的极性进行分类[1]。按照处理的粒度不同，情感倾向性分析主要包括四个方面的研究内容：词级情感倾向性分析、短语级情感倾向性分析、句子级情感倾向性分析和篇章级情感倾向性分析。词级情感分析包括识别候选情感词、判断候选情感词情感极性与强度以及构建情感字典[2]。短语级情感分析为根据情感词识别情感短语并判定情感极性与强度[3]。句级情感分析为识别句级观点持有人、评价对象以及判断句子的情感倾向[2][4]。篇章级情感分析为识别篇章对某一事物的观点[1]。

本文主要研究包括词和句子级情感倾向性分析。词级主要完成的任务是根据上下文抽

*

取出观点词，并判断观点词的褒贬极性，本文采用最大熵模型预测观点词的情感极性，然后通过最小割模型优化极性结果。句子级主要完成的任务是从文本中识别出观点句，再判断观点句的褒贬极性。本文将观点句切分成短句并用最大熵模型预测短句情感极性，然后用短句预测长句的情感极性。

本文的组织结构如下：第二部分介绍相关研究；第三部分介绍相关的模型；第四部分具体介绍词和句子的情感分析方法；第五部分分析实验结果，最后是结论和展望。

2 相关研究

词汇的情感倾向性分析研究主要包括两类：语义方法和统计方法[17]；语义的方法主要是通过一个现有的知识库，然后通过计算候选词和知识库里面的基准词的语义距离，进而得到候选词的情感倾向。在英文方面，Kamps, Marx, Mokken 和 Rijke 于 2002 年提出了基于 WordNet 的同义结构图计算候选词和知识库基准词的语义距离，进而判断候选词的情感倾向性[6]；在中文方面，朱嫣岚、闵锦、周雅倩于 2006 年提出了基于 HowNet 的词汇语义计算方法[7]；路斌、万小军、杨建武于 2007 提出了基于《同义词词林》来计算词汇的褒贬性[8]。统计方法主要是基于有监督和无监督的机器学习[5]，利用文本中词汇间的共现关系来计算词汇的倾向性；Peter D. Turney 和 Michael L. Littman[9]于 2003 年提出了基于搜索引擎的“NEAR”操作计算候选词和种子词之间的相关性，从而得到候选词的情感倾向性。Yu 和 Hatzivassiloglou[10]于 2003 年提出了一个基于种子词典的方法, 首先选择部分情感词作为种子词构建情感词典，通过计算候选词和种子词的共现概率来判断候选词的倾向性。

句级情感分析的研究方法主要是统计方法。Soo-Min Kim 和 Eduard Hovy[2]等人于 2004 年提出了情感倾向累乘模型、情感强度调和平均模型以及情感强度几何平均模型判断句子情感倾向性的方法。Hong Yu 和 Vasileios Hatzivassiloglou[10]于 2003 年提出了通过朴素贝叶斯模型来预测句子情感极性。McDonald[14]于 2007 年将句子情感极性分析转化成为情感极性序列化标注，该方法考虑到篇章中上下文对当前句子的影响，但是编码解码过程比较复杂。

本文对词的情感分析研究借鉴了文献[12]的思想，首先采用最大熵模型预测词的极性，然后采用最小割模型优化极性结果。句子情感分析借鉴文献[5]中关于最大熵模型在句子分类中的研究，本文将句子划分为短句并基于规则提取特征，然后采用最大熵模型预测短句极性，最后用短句极性预测长句极性。

3 相关理论和模型

本文用最大熵模型用来预测词和句子的极性，最小割模型优化词的极性结果。

3.1 最小割原理

文献[11]给出了流网络的定义。流网络 $G=(V,E)$ 的割 (S,T) 将 V 划分为 S 和 $T=V-S$ 两个不同的集合，满足条件 $s\in S,t\in T$，穿过割 (S,T) 的净流量定义为 $f(S,T)$，割 (S,T) 的容量定义为 $c(S,T)$ [11]，所以一个流网络的最小割就是这个流网络所有割中具有最小容量的割。

流网络 $G=(V,E)$ 的最小割的容量 $c(S,T)$ 等于其最大流 $|f|$ 值，文献[11]给出了证明，本文就不再赘述。1965 年 Ford 和 Fulkerson 给出了计算最大流的方法 Ford-Fulkerson 算法，文献[11]给出了详细介绍。其基本思想就是不断地寻找增广路[11]，当找不到增广路时，表

示流网络中没有流量可以扩展，此时流量达到了最大值。

3.2 最大熵模型

文献[15]基于信息熵理论建立了最大熵模型。在一定的限制条件下，如果这些限制无法确定唯一的系统分布，那么信息熵最大的分布是最优系统分布。最大熵模型被广泛用于自然语言处理中，文献[16]最早采用最大熵模型来处理自然语言问题。

4 词句情感分析

4.1 词级情感分析

词级情感分析首先构建领域情感词典，然后采用领域情感词典抽取出观点词并用最大熵模型预测观点词极性，最后构建词的加权无向图并用最小割模型优化极性结果。

4.1.1 领域情感词典构建

本研究以谭松波构建的数码情感语料[13]为实验基础，采用中科院 ICTCLAS 分词系统分词、专有名词识别并使用哈工大网络智能研究室情感词典过滤出情感词，得到原始领域情感词典，然后人工标注原始领域情感词典词的褒贬极性得到种子词并使用 HowNet 扩充种子词，扩充包括同义词、对义词、反义词；我们构建了一个训练词集[5]，将训练词集使用 HowNet 扩展同义词生成同义词特征并训练最大熵模型，再使用最大熵模型预测扩展词的极性并人工剔除非领域性词，得到领域情感词典。其流程图如图 1 所示。

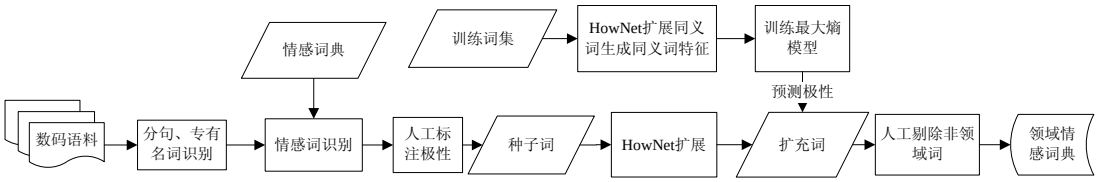

图 1 领域情感词典构建流程图

Fig.1 the Flow Chart of Building Domain Sentiment Lexicon

4.1.2 观点词抽取

首先用中科院分词系统 ICTCLAS 对测试语料分词，然后采用基于领域情感词典的方法、词性判断提取出观点词，我们假设“ ”、”“、【】、（）、{}、[ ]、《》这些符号中间的内容不会影响句子的情感，本文将这些符号中间的内容去掉。由于通用的分词系统不能很好地把特定领域一些专用的词汇准确分词，为此我们新增加了一个专有名词词典[18]，把一些公司名字、节目名称、艺人姓名等专有名词从文件里面识别出来，提高了分词的准确度。我们认为大多数观点词都是形容词，为了简单起见，我们只对形容词进行判断，对于每个形容词我们查找它是否在该领域的观点词词典里，如果不在则忽略，如果在，则抽取出该词以及该词附近的文本做答案中的第五部分。

4.1.3 观点词情感分类

观点词的情感分类首先通过HowNet扩展训练语料生成同义词特征并训练最大熵模型；对测试语料分词并识别出语料中的专有名词、人名等以提高分词正确率，然后采用领域情感词典提取观点词并用最大熵模型预测观点词情感极性，最后建立基于词汇的加权无向图并用最小割模型优化极性结果，具体过程如图 2 所示。

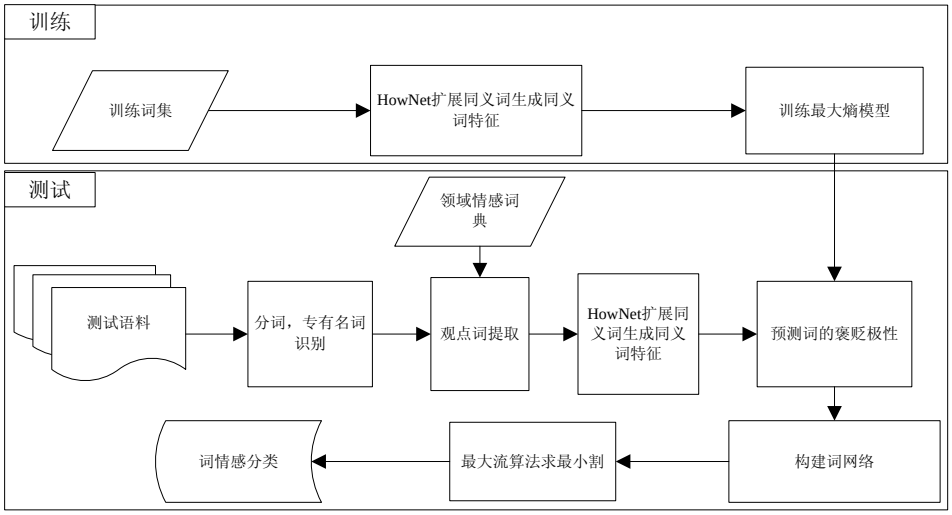

图 2 词情感分类流程图

Fig.2 the Flow Chart of Classification of Sentiment word

下面具体介绍最小割模型对结果的优化过程。

4.1.4 构建词的加权无向图

构建基于词汇语义关系的加权无向图，其中词语对应于图中的节点，边上的权重c_i是这两个节点(词)的同义词关系权重，权重d_i表示反义词关系权重，可以通过知识库得到，例如采用 WordNet 或者 HowNet 来计算词语语义关系来计算；引入两个节点：源点(s)和汇点(t)；其他节点到 s 和 t 的边上的权重(b_i)和(a_j)是这个节点到 s 和 t 的权重，本文通过最大熵模型预测词的褒贬极性，并将预测概率值作为该词节点到源点和汇点的权重构建加权无向图，如图 3 所示。

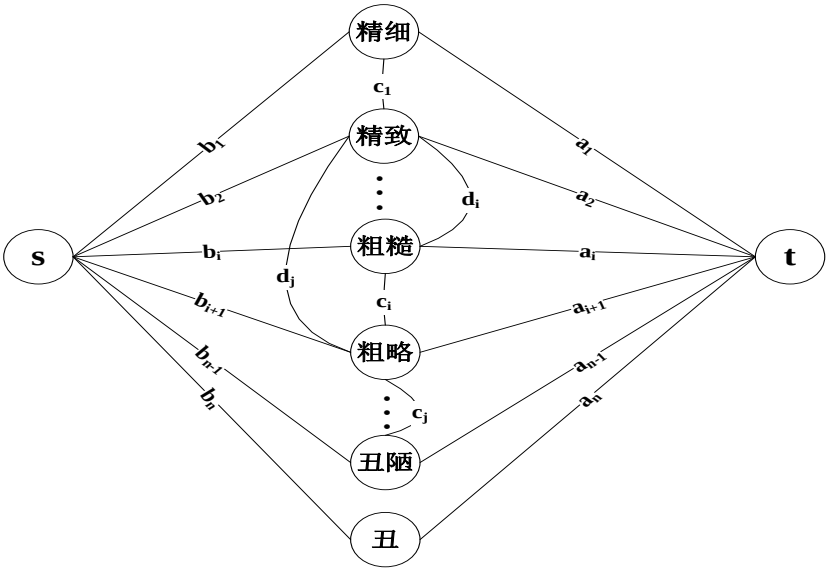

图 3 词汇加权无向图

Fig.3 A Weighted Undirected Graph of Words

其中，b_i和a_j 分别代表词节点到源点和汇点的权重，本文采用最大熵模型预测词的褒贬极性，并用预测概率值作为这个权重。d_i 和c_i 分别代表反义词关系权重和同义词关系权重，采用 HowNet 计算这个权重。

4.1.5 基于最小割模型的词褒贬极性分析

图论中最小割的二值划分是基于相似的元素总应该会被分到一个相同集合的假设[12]；词对应于无向图的顶点，一个二值划分等价于把这个词的加权无向图划分为两个子集 S 和 T，$s \in S, t \in T$，即从这个无向图中移除了一些边，使得这个连通的词加权无向图被分成两个不连通的联通子图。那么我们希望相似的元素被分到一个相同的集合，所以最好的分割就是把相似的元素被分割到相同集合中去，也就是移去的边的权重和最小，换句话来说就是求解一个最小割问题，被移去的边的权重表示如下：

$$W(S,T) = \sum_{u \in S, v \in T} w(u,v) \qquad (1)$$

其中 $w(u,v)$ 是词语 u 和 v 之间的权重，那么问题其实就是求 $\arg\min W(S,T)$。由于词的褒贬极性判断问题是一个二值划分问题，所以可以用最小割理论来处理，其次 $W(S,T)$ 达到最小，意味着我们的词的加权无向图从源点 s 到汇点 t 的概率达到最大，也就是破坏了最小量的语义关系和最大程度地保留了分类结果，所以这个分类是一个有效的分类。通过最小割分割图 3 得到图 4，如图 4 所示。

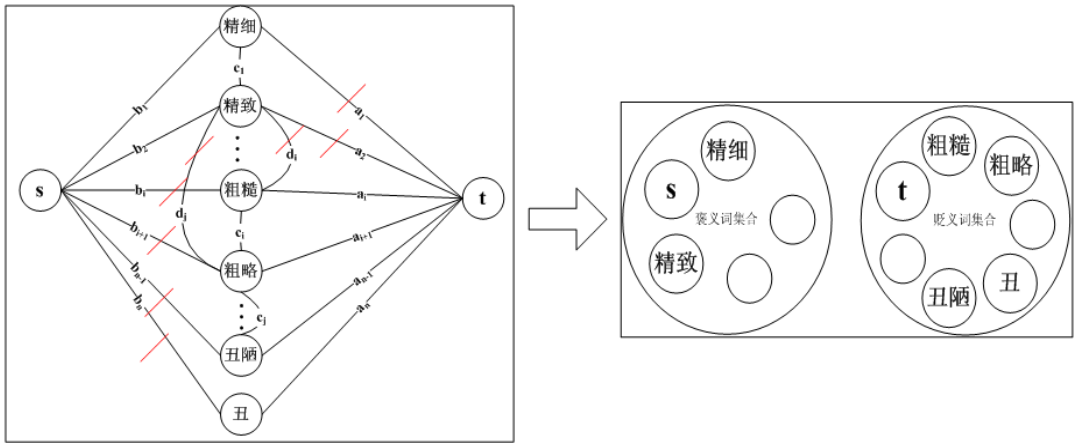

图 4 最小割图

Fig.4 A min-cut graph of networks of words

4.2 句级情感分析

本文主要采用最大熵分类的方法判断情感句类别，因为分类的方法能够融入比较多的特征，同时也可以获得比较高的准确率[3]；本文采用的语料包括 2000 篇褒义评价文章和 2000 篇贬义评价文章[13]。句级情感分析首先采用词典筛选观点句，然后将观点句切分成短句并用最大熵模型预测短句极性，最后通过短句预测长句极性。

4.2.1 观点句识别

抽样分析后发现，在整个测试语料中，观点句和非观点句的比例严重不平衡，所以将会对分类造成严重的影响；我们认为观点句里面基本上都包含了观点词，所以我们通过领域情感词典来筛掉非观点句，可以大大减少测试语料的不平衡给分类带来的影响。我们假设“ ”、”“、【】、（）、{}、[ ]、《》这些符号中间的内容不会影响句子的情感，本文将这些符号中间的内容去掉。

4.2.2 长句切分

实验表明长句分类的效果不是特别好，实验对比了长句分类和短句分类，发现短句分类效果更好，所以本文采用短句分类来处理。一个长句包含的情感相对短句较复杂，特征也比较多且难以准确提取，分类效果比较差；短句包含的结构简单、语义容易判断、特征容易抽取。本文主要采用的分割符有“，”、“：”、“；”、“。”；比如句子“这款笔记本外观造型比较时髦，但是显示器的显示效果总是感觉有点儿差，还有就是内存挺大的；所以总而言之来说还是比较实惠的”，按照我们的方法，这个句子会被分成四个子短句，分别是：

A 这款笔记本外观造型比较时髦

B 但是显示器的显示效果总是感觉有点儿差

C 还有就是内存挺大的

D 所以总而言之来说还是比较实惠的

4.2.3 特征获取

文献[5]给出了 3 个特征获取的方法，这些方法在 COAE2009 的数据上实验取得了比较好的效果。提取的特征主要包含词和词序列、情感词强度累加、句子的句型。词和词序列主要包括识别情感词，否定词和情感词周围的词序列；情感词强度累加指通过对识别的情感词的情感强度累加；句型特征包括问句、感叹句和长句。

4.2.4 长句情感倾向性计算

由于长句被分成了短句，长句情感极性由短剧情感极性决定，因此本文主要采用下面的决策思想：

（1） 如果一个长句的所有短句都是褒义，那么这个长句的倾向性就是褒义。

（2） 如果一个长句的所有短句都是贬义，那么这个长句的倾向性就是贬义。

（3） 如果一个长句里面既有褒义又有贬义，那么这个句子的倾向性是褒贬混合。

长句置信度计算采用下列方法，如果一个长句包含 n 个短句，每一个短句的置信度是 x_1、x_2、$x_3 \cdots x_n$，那么长句置信度 $p = \dfrac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n}$。句级情感褒贬分析的全部具体过程如图 5 所示。

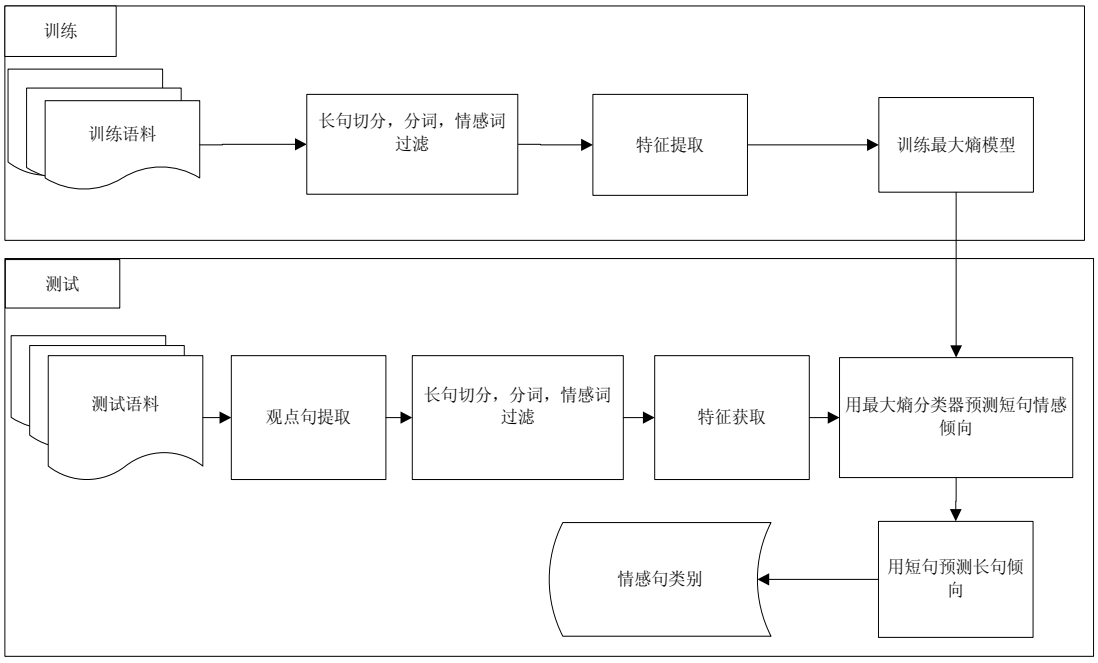

图 5 情感句褒贬极性分析流程图

Fig.5 the Flow Chart of Positive and Negative Polarities Analysis on Sentiment Sentences

5 实验与分析

词级情感分析实验结果如表 1 所示。

表 1 词级褒贬极性分析

Tab.1 Positive and Negative Polarities Analysis on Words

	Precision	Recall	F1	P@1000	Raccuracy
HITWI_D	0.3927	0.0956	0.1538	0.577	0.0956
HITWI_E	0.4528	0.1009	**0.1651**	0.596	**0.1009**
HITWI_F	**0.6125**	0.0936	**0.1624**	0.641	0.0936
Median	0.3430044	0.0947444	**0.1475933**	0.57126	0.0947444
Best	**0.6125**	0.1194	**0.1833**	0.674	**0.1194**

从表 1 的实验结果可以看出，电子数码领域和娱乐领域的准确度、召回率、Raccuracy 都比平均值高且接近最大值，F 值接近最大值，这主要是因为这两个领域的情感倾向容易判断，观点容易抽取。而在财经领域召回率和 Raccuracy 都比平均值低，原因是这个领域的观点词情感比较模糊以及领域情感词典不是很丰富所致。

句级情感分析实验结果如表 2 所示。

表 2 句级褒贬极性分析

Tab.2 Positive and Negative Polarities Analysis on Sentences

	Precision	Recall	F1	P@1000	Raccuracy
HITWI_D	**0.49268**	0.391297	**0.436174**	0.608	0.391297
HITWI_E	0.303493	0.2262	0.259207	0.278	0.2262
HITWI_F	0.204598	0.172147	0.186975	0.089	0.172147
Median	0.240815	0.3979462	0.27632415	0.290183	0.2554445
Best	**0.729751**	0.798097	**0.693304**	0.8	0.660324

从表 2 的实验结果可以看出，电子领域的准确度、P@1000、F 和 Raccuracy 值都大于平均值，主要是电子这个领域的特征比较明显容易提取，而召唤率小于平均值，这主要是领域情感词典不够丰富所致。娱乐和财经这两个领域的召回率和 Raccuracy 都小于平均值，原因是领域词典内容不够丰富、情感特征难以准确提取、分词也不够准确致使召回率不够高。

6 结果与展望

本文采用了最大熵模型和最小割模型来处理词和句子的褒贬极性分析问题。词级情感分析通过最大熵分类模型预测候选情感词的褒贬极性，然后构建基于词的加权无向图并采用最小割模型优化极性结果，该方法取得比较好的实验结果。句级情感分析将观点句切分成短句并提取短句的特征并使用最大熵模型来预测短句的褒贬极性，最后用短句情感极性预测长句的情感褒贬极性。下一步工作将丰富领域情感词典以及如何准确的提取句子的特征。

参 考 文 献

[1] Bo Pang, Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the ACL, 2004, pp.271-278.

[2] Soo-Min Kim, Eduard Hovy. Determining the sentiment of opinions. In Proceedings of COLING, 2004, pp.1367-1373.

[3] Peter D. Turney. Thumbs Up or Thumbs Down Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp.417-424.

[4] Y. Mao, G. Lebanon. Isotonic conditional random fields and local sentiment flow. In Proceedings of NIPS, 2006.

[5] 董喜双，关毅，李本阳，陈志杰，李生.基于最大熵模型的中文词与句情感分析研究.第二届中文情感倾向性分析会议，2009：1-8

[6] J. Kamps, M. Marx, R. J. Mokken and M. D. Rijke. Using WordNet to measure semantic orientation ofadjectives. In: Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, Lisbon, 2004, 1115-1118.

[7] 朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德. 基于 HowNet 的词汇语义倾向计算.中文信息学报，2006,20(1): 14-20.

[8] 路斌，万小军，杨建武，陈晓鸥. 基于同义词词林的词汇褒贬计算，第七届中文信息处理国际会议,2007,17-23.

[9] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems. 2003. 21 (4): 315 – 346.

[10] H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences [A], In: M. Collins and M. Steedman(eds). Sapporo. Japan: 2003.129-136

[11] Thomas H. Cormen Charles E. lerserson Ronald L. Rivest Clifford Stein. Introduction of Algorithms. Second Edition. China Machine Press.2006.396-419

[12] Fangzhong Su; Katja Markert. Subjectivity Recognition on Word Senses via Semi-supervised Mincuts. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 1–9, Boulder, Colorado, June 2009.

[13] 情感语料：http://www.searchforum.org.cn/tansongbo/corpus-senti.htm

[14] R. McDonald, K. Hannan, T. Neylon, et al. Structured models for fine-to-coarse sentiment analysis. In Proceedings of the 45th Association of Computational Linguistics, 2007, pp.435-439.

[15] T. Jaynes. Information Theory and Statistical Mechanics. Physics Reviews. 1957(106): 620-630

[16] Adam L. Berget, Stephen A. Della Pietra, Vincent J. Della Pietra. A Maximum Entropy Approach Natural Language Processing. Computational Linguistics, Vol. 22, No. 1. (1996), pp. 39-71.

[17] 何婷婷，闻彬，宋乐.词语情感倾向性识别及观点抽取研究.第一届中文情感倾向性分析会议，2008：1-8

[18] 专有名词词表：http://list.video.baidu.com/manhotlist/

基于多特征融合的文本情感分析研究

徐冰，吴建伟，鲍军威，朱少杰

哈尔滨工业大学机器智能与翻译实验室，哈尔滨，150001

E-mail:xb@mtlab.hit.edu.cn

摘要：近几年，文本情感分析已经成为自然语言处理的研究热点。在文本情感分析研究中，词语级和句子级的情感信息抽取是基础研究任务，具有重要的研究价值。通过研究发现，领域知识和上下文语境是情感信息抽取的两个重要影响因素，所以在 COAE2011 评测任务中成为重点考察对象。在此次评测任务中，本文采用基于多特征融合的方法研究了领域观点词抽取和评价搭配抽取。领域观点词抽取任务重点考察候选观点词的四种上下文特征，评价搭配抽取任务将词性、浅层句法等特征融入到条件随机场建模过程中。评测结果表明，本文提出的特征在相关任务中起到了较好的作用，其中评价对象抽取结果较接近于最大值。

关键词：文本情感分析；领域观点词；极性；评价搭配

Text Sentiment Analysis Based on Multi-feature Fusion

Xu Bing, Wu Jian-wei, Bao Jun-wei, Zhu Shao-jie

Machine Intelligence & Translation Laboratory, Harbin Institute of Technology, Harbin 150001

E-mail:xb@mtlab.hit.edu.cn

Abstract: In recent years, text sentiment analysis has become a hot issue of natural language processing. For the sentiment analysis, research on the basic tasks of word-level and sentence-level sentiment information abstraction is valuable. In COAE 2011, domain knowledge and context as key factors of abstraction will be considered carefully. This paper presents abstraction of domain opinion word and polarity collocation based on multi-features fusion. Four context features are used in task of domain opinion word abstraction, and then POS and shallow parsing features are integrated into modeling of Conditional Random Fields. Evaluation results show that the proposed features are effective in the relevant tasks. Especially, results of opinion attribute abstraction are close to the best value.

Keywords: Text sentiment analysis, Domain opinion word, Polarity, Polarity collocation

1 引言

随着网络上主观性文本的大量涌现，文本情感分析的研究越来越受到关注。文本情感分析研究的主要任务包括情感信息分类、情感信息抽取和情感信息的检索与归纳[1]；根据研究粒度的不同，还可分为词语级、短语级、句子级和篇章级的情感分析[2]。在文本情感分析的研究中，词语级和短语级的情感分析是研究基础，可以为句子级和篇章级的情感分析提供技术支持，而句子级和篇章级的情感分析可以为情感分析的进一步应用提供有效帮助[34]。

文本情感分析涉及机器学习、数据挖掘、信息检索等相关研究领域，并已在电子产品、影视娱乐和新闻等多个领域得到应用，该项研究具有重要的理论和实用价值。

近几年，文本情感分析评测开始在国内和国外的相关会议上不断涌现。从 2006 年开始，在 TREC、NTCIR 国际会议上出现了关于中、日、英三种语言的情感分析相关任务评测；

从 2008 年开始，在 CCIR 的国内会议上出现了针对中文的情感分析相关任务评测 [56]。2011 年依托 CCIR2011 会议举办了第三届中文倾向性分析评测，此次评测共包括 4 项评测任务，分别是领域观点词的抽取与极性判别、中文观点句抽取、评价搭配抽取和观点检索，这四项任务的关系是：首先，从给定的三个领域数据集（电子产品、影视娱乐和财经）中抽取观点词并判断极性；然后，从三个领域数据集中抽取出观点句并判断观点句的极性；进一步从得到的观点句中抽取评价搭配，即<观点句，评价对象，评价短语，极性>；最后结合领域知识和上下文语境对给定的查询对象进行观点检索。在此次评测中，重点考察“领域知识”和“上下文语境”这两个因素对倾向性分析的影响。

下面对我们参与的相关评测任务做以详细介绍。

2 领域观点词抽取与极性判别

评测任务 1 是领域观点词抽取与极性判别任务，评测指定的领域为电子产品、影视娱乐与财经领域。在评测中领域观点词是指不同领域之间所使用的不同的观点词以及在不同领域倾向性表达不同的相同观点词。由于上下文语境可能对领域观点词抽取的影响较大，所以本文采用了基于上下文特征的观点词抽取和极性判别方法。

2.1 基于上下文特征的观点词抽取

通过对语料的观察发现，在观点词的上下文中出现程度副词、特定标点符号、语气词和情感词这四种特征的概率比较高。因此，在抽取观点词时，以统计候选词上下文窗口内四种特征出现的比例来判断该词是否为观点词，并根据概率值对抽取结果进行置信度排序。

根据评测要求，分别对三个指定领域进行处理。具体方法是：首先，对发布的语料进行分词和词性标注处理；然后，选择所有形容词和动词作为候选观点词；进一步计算候选观点词上下文中出现的四种特征的比例，其计算方法如公式（1）所示。

$$kwr(w) = \frac{\#key\_word(w)}{\#context\_word(w)} \quad (1)$$

式中 kwr(w)——表示词 w 在语料中上下文窗口内四种特征出现的比例；

#key_word(w)——表示词 w 在语料中上下文窗口内四种特征出现的个数；

#context_word(w)——表示词 w 在语料中上下文窗口内词的个数。

特征选择：

（1）程度副词：非常、极、绝对、十分、完全、万分、无比等

（2）标点符号：问号（'！'）、叹号（'？'）等

（3）语气词：哦、啊、呀、嘛、吗、哎等

（4）情感词：爱、褒、崇拜、得意、感谢、恭喜、好、美等

2.2 观点词极性判别

观点词极性识别方法采用的是基于评价词词典的方法，如果在词典中查询不到抽取的观点词极性，就通过文献 [7] 提出的基于 Hownet 的词汇语义相似度计算方法进行判别。

具体方法是：

2 在 HowNet 的中文评价词词典中选择极性鲜明的极性词作为种子词；

3 选择抽取出的观点词的上下文窗口[-2，+2]中的实词与种子词之间进行语义相似度计算，计算公式如公式（2）和（3）所示；

$$positive\_sim(w)=\frac{\sum_{s\in positive\ seed\ set} sim(w,s)}{\#\ positive\ seed\ set} \quad (2)$$

$$negative\_sim(w)=\frac{\sum_{s\in negative\ seed\ set} sim(w,s)}{\#\ negative\ seed\ set} \quad (3)$$

4 采用公式（4）计算出观点词的极性。

$$polarity(sw)=sign(\sum_{w\in context\ of\ \ sw}(positive\_sim(w)-negative(w))) \quad (4)$$

式中 $sign(x)$——表示取 x 的符号。

3 评价搭配抽取

评测任务 3 是评价搭配抽取任务，在评测中评价搭配是指从观点句中抽取出评价对象和评价短语，并判别其评价极性。要求评价对象为评论针对的对象或对象的属性，越具体越好；评价短语为修饰成分和评价词语组合而成的评价单元，修饰成分指加强、减弱或置反观点的语言成分，可以是程度副词或否定词等。如果在观点句中包含多个评价搭配，要求全部抽取出来。

评价搭配抽取任务分为四个子任务：（1）评价对象抽取；（2）评价短语抽取；（3）评价对象和评价短语关系抽取；（4）评价搭配的极性判别。

下面对每个子任务的解决方法做以详细介绍。

3.1 评价对象和评价短语抽取

在评测中，我们将评价对象和评价短语的抽取任务合并，看作是一个序列标注问题，采用条件随机场模型进行建模。

条件随机场（Conditional Random Fields，CRFs）是一种用于序列标注的机器学习模型，它是基于概率结构的模型 8。该模型从隐马尔科夫模型（Hidden Markov models，HMM）和最大熵隐马尔可夫模型（maximum entropy Markov models，MEMM）的基础上发展而来。条件随机场被广泛用于序列标注、数据分割、组块分析等自然语言处理任务中，中文分词、命名实体识别、歧义消解等任务也经常使用。条件随机场在自然语言中有非常重要的应用，研究表明 8，条件随机场在序列标注问题上优于其他算法。

条件随机场模型的优点是可以引入多种有效的特征集，不需要考虑这些特征之间是否相互独立。当新的特征引入到模型中时，不需要对模型本身做任何修改。在抽取的过程中，分别引入了词特征、词性特征、浅层句法特征、否定词特征和评价词词典信息。

（1）分词、词性特征

分词、词性是在自然语言处理中经常使用的语法特征。在文本情感分析中，评价词往

往为形容词或副词，评价对象一般是名词，正确识别词性对文本情感倾向性分析具有很大的帮助。例如“操作系统使用起来非常繁琐，比Linux差远了”，进行分词词性标注后得到“操作系统/n 使用/v 起来/vf 非常/d 繁琐/a ，/wd 比/p Linux/x 差/a 远/a 了/ule”，评价对象“操作系统”的词性为名词，评价短语“非常”，“繁琐”的词性分别为副词和形容词，由此可见分词和词性特征对评价对象和评价短语的抽取非常有效。

（2）浅层句法特征

浅层句法分析（shallow parsing），也称为组块分析或者部分句法分析（partial parsing），可以用来识别句子中某些句法结构相对简单、功能和意义比较重要的成分，如主语、谓语、宾语等。浅层句法分析的结果并不是构建一棵完整的句法分析树，它的目的是简化自然语言处理分析的复杂度，以便提高分析的性能。

例如句子“佳能 A 系列的经典毋庸置疑”进行浅层句法分析后得到以下结果序列：

BNP[佳能/nz A/x 系列/n] 的/udeq 经典/n 毋庸置疑/vl

在上述的例子中，“佳能 A 系列”是一个评价对象，经过分词后该短语被处理为“佳能/A/系列”，浅层句法分析将其看作一个完整短语，所以融入浅层句法特征可以有效提高短语级评价对象的抽取效果。

（3）否定词特征

评价短语是由修饰成分和评价词组合而成的评价单元，修饰成分可以是程度副词或否定词等。词性特征可以识别出修饰成分中的副词，增加否定词特征可以有效地识别否定词修饰成分，该特征可以提高评价短语的抽取效果。

3.2 评价对象和评价短语关系抽取

采用 CRFs 模型抽取出评价对象和评价短语后，评价对象和评价短语之间的关系需要进一步确定。通过观察发现，在一般的观点句中修饰某评价对象的评价短语，与识别出的其他评价短语相比，在距离上与该评价对象最近。因此，针对评价对象和评价短语的关系抽取采用了最近匹配原则，即考察评价对象的上下文，选择在一个分句内与其最近的评价短语作为修饰该评价对象的评价短语。

3.3 评价搭配的极性判别

获得了评价对象和评价短语关系以后，将对该评价搭配进行极性判别。评价搭配的极性判别方法是采用与任务 1 相同的极性判别方法，并结合评价短语中出现否定副词的情况来最终确定其极性。

4 实验结果与分析

4.1 实验设置

在评价搭配抽取实验中，采用 CRF++-0.53 工具包来实现条件随机域模型的训练和测试。浅层句法分析工具为哈工大机器智能与翻译实验室研制开发，分词、词性标注工具是中科院开发的 ICTCLAS3.0，评价词词典来自知网的“情感分析用词语集”。

电子产品领域的训练语料采用 COAE 2008 发布的任务 3 语料，包括笔记本电脑、手机

和数码相机三种电子产品，财经领域和影视娱乐领域训练语料为手工从新浪、搜狐等网站收集并标注形成。训练语料统计情况如表 1 所示，测试语料情况详见评测大纲。

表 1　训练语料统计结果

Table 1 Statistical results of training corpus

	电子产品	影视娱乐	财经
句子数	5556	1302	1592
评价对象个数	5388	1411	1224
评价短语个数	4064	1974	1284

4.2 实验结果与分析

领域观点词和评价搭配抽取实验结果如下：

（1）　领域观点词抽取与极性判别结果

根据本文提出的基于上下文特征的领域观点词抽取方法和极性判别方法获得了如表 2 所示的评测结果。

从表 2 的评测结果中可以看到，基于上下文特征的领域观点词抽取方法在三个领域上效果基本一致，这说明该方法相对领域独立。抽取结果的精确率与召回率相比，精确率较好，召回率较差，其原因是由于候选观点词多以形容词为主，而副词、动词及名词都有可能成为领域观点词，考虑的较少。另外，极性判别方法相对简单，导致判别结果不够准确，还应基于更多的领域知识判别极性。

表 2 领域观点词抽取结果

Table 2 Results of domain opinion word extraction

标识		Precision @1000	Precision	Recall	F1	Raccuracy
单个领域	电子产品	0.502	0.254	0.076	0.1170	0.076
	影视娱乐	0.503	0.255	0.0773	0.1186	0.0773
	金融证券	0.502	0.261	0.0794	0.1218	0.0794
	Median	0.5712	0.3430	0.0947	0.1475	0.0947
	Best	0.674	0.6125	0.1194	0.1833	0.1194
宏平均	MI&Tlab	0.5023	0.2567	0.0776	0.1191	0.0776
	Median	0.5712	0.3429	0.0947	0.1478	0.0947
	Best	0.6567	0.4861	0.1136	0.1744	0.1136
微平均	MI&Tlab	0.5023	0.2567	0.0776	0.1191	0.0776
	Median	0.5712	0.3379	0.0947	0.1475	0.0947
	Best	0.6567	0.4681	0.1135	0.1744	0.1135

（2）　评价搭配抽取结果

基于条件随机场模型的评价搭配抽取和极性判别结果如表 3、表 4 和表 5 所示。

表 3 评价对象抽取结果

Table 3 Results of opinion attribute extraction

	标识	Precision @1000	Precision	Recall	F1	Raccuracy
宏平均	MI&Tlab	0.0553	**0.1229**	**0.0518**	0.0729	0.0518
	Median	0.0655	0.0742	0.0455	0.0543	0.0455
	Best	0.111	0.1339	0.0817	0.0916	0.0817
微平均	MI&Tlab	0.0553	**0.1209**	**0.1314**	0.1259	0.1028
	Median	0.0655	0.1012	0.0818	0.0834	0.0698
	Best	0.111	0.1598	0.1491	0.1447	0.1357

表 4 评价短语抽取结果

Table 4 Results of opinion phrase extraction

	标识	Precision @1000	Precision	Recall	F1	Raccuracy
宏平均	MI&Tlab	0.0296	**0.0715**	0.0295	0.0418	0.0238
	Median	0.0514	0.0553	0.0505	0.0467	0.0350
	Best	0.1183	0.0856	0.1006	0.0872	0.0664
微平均	MI&Tlab	0.0296	0.0514	0.0559	0.0536	0.0425
	Median	0.0514	0.0740	0.0619	0.0620	0.0520
	Best	0.1183	0.1174	0.1055	0.1024	0.0954

表 5 评价搭配抽取结果

Table 5 Results of polarity collocation extraction

	标识	Precision @1000	Precision	Recall	F1	Raccuracy
宏平均	MI&Tlab	0.0103	0.0234	0.0081	0.0121	0.0071
	Median	0.0272	0.0251	0.0204	0.0198	0.0151
	Best	0.0716	0.0396	0.0376	0.0347	0.0257
微平均	MI&Tlab	0.0103	0.0124	0.0135	0.0129	0.0108
	Median	0.0257	0.0331	0.0272	0.0273	0.0231
	Best	0.0716	0.0615	0.0468	0.0483	0.0433

从表 3 和表 4 的评价对象和评价短语的抽取结果中可以发现，基于条件随机场的一体化抽取方法比较有效，尤其是评价对象抽取的精确率和召回率，均超过评测平均值并接近

最大值，评价短语抽取结果在平均值上下。从表 5 的评价搭配抽取结果可以发现，由于极性判别结果采用任务 1 方法，可能导致评测结果较差。

5 结论

随着情感分析研究的不断深入，领域知识和上下文语境对情感分析结果的影响越来越受到关注。在评测中，领域知识和上下文语境的影响体现在领域观点词抽取和评价搭配抽取的任务上。本文提出了基于多特征融合的情感分析方法来解决相关问题，在评测中获得了较好的结果。但是，多特征融合的方法更多考虑的是上下文语境的影响，并没有充分考虑领域知识的因素，因此在下一步工作中，将对领域知识的构建进行研究，并考虑与情感信息抽取任务的有效结合。

参 考 文 献

1 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报.2010，21（8）：1834-1848

2 黄萱菁，赵军. 中文文本情感倾向性分析[J]. 中国计算机学会通讯. 2008，4(2)

3 B. Liu, M. Hu, and J. Cheng, Opinion Observer: Analyzing and Comparing Opinions on the Web[C], Proceedings of International World Wide Web Conference. Japan, May 2005: 342-351

4 AM Popescu,O Etzioni. Extracting Product Features and Opinions from Reviews[C], Proceedings of Empirical Methods in Natural Language Processing, 2005:339-346

5 赵军，许洪波，黄萱菁，等. 第一届中文倾向性分析评测技术报告. 第一届中文倾向性分析评测会议[C]. 2008:1-20.

6 许洪波，姚天昉，黄萱菁，等. 第二届中文倾向性分析评测技术报告. 第二届中文倾向性分析评测会议[C]. 2009:1-23.

7 刘群，李素建. 基于《知网》的词汇语义相似度计算. 第三届中文词汇语义学研讨会论文集，2002.

8 John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the Eighteenth International Conference on Machine Learning, 2001：282-289.

HIT_SCIR_OMS：情感分析系统

唐都钰，胡燊，赵妍妍，秦兵，刘挺
哈工大社会计算与信息检索研究中心，哈尔滨，150001
dytang@ir.hit.edu.cn

摘　要：情感分析是自然语言处理领域的重要问题。本文主要研究了 COAE 2011 中的两项任务，领域观点词的抽取和评价搭配抽取。针对任务一，首先自动构建观点词词典，继而生成依存句法路径库，最终使用句法路径匹配的方法抽取领域观点词；针对任务三，首先构建短语句法路径库，进而通过句法路径泛化、评价搭配合并等方法抽取评价搭配，同时赋予其极性。评测结果验证了我们方法的有效性，同时也为我们指明了未来的研究方向。

关键词：情感分析；观点词；评价搭配；依存句法；短语句法

HIT_SCIR_OMS: An Opinion Mining System

Duyu Tang, Shen Hu, Yanyan Zhao, Bing Qin, Ting Liu
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, 150001
dytang@ir.hit.edu.cn

Abstract: The research of sentiment Analysis is a important part in naturel language processing.This paper studies two tasks in COAE 2011, which are polarity word identification and polarity collocation extraction. For the first task, this paper automatically build polarity word lexicon, and then identify the domain polarity word via the method of matching denpendency syntactic relation route. For the third task, this paper construct the phrase syntactic relation library, and then extract polarity collocation and its polarity via the generalization of syntactic route, combination of polarity combination, etc. Evaluation results show that our method is effective, and it also indicates the direction of our research work in the future.

Keywords: Sentiment Analysis; Polarity Word; Polarity Collocation; Dependency Syntactic Relation; Phrase Syntactic Relation

1. 引言

情感分析又称意见挖掘，按照处理文本的粒度不同可以分为词语级、短语级、句子级和篇章级等[1]，通过结合文本挖掘、信息抽取、机器学习、自然语言处理等文本处理技术对主观性文本进行分析、处理和归纳。情感分析近几年持续成为自然语言处理领域研究的热点问题，可以广泛应用到很多的自然语言处理问题中，如信息抽取、自动问答、产品口碑等[2]。TREC 评测、NTCIR 评测以及前两届中文倾向性分析评测推动和加速了倾向性分析研究的发展。而随着文本倾向性分析研究的深入，国内外的研究发现领域知识和上下文语境(Context)对倾向性判别至关重要，吸引了国内外研究学者的广泛关注，并纷纷开展诸如领域观点词表构建、跨领域倾向文本分类等研究。

为了探索中文倾向性分析的新技术、新方法，推动中文倾向性分析理论和技术的研究及应用，COAE2011 为中文倾向性分析提供了一个很好的平台。本次评测包括下面四个子

任务：

(1) 领域观点词的抽取及极性判别

(2) 中文观点句抽取

(3) 评价搭配抽取

(4) 观点检索

哈工大社会计算与信息检索研究中心开发了 HIT_SCIR_OMS 情感倾向性分析系统，主要参与了 COAE 2011 的任务一和任务三。本文的第二部分介绍任务一领域观点词抽取的识别方法；第三部分介绍任务三评级搭配抽取及其倾向性判别方法；第四部分介绍评测结果和对结果的简单分析；第五部分介绍本文的小结。

2. 领域观点词抽取

2.1. 任务分析

本任务要求分别从电子产品、影视娱乐和财经三个不同领域数据集中抽取观点词，并判断该观点词的极性(褒义、贬义)。即考虑领域对观点词表达的影响，不同领域之间所使用的观点词不同，同一观点词在不同领域也有不同的倾向性。此外，观点词是表达对外评价的褒贬观点词，不是表达人物自身情绪的倾向词，也不包括评价短语。任务要求每个领域返回 2000 个观点词，并按照置信度降序排列。

本任务的难点在于如何从不同领域抽取观点词，并如何综合上下文信息对观点词赋予置信度。如，包含“圆滑”的两个句子：

(1) 该**相机的外形**边**角十分**圆**滑**。

(2) **他的性格很**圆**滑**。

句子(1)中“圆滑”修饰“外形边角”，表现褒义的情感；句子(2)中“圆滑”修饰“性格”，表达贬义的情感。在句子(1)中，“圆滑”表达作者的情感，我们称之为观点词；“外形边角”是观点词修饰的主体，我们称之为目标词。本文使用依存句法路径表示观点词和目标词之间的关系，使用依存句法路径匹配的方法抽取领域观点词，同时根据观点词的句法路径信息以及上下文信息进行置信度排序。

2.2. 方法介绍

针对领域观点词抽取的任务，我们首先构建一个大规模的观点词词典，并获取所有包含观点词的句子。继而对于每个包含观点词的句子，使用观点词和目标词之间的依存句法关系计算该观点词含有极性的可能，最终对抽取的观点词及其上下文进行置信度排序。

2.2.1. 构建观点词词典

观点词词典的构建共有三个步骤，如下所示：

Step1：借助知网情感分析用词语集(beta 版)中的中文正面评价词典和中文负面评价词典，形成原始观点词词典 D_1(包含观点词及其极性)。

Step2：使用 HowNet[4]对原始观点词词典 D_1 进行同义词扩展，扩展出的观点词沿袭原始观点词的极性，形成观点词词典 D_2。
Step3：合并原始观点词词典 D_1 和扩展观点词词典 D_2，手工过滤后得到最终的观点词词典 D。

通过观点词词典构建，最终获得观点词 9892 个，其中褒义观点词 5021 个，贬义观点词 4871 个。

2.2.2. 构建目标词词典

为了获取准确的依存句法路径库，我们针对电子产品、影视娱乐和财经三个领域分别构建目标词词典。这里我们认为目标词是名词，目标词词典的构建共有三个步骤，如下所示：
Step1：预处理。获取所有包含观点词的句子，对每个句子使用 LTP[5](Language Technology Platform)进行分词和词性标注。
Step2：对于每个领域，统计每个名词在该领域出现的频率，并按照该频率降序排列，获取所有词频大于 100 的名词，形成目标词词典 T_1。
Step3：对 T_1 进行手工过滤，获得最终的目标词词典 T。

通过对三个领域构建目标词词典，我们最终获得的目标词词典规模如表 1 所示，部分领域目标词及其出现频率如表 2 所示。

表 1 领域词典规模

领域	目标词词典规模	依存句法路径库规模
数码	276	8
娱乐	239	9
金融	171	10

表 2 目标词词典样例

领域	目标词样例
数码	产品(24465)、功能(12538)、价格(9977)、视频(8449)、手机(8103)、接口(7098)、外观(6625)、性能(6301)、机身(6018)、屏幕(5808)
娱乐	音乐(3592)、明星(2907)、电影(2500)、八卦(2241)、视频(2113)、电视剧(1679)、演唱会(1534)、导演(1442)、歌手(1421)、演员(1350)、节目(1258)
金融	财经(9577)、公司(9269)、市场(3959)、股价(3276)、基金(2920)、交易(2635)、经济(2135)、证券(1974)、港币(1913)、企业(1892)

2.2.3. 构建依存句法路径库

我们将观点词和目标词之间的句法关系用句法路径的形式表示，并使用 LTP 对句子进行依存句法分析。以“这是一个很漂亮的相机”为例，使用 LTP 进行句法分析的结果如图 1 所示。

[4] http://www.keenage.com/
[5] http://ir.hit.edu.cn/demo/ltp/

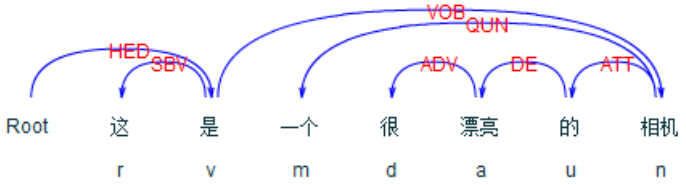

图 1 依存句法分析样例

Fig.1 Sample for Dependency Syntactic Relation

依存句法路径库构建共包含四个步骤，如下所示：

Step1：预处理。对各领域数据进行分句、分词、词性标注和依存句法分析。其中以短句为单位进行分句，即以“，”、“。”、“？”、“！”等作为分隔符。

Step2：为了确保生成句法路径库的准确性，减少依存句法分析的错误，我们仅保留同时包含观点词和目标词的句子，并要求句子中的词语长度小于 70。这是因为随着句子长度的增加，依存句法分析的准确率有所下降，而且长度大于 70 的句子占所有句子总量的 2%。

Step3：对与数据集中的每个句子，将句中的每个观点词和目标词进行匹配，并使用句法路径的形式表示二者之间的依存句法关系。以图 1 中的句子为例，“漂亮”是观点词，“相机”是目标词，二者满足 ATT 修饰关系，句法路径表示为“DE+ ATT+ back”，其中“back”表示目标词的位置在路径的最后面。

Step4：对于每个领域，统计每条依存句法路径出现的频率，并按照频率降序排列。最终，我们选取每个领域的前 10 条句法路径，手工过滤后形成各领域句法路径库。

经过以上四个步骤，我们分别获得三个领域的依存句法路径库，各领域规模如表 1 所示。

2.2.4. 领域观点词抽取和极性识别

基于观点词词典和依存句法路径库，领域观点词抽取主要包含四个步骤，如下所示：

Step1：预处理。同 2.2.3 中 Step1 的预处理操作相同。

Step2：遍历观点词词典，获取包含观点词的句子。

Step3：对每一个包含观点词的句子，获取该句中所有观点词和候选目标词(名词)之间的依存句法路径，若依存句法路径库中包含该路径，则认为该观点词在句中表现极性，抽取该观点词同时赋予其在观点词词典中的极性。

Step4：对于每个抽取出的观点词，获取其上下文信息。

通过上面四个步骤，我们获取了领域观点词及其上下文，在此基础上进行置信度排序。

2.2.5. 置信度排序

为了对所获取的观点词及其上下文进行排序，我们为每一个观点词及其上下文设置置信度，具体置信度设置规则如下：

(1) 置信度 C_0=2000。观点词的修饰对象在目标词词典中，观点词在上下文中是形容词。

(2) 置信度 C_0=1000。观点词的修饰对象在目标词词典中，观点词在上下文中不是形容词。

(3) 置信度 C_0=800。观点词的修饰对象不在目标词词典中，观点词在上下文中是形容词。

(4) 置信度 C_0=700。观点词的修饰对象不在目标词词典中，观点词在上下文中不是形容词。

在上述四条规则的基础上，最终置信度计算公式为： $C = C_0 + 0.0001 * \#token$

其中#token 为该观点词在抽取结果中总共出现的次数。

3. 评价搭配抽取

3.1. 任务分析

本任务关注上下文语境对观点识别和倾向性判断的影响。要求抽取出观点句中的评价对象、评价短语，并判断其倾向性。其中评价对象是指评论针对的对象或对象的属性；评价短语是指修饰成分，评价对象和评价词语共同组合成评价单元；修饰成分指加强、减弱或置反观点的语言成分，如程度副词和否定词等。

本任务的难点在于如何针对不同领域正确抽取评价对象，继而抽取评价搭配并同时赋予其极性。如，句子“这款相机拥有出色的画质”中，“画质”为评论针对的对象，在句中是评价对象；“出色”是评价对象的修饰成分，在句中是评价词语；“出色”和“画质”共同组成了评价搭配，并且在该上下文中表现极性为褒义。本文使用短语句法分析表示评价词和评价词语之间的句法关系，使用句法路径匹配的方法抽取评价搭配。

3.2. 方法介绍

针对评价搭配抽取任务，我们借助 2.2 中构建的观点词词典，获取所有包含观点词的句子。对于每一个句子，获取句中所有的候选评价词语和评价对象，并使用评价词语和评价对象之间的短语句法路径表示二者之间的关系，最终通过短语句法路径匹配抽取评价搭配，同时综合考虑其上下文信息对评价搭配赋予极性。

3.2.1. 构建短语句法路径库

短语句法路径库构建主要分为句法路径生成和句法路径泛化两个步骤。句法路径生成模块需要确定句子中的评价词语和评价对象，其中的评价词语来自 2.2.1 中的观点词词典，由于评价对象一般为名词(例如：“屏幕亮度”、“电池寿命”)，因此本文设定评价对象为名词，在短语句法中表示为 NN。

基于上述分析，若一个观点句中包含 n 个评价词语，m 个候选评价对象，则会产生 n*m 个候选句法路径。短语句法路径库构建共包含四个步骤，如下所示：

Step1：预处理。对各领域数据进行分句、分词、词性标注和短语句法分析。分句以短句为单位，即以“，”、“。”、“？”、“！”等作为分隔符，获取数据集 R。

Step2：对 R 中的每个句子，将任意评价词语和候选评价对象(NN)进行搭配，获取二者之间的短语句法路径。以句子“这是一个很漂亮的相机”为例，“相机”和“漂亮”之间的短语句法路径为“VA+ VP+ VP+ IP + CP+ NP- NP- NN- back”，其中“back”表示评价对象在句法路径的最后面。

Step3：句法路径泛化。某些句法成分标签表达相似的含义，我们用一个规范化的标签将它们代替。这里我们将“CD”和“NN”统一用“NN”表示；将“VV”和“VB”统一用“VB”表示。

Step4：对每个领域生成的句法路径按照出现频率降序排列，并选取前 10 条进行人工过滤，最终生成短语句法路径库。

3.2.2. 评价搭配抽取和极性识别

基于短语句法路径库，评价搭配抽取及其极性识别共包含四个步骤，如下所示：

Step1：预处理。与 3.2.1 中 Step1 的预处理步骤相同。同时遍历评价词词典，获取包含评价词的句子。

Step2：抽取评价搭配。将待处理句子中所有的名词(NN)作为候选评价对象，获取每个句子中任意评价词语和候选评价对象之间的短语句法路径，若短语句法路径库中包含该路径，则抽取该评价搭配。

Step3：合并评价搭配。我们分别对评价词语和评价对象进行了合并操作。即如果两个评价搭配具有相同的评价词语并且评价对象相邻，那么将这两个评价对象合并；类似地，如果两个评价搭配具有相同的评价对象同时评价词语相邻，则将评价词语进行合并。

Step4：合并修饰成分。我们人工构建了修饰词典，通过人工挑选修饰词种子、同义词扩展和人工过滤，最终生成修饰词典，共包含 140 个修饰词语，其中正面修饰词 103 个，反面修饰词 37 个。修饰词典样例如表 3 所示。对于一个评价搭配，若评价词语与修饰词典中的修饰词语相邻，则将修饰词语和评价词语进行合并。

Step5：对最终的评价搭配赋予极性。评价搭配的极性计算公式为：$C = C_{pol} * C_{mod}$，其中 C_{pol} 是评价词语在评价词典中的极性，褒义为+1，贬义为-1；类似的，C_{mod} 是修饰词语在修饰词典中的极性，正面修饰词为+1，反面修饰词为-1。

表 3 修饰词典样例

修饰词语类型	修饰词语样例
正面修饰词语	非常、分外、相当、格外、特别、更、更加、很、极其、略微
反面修饰词语	并非、并未、不够、从不、决不、绝不、绝非、没、不、没有

4. 评测结果

表 4 列出了观点词识别任务一的评测结果。评测采用 pooling 的方式，即每个提交结果提取前 1000 条组成评测池，经过人工判断后作为答案，评测指标为 P@1000、P(准确率)、R(召回率)、F1 和 Raccuracy 表示。

在实验中，我们使用 LTP 依存句法分析器完成句法分析任务。

表 4 观点词识别评测结果对比

标识	宏平均			微平均		
	P@1000	F1	Raccuracy	P@1000	F1	Raccuracy
HIT-SCIR	0.5737	0.1498	0.095	0.5737	0.1491	0.095
Median	0.571267	0.14788	0.09474	0.571267	0.147547	0.09474
Best	0.6567	0.1744	0.1136	0.6567	0.1744	0.1135

表 5 和表 6 列出了评价对象、评价短语以及总体极性识别的评测结果。通过标注完全答案集，采用自动评价方法进行评测，主要指标为 F 值，其他指标包括 P、R 等。

在实验中，我们使用 Charniak[3]短语句法分析器来完成句法分析任务。

表 5 各领域评价搭配抽取和极性识别评测结果

run-tag	domain	P@1000	Precision	Recall	F1	Raccuracy
HIT-SCIR	D	0.064	0.059657	0.032917	0.042425	0.032917
HIT-SCIR	E	0.013	0.014706	0.007813	0.010204	0.007813
HIT-SCIR	F	0.006	0.006501	0.009217	0.007624	0.009217

表 6 评价搭配抽取和极性识别评测结果对比

	宏平均			微平均		
run-tag	P@1000	F1	Raccuracy	P@1000	F1	Raccuracy
HIT-SCIR	0.027667	0.020584	0.016649	0.027667	0.034381	0.02758
average	0.02719	0.019833	0.015183	0.025778	0.027323	0.02309
max	0.071667	0.03478	0.025728	0.071667	0.048312	0.043298

通过分析评测结果，我们发现，任务一中的 P@1000 和 F 值都处于平均值水平，说明通过依存句法路径匹配可以解决部分观点词挖掘问题，尚有提升的空间；任务三中电子产品领域的准确率达到 5.9657%，F 值为 4.24%，而财经领域和影视娱乐领域 F 值均不足 2%，导致最终的宏平均和微平均 F 值不高。

5. 小结

HIT_SCIR_OMS 情感分析系统主要参与了 COAE2011 评测的任务一和任务三两项任务。针对任务一领域观点词抽取任务，我们通过自动构建观点词词典，生成依存句法路径库，最终使用句法路径匹配的方法抽取领域观点词；针对任务三评价搭配抽取任务，我们通过构建短语句法路径库、句法路径泛化、评价搭配合并等方法抽取评价搭配，同时赋予其极性。

此外，评测过程中，我们总结发现，评价对象(领域目标词)的抽取对领域观点词和评价搭配抽取起着至关重要的作用，准确地获取大量领域评价对象可以有效地提高领域观点词和评价搭配抽取的效果。同时，我们的方法还存在一些不足，如：领域观点词识别的准确率不高，这将是我们未来工作的重点。

参 考 文 献

[1] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报，2010.1834-1848.

[2] B.Pang, L.Lee. Opinion Mining and Sentiment Analysis[M]. 2008.

[3] Eugene Charniak. 2000. A maximum-entropy-inspired parser. In Proceedings of NAACL-2000, pages 132-139.

第三届中文倾向性分析评测 ISCAS-Opinion 系统报告*

韩先培，孙乐，江雪
中国科学院软件研究所，北京，100190
E-mail: xianpei@nfs.iscas.ac.cn

摘　要：本文介绍了中科院软件所参与第三届中文倾向性分析评测的 ISCAS-Opinion 系统。系统一共参与了两个任务，分别是领域观点词抽取和极性判别（任务 1）以及中文观点句抽取（任务 2）。针对倾向性分析的多领域问题，系统采用集成学习的方法，使用领域特定分类器的结果作为领域无关的特征，从而构建具有领域自适应性的中文观点句抽取系统。此外系统将领域观点词抽取建模为特征选择任务，使用卡方统计量（CHI）对领域观点词候选进行打分，并使用前 2000 个词作为领域观点词。实验结果表明，领域自适应的分类器仍然与有监督的分类器有较大的性能差距。

关键词：倾向性分析；情感词抽取

ISCAS-Opinion at COAE 2011

Xianpei Han, Le Sun, Xue Jiang
Institute of Software, Chinese Academy of Sciences, Beijing 100190
E-mail: xianpei@nfs.iscas.ac.cn

Abstract: This paper describes the opinion analysis system — ISCAS-Opinion for the COAE 2011 evaluation. Our system participate two tasks: the domain opinion lexicon building task and the sentence sentiment classification task. Specifically, our system employs ensemble classification method for the sentence sentiment classification task: first classify each sentence using the domain-specific lexicon based classifier and the domain-specific self-learning classifier, then the classification result is used as domain independent feature of the meta-SVM classifier. For the domain opinion lexicon building task, our system employs the text feature selection method – we use the Chi-squared statistics to rank all opinion word candidates, then the top 2000 candidates are used as finally opinion words. Experimental results show that our method can achieve competitive performance.

Keywords: Sentiment Analysis; Opinion Lexicon Building

1　引言

第三届中文倾向性分析评测(COAE 2011)共有四个任务：领域观点词抽取与极性判别、中文观点句抽取、评价搭配抽取和观点检索。我们参与了其中两个任务：领域观点词抽取和极性判别（任务 1）以及中文观点句抽取（任务 2）。领域观点词抽取和极性判别任务的目标是从特定领域的文本数据集中抽取观点词，并判断抽取出观点词的极性（包括褒义和贬义）。例如，给定如下三个句子：

1) *亮度适中，色彩鲜艳。*
2) *音响效果不好。*
3) *亚马逊计划推出电子书。*

*本文受国家自然科学基金项目（编号：90920010，61100152，60736044 和 61003117）资助，并特此感谢中科院自动化所刘康提供 digital 领域标注语料

系统需要抽取出上述句子中的三个观点词***适中***、***鲜艳***和***不好***，并判断它们的极性依次为*褒义*、*褒义*和*贬义*。

中文观点句抽取任务的目标则是从文本数据集中识别出所有表达观点的句子及其所表达观点的总体极性（褒义、贬义或混合观点）。具体的，给定上面三个句子，系统需要识别出句子 1 和句子 2 是观点句，并判别这两个句子的极性分别是褒义和贬义。

通常中文观点句抽取被建模为文本分类的任务：给定一个句子，系统根据训练语料构建分类器，并将新来的句子划分到无观点、褒义、贬义和混合观点（根据任务的不同，类别的设定可能不一致）等多个类别中去。但是与传统文本分类不同的是，倾向性分析常常面临多领域的问题：也就是使用特定领域语料训练的分类器通常难于在另一个领域取得良好的分类性能。多领域问题使得倾向性分析算法需要具有一定的领域自适应性。不幸的是，考虑到领域之间特征分布的差异，构建领域自适应的分类器通常是一个极为困难的任务。

为解决上述领域自适应性问题，我们的中文观点句抽取系统采用集成学习的方法：首先使用领域特定的无监督分类方法（分别是基于情感词典的分类方法和基于自学习的方法）对领域内的句子进行分类，并使用这些分类结果作为样本的特征。上述过程可将领域相关的特征（如领域特定的实体和观点词）转换为领域无关的特征（情感词典分类结果和自学习方法分类结果）。在此基础上，系统使用单领域的已标注语料进行训练，从保证构建的中文观点句抽取系统具有一定的领域自适应性。

基于中文观点句抽取结果，我们的系统将领域观点词抽取和极性判别建模为显著特征选择任务。具体的，对每一个领域的文本数据集，系统首先根据词性选择情感词的候选，然后计算每一个候选词相对每一个类别的卡方统计量（CHI）来对其进行打分，最后打分在前 2000 位的情感词候选被认为是最终的领域情感词。

本文按如下方式进行组织：在第二节描述了系统所采用的句子级观点分类方法，第三节中介绍了系统的领域观点词抽取方法，第四节介绍相关实验及其结果，第五节中对本文工作进行了总结。

2　中文观点句抽取

在 ISCAS-Opinion 系统中，我们将中文观点句抽取任务建模为文本分类任务：给定一个句子，系统通过将其划分到无观点、褒义和贬义三个类别中（这里不考虑混合观点）来实现观点句抽取。在 COAE 2011 评测中，系统需要对三个领域（包括 digital 领域、entertainment 领域和 finance 领域）的文本进行观点句抽取，而我们只有 digital 领域的训练语料，因此所构建分类器的领域自适应性对最终的观点句抽取性能至关重要。

考虑到文本层面的特征往往是领域特定的（如实体和观点词往往都是领域特定的），因此直接使用文本层面特征通常难于取得良好的领域自适应性能。为克服上述问题，ISCAS-Opinion 系统采用集成学习的策略：首先构建基于情感词典的分类器和自学习的领域特定分类器，并使用这些分类器的分类结果作为句子的特征表示，从而使得特征空间是领域无关的，基于上述特征表示，系统使用单领域的训练语料也可构建具有自适应特性的分类器。具体方法如以下所述。

2.1 基于情感词典的分类器

观点分类的最基本方法是基于情感词典的分类方法，其基本出发点是：如果一个句子中褒义词所占的权重大，则这个句子表达的观点有很大可能是褒义，同理如果句子中贬义词所占的权重大，则这个句子表达的观点有很大可能是贬义。

基于上述出发点，给定褒义词词典 D_p 和贬义词词典 D_n，系统判别一个句子 $S = w_1w_2\ldots w_n$ 的观点是褒义还是贬义的概率 T^p 和 T^n 计算如下：

$$T^p = \sum_{w_i \in D_p} f(w_i, S) \Big/ \sum_{w_i \in D_n \cup D_p} f(w_i, S)$$

$$T^n = \sum_{w_i \in D_n} f(w_i, S) \Big/ \sum_{w_i \in D_n \cup D_p} f(w_i, S)$$

上述公式中 *f(w, S)*是情感词 *w* 在句子 *S* 中的重要性打分。

从上面公式可以看出，基于情感词典的分类器的关键在于如何计算情感词的权重 *f(w, S)*。词权重的直观计算方法是使用词 *w* 在句子中的出现次数，但是这种方法往往不能取得很好的效果。在本文中我们提出一种基于管辖长度的重要性打分算法，其具体步骤如下:

(1) 首先，系统根据词性抽取出句子中的评价对象候选和情感词候选，其中评价对象候选为句子中的名词性成分（词性包括 Ng，n，nr，ns，nt，nz，vn，an，r 和 l）以及实体词（词性为 LOC，ORG 和 PER），情感词候选为句子中的形容词性成分（词性包括 a，ad，an，Ag，d，v，vd 和 i）。例如，给定下面的句子：

1) 亮度适中，色彩鲜艳。

系统抽取其评价对象候选为{*亮度，色彩*}，情感词候选为{*适中，鲜艳*}。

(2) 根据上面抽取出的评价对象候选和情感词候选，系统基于词语之间的搭配关系选择每一个情感候选词的评价对象。具体的，对每一个情感候选词，系统选择与其有最高搭配频次的评价对象候选作为其评价对象，其中搭配频次使用 Sogou 搭配词典[6]（[1]）进行获取。在上述例子中，根据 Sogou 中文搭配词典，候选词之间的搭配频次如下：

{*亮度-鲜艳 = 0；亮度-适中 = 23；色彩-鲜艳 = 4653；色彩-适中 = 0*}。

基于上述搭配频率，系统选择*亮度*作为*适中*的评价对象，*色彩*作为*鲜艳*的评价对象。

(3) 根据评价对象和情感词候选的搭配关系，一个情感词的权重计算方式如下：

$$f(w, S) = Length(w) + Length(c(w))$$

其中 *c(w)*是词 *w* 的评价对象搭配词，*Length(w)*是词 *w* 包含的字数。

[6] http://www.sogou.com/labs/dl/r.html

基于上述权重计算方法，基于情感词典的分类方法输出如下结果作为元分类器的特征：分类结果，T^p 和 T^n。

2.2 自学习分类器

除基于情感词典的分类方法之外，系统还采用了自学习分类器来获取额外的元分类器特征。系统采用文献[2]中提到的自学习策略，对每一个领域，系统选取基于情感词典的分类器的分类结果中具有前 20%置信度的褒义句子和前 20%置信度的贬义句子作为训练样本，并利用这些训练样本对每一个领域构建一个特定的文本分类器。然后这些分类器的结果（包括分类结果、褒义概率和贬义概率）被输出为元分类器使用的特征。

具体的，系统采用最大熵分类方法，每一个句子的特征包括词、命名实体和具有搭配关系的词对。同时我们还使用基于卡方统计量的文本特征选择方法（[3]）进行了特征选择。

2.3 基于 SVM 的集成分类

经过上述两个步骤的处理，三个领域（digital 领域、entertainment 领域和 finance 领域）的句子都被表示为一个具有六个特征的向量：

{基于情感词典的分类结果，T^p，T^n，自学习分类器分类结果，自学习分类器褒义概率，自学习分类器贬义概率}。

相比于文本特征，我们认为上述特征在领域之间的分布具有更高的一致性，这也就保证学习出来的分类器具有更高的领域自适应性。基于上述观察，系统使用 digital 领域的标注语料，采用 SVM 算法（使用 LibSVM 工具包[7]，基于多项式核）构建了元分类器。

最终，元分类器将三个领域内的所有句子分类到{褒义、贬义、无观点}三个类别上，并输出褒义句子和贬义句子作为观点句抽取结果，所有观点句按照其分类概率进行排序。

3 领域观点词抽取

在 ISCAS-Opinion 系统中，我们将领域观点词抽取看成是一个文本分类特征选择的任务：对每一个领域，系统计算每一个观点候选词与特定观点类别（褒义类别或贬义类别）的相关度，并将排在前 2000 位的词语作为该领域的观点词。上述观点词抽取方法的关键在于计算观点词候选与特定观点类别的相关性，并通常可以采用文本特征选择的方法([3])。在 ISCAS-Opinion 系统中，我们使用卡方统计量（CHI）来对情感候选词进行打分，给定一个特定观点类别 *c*，情感词 *w* 的打分由如下公式计算：

$$CHI(w,c)=\frac{N\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

其中 A 表示包含词 *w* 且属于类别 *c* 的文本频率；B 表示包含词 *w* 但不属于类别 *c* 的文

[7] www.csie.ntu.edu.tw/~cjlin/libsvm/index.html

本频率；C 表示不包含词 *w* 但属于类别 *c* 的文本频率；D 表示既不包含 *w* 也不属于类别 *c* 的文本频率；N=A+B+C+D 为总的文本数。

基于上述打分，系统选取每个领域的前 2000 个词作为领域的情感词，并选择包含该词且具有最高同观点类别分类概率的句子作为其句子上下文。

4 实验

本文在 COAE 2011 的标准评测数据集上进行了测试，其中任务 1 使用的语料共包括三个领域的 44245 篇文本（其中 digital、 entertainment 和 finance 领域分别包含 14799、14904 和 14542 篇文本），任务 2 使用的语料是从任务 1 语料中抽取出来的文本子集，共计 6000 篇文本（其中 digital、 entertainment 和 finance 领域各 2000 篇）。

在中文观点句抽取（任务 2）中，ISCAS-Opinion 系统对 digital 领域进行观点句抽取时，除使用第二小节中所使用的六个特征之外，还额外构建了基于 digital 领域标注语料的有监督分类器，并将其分类结果（分类结果，褒义概率和贬义概率）也放入最终的元分类器特征中。具体的总体性能如下表 1 所示。

	Precision	Recall	F1	P@1000	Raccuracy
ISCAS-Opinion	0.267905	0.43593	0.322069	0.377667	0.311301
Average	0.2408149	0.406039	0.276324	0.290183	0.2554445
Best	0.5346933	0.723411	0.541377	0.532	0.4945113

表 1. 中文观点句抽取总体性能

从表 1 中可以看出，ISCAS-Opinion 能够取得比平均值稍好的性能，这也表明集成学习具有一定程度上的领域自适应性。

为了分析领域自适应分类器与单领域分类器的性能差距，表 2 中展示了 3 个领域的单独中文观点句抽取性能。

	Precision	Recall	F1	P@1000	Raccuracy
ISCAS-Opinion_D	0.462408	0.554792	0.504404	0.72	0.514094
ISCAS-Opinion_E	0.213675	0.366151	0.269865	0.264	0.245728
ISCAS-Opinion_F	0.127632	0.386847	0.191939	0.149	0.174081

表 2. 中文观点句抽取分领域性能

从表 2 中可以看出，有训练语料的领域（ISCAS-Opinion_D）要远远好于使用自适应分类器的领域（ISCAS-Opinion_E，ISCAS-Opinion_F），这也展示了构建领域自适应机器学习技术的难度。

在领域观点词抽取和极性判别（任务 1）中，系统采用了特征选取的方法，系统性能如表 3 中所示。从表 3 中可以看出，基于特征选取的方法能够取得与平均值相近的性能。这也表明，使用简单的方法也可能取得一个很好的观点词抽取起点。

	Precision@1000	Precision	Recall	F1	Raccuracy
ISCAS-Opinion	0.5597	0.271	0.0819	0.1258	0.0819
Average	0.5712	0.343	0.09474	0.1479	0.09474
Best	0.6567	0.486	0.1136	0.1744	0.1136

表 3. 中文观点句抽取分领域性能

5 总结

本文介绍了中科院软件所参与第三届中文倾向性分析评测的 ISCAS-Opinion 系统。系统一共参与了两个任务，分别是领域观点词抽取和极性判别（任务 1）以及中文观点句抽取（任务 2）。针对倾向性分析的多领域问题，系统采用了集成学习的方法：首先使用领域特定的无监督分类方法(基于词典的分类方法和基于自学习的分类方法)对领域内的句子进行分类，并使用这些分类结果作为元分类器的特征，从而将领域相关的文本特征转换为领域无关的特征，使得构建的中文观点句抽取系统具有领域自适应性。系统将领域观点词抽取的任务建模为特征选择的任务，使用卡方统计量（CHI）对领域观点词候选进行打分，并选取具有最高分值的前 2000 个结果作为领域情感词。实验结果表明，领域自适应的分类器仍然与有监督的分类器有较大的性能差距。

参 考 文 献

[1] 多策略融合的搭配抽取方法。王大亮，涂序彦，郑雪峰，佟子健，清华大学学报(自然科学版)，第 48 卷，第 4 期，2008 年

[2] Songbo Tan, Yuefen Wang, Xueqi Cheng: Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. SIGIR 2008: 743-744.

[3] Forman, G., An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research, 2003. 3: p. 1289--1305.

中文评论文本观点抽取方法研究

朱艳辉，徐叶强，王文华，鲁琳，杜锐，邓程

湖南工业大学计算机与通信学院，株洲，412008

E-mail: swayhzhu@163.com

摘 要：对网络评论文本自动的进行观点抽取成为了近年来自然语言处理领域研究的热点问题之一。本文在分词方面，提出了一种基于多重分词系统构建自定义词典识别专业领域生词的方法，有效减低了分词错误率。利用基础观点词词典、网络观点词词典、连词词典，提出了一种基于互信息和多重词典的领域观点词抽取与极性判别方法。在评价对象的抽取方面，提出了一种基于规则和统计相结合的评价对象识别算法；在评价对象与评价短语搭配识别方面，提出了基于 SVM 搭配识别算法。

关键字：中文评论文本；观点抽取；评价对象；极性判别。

Research On Opinion Extraction of Chinese Review

Zhu Yanhui, Xu Yeqiang,Wang Wenhua, LuLin, DuRui, DengCheng

School of Computer and Communication, Hunan University of Technology, Zhuzhou,412008

E-mail: swayhzhu@163.com

Abstract: In recent years,the extraction of opinion to the net's Chinese review has become the one of a hot issue to deal with the research on natural language processing.In the split words,a mothed that the recognition of special domain words based on multiple split words systems creating define dictionary was advanced in this paper.The mothed is effective going down the error rate of split words.Another method was advanced,which is the extraction of the domain opinion words and polarity judgment based on mutual information and multiple dictionary by use the baseic opinion dictionary,the network opinion dictionary and conjunction dictionary.In the extraction of opinion target,the algorithm that the recoganition of the opinion target by integrate the regulation and the statistics. In the polarity judgment and match recoganition of between the attribute feature and the opinion word,the algorithm that the match recognition based on the SVM was advanced.

Keywords: Chinese Review; Opinion Extraction; Opinion Target; Polarity Judgment

1 引言

随着电子商务和 web 2.0 应用的发展，越来越多的消费者在购买和使用产品之后，在电子商务网站、论坛、博客发表对产品的观点态度，评论包含了用户对产品的特征、功能、性能等的看法。潜在客户在购买产品之前总会咨询别人对产品的意见从而做出明智的购买决定，商家也可以根据用户的评论来规划销售策略，人工的去浏览海量评论文本是费时和低效的，并且还有滞后性和片面性。近年来，如何对大量的非结构化的网络评论文本自动的进行观点抽取成为了一个研究热点。

在本次 COAE2011 评测中，我们组队报名参加了任务 1，2，3 的评测。最终完成了任务 1 和任务 2 并提交了评测结果，但由于时间关系，任务 3 未能全部完成以致最终未能提交结果。本文将报告我们在这三个任务上所做的研究工作。

2 领域观点词抽取与极性判别

在领域观点词抽取与极性判别方面，我们利用基础观点词词典、网络观点词词典、连词词典，提出了一种基于互信息和多重词典的领域观点词抽取与极性判别方法。

2.1 构建观点词词典

2.1.1 构建基础观点词词典

词语的情感倾向分为两类：褒义和贬义，观点词语的情感倾向权值是指该词语的主观色彩强度，主观色彩越强情感倾向权值分数越大。在中文词语集中，有很多词本身就具备很强的主观色彩，比如“喜欢”,“漂亮”,“幸福”等褒义词语在任何主题的文章中，如果不考虑否定前缀词或其他一些副词连词的影响，它们都带有很强的褒义色彩；而“厌恶”,“残酷”,“暴力”等是带有贬义倾向的词语。我们把这些具备跨领域能力的观点词称为基础观点词。我们以 HowNet[1]发布的观点词语集为基础，通过人工挑选去掉一些不太常用或者情感倾向不很明显的词语，比如:“雩 ”、“怿” 、“凹凸” 等词语，得到褒义词 3219 个,贬义词 2905 个,共 6196 个基础观点词。以刘群提出的语意相似度计算公式[2]为基础，将观点词语按 Google 搜索返回的 hits 数进行排序，选择 hits 数最高的 15%词作为种子词（实验证明，种子词数量为总观点词数量的 15%时观点倾向性判断准确率达到峰值[3]），提出基础观点词语的情感倾向权值计算方法，如式（2-1）所示：

$$\text{SO-IR(w)} = \frac{\sum_{i}^{M}\text{Similarity(Key\_p}_i\text{,w)}}{M} - \frac{\sum_{i}^{N}\text{Similarty(Key\_n}_i\text{,w)}}{N} \qquad (2\text{-}1)$$

式中 SO-IR(w)表示观点词语 w 的情感倾向值，以 0 作为默认阈值，最终倾向值大于阈值为褒义，小于阈值为贬义。SO-IR(w)的数值表示 w 的倾向性强度，值越大倾向性强度越强。观点词的倾向性权值大小由这个词与种子词的语义关联的紧密程度有关，这里的种子词是指褒贬态度非常明显、强烈，具有代表性的词语。与褒义种子词联系越紧密，则词语的褒义倾向越强烈。与贬义种子词联系越紧密，则词语贬义倾向越明显。式中 Key_p 表示褒义种子词，M 为褒义种子词数目，Key_n 表示贬义种子词，N 为贬义种子词数目，M 和 N 可以相等,也可以不等。计算后去掉分类判断不正确的词语,得到正确的褒义词数 2807 个，贬义词 2474 个 ，总词语数 5281 个。具体请参看文献[3]。

2.1.2 构建网络观点词词典

由于网络上的评论文本口语化，仅仅利用书面性的观点词汇并不能满足需要，所以还需构建网络用语词典。采用两种方式构建网络观点词词典，一是利用现有的《中国网络语言词典》，挑选观点词语，挑选了83个观点词语。二是利用本次评测提供的评测语料，人工随机浏览评论语料300篇，抽取具有观点倾向性的网络词语，整理了36个观点词语。部分网络观点词词典如表2-1所示。

2.1.3 构建连词词典

在中文的表达方式中，连词起到连接词与词、短语与短语以及句与句的作用。如果一个连词连接的两个词语中，有一个是带有情感倾向的词语，那么可以判定与它连接的另

外一个词语也带有一定的情感倾向。利用这一方法在已知的部分情感特征词语的基础上进一步查找未知的情感特征项。收集这一类相关的中文连词，构建连词词典。

表 2-1 网络观点词词典部分词语表

Tab.2-1 Partial list of network opinion lexicon

网络观点词	顶。 NB 牛逼 给力 亚克西 杯具 SB 垃圾 抓狂 晕死 追捧 神马都是浮云 木有 有米有 蒜你狠 砖家 聚划算 斑 竹 大虾 灌水 潜水 拍砖 刷屏 楼主

能够用来抽取文本情感特征的连词只有 3 类：转折，递进和并列。这 3 类连词在中文文本中对文章的主观情感有较大的影响而且有着各自不同的作用。

转折连词连接的 2 个情感词，它们具有相反的情感倾向；递进连词连接 2 个情感倾向相同的情感词语，并且连词后的情感倾向要强于连词前的情感倾向。同时递进连词与转折连词一样，不同的递进连词有不同的语气程度；并列连词连接 2 个情感倾向相同的词语，与前面两类连词不同的是，并列连词连接的两词语情感权重相同。

根据连词的上述特性使用转折，递进和并列 3 类连词构建连词词典，如表 2-2 所示。

表 2-2 整理得到的三类连词集

Tab2-2 Three types of organized conjunction collections

连词类型	连词
转折	但、可是、然而、不过，却、但是、偏偏、只是、不过、至于、不料、岂知
递进	而且，更加，甚至，不如、不及，乃至，并且，乃至，况、况且
并列	和、跟、与、同、及、何况

2.2 基于多重分词系统构建自定义词典识别专业领域生词

ICTCLAS 分词系统[3]的词库是通用性词库，分词之后会对每一个词的词性做出标注，同时也会对一些专业领域的词汇造成分词错误，例如：靓丽分成了“靓/x 丽/g”,“低功耗”分成了“低/a 功/n 耗/v”等，为了降低分词错误对观点词识别的影响，提出了一种利用多重分词系统分词重现的方法来识别专业领域生词。

极易分词、庖丁分词是用 java 语言实现的开源包，对专业领域分词相对准确，例如：“低功耗”就分成“低功耗 ”，但是分词后没有标注词性。

N-Gram 是最为常用的统计语言模型，以马尔科夫模型为理论基础。在语言模型的构造中，以字、词、词性或词义等作为 N-Gram 的统计单元。本文选择 ICTCLAS 分词后的词语作为统计单元，对分词后的词语进行 2-Gram 和 3-Gram 处理。例如一个分词后的句子：“这/rzv 款/q 手机/n 对/p 蓝/a 牙/n 的/ude1 实现/v 不好/a 。/wj”，可以得到“这款”、“款手机”、“手机对”、“对蓝”、“蓝牙”、“牙的”、“的实现”、“实现不好”8 个 2-Gram 项，“这款手机”、“款手机对”、“手机对蓝”、“对蓝牙”，“蓝牙的”、“牙的实现”、“的实现不好”7 个 3-Gram 项。利用极易分词、庖丁分词分别对同一篇评论文本分词，用 2-Gram 和 3-Gram 处理的结果与两者分词的结果比较，如果 2-Gram 或 3-Gram 在极易分词和庖丁分词中同时出现，则将该 Gram 项作为自定义词库中的词。处理步骤如下：

（1） 原始语料 SetSource_txt，利用 ICTCLAS 对 SetSource_txt 分词结果为 ictclas_split_txt。

（2） 对 ictclas_split_txt 进行 2-Gram 和 3-Gram 处理，结果为 gram_txt。

（3） 对 SetSource_txt，利用极易分词处理，结果为 jiyi_split_txt。

（4） 对 SetSource_txt，利用庖丁分词处理，结果为 paoding_split_txt。

（5） 将 gram_txt 与 jiyi_split_txt 比较，将相同的词放入 jiyi_compared_txt 中。

（6）将 gram_txt 与 paoding_split_txt 比较，将相同的词放入 paoding_compared_txt 中。

（7）将 jiyi_compared_txt 与 paoding_compared_txt 比较，将相同的词作为结果 result_txt。

将 result_txt 和网络观点词词典合并成一个文件，作为自定义词库。将此词库放入 ICTCLAS 分词系统中，对原始语料重新分词，这样大大提高了分词的准确率，并且结果中带有词性标注，对自定义词库中的词，标注为/un。

将 result_txt 中的自定义词与 2.1 节中的基础观点词词典比较，把相同的词或包含基础观点词词典中的词作为自定义候选观点词集。

2.3 基于互信息和多重词典抽取领域观点词

2.3.1 基于多重词典构建候选观点词集

基于多重词典构建候选观点词集见算法 1 和算法 2：

算法 1 基于基础观点词词典、网络观点词词典、自定义候选观点词集构建候选观点词集

// ci表示已经抽取的观点词，初始为空
输入：文档di，ci{ }
输出：观点词集ci{ w1,w2,w3,…,wk}
Begin
1)使用2.2节方法对语料di分词
2)对于分词后的每一词语wi
if wi属于基础观点词词典 or wi属于网络观点词词典 or wi属于自定义候选观点词词典
将wi加入到ci中
end if

一般来说，大多数形容词都是观点词，因此将语料分词后的形容词以及算法1和算法2的结果集作为最终的候选观点词集。

2.3.2 基于互信息抽取领域观点词

Turney[5]利用具有 NEAR 操作符的搜索引擎 Altavista，通过对一个词附近的 10 个词语的窗口内同时出现另一个词的 hits 数来

算法 2 基于连词词典扩展候选观点词集

// c_i表示算法1中抽取的观点词， $O(w)$ 设为词w的情感权重
输入：文档d_i，$c_i\{w_1,w_2,w_3,\ldots,w_k\}$
输出：观点词集$c_i\{ w_1,w_2,w_3,\ldots,w_m\}$
Begin
1）在文本中查找连词
2）if 存在连词conj(w)
（1） 在conj(w)所在句中查询两个相同词性的词语w_1,w_2
（2） if w_1,w_2中任意一个存在于c_i
//假设w_1存在c_i中
（i）w_2加入c_i
（ii）if conj(w)是并列连词
赋予w_1,w_2相同的情感权重
else if conj(w)是递进连词
if w_2在conj(w)前
$O(w_2) = O(w_1)/2$
else $O(w_2) = O(w_1)*2$
else if conj(w) 是转折连词
$O(w_2) = - O(w_1)$
endif
3）循环执行步骤2，直到文本结束
End

计算两个词语的 PMI，进而将与各褒义词的 PMI 之和，减去与各贬义种子词的 PMI 之和后所得的结果作为情感倾向权重，如式（2-2）所示。相对 AND 操作符，NEAR 操作在度量词语之间语义关联性强度效果更好。

$$SO\text{-}PMI(w)=\log_2\left(\frac{hits(w \quad NEAR \quad K_p)*hits(K_n)}{hits(word \quad NEAR \quad K_n)*hits(K_p)}\right) \qquad (2\text{-}2)$$

其中：

K_p ={耐磨　*OR*　高清晰　*OR*　…　*OR*　低功耗}，表示正面种子词集，

K_n ={无光泽　*OR*　失真　*OR*　…　*OR*　坏点}，表示负面的种子词集，

$hits(w \quad NEAR \quad K_p)$ 表示词 w 与种子词在共现时搜索引擎返回的记录数，

$hits(K_n)$ 表示种子词单独搜索返回的记录数，当种子词确定后，该值只需要计算一次即可。

针对 2.2 节中加入自定义词库分词后的结果，通过词频统计及人工整理的方式，在本次大赛给定的三个领域中，分别确定正面种子领域观点词、负面种子领域观点词各 40 个。对 2.3.1 的候选观点词集利用公式 2-2 计算每个词的 SO-PMI 值，由于待处理的是中文文本，因此我们利用 google 搜索引擎的 NEAR 操作返回 hits 数，但在实际处理过程中，因为数据量太大，使用 google 搜索引擎效率很低，因此我们利用两个词是否在评测语料的同一文本中同现方法来进行计算，即如果 w 与 Kp 在一个文本中出现，则 $hits(w \quad NEAR \quad K_p)$ 的值为 1，否则为 0，最后按 SO-PMI 值进行排序，取绝对值最大的 2000 个词作为任务 1 的评测结果。

2.4 评测结果分析

表 2-3 COAE2011 任务一的评测结果

Tab2-3 Results of the task 1 of the COAE2011

	Precision	Recall	F1	P@1000	Raccuracy
hut_D*	0.352539	0.5821	0.439128	0.502	0.404686
hut_E*	0.172962	0.540277	0.262036	0.205	0.197722
hut_F*	0.070588	0.510638	0.124031	0.084	0.092843
Median	0.2408149	0.397946	0.276324	0.290183	0.25544452
Best	0.729751	0.798097	0.693304	0.8	0.660324

评测结果表明，digital 领域的准确率和召回率较好，而 entertainment 和 finance 领域的准确率不理想，finance 领域的最差。究其原因，在领域种子词的选择上可能有所偏差，如 finance 领域选择了“上扬、回升、反弹、大跌、暴跌、攀升”等表示客观变化的词作为了种子观点词，一定程度上影响了结果的准确性。

3 观点句抽取与倾向性分析

3.1 利用观点词抽取候选观点句

将 2.2 节中的自定义词库加入 ICTCLAS 分词系统中，对大赛给定的句子集进行分词。将分词结果与第 2 节中的领域观点词比较，若句子中包含一个或多个领域观点词，则将该句作为观点句的候选句，并且在句子中标记领域观点词。

3.2 过滤非观点句

不是所有包含观点词的句子，都是观点句。例如:“数据显示，今天的股市下跌了 0.1%”，虽然“下跌”是一个观点词，但是整个句子只是在陈述一个下跌的事实，并不是观点句。因此，利用两条规则来过滤客观句和确定观点句。

规则 1：利用客观表示词过滤客观句

收集“数据显示”、“调查表明”等词构建客观表示词词典，将词典作为自定义词库加入 ICTCLAS 分词系统对观点句的候选句进行分词。利用客观表示词词典与分词后的句子匹配，将出现客观表示词的句子过滤。从而得到新的观点句的候选句集。

规则 2：利用主观表示词进一步确定观点句

收集“认为”、“觉得”等词构建主观表示词词典，将词典作为自定义词库加入 ICTCLAS 分词系统对规则 1 中得到的候选句进行分词。根据主观表示词词典，标注候选句中出现主观表示词的词频。采用公式（3-1）进一步确定观点句。

$$P = Object_f * a + Opinion_f * (1 - a) \quad (3\text{-}1)$$

其中：P 为最终的权值，也是任务 2 中的置信度标识。$Object_f$ 表示句子中出现主观表示词的个数，同一个词出现多次，那么相应的次数也算是多次。$Opinion_f$ 表示观点词出现的次数，同一个词出现多次，那么相应的次数也算是多次。a 是一个可调因子。

根据实验给 P 的值设定阈值，大于阈值的为观点句，得到最终的观点句集。

3.3 判别观点句极性

若观点句中包含的观点词均为正面（根据任务 1 的结果），则标注此观点句为正面；若观点词均为负面，则标注观点句为负面；若同时包含正面观点词和负面观点词，则标注为中性。

3.4 评测结果分析

表 3-1COAE2011 任务二的评测结果

Tab3-1 Results of the task 2 of the COAE2011

	宏平均				
	Precision	Recall	F1	P@1000	Raccuracy
hut	0.1986963	0.54433833	0.275065	0.2636667	0.23175033
Median	0.2408149	0.40603852	0.276324	0.2901833	0.25544452
Best	0.5346933	0.72341067	0.541377	0.532	0.49451133
	微平均				
	Precision	Recall	F1	P@1000	Raccuracy
hut	0.25556	0.58367	0.355475	0.2636667	0.34869279

Median	0.3152518	0.45053215	0.357871	0.2901833	0.36317024
Best	0.654448	0.775397	0.639614	0.532	0.61142533

从评测结果可以看出，我们的准确率略低于平均水平，召回率高于平均水平，离预期目标仍然有较大差距。主要原因在以下几方面：一是仅用任务 1 结果中的领域观点词抽取观点句，具有一定的局限性。因为任务 1 的结果准确性对此有很大的影响。二是客观表示词和主观表示词的收集不全面，导致结果有偏差。三是观点句的极性是仅仅根据观点词来判定的，对准确率有一定的影响。

4 评价搭配抽取

4.1 基于规则和统计方法抽取评价对象[6]

4.1.1 设计词性序列模板构建候选评价对象集

为了获得候选评价对象集，设计了一个词性序列模板。首先对第一届中文倾向性分析评测[7]会议中任务三的 8177 条标准答案进行分词，获得评价对象的词性序列并统计出现次数，把出现次数大于 99 的词性序列作为词性序列模板，如表 4-1 所示。词性序列模板不仅包含了名词和名词短语，还包含了动词和动词短语。

表 4-1 词性序列模板

Tab.4-1 Part of speech sequence template

词性序列模板	例子	词性序列模板	例子
n	外观/n	v+n	显示/v 效果/n
n+n	来电/n 铃声/n	vi+n	拍照/vi 功能/n
v	操作/v	n+v	系统/n 反应/v
vn+n	通话/vn 质量/n	n+vn	键盘/n 设计/vn

评论文章中，评价对象会频繁出现，所以，候选评价对象在语料中出现次数越多，它是评价对象的可能性就越大。首先利用词性序列模板对语料进行匹配获得符合模板的词条，并统计在语料中的出现次数，去除那些出现特殊符号（#、|、&、@、>、<、\、〃、^、_、´、*、◆、δ、~等）的词条，把出现次数大于 4 的词条作为候选评价对象集。

4.1.2 非评价对象的过滤

不是所有的名词、名词短语、动词以及动词短语都是评价对象，仅利用设计的词性序列模板把所有满足模板的词条作为评价对象必然引入很多噪声。下面利用统计技术和自然语言处理技术，提出三条条规则对候选评价对象集进行过滤。

规则 1：利用互信息过滤

实验发现有些词汇虽然在评论语料中出现频繁，例如：“时候”，“问题”等，但这些词汇不是真正的评价对象，通过量化领域相关性对候选评价对象进行过滤，而互信息 PMI 值提供了这样一种量化方法。从语料中手工挑选 20 个典型的评价对象组成领域性种子词集 Seeds。PMI-IR 的计算公式如下：

$$PMI\text{-}IR(w_1) = \sum_{w \in Seeds} \log_2 \frac{hits(w_1 \& w)}{hits(w_1)hits(w)} \qquad (4\text{-}1)$$

hits(w1)、hits(w)是词条 w1、w 在 Google 中利用双引号技巧精确匹配后返回的页面数。hits(w1 & w)是词条 w1 和 w 在 Google 中精确匹配并同时出现的页面数。候选评价对象的

PMI-IR 值越高，它是真正的评价对象的概率就越大，例如：“通话音质”的 PMI-IR 值为 -180.79，“地方”的 PMI-IR 为-209.44，尽管前者在语料中仅仅出现了 18 次，而后者在语料中出现了 684 次，但“通话音质”是评价对象而“地方”不是。设定一个阈值，如 4-2 式所示，PMI-IR 值大于阈值的为评价对象。

$$IsAttri\_ByPMI(w_1)=\begin{cases}Yes & PMI\text{-}IR(w_1)\geq\alpha\\ No & PMI\text{-}IR(w_1)<\alpha\end{cases} \quad (4\text{-}2)$$

其中，α 值根据实验结果进行选取。

规则 2：候选评价对象中观点词、单字的动词及特殊动词的过滤。

如果一个词语是主观词或者观点词，则它肯定不是一个评价对象。基于这样的观点，利用 2.1 节的基础观点词词典对候选评价对象中的观点词进行过滤。

同时根据语言特性，单个的动词作为评价对象的可能性也比较小，为了提高准确率，所以对单字的动词也进行过滤。

在句子中，一些符合模板的词语序列，不可能构成评价对象，最容易出现歧义的模板是“v+n”，“n+v”，“vi+n”，例如：“这/rzv 款/q 手机/n 支持/v 红外/n 。/wj”，这句话中，“手机支持”、“支持红外”就不是评价对象。对于这种情况，整理了一些不能构成评价对象的动词，对包含这些动词的短语进行过滤，如表 4-2 所示。

表 4-2 不能构成评价对象的动词

Tab.4-2 Verbs that cannot form attribute words

形式动词	给与、进行、有、无、可以
趋向动词	到、上来、下来、进来、出来、回来、过来、起来、去、上去、下去、进去
心理动词	打算、喜欢、希望、害怕、担心、讨厌、愿意
助动词	能、要、会、带、放、当、用、让、看、允许、应当、该、能够、敢、可能、必须、支持
系动词	成为、当做、是、为
变化动词	下降、降低、增加、升高

规则 3：利用候选评价对象所在句子中形容词和动词性惯用语的同现比例过滤。

如果一个候选评价对象是真的评价对象，通常句子中会同时出现评价词来表达观点。所以，在句子中出现候选评价对象的情况下，同时出现形容词和动词惯用语的比例越高，那么候选评价对象是真的评价对象的可能性就越大。设定： times_with_adjorvl = 出现候选评价对象并同时出现形容词或动词惯用语的句子个数，times = 出现候选评价对象的句子数，Ratio = times_with_adjorvl / times，过滤方法如 4-3 式所示。

$$IsAttri\_ByRatio(w)=\begin{cases}Yes & Ratio\geq\beta\\ No & Ratio<\beta\end{cases} \quad (4\text{-}3)$$

其中，β 值根据实验结果进行选取。将过滤后的候选词集作为领域评价对象词典。

4.2 评价短语抽取

(1) 首先构建程度词词典，选择 HowNet 的程度词表，依据程度词的修饰程度分成高、中、低三个级别，其中高级别程度词 88 个，中级 29 个，低级 36 个[1]。将这三类程度词的修饰程度权值按低、中、高的次序确定为三个区间（0-0.4，0.3-0.7， 0.6－1），其中修饰权值大于 0.5 的程度词加强观点词的情感倾向，小于 0.5 的词弱化观点词的情感倾向。

(2) 然后构建否定词词典。否定词将使观点词在句子中的极性取反。

(3) 使用哈尔滨工业大学开发的 DeParser 句法分析器[9]对语句进行句法分析，提取由修饰词与任务 2 抽取的观点词组成的评价短语，并计算其情感倾向权值，详见文献[8]。

4.3 基于支持向量机的搭配识别算法

首先对评测语料中的每一条句子生成所有三元组搭配(笛卡尔乘积)，三元组结构为(评价对象，评价短语，0/1)，其中 1 表示搭配，0 表示不搭配。使用支持向量机搭配识别算法，利用抽取好的特征文件，训练出一个语言模型，并根据该语言模型进行二元分类。 使用支持向量机进行搭配识别，特征选择是研究的重点，我们研究确定的部分特征如下：

(1) 评价对象周围的 n 个词的词性

(2) 评价短语周围的 m 个词的词性

(3) 评价短语的字个数

(4) 评价对象和评价短语的距离

(5) 评价对象和评价短语的顺序（0 表示评价对象在评价短语前面，1 表示评价对象在评价短语后面）

(6) 评价对象和评价短语之间是否有标点符号（1 表示有，0 表示没有）

(7) 评价对象和评价短语之间是否还有特征词（1 表示有，0 表示没有）

(8) 评价对象和评价短语之间是否还有评价短语（1 表示有，0 表示没有）

(9) 句子的长度

(10) 评价短语窗口内是否存在其他评价短语

5 结论与展望

在本次评测中，我们完成了任务 1 和任务 2 并提交了评测结果，但由于时间关系，任务 3 未能提交结果。通过评测对比，结果还有较大的提升空间，造成准确率不太高的原因主要有：

（1）在分词上，借助极易分词、庖丁分词和 N-gram 来降低 ICTCLAS 分词系统在专业领域分词的误差，有一定的局限性。而分词的准确性是评测中每个任务的基础。

（2）对于任务 1，由于在领域种子词的选择上可能有所偏差，如 finance 领域选择了“上扬、回升、反弹、大跌、攀升”等表示客观变化的词作为了种子观点词，一定程度上影响了结果的准确性。

（3）任务 2 的计算方法依赖于任务 1 的结果，所以任务 1 的结果好坏直接影响了任务 2。

（4）所采取的基于互信息和多重词典的领域观点词抽取与极性判别方法，还存在一定的局限性，需研究更多更有效的方法。

参 考 文 献

[1] 董振东, 董强.知网. http://www.keenage.com.

[2] 刘群 ,李素建.基于《知网》的词汇语义相似度的计算[C] .第三届汉语词汇语义学研讨会 ,台北 ,2002.

[3] 柳位平，朱艳辉，栗春亮等. 中文基础情感词词典构建方法研究[J] .计算机应用, 2009, 29(11)

[4] ICTCLAS. http://ictclas.org/

[5] Peter . Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. 2002. 417-424.

[6] Yanhui ZHU, Ping WANG, Zhihui Wu, ZhiQiang WEN. Research on Extracting Semantic Orientation of Chinese Text Based on Multi-algorithm[J]. Communications in Computer and Information Science. 2011：461-469

[7] 赵军,许洪波,黄萱菁,谭松波,刘康. 中文倾向性分析评测技术报告[C]. 第一届中文倾向性分析评测论文集. 北京，2008:1-20.

[8] XU Ye-qiang, ZHU Yan-hui, Wang Wen-hua, GAO Li-chun. A dynamic adjustment algorithm research of sentiment word weight based on context[J]. 2011 3rd IEEE International Conference on Computer Research and Development(ICCRD 2011). 2011.3（2）:19-22

[9] 哈工大信息检索研究中心(HIT CIR)语言技术平台共享资源. http://ir.hit.edu.cn/demo /ltp/Sharing_Plan.htm.

评价对象、短语、搭配关系抽取及倾向性判断

朱圣代，徐向华，叶正

（杭州电子科技大学 计算机学院 云计算实验室 浙江 杭州 310018）

E-mail：zhushengdai@yahoo.com.cn, {xhxu, zye}@hdu.edu.cn

摘要：观点挖掘近年来已经成为自然语言处理领域的热点问题，本文对观点挖掘的几项关键技术---评价对象、评价短语、主观性关系抽取、倾向性判断进行了研究。在评价对象抽取阶段，通过统计得到所有的名词和名词短语作为候选，然后结合词频，词共现等特征进行过滤得到最终的评价对象；在评价短语抽取阶段，使用基于观点词词典的匹配方法，并把观点词前面的副词也作为评价短语的一部分；在搭配关系抽取阶段，目的是抽取评价对象和评价短语的关联关系，采取的方法是将在句中距离评级对象最近的评价短语作为该短语的评级短语；在情感倾向分析阶段，通过将情感句进行分类，然后制定规则进行无监督的倾向性判断。

关键词：观点挖掘；评价对象；评价短语；主观性关系；倾向性判断

Target, Phrase and Subjective Relationship Extraction And Sentiment Classification

ZHU Sheng-dai, XU Xianghua，YE Zheng

(Cloud Lab ， School of Computer Science and Technology，Hanzhou Dianzi University ， Hanzhou, Zhoujiang, 310018)

E-mail：zhushengdai@yahoo.com.cn, {xuxh, zye}@hdu.edu.cn

Abstract: In recent years, opinion mining has become a hot issue in natural language processing field. This paper carries out a research on several key technologies of opinion mining, such as the object of evaluation, evaluation phrases, subjective relationship extraction and orientation judgment. In the phase of evaluating objects, all the nouns and noun phrases collected through statistics are regarded as candidates and then we associate them with word frequency and co-occurrence to filtrate and finally gets the object of evaluation. In the phase of evaluation phrases extraction, the writher applies a matching methodology which is based on affection and takes the adverbs which precede the affective words as a part of evaluating phrases. The goal in the phase of subjective relationship extraction is to extract the incidence relation between evaluation object and evaluation phrases and the solution that we have adopted is to take the evaluation phrase which is close to the evaluation object as rating phrases. In the phase of orientation judgment, we classify the affective sentences and then develop rules for non-supervised orientation judgment .

Keywords: Opinion Mining; Target; Evaluation phrase; Subjective Relationship; Orientation Judgment

1 引言

近年来，观点挖掘（Opinion Mining）受到了很多学者的关注[7][8],它是一个非常新颖且有应用价值的课题，比如：问答系统，客户关系管理，产品信誉度分析等等。同时，观点挖掘也产生了许多具有挑战性的相关子方向。例如：领域观点词的抽取，旨在识别领域对观点词倾向性的影响；文本主客观分类，旨在识别文本单元的主客观性。

本文致力于研究主管句中的评价搭配抽取任务，考虑上下文对词语倾向性的影响，抽取被评价对象、评级短语，并判断倾向性。该任务可分为四个主要阶段：（1）自动识别观点句中的评价对象；（2）自动识别句中的评价短语；（3）识别抽取评价对象以及评价短语之间的主观性关系；（4）判断主观句中评价对象的情感倾向性。例如：对于某一评论“这款相机资源占用率低、看图快速且具备不错的人物照片筛选功能。”，系统首先识别评论中的被评价对象（如： “资源占用率”,“看图”,“人物照片筛选功能”）以及评价短语（如：“低”,“快速”,“不错的”),然后结合评价对象和评价短语之间的词共现和句中距离特征，抽取句子的主观性搭配关系，最后分析评价对象的情感倾向性，即“资源占用率，低，褒义”,“看图，快速，褒义”,“任务照片筛选功能，不错的，褒义”。

关于评价对象的抽取，Liu Bing[9]等借助词性标注与关联规则挖掘提取产品评论中的产品属性，然后针对产品属性进行观点摘要及倾向性分析,Wu Yuanbin[10]将短语依存分析引入到产品评论的特征识别及评论分析中.*Guo* Hongli[11]等利用多级潜在语义关联分析进行产品特征的分类和评论分析研究，Zhuang[12]等对用户生成电影评论进行观点挖掘[15]，采取基于多知识的方法进行评论特征选取，并记录特征词与观点词直接的依存语法图，最后使用不同的评论特征分别进行倾向性分析。Liu和Zhuang的方法都通过观点特征选取的方法来进行观点分析和倾向性计算。因此观点特征选取正确与否至关重要，直接影响到后续工作的结果。

关于评价对象和评价短语的主观性关系抽取，B Liu [3]的方法是首先找出产品特征词，然后将离产品特征词最近的形容词作为评价词。这种方法把评价词局限于形容词,但动词、名词以及副词也可以作为评价词。Soo-MinKim[4]的方法是首先找出评价词，然后借助FrameNet分析句子的语义结构来找出评价对象。然后对句子的各部分进行语义角色标记，最后将对应的角色映射为评价对象。这种方法受限于外部资源：FrameNet，如果无法在FrameNet找出相应的或者语义相近的评价词，那么就无法找出评价对象。章剑锋[6]等提出的基于最大熵机器学习的方法将评价词周围的词和词性作为特征，特别是将程度副词加入特征，以此来帮助判断评价词的主观性，但是需要大量的预料标注，缺乏适应性。

关于判断主观句中被评价对象的情感倾向性，前人的工作中所提出方法能取得较好效果，如刘鸿宇[15]等人首先分析主观句的结构，将其分为四类；继而针对各类制定相应的倾向性判断规则，最终基于无指导的方法完成评价对象在的倾向性判断，他们的准确率已达到了97%以上。

本文使用的无监督的方法进行评价对象、评价短语、主观性关系的抽取和倾向性分析。评价对象的的抽取上，使用基于词频的抽取名词和名词短语作为候选，同时加入PMI过滤技术。在评价短语抽取部分，评价词一般都是形容词，动词或者副词，他们的数目一般是不变的，并且是有限的，所以这里采用建立情感词典的方式，然后对于需要处理的文本来匹

配这些词，另外还将评价词前的副词加入评价短语。在主观性关系抽取上，从评价对象抽取与评价短语抽取模块，抽取的评价对象以及评价短语，它们都只是候选，本文找出距离评价对象最近的评价短语最为该评价对象的评价短语。在情感倾向分析上，将情感句分为四类，对每类分别用不同的规则来判定情感倾向。

2 基于统计的评价对象抽取

基于关联规则的评价对象抽取在Hu[1],Liu[9]中首次提到，中文评论的评价对象抽取如刘鸿宇[15]提出的基于句法路径的评价对象提取是对关联规则的改进。本文的使用基于词频的评价对象抽取技术，和上述技术基本类似，对于给定语料，首先对其分词、词性标注，然后提取其中的名词和名词短语，过滤词频低于阈值的名词或名词短语，词频过滤主要考虑到评价对象大都是在评论中多次出现的，一些不相关的名词或者名词短语很少在评价对象中出现，而且那些低词频的评价对象是用户不太关系的评价对象，可以被过滤掉。本文还过滤掉单个字的情况，因为经过观察，单个字几乎不可能成为评价对象。然后再进行PMI算法筛选得到最终的评价对象。

本文采用PMI(Pointwise Mutual Information)指标来量化词A和词B的关系，计算两词的PMI的公式如下：

$$PMI(A，B)=\log_2\frac{hits(A，B)}{hits(A)*hits（B)} \tag{1}$$

从公式中可以看出，PMI 值的计算使用了统计的思想，其基本假设是：两个对象共现的次数越多，则它们之间的联系也就越大。PMI 值计算的难点在于大规模文本集合的获取，理论上讲，文本数越多，则统计效果越明显，PMI 值的计算也应该越准确。理论上讲语料库越大，所得到的结果越准确。本文采用雅虎的搜索结果作为语料库，对于不同的领域选取不同的代表词，比如数码领域选取“手机”作为代表词，娱乐领域选取“娱乐”作为代表词，金融领域选取“金融”为代表词等，计算代表词语候选评价对象的PMI值，选取合适的阈值，过滤掉低于阈值作为最终的评价对象。具体步骤如下：

(1)分词，词性标注；
(2)对于“的”之后的词不管词性如何词性一律改为名词，对于“地”之前的词不管词性如何词性一律改为副词，然后再去掉句中“的”和“地”这两个字；
(3)对于单个名词，直接抽取；
(4)对于“NN+NN”,“NN+NN+NN”，“JJ+NN”等形式直接抽取，其中“NN”代表名词，“JJ”代表形容词；
(5)统计抽取出来的名词或名词短语出现的句子数目，过滤掉低于设定支持度的名词或名词短语；
(6)经观察，评价对象基本不可能是单字的情况，因此过滤掉单个字的情况；
(7)计算领域代表词和候选评价对象的PMI值，设定合适的阈值，过滤掉PMI低于阈值的候选评价对象得到最终的评价对象；

(8)对于句中出现评价对象的句子就认为它是主观句，进行后续任务的处理，对于句中没有出现评价对象的句子认为它是非主观句过滤掉。

3 基于评价词典匹配的评价短语抽取

本文的评价词典使用的是 WordNet 中文观点词典，使用的匹配方法是首次匹配方法，对于分词后的单词串，提取“JJ”,“JJ+JJ”,“JJ+JJ”,“JJ+JJ+JJ”等形式的单词或短语，查询它们是否在观点词典中出现，如果它出现，并且前面的词不是副词则把它作为评价短语；如果它出现并且前面的词是副词则把副词和观点词一起作为评价短语。

由于时间仓促，本文采取的基于分词的首次匹配的方法不是理想的方法，理论上采用序列最大匹配的原则来进行匹配效果会更好。

4 搭配关系抽取和倾向性判断

在评价对象和评价短语抽取后，需要对评价对象搭配合适的评价短语本文采用的规则具体如下：

(1)如果句子没有评价对象，认为它是非主观句，过滤掉这条句子；

(2)如果句子既有评价对象，又有评价短语，选取距离评价对象的最近的评价短语作为该评价对象的评价短语，得到(评价对象，评价短语)组合；

(3) 如果句子含有评价对象，但是没有评价短语，选取距离该评价对象 5 个单词内的，最近的，并且具特定词性组合的短语作为该评价对象的评价短语，短语的词性组合为“JJ”“JJ+JJ”、“JJ+JJ+JJ”,如果该短语的前面是副词，那么把这个副词也加入到这个评价短语中。

在倾向性判断模块，首先构建了情感词典，本文采用的情感词典主要来源于三部分：WordNet 的中文情感词典和中文观点词典；李军[2]所总结的中文情感词典；Hu[12]所采用的英文情感词典用谷歌翻译成中文。将三部分的情感词过滤掉重复的部分得到最终的情感词典。在情感句的判别方法上，本文采用了刘鸿宇[9]的分治的策略，根据情感句的结构将其分为四类；继而针对各类制定相应的倾向性判断规则，最终基于无指导的方法完成评价对象的倾向性判断。主观句的类型分为三类，具体定义以及相应的情感判断规则如下：

类别一：句子带有明细的倾向性，即在情感词典中找到的带有一种倾向性(褒义或贬义)的情感词明显多于宁一种带有另一种倾向性情感词的数目，那么句中所有的评价对象的情感倾向为情感词多的情感倾向。

类别二：句中含有的情感词褒义和贬义的数目相等，那么针对句中的每个评价对象选取最近的情感词的情感倾向的为它的情感倾向。

类别三：句中没有情感词但是句子有评价对象，那么句子的极性有限与当前句子的前一个句子的极性相同，如果前一个句子没有极性，那么与离当前句子最近的有极性的句子的极性相同，句中的所有评价对象的极性为句子的极性。

5 试验结果与分析

本系统参加了第三届中文倾向性分析评测，在评测中成绩一般。此次评测的语料主要涉及数码，娱乐，金融三个领域，表 1 给出了评价对象抽取的性能指标。

表 1：评价对象正确

Run-tag	Domain	P@1000	Precision	Recall	F1	Raccuracy
Hdu	D	0.172	0.143681	0.06596	0.090414	0.06596
Hdu	E	0.016	0.038741	0.010417	0.016419	0.010417
Hdu	F	0.008	0.011976	0.012289	0.01213	0.012289
宏平均		0.065333	0.064799	0.029555	0.040595	0.029555
微平均		0.065333	0.115865	0.054073	0.073735	0.054073
所有结果平均宏平均		0.065524	0.074285	0.045504	0.054307	0.045504
所有结果最大宏平均		0.111	0.133933	0.081763	0.091606	0.081763
所有结果平均微平均		0.065524	0.101236	0.081829	0.083421	0.069819
所有结果最大微平均		0.111	0.159847	0.149071	0.144701	0.135726

从上表结果可以看出，本文的评价对象抽取部分总体平均结果接近所有结果的平均值，但是与最好的结果还有一定的差距，但是本文在领域 D 的结果明显高于其他领域的结果，领域 D(数码)的结果接近于所有结果的最好值，领域 E(娱乐)的远远低于领域 D 的结果，领域 F(金融)的结果最差，说明本系统在领域 D 达到了较理想的性能，但是缺乏领域的适应性。究其原因大致有三：(1)领域 D 是数码产品领域，评价大都是针对产品本身的评价，评价对象大都为描述产品本身或者本身的一部分，词性特征明显，大多为名词或者名词组合，而且虽然数码产品种类众多，但是大都具有与本文选取的领域特征词“手机”具有类似的评价对象，因此本文的方法能取得较好的结果。(2)领域 E 是娱乐领域，用户关注的内容繁多，评价的内容五花八门，评价针对的对象也具有不确定性，很难选出具有领域代表性的词。(3)领域 F(金融领域)更是一个特殊的领域，评价对象很多不是名词或名词短语，领域代表词很难去确定。

表 2:评价短语正确

Run-tag	Domain	P@1000	Precision	Recall	F1	Raccuracy
Hdu	D	0.083	0.075824	0.034809	0.047714	0.034809
Hdu	E	0.012	0.029056	0.007813	0.012314	0.007813
Hdu	F	0.007	0.010479	0.010753	0.010614	0.010753
宏平均		0.034	0.038453	0.017791	0.024327	0.017791

微平均		0.034	0.062487	0.029162	0.039765	0.029162
所有结果平均宏平均		0.051429	0.055383	0.05059	0.046795	0.035003
所有结果最大宏平均		0.118333	0.085672	0.100699	0.087237	0.066468
所有结果平均微平均		0.051429	0.074015	0.061917	0.062064	0.052096
所有结果最大微平均		0.118333	0.117425	0.105575	0.10248	0.095492

表 3：评价对象、短语、极性都正确：

Run-tag	Domain	P@1000	Precision	Recall	F1	Raccuracy
Hdu	D	0.049	0.031868	0.01463	0.020054	0.01463
Hdu	E	0.003	0.007264	0.001953	0.003079	0.001953
Hdu	F	0.001	0	0	0	0
宏平均		0.017667	0.013044	0.005528	0.007765	0.005528
微平均		0.017667	0.025207	0.011764	0.016041	0.011764
所有结果平均宏平均		0.02719	0.025047	0.020403	0.019833	0.015183
所有结果最大宏平均		0.071667	0.039639	0.037671	0.03478	0.025728
所有结果平均微平均		0.025778	0.033019	0.027191	0.027323	0.02309
所有结果最大微平均		0.071667	0.0616	0.046856	0.048312	0.043298

表 2 和表 3 分别给出评价短语抽取的结果和评价对象、短语、极性都正确的结果，表 2 的结果与表 1 类似，从表 3 的最终结果可以看出领域 D 的结果仍然远远高于领域 E 和领域 F 的结果，可见领域 E 和领域 F 的复杂性高于领域 D，因此本文中所使用系统的领域适应性有待提高。

本系统的总体不够理想，原因有三：(1)第一次参加评测，经验不足，对于评测的流程也不够熟悉；(2)由于时间紧张，调研和实验的准备不够充分，实验的细节处理也有待提高；(3)系统应该考虑在不同领域的差异性做出一定的修改。

6 结论

本文实现了一个评价关系抽取系统，可分为评价对象抽取、评价短语抽取、主观性抽取和倾向性判断四个部分，系统在 COAE2011 的评测中取得成绩一般。由评测的总体结果可以看出，观点挖掘技术目前还处于初级阶段，因此还有很广阔的研究空间。如：如何使系统具有更高的移植性和适用性，如何挖掘出更多的主观句等等，都将成为我们下一步的工作。

参考文献

[1] Hu M．Liu B．Mining Opinion Features in Customer Reviews[C]．In AAAI，2004．755—760．

[2] 李军，中文评论的褒贬义分类研究［thesis］清华大学 2008

[3] Liu Bing, Hu Minqing, Cheng Junsheng. Opinion Observer: Analyzing and Comparing Opinions on the Web[C]//Proceedings ofthe 14th Internationa1 Conference on World Wide Web. [S. l.]:IEEE Press, 2006: 221-229.

[4] Kim Soo-Min, Hovy E. Extracting Opinions, Opinion Holders, andTopics Expressed in Online News Media Text[C]//Proc. of Conf. ofAssociation for Computational Lingustics. [S. l.]: IEEE Press, 2007:318-327.

[5] 章剑锋，张奇，吴立德，等中文观点挖掘中的主观性关系抽取[J]，中文信息学报. 2008, 22(2), 55-59.

[6] 刘鸿宇，赵妍妍，秦兵，刘挺，评价对象抽取及其倾向性分析[J]，中文信息学报. 2010.

[7] W．Xi，J．Lind，E．Brill，I earning Effective Ranking Functions for Newsgroup Search[A]．In Proceedings of the 27thAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]．Sheffield，U．K．25—29 July 2004．394—4O1

[8] C．Macdonald and I．Ounis．Combining Fields in KnownItem Email Search：A．In Proceedings of SIGIR 2006[C]．Seattle，USA：2004．675—676．

[9] HU Minqing，LIU Bing．Mining and summarizing customer reviews[C]//Pr0ceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining．New York：ACM Press，2004：168-177．

[10] WU Yuanbin，ZHANG Qi，HUANG Xuanjing．Phrase dependency parsing for opinion mining[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNI P2009)：Volume 3．New York：ACM ，2009：1533—1541

[11］GUO Hong-lei，ZHU Huijia，GUO Zhili．Product feature categorization with multilevel latent semantic association[C]//Proceeding of the 18th ACM Conference on Information and Knowledge Management(CIKM2009)．New York：ACM ，2009：1087—1096．

[12] ZHUANG L，JING F，ZHU X Y．Movie review mining and summarization[C]//Proceeding of the 18th ACM Conference on Information and Knowledge Management(CIKM 2006)．New York：ACM ，2006：43—50．

基于 CRF 模型的半监督学习迭代观点句识别研究*

丁晟春，文能，蒋婷，孟美任
南京理工大学经济管理学院信息管理系，南京 210094
E-mail: todingding@163.com

摘　要：近年来中文情感倾向性分析受到越来越多研究者的关注。研究者从不同粒度层次对中文文本进行情感极性的分析研究，本文从句子级的角度进行了中文文本的情感倾向分析，提出以 HowNet 中的情感词表为种子情感词集，采用基于 CRF 模型的半监督学习迭代方法获取大量评价词，进一步自动识别测试集中的观点句，并依据中文词间的语义规则判断观点句表达的总体极性，实现对 COAE2011 中任务 2 –观点句识别。

关键词：CRF；观点句；半监督；情感倾向性

Semi-supervised Bootstrapping Based on CRF of Sentiment Sentence Recognition

SHENGCHUN DING, NENG WEN, TING JIANG, MEIREN MENG
Department of Information and Management of Nanjing University of Science & Technology, Nanjing 210094
E-mail: todingding@163.com

Abstract: During recent years, sentiment analysis about text in Chinese is becoming more and more popular in academic research. In this paper, sentiment analysis is processed on sentence level. Sentiment words published by HowNet is used as the original evaluated-word set, a large amount of evaluated-words are obtained by semi-supervised bootstrapping based on CRF model. Then sentiment sentence can be recognized by evaluated-words, and the polarity of sentiment sentence can be judged by the designed semantic rules.

Keywords: CRF; Sentiment sentence; Semi-supervised; Sentiment analysis

1　引言

随着网络上各类贴吧、论坛、博客、评论网站等的兴起，网络成为了一个供大众评论交流的大平台，出现了大量的主观性文本。这类文本中包含了大量有价值的观点信息，如社会舆论导向、网民焦点话题、经济发展趋势、用户体验及消费经验等等。由此，中文文本的情感倾向性分析成为观点挖掘领域学术研究热点，涵盖了自然语言处理、数据挖掘、信息抽取、机器学习等多个领域，获得多领域学者的关注。另外，由于这些大量有价值的观点信息具有非常重要的现实意义，由此也成为众多互联网企业关注的焦点。

本文主要探讨参加了 COAE2011 中任务 2-观点句的抽取，要求从每个领域的测试集中自动识别出所有观点句并判断其表达的总体极性（褒义，贬义，或混合观点）。任务 2 中的观点句是指包含了对其他对象的评价的句子，不包括内心自我情感（比如“心情”），因此，

*作者简介丁晟春，女，1971 年生，南京理工大学信息管理系，副教授，主要研究方向：WEB 数据挖掘、信息检索、信息系统开发。作者简介：文能，女，1987 年生，南京理工大学信息管理系，情报学硕士在读。基金项目：教育部人文社会科学研究规划基金项目“基于语义的电子商务产品主/客观信息提取研究”（09YJA870015），项目负责人丁晟春；江苏省研究生科研创新计划“基于领域本体_CRF 的商品主观评价”（CX10S-001R），项目负责人文能。

本文以是否含有这类用来对其他对象进行评价的词为判断句子是否为观点句的依据，这类词在本文中被定义为评价词。本文以 hownet[1]中的情感词表为种子评价词集，并将该评价词集作为模型的情感特征，采用基于 CRF 模型半监督学习迭代方法获取大量评价词。最后利用评价词集，采用模式匹配的方法自动识别测试集中的观点句，并依据中文词间的语义规则判断观点句表达的总体极性。

2 基于 CRF 模型半监督学习方法构建评价词集

条件随机场(Conditional Random Fields,CRFs)[2]最早由 Lafferty 等人于 2001 年提出的，结合了最大熵模型和隐马尔可夫模型的特点，可以看成是一个无向图模型或马尔可夫随机场，它是一种用来标记和切分序列化数据的统计模型。

CRF 模型不同于产生式模型，它可以使用丰富的、彼此重叠的观察序列的特征，而且不需要很严格的前提假设；同时，不同于最大熵马尔可夫模型等概率模型，不对单个标记归一化，而是在整个观测序列求解一个最优的标记序列，避免了标记偏置问题。

本文采用灵玖软件[3]对测试集进行批量处理，获取测试集的分词及词性结果。本节将介绍以 Hownet 中情感词表为种子评价词集，采用基于 CRF 模型的半监督迭代学习方法构建评价词集。由于某些评价词具有领域特征，例如金融领域的评价词具有很强领域特征，对股指等的描述，因此本文针对三个领域测试集（数码、金融、娱乐）分别构建评价词集。

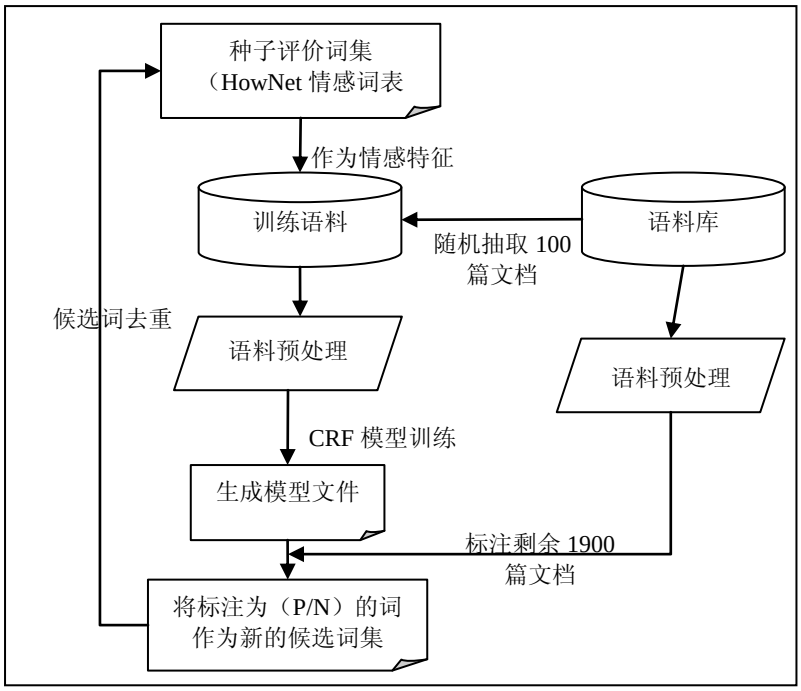

图 1 基于 CRF 模型的半监督迭代的评价词集构建流程

每个领域经分词后测试集的基于 CRF 模型的半监督迭代的基本流程如下：

（1）每次随机选择 100 篇语料作为训练语料，剩余的 1900 篇语料作为标注语料。

（2）对语料预处理，通过对语料的文本特征分析，发现单句中存在很多与判断是否为

观点句不相关的词（停用词，量词等），并且这些词的存在会影响模型对上下文内容语义特征的判断的准确性。因此本文借助相关词表对经分词后的语料中这类词予以剔除，作为模型的训练语料和模型标注语料。

（3）将 hownet 中情感词表为种子评价词集，并将评价词集作为情感特征，结合词、词性构造 CRF 模型的特征集，评价词集的特征描述如表 1，特征模板如表 2。

（4）使用外部工具包 CRF++-0.53[2]训练 CRF 模型，将生成新的模型文件对标注语料进行情感特征分类，分为三类：褒义（P）、贬义（N）、客观（O）。

（5）将标注为（P/N）词作为候选评价词，在评价集中查找，如果该词不存在与评价集中，则作为新的种子评价词添加到评价词集中，并记录其情感倾向。

（6）重复上述过程，直至不再产生新的评价词，最终获得评价集。

表 1 特征集描述

特征类别	特征标记	特征描述
词特征（W）	W	词
词性特征（P）	Pos	词性
情感特征（S）	RP	当前词为 Hownet 中正面评价词语
	RN	当前词为 Hownet 中负面评价词语
	SP	当前词为 Hownet 中正面情感词语
	SN	当前词为 Hownet 中负面情感词语
	P	当前词为新褒义评价词
	N	当前词为新贬义评价词
	O	当前词为中性词

表 2 特征模板

特征类型	特征表示	描述
原子特征	P_nP_{n+1}, W_nW_{n+1}, S_nS_{n+1} n=-2,-1,0,1	单个特征
复合特征	$W_nP_nS_n$, W_nS_n, P_nS_n n=-2,-1,0,1	多个特征组合

a. n 表示在文本中观察词相对当前词的位置，n 的值观察值与当前词之间间隔词的个数。当 n>0，观察词的位置在当前词的前面；当 n<0，表示观察词的位置在当前词的后面。

3 基于评价词集的观点句的识别

3.1 判断观点句极性的规则

评测组已经把句子作了划分，以“\r\n”为句子的结束符，所以本文的观点句的识别是判断以“\r\n”结束的句子是否包含了对其他对象的评价，为方便说明后面都用“单句”表示。本文提出将单句中是否存在评价词作为判断单句是否为观点句的依据。对语料的文本特征进行分析，发现单句存在下述几个语义特征：

1.单个评价词的处理

单句中只有一个评价词，则该句子被判定为观点句，单句的情感倾向与评价词的一致。

2.多个评价词处理

单句中存在多个评价词，则该句子被判定为观点句。如果这些评价词都表示同类情感倾向（褒\贬），则这类单句情感倾向与评价词的（褒\贬）一致；但如果存在多个不同情感倾向，这类单句的情感倾向为“或褒或贬”。

3.否定词处理

有些单句中用否定词（不，没有等）与评价词一起表达对其他对象的评价。本文对于判定含有否定词和评价词的单句的组合情感倾向定义了如下规则：

如果单句含有否定词，则将其与紧邻其后的评价词合并，组合评价词情感倾向为评价词的情感的反面。

如果单句含有多个否定词且否定词间没有评价词，将多个否定词组合，根据“否定之否定为肯定”判断否定词组合的情感倾向。然后将否定词组合与紧邻其后的评价词合并，如果否定词组合为褒义，组合评价词情感倾向为评价词倾向；如果否定词组合为贬义，组合评价词情感倾向为评价词的反面。

只含有一个组合评价词的单句情感倾向与组合评价词的情感倾向一致，而含有多个组合评价词的单句情感倾向判断与含有多个评价词的单句（见本节 2.多个评价词处理）处理方法相同。

4.连词处理

连词对观点句的情感倾向有重要的作用，连词可以被分为三类：并列、递进、转折。对含有这三类连词的观点句情感倾向判断规则如下：

如果连词两端是并列（如“而且”，“既……又……”等），这类连词两端的子句情感倾向是相同，所以含有这类连词的观点句的情感倾向为观点句中所包含评价词的情感倾向。

如果连词是递进（如“不但……而且……”，“尚且……何况……”），这类连词用来加强观点句情感倾向的程度，观点句的情感倾向为其所包含评价词的情感。

如果是转折（如“但是”，“即使”），这类连词一般被用来表示子句情感的转折。含有这类连词的观点句中有多个评价词，如果评价词情感倾向不一致，有褒有贬，则这个观点句情感倾向为“或褒或贬”；如果评价词的情感倾向一致，则这个观点句情感倾向与评价词一致。

3.2 观点句提取的规则

由于单个观点句中可能含有多个评价词，本文以观点句中最后一个评价词为中心，取该评价词前后的范围的共 20 个词作为提交的观点句结果。具体规则如下：

- **规则 a**：如果观点句的长度小于等于 20 时，取整个句子作为提交的观点句结果。
- **规则 b**：如果观点句的长度大于 20 时，就以观点句中最后一个评价词为中心，确定观点句的取词范围。

规则 b.1：如果评价词在句中的位置大于 20，就取关键词在内向前的 20 个词作为提交的观点句结果；

规则 b.2：如果评价词在句中与句首位置为小于 10，就从句首向后包括评价词在内的 20 个词作为提交的观点句结果；

规则 b.3：如果评价词在句中与句尾位置小于 10，就从句尾向前包括评价词在内的 20 个词作为提交的观点句结果；

规则 b.4：如果评价词与句首句尾位置均大于 10，就以评价词为中心，向前向后各取 10 个词作为提交的观点句结果。

3.3 置信度的计算

置信度（confidence）:在关联规则中常用来衡量关联规则强度的指标。一个规则 X->Y 的置信度是指“既包含了 X 又包含了 Y 的事务的数量占所有包含了 X 的事务的百分比”。可以看做是条件概率 pr(Y|X)的一个估计。任务 2 中要求将观点句按照置信度排序输出，本文对观点句识别的置信度计算以观点句中出现的评价词为依据，并假设观点句中评价词的出现是独立的，与其他评价词不相关。因此，第 j 个观点句分类结果置信度 Pj（j=1,2,……,n）计算方法为:

（1）记 Xi 表示观点句中出现的评价词（i=1,2,……,n），观点句中可能出现 n 个评价词。

（2）记 Yi 表示句子的分类结果（Y1=褒义，Y2=贬义，Y3=或褒或贬）

Pj=P(Yi|X1,X2,…….,Xn)= 1/n(P(Yi|X1)+ P(Yi|X2) +……+P(Yi|Xn))，其中 Yi 取值为观点句的情感倾向。

P(Yi|Xi)=count(Yi,Xi)/count(Xi)

其中 P(Yi|Xi)表示在评价词集中评价词 Xi 表示情感倾向 Yi 的概率；count(Yi,Xi)表示在评价词集中 Xi 表示情感倾向 Yi 的个数，count(Xi)表示在评价词集中 Xi 表示所有情感倾向的个数。

4 评测与分析

COAE2011 任务 2 观点句识别的评测语料分为三个领域：digital，entertainment 和 finance。本文采用了基于 CRF 模型的半监督迭代方法分别三个领域的测试集进行了观点句识别。由于有些领域的评价词具有较强的领域性，如 finance 领域的评价词领域局限性很强。因此该方法在三个领域获得的结果有些差别，如表 3 所示。

表 3 观点句识别结果

	Precision	Recall	F1	P@1000	Raccuracy
8-NJUST_D	0.135667	0.120155	0.127441	0.146	0.120155
8-NJUST_E	0.068022	0.149715	0.093543	0.071	0.070789
8-NJUST_F	0.040834	0.090909	0.056355	0.047	0.040619
Median	0.24081488	0.397946	0.276324	0.290183	0.255444517
Best	0.729751	0.798097	0.693304	0.8	0.660324

表 4 修正后 digital 测试集观点句识别结果

Data	Precision	Recall	F1
Digital	0.5839	0.55859	0.576245
Median	0.24081488	0.397946	0.276324
Best	0.729751	0.798097	0.693304

表 3 中各领域观点句识别结果非常低，是由本文对观点句提取的方法与 COAE2011 评测的差异导致的。在 COAE2011 评测中任务 2 提交结果提取的是整个观点句，但本方法是按照前面说的观点句提取规则提取作为提交结果，对于句子长度大于 20 个词的句子提取的

是观点句中包含评价词在内的 20 个词的子句作为提交结果，但评测任务对子句是不能进行匹配的，所以导致本系统的评测结果出现了一些问题。因此，本文对观点句提取方法按照 COAE2011 任务 2 的标准的修正后，针对 digital 领域得到了表 4 中的结果。在 digital 领域，COAE2011 中正确观点句为 5675 个，本系统共识别出观点句 5429 个，其中与 COAE2011 给出的正确答案一致的（注：句子识别和极性的判断都一致）的结果有 3170 个观点句。表 4 中的结果可以看出在 digital 领域对观点句的识别明显高出平均值，说明了本文采用基于 CRF 模型的半监督迭代获取评价词集，并在此基础上进行观点句极性判断的规则设计是有效的。

5 结论

本文介绍了参与 COAE2011 任务 2 观点句识别所采用的方法，主要说明了所采用的流程及用于观点句极性判别的规则的设计。由于观点句提取方法的差异影响评测的结果，按照 COAE2011 任务 2 中观点句提取的原则修正后，与正确结果对比后，说明了本方法在观点句识别上的有效性。

参 考 文 献

[1] http://www.keenage.com/html/c_index.html.

[2] J Lafferty, A McCallum, F Pereira, “Conditional random fields: Probalistic models for segmenting and labeling sequence data,” Proc. ICML, 2011, PP. 282-289.

[3] http://www.lingjoin.com.

基于词典的中文倾向性分析报告

张成功[1,2]，刘培玉[1,2]，朱振方[1,2]，杨玉珍[1,2]，方明[1,2]

（1．山东师范大学信信息科学与工程学院，山东 济南 250014；2．山东省分布式计算机软件新技术重点实验室，山东 济南 250014）

邮箱：zcg870108@163.com

摘 要：文本倾向性分析成为近几年自然语言处理领域研究的热点，领域知识和上下文语境对倾向性判别至关重要。本文开发的基于词典的中文倾向性分析系统，根据建立的观点词词典完成了评价搭配抽取及观点句抽取，并且以修饰词语和评价词语组合而成的评价短语作为倾向性计算的基本单元，评价对象及观点句的倾向性均可由评价短语的倾向性得出。评测结果显示，本文开发的系统参加任务二和任务三取得了不错的效果。

关键词：观点词；评价短语；评价对象；评价搭配；倾向性分析

Chinese Orientation Analysis Based on Lexicon

ZHANG Cheng-gong [1,2],LIU Pei-yu[1,2],ZHU Zhen-fang[1,2], YANG Yu-zhen[1,2],FANG Ming[1,2]

1.School of Information Science and Engineering, Shandong Normal University, Jinan ,Shandong,250014；

2．Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan，Shandong ,250014

E-mail:zcg870108@163.com

Abstract：Text Orientation Analysis has become a hot issue in nature language processing in the past few years, domain knowledge and context are important to orientation analysis. This paper developed a Chinese Orientation Analysis System, extracted evaluating collocation and sentiment sentence, combined modifier words with evaluating words to form evaluating phrases, the phrases were as the basic unit for orientation analysis, the orientation of target and sentiment sentence are denoted by the orientation of evaluating phrases. Evaluating results show that the effect is good.

Keywords: Sentiment Word; Sentiment Sentence ; Evaluating Phrases ; Evaluating Collocation ; Orientation Analysis

1 引言

随着 Web 2.0 时代的到来，Internet 不仅成为越来越多人获取信息的重要来源，也成为人们表达自己观点和情感的重要平台，随之而来的是网络上出现了大量的主观性文本，比如网站的评论信息以及各大论坛、博客、微博上人们表达观点的帖子等。如何在浩如烟海的信息中提取有价值的信息成为研究的热点，文本倾向性分析应运而生。

文本倾向性分析就是挖掘评论文本中评价者对评价对象的观点看法，进而判断评价者的观点看法是属于积极的或消极的[1]。文本倾向性分析不仅可以给个人购买产品提供一定的参考，还可以让企业了解产品在用户中的口碑以及时调节营销策略，还可以帮助政府相关部门正确把握民众的舆论导向，及时调整政策，抵制恶意的言论，鼓励有建设性的意见，维护社会的稳定。

文本倾向性分析已经成为自然语言处理的一个研究热点，越来越多的国际会议针对倾向性分析开展评测任务，国内针对中文的倾向性分析虽然起步较晚，但也取得了一系列成就，中文信息学会举办的中文文本倾向性评测(COAE)为国内学者、科研机构提供了一个很好的学习和交流的平台，该评测任务已经成功举办两届，今年是第三届。本次评测共设置了四个子任务，如表 1 所示。

表 1 表 1 COAE 2011 评测任务设置

Tab.1 The Subtasks in COAE 2011

任务号	类型	任务名称	任务说明
任务 1	要素级	领域观点词的抽取与极性判别	考虑领域对倾向性的影响，识别给定的三个领域的观点词，并判断极性。
任务 2	句子级	中文观点句抽取	从三个领域数据集随即抽取一定比例的数据经过自动分句构成测试集，从中抽取所有观点句并判别观点极性。
任务 3	要素级	评价搭配抽取	考虑上下文语境对词语倾向性的影响，识别评价短语，抽取评价对象并判别其极性。
任务 4	篇章级	观点检索	面向特定对象的中文文本观点检索。

针对此次评测任务，我们开发了基于词典的中文倾向性分析系统，完成了任务二和任务三。

2 基于词典的评价搭配及观点句抽取

任务二是中文观点句的抽取，任务三是评价短语和评价对象的抽取，在基于词典的中文倾向性分析中两个任务是密不可分的，都是基于已有的观点词词典进行的情感信息抽取，故将两个任务放到一起介绍。

针对任务二、三开发了基于词典的中文倾向性分析系统，利用预先建立好的观点词词典来识别观点句、评价对象以及评价短语等情感信息，具体的流程如下：

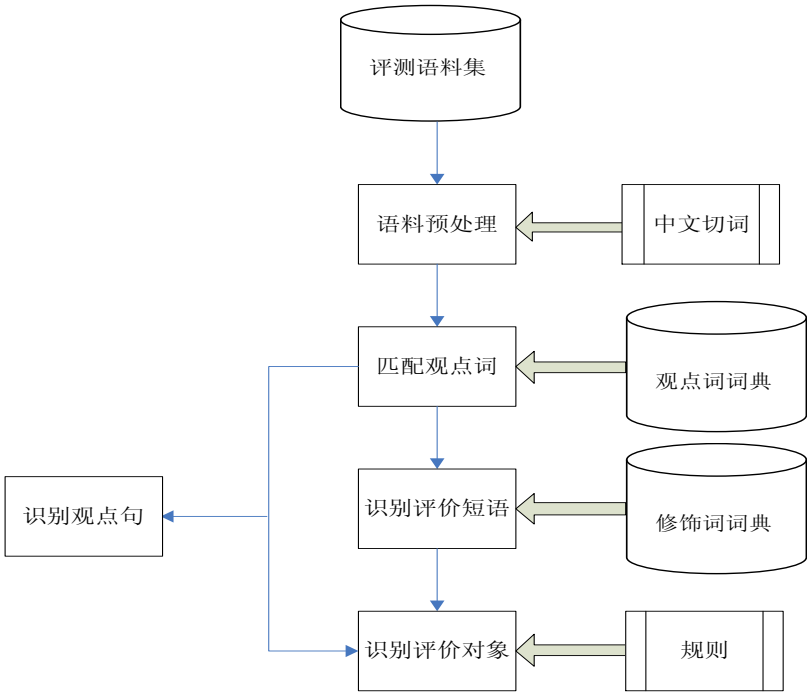

图 1 基于词典的评价搭配及观点句抽取

Fig.1 Extraction of Evaluating Collocation and Sentiment Sentence Based on Lexicon

2.1 语料预处理

因为给定的数据集是已经经过分句处理的，故我们只需对语料进行切词处理。系统使用灵玖 LJParser 软件进行切词，LJParser 采用条件随机场（Conditional Random Field,简称 CRF）模型，分词速度快，准确率接近 99%。

2.2 评价搭配抽取

文献[2]将情感分析归纳为三项递进的研究任务，其中情感信息的抽取作为最底层的任务，旨在抽取情感评论文本中有意义的信息单元，将无结构化的情感文本转化为计算机容易识别和处理的结构化文本，继而公情感分析上层的研究和应用服务。其中有价值的情感信息单元主要有评价词语、评价对象以及组合评价单元中的评价短语和评价搭配等。

2.2.1 评价短语的识别

评价短语是指修饰成分和评价词语组合而成的评价单元，比如“不很喜欢”。修饰成分是指加强、减弱或置反观点的语言成分，主要是否定副词和程度副词。为了能够准确的识别评价短语，我们建立了一个修饰词词表，选取不、不必、不要、别等 31 个否定副词，采用蔺璜[14]对程度副词的分类，将程度副词分为相对程度副词和绝对程度副词，如表 2 所示。

表 2 表 2 程度副词分类表

Tab.2 Classification Table of Degree Adverbs

程度副词	相对程度	极量	最 最为
		高量	更 更加 更为 更其 越 越发 备加 愈加 愈 愈发 愈为 愈益 越加 格外 益发 还
		中量	较 比较 较比 较为 还
		低量	稍 稍稍 稍微 稍为 稍许 略 略略 略微 略为 些微 多少
	绝对程度	极量	极 极为 极其 极度 极端 至 至为 顶 过 过于 过分 分外 万分
		高量	很 太 挺 怪 老 非常 特别 相当 十分 好 好不 甚 甚为 颇 颇为 异常 深为 满 蛮 够 多 多么 殊 特 大 大为 何等 何其 尤其 无比 尤为 不胜
		中量	不大 不太 不很 不甚
		低量	有点 有点儿 有些

本系统将将观点词与其修饰词构成评价短语作为极性计算的基本单元，给出了极性强度的计算公式，如表 3 所示。

表 3 表 3 评价短语的极性计算

Tab.3 Computation of Polarity Phrases

评价短语	强度计算公示	例句	强度
S=PW	E(PW)	1.1.1 她长的漂亮	0.8
S=NA+PW	E(PW)*E(NA)	她长的不漂亮	-0.64
S=NA+NA+PW	E(PW)*E(NA) *E(NA)	她长的不是不漂亮	0.512
S=DA+PW	E(PW)+(1-E(PW))*L(DA) 若 PW 是正面的 E(PW)+(-1-E(PW))*L(DA) 若 PW 是负面的	她长的很漂亮 这间房间很旧	0.94 -0.94
S=DA+DA+PW	E(PW)+(1-E(PW))*L(DA_1)+ [1-E(PW)-(1-E(PW)*L(DA1)]*L(DA_2)	她长的十分很漂亮	0.982
S=NA+DA+PW	E(PW)+(1-E(PW))*(L(DA)-0.2)	她长的不很漂亮	0.9
S=DA+NA+PW	E(PW)*E(NA)+(-1-E(PW)*E(NA)*L(DA)	她长的很不漂亮	-0.864

其中 NA 表示否定副词， DA 表示程度副词，PW 表示极性词，S 表示由修饰词和极性词组成的评价短语，E(PW)和 E(NA)分别代表评价词语和否定词的极性强度。对修饰词的强度设定为：E(NA)为-0.8 而不是-1，因为否定副词修饰中心词时并不是对极性词极性的简单置反，例如“不漂亮”和“丑”并不是等价关系，L(DA)分别定义为 0.9，0.7，0.5 和-0.5。

用评价短语作为基本的计算单元，评价对象以及整个观点句的极性都可以用评价短语的极性来进行表示。

2.2.2 评价对象的抽取及倾向性分析

评价对象一般为名词、动名词、名词性短语或中英文字符串，如屏幕、设计、屏幕的分辨率、5230 等。目前抽取评价对象主要有两种方法：一种是基于机器学习模型进行训练获得模型；另一种是根据句法结构获得候选特征集合，然后利用规则进行进一步筛选。

2.2.2.1 评价对象的抽取

一般情况下，评价对象是名词、动名词、英文字符串或名词性短语，名词性短语一般是以名词为中心，在名词的两边会是名词、动词、动名词或者“的”等，我们总结了名词性短语的模板有 n+n、n+n+n、n+n+vn、n+v+n、n+vn、n+vn+n、n+x、n+的+n、n+的+vn、nl、nrf、nz、nz+的+n、v+n、vn、vn+n 等。

系统中设计了评价对象抽取算法，首先识别出观点句中的观点词，设置一个动态的滑动窗口，在窗口范围内分别往前、后找与观点词距离最近的名词，然后以名词为中心扩展得到名词性短语，凡是与名词相连的名词、动词、动名词、动词、的、英文字符串都与该名词合并；若窗口范围内没有找到名词，则找动名词或英文字符串，也以相同的方式进行扩展。

2.2.2.2 倾向性分析

系统中用评价短语的倾向性代表其修饰的评价对象的倾向性，评价短语的极性强度计算用上表 3 中的公式计算。例如“国外 MP3 音质非常出色，但功能比较单一”这一句，评价对象和评价短语组合为“国外 MP3 音质-非常出色”“功能-比较单一”，两个评价短语的强度分别是 0.94 和-0.9，故评价对象的极性分别为 1 和-1。

2.2.3 观点句的抽取及极性判定

观点句是指在表达过程中带有情感和观点的句子，是作者对于某个评价对象或事件进行评价的句子。近年来观点句以及评价对象的抽取受到了很多学者和研究机构的重视，国内外也相继开展了许多相关研究领域的评测，例如 TREC Blog Track、NTCIR、COAE 等评测。

观点句抽取作为一个比较新的研究领域，还处于探索阶段，目前的分类算法仍比较单一，大致可分为三类：(1)基于词典的方法。利用预先建立好的词典，统计文本中出现的词语是否具有情感信息，进而判断其主客观性；(2)基于统计的方法。利用训练好的数据，采用某种机器学习方法(如 SVM,最大熵)，判断新数据应该划分为观点还是非观点；(3)基于图的方法。利用求最小分割的方法把文本在句子级别上切分为观点和非观点两个部分。

2.2.3.1 观点句的抽取

本次评测给定的数据集包括三个领域，分别是电子产品、财经、影视娱乐，而且多是评论性句子，评论句中的观点句一般都是阐述评论者对某个评价对象或某个事件的观点的句子。因此，我们把句子中有无观点词作为该句是否是观点句最重要的判断依据，而把有无评价对象作为一个补充，即一个句子中如果既有观点词也有评价对象，则我们认为该句必然是观点句，若只有观点词没有找到评价评价对象，我们认为该句是疑似观点句。

2.2.3.2 观点句极性判定

上面已经提到以评价短语为基本的计算单元来计算评价对象的极性，观点句的极性计算也是基于评价短语来计算的，如公式(1)所示。

$$E(Sentence)=\frac{1}{m+n}(\sum_{i=1}^{n}E(S_i)+\frac{1}{2}\sum_{j=1}^{m}E(S_j)) \quad (1)$$

其中$E(Sentence)$代表一个句子的极性强度，$E(S_i)$表示该句子中的与评价对象同时出项的评价短语的极性强度，$E(S_j)$表示没有与评价对象同时出现的评价短语的极性强度。我们认为没有与评价对象一起出现的评价短语对观点句极性判别的贡献度较小，因此将其强度减半处理。

2.2.3.3 置信度计算

置信度是指某一句话被判定为观点句的确定程度，在本系统中我们仍然用评价短语作为基本计算单元，只是我们不区分评价短语极性的正负，统一用绝对值代替，因为不管出现正面观点词还是负面观点词都会被判定为观点句。故置信度的计算用公式(2)表示。

$$C(Sentence)=\frac{1}{m+n}(|\sum_{i=1}^{n}E(S_i)|+\frac{1}{2}\sum_{j=1}^{m}|E(S_j)|) \quad (2)$$

3 评测结果分析

3.1 任务二结果分析

表 4 表 4 任务二结果分析

表 5 Tab.4 Analysis of Task 2

	宏平均				
	Precision	Recall	F1	P@1000	Raccuracy
SDNU	0.198088	0.551367	0.279791	0.294333	0.273408
Median	0.240815	0.406039	0.276324	0.290183	0.255445
Best	0.534693	0.723411	0.541377	0.532	0.494511
	微平均				
	Precision	Recall	F1	P@1000	Raccuracy
SDNU	0.262407	0.624764	0.369585	0.294333	0.369981
Median	0.315252	0.450532	0.357871	0.290183	0.36317
Best	0.654448	0.775397	0.639614	0.532	0.611425

从返回的评测结果来看，我们提交的任务二的结果基本在平均水平以上，而且利用文中提出的倾向性判别方法判别出的倾向性的准确率很高，证明了文中方法用于倾向性判别的有效性。

从结果来看，我们提交的结果召回率不错，准确率偏低，主要原因是我们采用的是基于词典的观点句识别，认为观点词和评价对象同时出现时候，该句为观点句，而事实上有很多句子出现了观点词和评价对象，但却是客观句。

从提供的标准答案来看，标准答案中标注的观点句采用的标准也不统一，例如：

(1) 配置方面，戴尔 10 采用 530 处理器，基于 945 主板芯片组，集成 500 显示核心，配置 1 内存，160 硬盘，除了处理器其他方面与普通上网本无异。　3

(2) 接口方面，戴尔 10 配有 3 个 20 接口，、视频接口，1 组音频输入输出接口，45 网络接口以及 5 合 1 读卡器的配置方便用户日常的使用。　1

因此，制定一个明确的标准，给观点句一个明确的定义或者规定拥有怎样特征的句子为观点句将会对推动观点句的抽取有很大的帮助。

3.2 任务三结果分析

表 6 表 5 任务三结果分析

表 7 Tab.2 Analysis of Task 3

评价对象正确：					
	宏平均				
	Precision@1000	Precision	Recall	F1	Raccuracy
SDNU	0.023667	0.017598	0.018828	0.018192	0.018828
average	0.065524	0.074285	0.045504	0.054307	0.045504
max	0.111	0.133933	0.081763	0.091606	0.081763
评价对象正确：					
	微平均				
	Precision@1000	Precision	Recall	F1	Raccuracy

SDNU	0.023667	0.02188	0.032622	0.026193	0.026097
average	0.065524	0.101236	0.081829	0.083421	0.069819
max	0.111	0.159847	0.149071	0.144701	0.135726
评价短语正确：					
	宏平均				
	Precision@1000	Precision	Recall	F1	Raccuracy
SDNU	0.007333	0.010677	0.023185	0.014621	0.009207
average	0.051429	0.055383	0.05059	0.046795	0.035003
max	0.118333	0.085672	0.100699	0.087237	0.066468
评价短语正确：					
	微平均				
	Precision@1000	Precision	Recall	F1	Raccuracy
SDNU	0.007333	0.012531	0.018683	0.015001	0.013938
average	0.051429	0.074015	0.061917	0.062064	0.052096
max	0.118333	0.117425	0.105575	0.10248	0.095492
评价对象、短语、极性都正确：					
	宏平均				
	Precision@1000	Precision	Recall	F1	Raccuracy
	0.001667	0.002701	0.006293	0.00378	0.002726
average	0.02719	0.025047	0.020403	0.019833	0.015183
max	0.071667	0.039639	0.037671	0.03478	0.025728
评价对象、短语、极性都正确：					
	微平均				
	Precision@1000	Precision	Recall	F1	Raccuracy
	0.001667	0.002718	0.004053	0.003254	0.002966
average	0.025778	0.033019	0.027191	0.027323	0.02309
max	0.071667	0.0616	0.046856	0.048312	0.043298

在评价短语抽取方面，主要存在以下问题：

(1) 系统中考虑的修饰成分主要是程度副词和否定副词，并没有考虑其他的修饰成分。

(2)我们抽取出的评价短语是将修饰评价词语的修饰词(包括程度副词和否定副词)抽取出来，将修饰词和评价词语重新组合成评价短语的形式，如“这款相机外观还是非常时尚的”，答案集中抽取的评价短语是“还是非常时尚的”，而我们认为的评价短语是“还是非常时尚”。

(3) 答案集中有许多评价词后面带了一个“的”字，而我们的结果中没有。例如“作为经常忙于商务的人士对上网本的认可度相对较高，不过大部分都还没有购机的计划”这一句我们抽取出的评价短语为“较高”，而答案集中给出的是“较高的”。

在评价对象抽取方面，主要存在的问题有：

(1) 由于任务三是在任务二抽取的观点句的基础上进行的评价对象的抽取，因此任务二的准确率会较大的影响任务三的准确率。

(2) 系统在 D 领域抽取的效果较好，而在 E、F 两个领域效果较差，原因是系统抽取评价对象时是主要基于产品属性的，没有考虑像命名实体识别、带引号或书名号的评价对象。

(3) 对于切词系统中不能正确切分的词语，如“性价比”等无法处理。

与观点句抽取一样，如何准确的界定评价短语和评价对象是一项重要而艰巨的任务。

4 总结

本文开发了一种基于词典的中文倾向性分析系统，完成了评价搭配、观点句的抽取以及倾向性判别的相关工作，取得了一定的成绩，下一步将完善领域词典的构建，更加深入的研究评价对象、评价短语以及观点句的特征，提高情感信息抽取的准确率。

参考文献

[1] 张博. 基于 SVM 的中文观点句抽取[D]. 北京邮电大学，2011 年 3 月.

[2] 赵妍妍,秦兵,刘挺 . 文本情感分析综述[J].软件学报,2010,21(8):1834-1848.

基于特征扩展的领域情感分析系统

宋施恩，樊兴华，赵静，魏平杰
重庆邮电大学，重庆，400065
E-mail: song_shi_en@126.com

摘　要：情感倾向性分析具有广泛的应用价值，已经成为当前自然语言处理的研究热点之一。本文主要介绍重庆邮电大学中文信息处理研究所在 COAE2011 中完成任务 1，2，3，4 的情况。任务一利用知网发布的评价词词典，从三个领域中各选取 50 对种子词，然后采用知网的相似度计算和 PMI 相结合的方法找出观点词。任务二利用四种情感资源对观点词表和情感词表进行三步特征扩展，将观点句中出现在情感词表和观点词表中的词语分配不同的权重，通过求和进行极性判断。任务三以评价词为中心，在一定窗口内选取候选评价对象，完成评价关系的抽取。任务四借助 Lemur 检索工具，采用基于 KL 距离语言模型检索方法返回的结果，结合情感强度计算置信度。

关键词：情感倾向性分析；观点词；评价关系；KL 距离

Domain Sentiment Analysis Based on Feature Extension

Song Shien, Fan Xinghua, Zhao Jing, Wei Pingjie
Chongqing University of Posts and Telecommunications, Chongqing , 400065
E-mail: song_shi_en@126.com

Abstract: Sentiment Orientation analysis has a wide value of application, which has become one of the hot research topics in Natural Language Processing. This paper mainly introduces how the Chinese Information Processing Research Institute of Chongqing University of Posts and Telecommunications complete the task 1,2,3,4 in COAE2011. For task one, we use the Evaluation word dictionary released by HowNet to select 50 sets of seed words from each field. Then we combine the HowNet similarity computation and PMI to find out view words. For task two, we use four types of emotional resources to expanse the features of view vocabulary and emotional vocabulary. We give different weight to the words in the view vocabulary and emotional vocabulary and sum them to judge the polarity of the view sentences. For task three, we select the closest noun or gerund around the evaluation words as Evaluation object to complete the extracting of evaluation relationship. For task four, we combine the result of Lemur which based on KL divergence model and Emotional intensity to calculate the degree of confidence.

Keywords: sentiment orientation analysis; view words; evaluation relationship; KL divergence

1　引言

近年来随着 web2.0 的飞速发展，网络已成为人们交流意见和发表观点的重要平台。每时每刻都有大量的用户利用论坛、微博等网络渠道发布观点，单纯的靠人工处理难以应对收集和处理这些海量信息的要求，这就催生了情感分析技术。近年来，情感分析已经成为自然语言文本处理领域的研究热点之一，它在市场调查以及预测分析、民意调查、问答系统、信息检索等领域有着广阔的应用空间。

情感分析一般来说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。主要是通过抽取情感文本中有价值的情感信息来识别主观性文本以及计算文本中主观性极性

或者来进行观点挖掘。

在这次 COAE2011 中，我们完成了四个任务，任务一为领域观点词的抽取与极性判断；任务二为中文观点句的抽取；任务三为评价搭配抽取；任务四为观点检索。下面分别探讨四个任务的国内外研究状况和我们的工作，最后对我们的工作进行了总结。

2 领域观点词极性判断

词汇级的情感分析是情感分析的基础，观点词的识别和极性判断的准确性直接影响到后面的任务。这次评测与上两届不同的是，将领域知识对倾向性的影响融入到任务中，使任务增加了难度。下面在 2.1 中主要介绍国内外词汇级情感分析的主要研究方法；在 2.2 中介绍我们完成任务 1 的情况。

2.1 国内外研究状况

词汇级的情感分析研究方法主要有两种，一种是基于语料库的情感词的识别和极性判断；一种是基于词典的情感词的识别和极性判断。

基于语料库的方法主要是利用大规模的语料库的统计特性，观察一些现象来挖掘语料库中的情感词并判断极性。Hatzivassiloglou 和 McKeown[1]利用连词连接的两个形容词往往存在一定的关联性的特性，发掘大量的观点词。Turney 和 Littman[2]通过选取种子词集，提出了点互信息的方法判别某个词语是否是情感词语。Yu 和 Hatzivassiloglou[3]利用极性较强的形容词构建一个种子词集，通过计算新词和种子词集中的词语的共现率来判断新词的语义倾向。基于语料库的方法简单易行，但是情感词语的分布等现象不容易挖掘。

基于词典的方法主要是使用词典中词语之间的词义联系来识别情感词语和判断极性。一般国外的词典采用的是 WordNet，国内使用的是 HowNet，计算新词与基准词的语义距离，进而判断新词的极性。朱嫣岚等[4]提出了一种基于 HowNet 的词汇语义倾向计算方法。也有利用其它资源如路斌等[5]采用《同义词林》来计算词汇语义倾向性。

2.2 我们的工作

这次评测任务融入了领域知识和上下文环境。我们的工作主要分为两个部分，第一部分是构建领域种子情感词集；第二部分是采用组合的方法来对候选词进行语义倾向性识别和极性判断。

在构建领域种子情感词集的过程中，我们提取每个领域数据集中的动词和形容词（在任务中我们发现观点词主要为形容词和动词），分别与知网发布的正负评价词表取交集，然后人工从中选取领域性和倾向性较强的观点词对，每个领域各取 50 组。

在完成第二部分时，我们发现形容词为观点词的概率要大于动词，“的”字结构中形容词多为观点词，连词连接的两个形容词多为观点词且两个词语相关联，否定词以及程度副词修饰的动词多为观点词。

根据这几个特性，第二部分工作主要包括如下步骤：

（1）首先提取每个领域数据集中的“的”字结构的形容词与连词连接的形容词，利用知网来计算这些词语的语义倾向性，根据阈值选取词语，将其添加到种子词集中，构成新的种子词集 C_1。

（2）然后将阈值外的词语和剩下的形容词作为新候选词集，利用 C_1 使用 SO-PMI 的

方法计算词语的语义倾向性，根据阈值选取词语。

（3）对于否定词以及程度副词修饰的动词以及上一步阈值外的词语，利用初始领域种子情感词集使用知网来计算这些词语的语义倾向性，根据阈值选取词语，将其添加到领域情感种子词集中，形成新的种子词集 C_2。

（4）对于上一步阈值外的词语和剩下的动词，利用 C_2 使用 SO-PMI 的方法计算词语的语义倾向性。根据阈值选取词语。

（5）根据上面处理的先后顺序分别赋予 0.9,0.7,0.6,0.5 的置信度，对初始构建的领域种子情感词集赋予 1 的置信度，按置信度由高到底从每个领域选取 2000 词语提交。

3 观点句抽取和极性判断

观点句的抽取实际上是要完成主观句子的识别即主客观句分离，而极性判断实际上是完成句子级的情感分类。

3.1 国内外研究状况

关于观点句抽取和极性判断主要可分为两种思路：基于情感知识的方法以及基于机器学习的方法。

基于情感知识的方法主要是利用情感词典或者领域词典以及一些带有极性的组合评价单元来进行计算。基于机器学习的方法主要是选取大量有意义的特征来完成分类任务。有利用情感词或组合单元来进行加权求和，如文献[6, 7]。对于机器学习方法，主要是选取一些词语位置特征、词性特征、情感词特征等，用一些分类器如 NB，ME，SVM 来进行分类。

3.2 我们的工作

这部分工作中，我们借鉴了重庆邮电大学樊兴华、王鹏提出的一种基于扩展的文本倾向性分析方法[8,9]，并在实验中取得一定得效果。在此基础上利用了四种情感资源，分别是从任务一中选取的领域观点词构成的词表 T_1，知网的情感词表 T_2，程度副词表 T_3，否定词表 T_4。通过这四种资源对情感特征进行扩展，形成高效的组合单元，来进行加权求和完成观点句抽取和极性判断。

特征扩展的主要步骤如下：

（1）主要是针对上下文环境中出现的“否定词+倾向性词汇”的结构进行处理。

输入预处理处理的语料，T1，T2，T3，T4

Step1

Step1.1 读入预处理的句子 s

Step1.2 逐个读取 s 中切分后的词语 t,扫描否定词表 T4，如果 t 属于否定词表 T4，设置以 t 为起始，大小为 n 的检测窗口

Step1.3 扫描窗口内的每个词语 w,如果 w 属于 T1 或 T2，将 t 与 w 形成一个组合单元，加入到相应的词表中，该组合单元设定为（-1）*w 的倾向值

Step1.4 在 s 中插入组合单元作为扩展特征

Step1.5 在 s 中删除 t,删除 w

Step1.6 读取下一个句子转 Step1.2，如果所有句子都已处理完毕，输出经过 step1 处理

后的重构语料，经过 step1 处理后的倾向性词表,进入 step2

（2）程度副词对倾向性分析有着重要的作用，当一个倾向性的词被程度副词修饰时，它的倾向性强度会被明显增强或者减弱。主要是针对“程度副词+倾向性词语”的结构进行处理。

Step2

Step2.1 读入预处理的句子 s

Step2.2 逐个读取 s 中切分后的词语 t,扫描程度副词表 T3，如果 t 属于程度副词表 T3，设置以 t 为起始，大小为 n 的检测窗口

Step2.3 扫描窗口内的每个词语 w,如果 w 属于 T1 或 T2，将 t 与 w 形成一个组合单元，加入到相应的词表中，该组合单元设定为 t 的强度值*w 的倾向值

Step2.4 在 s 中插入组合单元作为扩展特征

Step2.5 在 s 中删除 t,删除 w

Step2.6 读取下一个句子转 Step2.2，如果所有句子都已处理完毕，输出经过 step2 处理后的重构语料，经过 step2 处理后的倾向性词表,进入 step3

（3）主要是针对“否定词+否定词+倾向性词汇”、“否定词+程度副词+倾向性词汇”的结构进行处理。

Step3

Step3.1 读入预处理的一个句子 s

Step3.2 逐个读出 s 切分后的特征 t 扫描倾向性列表，如果 t 属于，设置以 t 为基准，向前开大小为 n 的窗口

Step3.3 扫描窗口内的每个词语 w，如果 w 属于倾向性列表或者程度副词表或者否定词表,将 t 与 w 组合为一个 phrase 加入到相应的倾向性列表中,该 phrase 的倾向性值为其中 n 为否定结构在 phrase 中的出现次数，表示 t 和 w 中倾向性值较大的一个

Step3.4 在 s 中插入该语义结构 phrase 作为扩展特征

Step3.5 在 s 中删除组合前的 t 与 w

Step3.6 读取下一个句子，转 Step3.2，如果所有句子都处理完毕，则输出经过 step3 处理后的重构语料和经过 step3 处理后的倾向性词表,结束。

对于经过三步特征扩展的句子，和重构后的领域观点词表 T1 和情感词表 T2，我们采用加权求和的方法来判别句子极性。相对于单纯的用主观词表去加权求和，我们引入情感词表资源来进行辅助判断。即对主观词表 T1 中出现的词，和情感词表 T2 出现的词语赋予不同的权重，其中主观词表中的词语权重大于情感词表的词语的权重。

4 评价搭配抽取

4.1 国内外研究状况

评价搭配抽取主要是指评价短语和评价对象的抽取。有基于模板的方法，大部分学者，如 Bloom 等人[10]利用手工构建的句法规则来获取评价搭配；国内姚天昉等人[11]基于依存句法分析总结出一些匹配规则来用于评价搭配抽取。也有机器学习的方法，如樊娜等[12]利用最大熵模型的来进行主观关系的提取。

4.2 我们的工作

由于时间紧迫，我们采用一种最基本的方法。

处理过程如下：

输入某领域预处理语料和领域观点词表

Step1

Step1.1 筛选只包含观点词的句子

Step1.2 利用任务二的三步特征扩展的方法来获取评价短语及其极性，重构领域观点词表。经过三步特征扩展后，评价短语和候选评价对象之间的词语距离减小。

Step2

Step2.1 读入一个句子 s

Step2.2 查找句子中包含在领域观点词表的词语 w，以该词为中心，开大小为 m 的窗口，查找最近的名词或者动名词，作为它的直接评价对象。

Step2.3 根据评价词语和评价对象之间的词语距离来计算置信度

Step2.4 读入下一个句子，转 Step2.2。如果所有句子都处理完，输出相应的结果，并根据结果中的置信度，进行降序排序。

5 观点检索

观点检索有别于传统的检索，它是根据搜索主题词来查找包含用户观点的文本。

5.1 国内外研究状况

2006 年 TREC 引入了 Blog Track[13],使得很多学者致力于观点检索的研究之中。它要求兼顾文档的相关性和观点的倾向性，一种最直接的方式是使用线性加权函数来融合两部分的得分如文献[14]。

5.2 我们的工作

在这部分工作中我们采用了统计语言建模 IR 模型之中的 KL 距离模型，它的思想是查询对应某种语言，每篇文档对应某种语言，查询语言和文档语言的 KL 距离作为相关度度量。

主要步骤如下：

（1）首先利用百度百科和维基百科，找出主题词的相关词条，将这些词条作为候选扩展词集，采用计算点互信息的方法，对将要检索的 20 个主题词进行查询扩展。

（2）对于扩展后的主题词借助 lemur 检索工具，采用 KL 距离模型，计算文档与查询的相关度，取返回的前 1000 个结果。

（3）统计相关文档中观点词的个数，通过观点词的个数来计算情感得分。

（4）采用线性加权函数来调整两部分的得分。

（5）根据上一步的得分计算置信度。

6. 实验结果与分析

我们针对每组任务都提交了一组结果。

表 1 任务一、二、三评测结果

Tab.1 Task 1 2 3 Evaluation Results

TASK	标识	宏平均				微平均				P@1000 宏平均	P@1000 微平均
		Precision	Recall	F1	Raccuracy	Precision	Recall	F1	Raccuracy		
TASK1	CQUPT	0.2533	0.0766	0.1176	0.0766	0.2533	0.0766	0.1176	0.0766	0.5063	0.5063
	Median	0.342993	0.09474	0.14788	0.09474	0.337933	0.09474	0.147547	0.09474	0.571267	0.571267
TASK2	CQUPT	0.216499	0.428132	0.257968	0.255901	0.265283	0.433845	0.329244	0.372676	0.307667	0.307667
	Median	0.240815	0.406039	0.276324	0.255445	0.315252	0.450532	0.357871	0.36317	0.290183	0.290183
TASK3	CQUPT	0.016075	0.027016	0.020157	0.015352	0.021505	0.028667	0.024575	0.02412	0.016333	0.016333
	Median	0.025047	0.020403	0.019833	0.015183	0.033019	0.027191	0.027323	0.02309	0.02719	0.025778

从实验结果分析，任务一的结果距离平均水平还有一定的差距，一方面，可能是选取的种子词质量不好，导致后续工作不理想；另一方面，可能是由于将情感词当作观点词来处理，上下文环境和领域特性考虑不足，导致准确率和召回率普遍偏低。任务二、任务三接近于平均水平，证明了三步特征扩展的有效性，但由于后续选择的分类和处理模型过于简单，以及受任务一选取的观点词的影响，导致结果偏低。

表 2 任务四评测结果

Tab.2 Task 4 Evaluation Results

标识	宏平均			微平均			P@N	
	MACRO_P	MACRO_R	MACRO_F1	MICRO_P	MICRO_R	MICRO_F1	P@5	P@10
CQUPT	0.0881338	0.0277143	0.0335662	0.0210123	0.0877269	0.033904	0.55	0.5
MEDIAN	0.088134	0.027714	0.033566	0.021012	0.087727	0.033904	0.55	0.5

任务四达到平均水平，说明查询扩展和线性加权的方法用于观点检索的有效性，但是参数的选择还是存在一定的随机性。

7. 总结

本文主要介绍了重庆邮电大学中文信息处理研究所参加第三届中文倾向性分析评测的情况和具体方法。根据要求，完成了观点词抽取和极性判断、观点句抽取和极性判断、评价关系抽取、观点检索四个任务。任务一采用知网的相似度计算和 PMI 相结合的方法找出观点词。任务二利用四种情感资源对观点词表和情感词表进行三步特征扩展，将观点句中出现在情感词表和观点词表中的词语分配不同的权重，通过求和进行极性判断。任务三以评价词为中心，在一定窗口内选取候选评价对象，完成评价关系的抽取。任务四借助 Lemur 检索工具，采用基于 KL 距离语言模型检索方法返回的结果，结合情感强度计算置信度。在完成实验任务的过程中，时间较紧，做得仓促，方法比较简单，且存在很多不足。比如在分词之前没有对语料进行预处理，导致引入一部分数据噪音等等。在系统的实现过程中，我们发现有很多地方需要进一步的完善，今后我们将继续在中文倾向性分析领域做更深入的研究工作。

参 考 文 献

[1] Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives. In: Proceedings of EACL-1997. Morristown: ACL, 1997. 174-181.

[2] Turney P, Littman ML. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems(TOIS), 2003, 21(4):315-346.

[3] H.Yu and V. Hatziassiloglou Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences [A]. In: M. Collins and M. Steedman (eds): Proceed of EMNLP-03. 8th Conference on Empirical Methods in Natural Language Processing [C]. Sapporo. Japan:2003. 129-136.

[4] 朱嫣岚，闵锦，周雅倩，黄萱菁，吴立德，基于 HowNet 的词汇语义倾向计算[J].中文信息学报，2006，20(1):14-20.

[5] 路斌，万小军，杨建武，陈晓鸥，基于同义词词林的词汇褒贬计算.In: Proceedings of the 7th International Conference on Chinese Computing, wuhan,2007,17-23.

[6] Hu MQ, Liu B. Mining and Summarizing Customer Reviews. In: Proceedings of KDD-2004. 2004. 168-177.

[7] Turney P. Thumbs up Or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews: Proceedings of ACL-2002. 2002. 417-424.

[8] 樊兴华，王鹏，一种基于扩展的文本倾向性分析方法[J].计算机工程与应用，被录用.

[9] 王鹏，基于扩展的两步文本倾向性分析方法研究[D]. 重庆邮电大学，2011.

[10] Bloom K, Garg N, Argamon S. Extracting Appraisal Expressions. In: Proceedings of HLT-NAACL. 2007. 308-315.

[11] Yao TF, Nie QY, Li JC. An Opinion Mining System for Chinese Automobile Reviews. Frontiers of Chinese Information Processing, 2006,260-281.

[12] 龚娜，蔡皖东，赵煜，基于最大熵模型的观点句主观关系提取[J].计算机工程，2006，36(2):4-6.

[13] Ounis I, Rijke MD, Macdonald C, Mishne G, Soboroff I. Overview of the TREC-2006 Blog Track. In: Proceedings of TREC. 2006，20(1): 14-20.

[14] Zhang W, Yu C, Meng WY. Opinion Retrieval from Blogs. In: Proceedings of CIKM. 2007.