

基于有监督学习方法的多文档文本情感摘要*

李艳翠^{1,2}, 林莉媛^{1,3}, 周国栋^{1,3}

(1. 苏州大学计算机科学与技术学院, 江苏 苏州 215006; 2. 河南科技学院信息工程学院, 河南 新乡 453003;
3. 苏州大学自然语言处理实验室, 江苏 苏州 215006)

摘要: 本文研究有监督学习方法在多文档文本情感摘要中的应用。利用从亚马逊中文网和亚马逊英文网上收集的产品的评论语料, 抽取文本内特征、PageRank 特征、情感特征和评论质量特征, 基于有监督方法进行多文档文本情感摘要抽取。实验结果表明有监督学习方法比无监督学习方法在 ROUGE 值上有显著的提高, 情感特征和评论质量特征均有助于文本情感摘要。

关键词: 情感摘要; 评论质量; 情感特征; 有监督学习; 最大熵分类器

中图分类号: TP391

文献标识码: A

Multi-Document Opinion Summarization based on Supervised Learning

Method

LI Yancui^{1, 2, 3}, LIN Liyuan^{1, 3}, ZHOU Guodong^{1, 3}

(1. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China;

2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, Henan 453003, China; 3. Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu 215006,

China)

Abstract: This paper investigate the application of supervised learning methods in Multi-Document opinion summarization. We use the corpus collected from Amazon.cn and Amazon.com, extract text features, PageRank feature, opinion features and reviews quality features, and extract Multi-Document opinion summarization based on supervised learning method. Experimental results show that the ROUGE values are significantly improve by using supervised learning method than that unsupervised learning method. The opinion features and reviews quality features are helpful for summarization.

Key words: opinion summarization; comments quality; emotional features; supervised learning; maximum entropy classifier

1 引言

电子商务逐渐改变了人们的购物方式, 许多电子商务网站如亚马逊、淘宝和京东等, 不仅是电子商品展示与交易平台, 而且允许用户对商品发表评论。这些评论可以给潜在用户提供购物参考, 同时可以帮助生产商分析、了解产品的市场反映。然而, 一件热门的商品往往有成百上千条评论, 并且评论中存在一些观点偏激甚至文不对题等质量差的评论。一次性阅读完这些评论费时、费力, 文本摘要可以帮助用户快速有效的阅读评论, 但文本摘要主要针对语言严谨、文档结构规范、陈述客观事实的科技文献和新闻等。用户评论文本简短、风格多样化、结构灵活松散、内容带有主观性, 文本情感摘要 (Opinion Summarization) 就是对用户评论的观点和情感进行归纳、总结以帮助用户消化这些评论文本的情感信息。文本情感摘要可以帮助用户更好的理解网络上大量的情感信息, 并且可以给搜索引擎、问答系统、话题检测与跟踪提供支持等。

在文本情感摘要的研究中, 根据输出的不同可以将其分成两类: 第一类是输出产品的各方面

*收稿日期: 2014-6-15 定稿日期: 2014-7-28

基金项目: 国家 863 计划前沿技术研究类项目 (2012AA011102); NSFC 面上项目(61273320)

作者简介: 李艳翠(1982—), 女, 博士研究生, 讲师, 主要研究方向为自然语言处理; 林莉媛(1987—), 女, 硕士研究生, 主要研究领域为自然语言处理; 周国栋(1967—), 男, 博士, 教授, 博士生导师, 通讯作者, 主要研究领域为自然语言处理、多语言跨文本信息抽取。

特征信息，比如评价对象 (Opinion Target)、评价词 (Opinion Word)、评价持有者 (Opinion Holder) 等信息^[1,2]；第二类是从评论语料中抽取一系列有序的能够代表评论广泛意见的句子^[3-6]。目前，对于第二类文本情感摘要的研究相对较少，语料较贫乏，本文主要研究后者。

目前，文本情感摘要的研究主要集中在无监督学习方法的研究上，文献[9]手工标注了各 30 个主题的中文多文档评论语料。文献[9]中采用基于情感的 PageRank 模型从产品评论语料中抽取一系列有序的能够代表评论广泛意见的句子构建文本情感摘要。实验结果可以看出该方法取得了一定的效果，能够得到具有一定总结性的摘要，说明情感信息对文本情感摘要有一定的帮助。参考文献[9]的标注方法，本文手工标注了 30 个主题的英文多文档评论语料，在英文评论语料上的实验结果表明评论质量对情感摘要有重要的影响，能够有效的提高自动情感摘要的准确率。传统的基于无监督学习的文本情感摘要无法很好的融合评论文本中的主题相关性、情感相关性和评论质量信息等。

文献[10-12]采用有监督学习的方法在文本摘要上都取得了不错效果。因此，本文将文本情感摘要看成是一个二元分类问题，即将句子分为情感摘要句和非情感摘要句，将主题、情感和评论质量作为特征加入到机器学习方法中研究有监督学习方法在文本情感摘要中的应用。本文采用文献[9]中所标中文语料及本文新标英文语料，其中中英文产品评论各 30 个主题，采用有监督的方法研究文本情感信息和评论质量信息对文本情感摘要的影响。

本文第二部分语料标注情况；第三部分介绍文本情感摘要系统框架；第四部分介绍实验设置，对实验结果进行分析与比较；最后总结全文。

2 语料介绍

由于多文档文本情感摘要的研究较少，语料不足，本文采用文献[9]中所标中文多文档文本情感摘要语料和参考文献[9]标注方法所标注的英文多文档文本摘要语料。本节简要介绍语料的来源及对语料的处理、标注方法和标注一致性。

2.1 语料来源及处理

语料来源是分别从亚马逊中文网¹和亚马逊英文网²收集的 30 个主题 (产品) 的评论，每个主题中的评论均包含褒义评论和贬义评论。中文评论和英文评论中均包括了电子产品、书籍、影视和生活用品等的评论。中文评论中每个主题有 200 篇评论，包括评论的内容和作者的打分；英文评论中每个主题有 500 篇评论，包括评论的内容、作者打分以及其他用户对该评论的投票信息 (即有多少人认为该评论有用)。

在进行自动情感摘要前需要对自然语言文本进行预处理。根据需要先对原始语料进行句子识别，对语料中的每个主题的多文档集合以句子为单元进行分句处理。分句处理后整个段落被分为一个个的单句，每个单句一行。对于中文文本，句子中词与词之间没有明显的分词信息，故采用中科院开发的 ICTCLAS³分词工具对收集的 30 个主题的中文语料进行分词。

2.2 标注方法

标注文本情感摘要的标准是选择观点和内容在整个评论中出现频率最高、覆盖面最广的评论语句。标注不考虑非产品评论，例如：“评论亚马逊的客服太差”，“快递很慢”等。标注中文语料时，从每个主题中抽取 110 个单词左右的原文本作为摘要^[9]。标注英文语料时，从每个主题中抽取 120 个左右的单词原文本作为摘要。图 1 和图 2 给出了中文语料“KANSOON 凯速 KA05 型 静音双轮健腹轮 美腹瘦腹腹肌轮 带 1cm 加厚防滑垫”和英文语料“Kingston 8 GB Class 4 SDHC Flash Memory Card SD48GB”的文本情感摘要中一名标注者的标注结果。

健身轮非常轻巧，配的垫子挺厚的，就是有点小，试用了一下，轴承非常顺滑，一点都不卡

¹ <http://www.amazon.cn>

² <http://www.amazon.com>

³ [http://ictclas.org/\[EB\]](http://ictclas.org/[EB])

阻，所以用起来很舒服。塑料感蛮强的，不过应该不会一下子就没用的，外圈不会很硬，不伤地板挺好，使用起来也没有声音，生命在于运动嘛！有点令人失望，东西并不怎么样，做工很粗糙，又小，看上去不值这个价钱，我觉得垫子不怎么小，但是味道特大，不过在室外放几天也就好了，包装盒上只有英文说明，没有厂址、厂名、电话，给人感觉就是一个三无产品。安装方便，快递速度很不错。

图 1 中文标注示例

The speed rating and its small capacity and its low price make this an ideal choice for picture frames and some of the less expensive cameras and camcorders. This gives us a massive amount of storage capacity and has worked without a hitch. Speed is good enough and works great. Nice and quick data transfer, love the plastic case it comes with, and the locking feature on the side of the card itself. A quality product at a good price. One thing you need to consider is that it's a SDHC card, that won't work on all devices and won't be read by a standard SD card reader, you'll need a SDHC reader. Great card, very reliable with more than enough space for occasional RAW shooting.

图 2 英文标注示例

2.3 语料统计

标注结束后，对语料进行统计，表 1 给出了中英文原始句子数、标注句子数、原始单词数、标注单词数及压缩比统计结果。表 1 中原文句子数是指语料中 30 个主题的所有句子数目；标注句子数指 30 个主题人工标注的平均句子数；句子压缩比是标注句子数与原文句子数的比值。

表 1 压缩比统计结果

语言	原文句子数	标注句子数	句子压缩比	原文单词数	标注单词数	单词压缩比
英文	250516	688	0.27%	1920973	11059	0.58%
中文	11594	502	4.33%	213934	10489	4.9%

表 2 给出了人工标注的一些统计数据。表 2 中，英文 1、2、3 表示标注英文语料的 3 名标注者，中文 1、2、3 表示标注中文语料的 3 名标注者。总句子数是指 30 个主题中每名标注者抽取的句子总数，平均句子数是指每个主题的平均句子数。总单词数是指 30 个主题中每名标注者抽取的单词总数，平均单词数是指每个主题的平均单词数。

表 2 人工标注数据统计值

标注者	总句子数	平均句子数	总单词数	平均单词数
英文 1	238	7.93	3637	121.23
英文 2	199	6.63	3276	109.2
英文 3	251	8.37	4146	138.2
中文 1	176	5.87	3410	113.67
中文 2	159	5.3	3770	125.67
中文 3	167	5.57	3309	110.3

2.4 标注一致性

人工摘要具有很强的主观性，由于标注者对语义理解的不同以及知识背景的不同，使标注结果存在一定的主观性差异。表 3 给出了人工标注的 ROUGE 值。由表 3 中 ROUGE 结果可以看出标注者们对文本内容的概括相对一致，其抽取的文本的一致性较高。

表 3 人工标注 ROUGE 值

语言	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
英文	0.409	0.136	0.383	0.173	0.154
中文	0.436	0.203	0.408	0.185	0.171

图 3 的例子也说明了标注者的一致性，对于“KANSOON 凯速 KA05 型 静音双轮健腹轮 美腹瘦腹腹肌轮 带 1cm 加厚防滑垫”的评论，两名标注者都关注到了“健身轮非常轻巧，配的垫子挺厚的，就是有点小，试用了一下，轴承非常顺滑，一点都不卡阻，所以用起来很舒服。”，说明本文所用标注集具有一定的可信度。

中文 1: 挺好的, 无声静音, 正在持续锻炼中。健身轮非常轻巧, 配的垫子挺厚的, 就是有点小, 试用了一下, 轴承非常顺滑, 一点都不卡阻, 所以用起来很舒服。质量还行, 方便实用, 若能坚持, 应对健腹还是会有点效果。收到后我也试了一下, 确实很能锻炼腹肌, 商品是那种比较轻的塑料做的, 感觉很结实, 价钱也比超市便宜。锻炼几天才评论, 锻炼的效果不错, 锻炼后第二天腹部微痛。

中文 2: 健身轮非常轻巧, 配的垫子挺厚的, 就是有点小, 试用了一下, 轴承非常顺滑, 一点都不卡阻, 所以用起来很舒服。塑料感蛮强的, 不过应该不会一下子就没用的, 外圈不会很硬, 不伤地板挺好, 使用起来也没有声音, 生命在于运动嘛! 有点令人失望, 东西并不怎么样, 做工很粗糙, 又小, 看上去不值这个价钱, 我觉得垫子不怎么小, 但是味道特大, 不过在室外放几天也就好了, 包装盒上只有英文说明, 没有厂址、厂名、电话, 给人感觉就是一个三无产品。安装方便, 快递速度很不错。

图 3 一致性标注示例

3 基于有监督学习的多文档文本情感摘要系统

3.1 系统框架

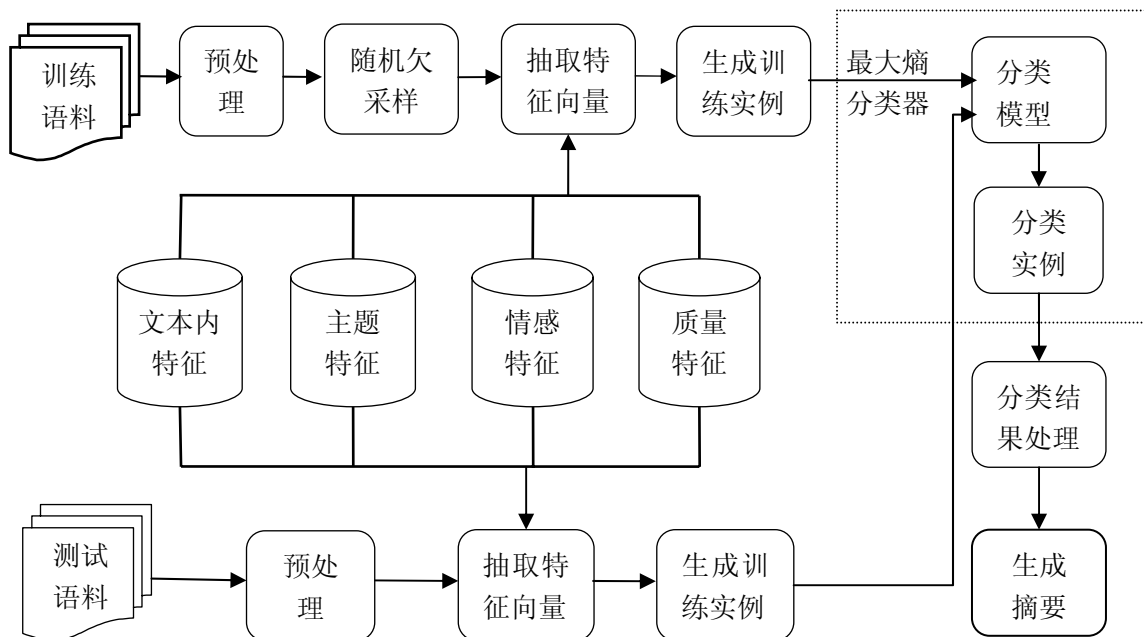


图 4 基于有监督学习的文本情感摘要框架

图 4 给出了有监督学习的文本情感摘要框架。训练语料首先进行预处理，在预处理训练语料后，训练文本中正类（摘要类）为人工标注的句子，负类（非摘要类）为文本中去除正类的句子，且这些句子与正类中的句子的相似度小于一定的阈值，本文实验中中文设定为 0.65，英文设定为 0.75。由于负类文本数明显大于正类文本数，而样本分布的不平衡往往会使传统的机器学习分类方法在分类过程中严重偏向多样本类别，从而导致分类的性能急剧下降。因此，本文对训练语料进行了随机欠采样，随机选择与正类样本数目一样的负类样本进行实验。因为随机欠采样存在一定的偶然因素，所以本文报告的结果是进行 20 次实验取平均值。本文抽取文本内特征、主题特征、情感特征和质量特征（英文）生成训练实例，然后利用最大熵分类器得到分类模型。测试文本首先进行预处理，抽取特征向量，然后生成测试实例，分类器根据训练好的模型对测试实例进行分类，最后根据分类结果生成文本情感摘要。

3.2 特征选择

在机器学习的分类方法中，特征选择是关键的一步。本文选用了四种特征构建句子的特征向量，分别为文本内特征、PageRank 特征、情感特征和质量特征。中英文参数的设定是在做无监督方法是通过实验确定的，具体请参考文献[9]。

- 文本内特征：1) 基于词的 Unigram 特征，特征的权值为句子中词的词频数；2) 句子长度特征，特征权值为：句子单词数/5。
- PageRank 特征：特征权值为文本内句子通过基础 PageRank 算法计算得到的打分值。此特征参考文献[13]，将每句话视为一个节点，句子之间的文本余弦相似度为边之间的权重，从而构建网络，并计算每句话的 PageRank 值。
- 情感特征：选用词计数方法加入“opinion”特征作为情感特征，中文中特征权值为 1.5，英文中特征权值为 2.0。词计数（Term-counting）的方式：具体来讲，如果句子 s 包含情感词，则认为该句带有情感；如果 s 不包含情感词，则认为该句没有情感。
- 质量特征：本文参考文献[14]中使用“helpful”信息评价评论的质量，“有用率” h 大于 0.6 的评论中的句子添加评论质量特征“helpful”，特征权值设定为：

$$h + \frac{\log(\text{总评论人数})}{2}。$$

4 实验结果与分析

本文实验中使用的语料第 2 节介绍的 30 个主题的中文评论和 30 个主题的英文评论，人工标注的结果作为评价标准，评测工具使用 ROUGE-1.5.5，评测指标是 ROUGE-1、ROUGE-2、ROUGE-L、ROUGE-W 和 ROUGE-SU4。中文抽取 110 个单词左右的句子构建摘要，英文抽取 120 个单词左右的句子构建摘要。监督学习使用 3.1 中描述的特征，采用 MALLET 机器学习工具包⁴中的最大熵分类器，分类算法的所有参数都设置为默认值。词计数方法使用的情感词中文选用实验室已经收集并标注好的中文情感词集，其中正面情感词数量为 846，负面情感词数量为 809；英文来自于 MPQA⁵的情感词集。

实验中，将 25 个主题的评论文本作为训练语料，剩余的 5 个主题的评论文本作为测试语料，进行 6 组实验，每组实验的测试语料都不同，得到 30 个主题的文本情感摘要。由于训练文本中负类是随机选取的，所以实验中有监督学习方法的实验结果为 20 组实验结果取平均值。表 4 和表 5 分别为中文和英文的实验结果，表 4 和表 5 中：

Unigram 表示只采用基于词的 Unigram 特征的实验结果

⁴ <http://mallet.cs.umass.edu/>

⁵ http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Uni_Len 表示只采用文本内特征的实验结果

AddTopic 为在使用文本内特征的基础上添加了 PageRank 特征的实验结果

AddOpinion 指的是在已经使用了文本内特征和 PageRank 特征的基础上，进一步添加情感特征的实验结果

AddHelpful 为进一步添加评论质量特征的实验结果

PageRank 指的是文献[9]所用传统基于主题的非监督 PageRank 算法的实验结果

Bi-Rank 指的是文献[9]提出的融合主题和情感信息的基于双层图模型的 PageRank 算法实验结果

PageRank_{Term, helpful} 是指融合主题信息、情感信息和评论质量信息的 PageRank 方法

Random 指每个主题中随机抽取句子作为该主题的文本情感摘要，由于结果存在随机性，所报告的结果是随机抽取 20 次的平均值的结果

Human 为人工抽取每个主题的文本情感摘要的结果

表 4 中文有监督学习的文本情感摘要结果

Method		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
文献[9]方法	PageRank	0.367	0.101	0.318	0.141	0.133
	Bi-Rank	0.386	0.121	0.345	0.155	0.146
本文方法	Unigram	0.377	0.113	0.333	0.146	0.131
	Uni_Len	0.386	0.121	0.342	0.150	0.137
	AddTopic	0.392	0.126	0.349	0.152	0.142
	AddOpinion	0.403	0.142	0.358	0.158	0.152
	Random	0.335	0.066	0.301	0.129	0.105
	Human	0.436	0.203	0.408	0.185	0.171

表 5 英文有监督学习的文本情感摘要结果

Method		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU4
文献[9]方法	PageRank	0.330	0.054	0.291	0.126	0.108
	Bi-Rank	0.368	0.084	0.335	0.146	0.136
	PageRank _{Term,helpful}	0.368	0.106	0.340	0.154	0.136
本文方法	Unigram	0.347	0.070	0.310	0.136	0.115
	Uni_Len	0.354	0.076	0.319	0.141	0.120
	AddTopic	0.364	0.084	0.330	0.146	0.127
	AddOpinion	0.377	0.090	0.342	0.153	0.137
	AddHelpful	0.386	0.111	0.353	0.163	0.143
	Random	0.312	0.041	0.285	0.122	0.091
	Human	0.409	0.136	0.383	0.173	0.154

从表 4 和表 5 的实验结果来看，由于没有考虑句子间的主题相关性和情感信息，Random 的实验效果在中英文上的效果均不理想。文献[9]所用考虑了句子间的主题相关性的 PageRank 的实验结果在 ROUGE-1、ROUGE-2、ROUGE-L、ROUGE-W 和 ROUGE-SU4 上优于 Random。而文献[9]提出的融合主题和情感信息的基于双层图模型的 PageRank 算法，既考虑了句子间的主题相关性又考虑了句子的情感信息，在中英文的实验上均表明实验效果比 PageRank 也有明显的提高。

文献[9]用的是无监督的方法，本文用的是有监督学习的方法。比较表 4 和表 5 中的 Bi-Rank

和 AddOpinion 的中英文情感摘要结果,可以看出 ROUGE-1、ROUGE-2、ROUGE-L、ROUGE-W 和 ROUGE-SU4 的实验结果均有所提高。其中,表 4 中,本文 AddTopic 比文献[9]最好结果 Bi-Rank 在 ROUGE-1 上高 1.7 个百分点,表 5 中,AddOpinion 比 Bi-Rank 在 ROUGE-1 上提高了 0.9 个百分点。表 5 中 PageRank_{Term,helpful} 和 AddHelpful 均为添加评论质量特征的实验结果,有监督学习方法在 ROUGE-1 的表现上比无监督学习方法高了 1.6 个百分点。可见,采用有监督学习的方法得到的文本情感摘要的效果优于无监督学习方法的实验效果。

比较表 4 中的最好结果 AddTopic 和人工抽取的摘要结果,可知,用有监督的方法抽取的摘要 ROUGE-1 值比人工抽取的摘要结果低 3.3%,表 5 中的最好结果 AddHelpful 中 ROUGE-1 值仅比人工抽取的情感摘要低 2.3%。由实验结果可知,有监督学习方法能更有效的利用主题信息、情感信息和评论质量信息,获得的文本情感摘要更接近人工摘要的结果。

表 4 和表 5 中的数据显示,当考虑了句子间的主题相关性时,文本情感摘要的结果在中文和英文语料上都有了一定的提高。当使用了情感特征时,中文和英文的摘要结果在 ROUGE-1 上都提高了约 1 个百分点,这说明情感信息是情感文本的一个重要内容,其对文本情感摘要有着重要的影响。从表 5 可以发现,当进一步使用评论质量特征时,英文文本情感摘要的结果与 AddOpinion 相比在 ROUGE-1 上提高了 0.9 个百分点,而其他几个指标也有相应的提升,这表明质量好的评论其涵盖的信息更多,更能为读者提供帮助,同时,读者也更相信质量好的评论。

5 结语

本文研究了有监督学习方法在文本情感摘要中的应用,使用最大熵分类方法分别作用在中文和英文多文档文本上,抽取情感摘要,并与无监督学习方法进行了对比。由实验结果可以看出有监督学习方法与无监督学习方法相比在 ROUGE 值上有显著的提高,这说明在文本情感摘要中,有监督学习方法能够更有效的利用情感文本的主题信息、情感信息和评论质量信息。同时,实验结果说明了情感信息和质量信息可以帮助文本情感摘要,这表明情感文本的主题信息和情感信息密切相关,在情感摘要中不能忽略情感的重要性,同时质量高的情感文本能更有效的帮助读者进行阅读。

参考文献

- [1]. Hu M., Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.
- [2]. Titov I., McDonald R. A joint model of text and aspect ratings for sentiment summarization [J]. Urbana, 2008, 51: 61801.
- [3]. Carenini G., Cheung J. C. K., Pauls A. Multi-document summarization of evaluative text [J]. Computational Intelligence, 2013, 29(4): 545-576.
- [4]. Carenini G., Cheung J. C. K. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality[C]//Proceedings of the Fifth International Natural Language Generation Conference. Association for Computational Linguistics, 2008: 33-41.
- [5]. Lerman K., Blair-Goldensohn S., McDonald R. Sentiment summarization: evaluating and learning user preferences[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 514-522.
- [6]. Lerman K., McDonald R. Contrastive summarization: an experiment with consumer reviews[C]//Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers. Association for Computational Linguistics, 2009: 113-116.
- [7]. Nishikawa H., Hasegawa T., Matsuo Y., et al. Opinion summarization with integer linear programming formulation for sentence extraction and ordering[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 910-918.
- [8]. Wang D., Liu Y. A pilot study of opinion summarization in conversations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 331-339.
- [9]. 林莉媛, 王中卿, 李寿山, 周国栋. 基于 PageRank 的中文多文档文本情感摘要[J]. 中文信息学报. 2014, 28(2):85-90.
- [10]. Liu F., Liu F., Liu Y. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion[C]//Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. IEEE, 2008: 181-184.
- [11]. Wong K. F., Wu M., Li W. Extractive summarization using supervised and semi-supervised

- learning[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 985-992.
- [12]. Li C., Qian X., Liu Y. Using supervised bigram-based ILP for extractive summarization[C]//Proceedings of ACL. 2013: 1004-1013.
- [13]. Shen D., J. Sun, H. Li, Q. Yang and Zheng Chen.Document Summarization using Conditional Random Fields[C] // In Proceeding of IJCAI-07.
- [14]. Hong Y., Lu J., Yao J., et al. What reviews are satisfactory: novel features for automatic helpfulness voting[C]//Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 495-504.