中文文本情感倾向性分析

黄萱菁　　赵　军

引言

大约在两年半前，《新华网》、《环球时报》等大众媒体纷纷转载了英国《新科学家》杂志的一则报道，英国 Corpora 软件公司开发了一套名叫“感情色彩(Sentiment)”的软件[1]，它能判断报纸刊登的文章对一个政党的政策是持肯定态度还是否定态度、或者网上评论文章是称赞还是贬低一种产品，以帮助政府和一些大公司全面了解公众舆论对他们的看法。

这则报道之所以引起了舆论的广泛关注，是因为它介绍的是一个非常新颖而又很有价值的研究方向。所谓文本情感倾向性分析，就是对说话人的态度（或称观点、情感）进行分析，也就是对文本中的主观性信息进行分析。由于立场、出发点、个人状况和偏好的不同，民众对生活中各种对象和事件所表达出的信念、态度、意见和情绪的倾向性必然存在很大的差异。这种差异尤其体现在论坛、博客等反映草根观点的网络媒体上。

长期以来，要了解关于某个问题的报道是正面的还是反面的，是消极的还是中立的，往往需要求助于调查公司。这些公司的员工仔细阅读有关某个机构、个人、事件或问题的所有文字，然后就这些评论的态度做出反馈。这不仅耗费大量人力和财力，而且过程相当缓慢。由此可见，这一过程的自动化，具有很好的商业应用前景。

文本情感倾向性分析属于计算语言学的范畴。在计算语言学以及相关领域，研究人员以前普遍关注的是客观性信息的分析和提取，对主观性信息分析与提取的研究尚处于起步阶段，存在很多问题需要进行全面的探索。这项研究涉及到计算语言学、人工智能、机器学习、信息检索、数据挖掘等多方面研究基础，因此文本情感倾向性分析也具有重要的学术研究价值。

总体来看，情感倾向分析的研究大致可以分成词语情感倾向性分析、句子情感倾向性分析、篇章情感倾向性研究、海量信息的整体倾向性预测四个研究层次。接下来将介绍在各个层次所取得的研究进展，之后是情感倾向性分析标准语料库的建设和系统评测，最后是本文的结论。

词语情感倾向性分析

对词语的情感倾向研究是文本情感倾向分析的前提。具有情感倾向的词语以名词、动词、形容词和副词为主，也包括人名、机构名、产品名、事件名等命名实体。其中，除部分词语的褒贬性（或称为极性，通常分为褒义、贬义和中性三种）可以通过查词典[2]的方式得到之外，其余词语的极性都无法直接获得。而词语的情感倾向除了极性之外，还包括倾向性的强烈程度。例如，“谴责”的强度就远远超过了“批评”和“指责”，而这种强度是很难由词典编撰者用人工的方式量化的。另外，词语的极性往往取决于特定的上下文环境，例如，“骄傲”在表示

1 http://www.corporasoftware.com/products/sentiment.aspx

2 例如，General Inquirer [Stone,1966]，知网：heep://www.keenage.com

自豪概念时，是褒义词，而在表示自满概念时，则是贬义词。

词语情感倾向分析包括对词语极性、强度和上下文模式的分析。其分析结果甚至可以写入到语义词典中，如北京大学计算语言学研究所基于人民日报基本标注语料库的真实文本实例进行统计归纳，从而得到词语的情感倾向，然后在现代汉语语法信息词典中形式化[王治敏 2004]。词语情感倾向分析目前主要有三种方法：

①由已有的电子词典或词语知识库扩展生成情感倾向词典：英文词语情感倾向信息的获取主要是在 WordNet[3]和 General Inquirer 的基础上进行的[Hatzivassiloglou,1997] [Wilson2005]；而中文词语情感倾向信息的获取依据的主要有 HowNet[朱嫣岚 2006]。这种方法的主要思想是：给定一组已知极性的词语集合作为种子，对于一个情感倾向未知的新词，在电子词典中找到与该词语义相近、并且在种子集合中出现的若干个词，根据这几个种子词的极性，对未知词的情感倾向进行推断。这种方法对种子词数量的依赖比较明显。

②无监督机器学习的方法：这种方法与第①种方法类似，也是假设已经有一些已知极性的词语作为种子词，对于一个新词，根据它和种子词的紧密程度对其情感倾向性进行推断。不同的是，第①种方法的词语紧密程度度量是以词典信息为依据判断，而这种方法是根据词语在语料库中的同现情况判断其联系紧密程度。根据[Turney, 2002&2003]的经典方法，假设以“真”、“善”、“美”作为褒义种子词，“假”、“恶”、“丑”作为贬义种子词。则任意其它词语的语义倾向 SO 定义为与各褒义种子词 PMI(点态互信息量)之和，减去与各贬义种子词 PMI 之和。SO 的正负号就可以表示词语的极性，而绝对值就代表了强度。词语 A 和 B 的 PMI 定义为它们在语料库中的共现概率与 A、B 概率之积的比值。这个值越高，就意味着相关性越大。有趣的是，PMI 计算可通过搜索引擎进行，计算 A 的概率，可以把 A 当作查询送给搜索引擎，那么返回的 Hits 值（含有 A 的页面数）和总的索引页面数的比值，就可以认为是 A 的概率。要计算 A 和 B 的共现概率，只要把 A 和 B 同时送给搜索引擎就可以了。这种方法同样存在着对种子集的依赖性比较强的问题，而且噪声比较大。

③基于人工标注语料库的学习方法：首先对情感倾向分析语料库进行手工标注。标注的级别包括文档集的标注（即只判断文档的情感倾向性）、短语级标注和分句级标注。在这些语料的基础上，利用词语的共现关系、搭配关系或者语义关系，以判断词语的情感倾向性。这种方法需要大量的人工标注语料库，典型的工作如 Wiebe 利用词语的搭配模式发现在主观性文本中的倾向性词语及其搭配关系[Wiebe,2001]。

不可不提的是香港城市大学语言资讯科学中心在 LIVAC 共时语料库上进行的名人信誉分析研究。他们选择泛华语地区有代表性的中文媒体，对相应的新闻报道进行深层次的人工标注，对并在该语料库上开展中文文章正负两极性自动分类的研究，通过人物褒贬指数的计算，发布京港台双周名人榜，并用－10 到 10 之间的数表示名人在三地报章的信誉度[T'sou et al. 2005]，例如某段时间内，陈水扁在中港台三地的信誉度分别是-10、-6.2 和-4.6。

句子情感倾向性分析

词语情感倾向分析的处理对象是单独的词语或者实体，而句子情感倾向性分析的处理对象则是在特定上下文中出现的语句。其任务就是对句子中的各种主观性信息进行分析和提取，包括对句子情感倾向的判断，以及从中提取出与情感倾向性论述相关联的各个要素，包括情感倾向性论述的持有者、评价对象、倾向极性、强度，甚至是论述本身的重要性等。例如，对于例句“XXX

[3] WordNet

绝不是一款能放心开下公路的 SUV。当然，在公路上它的表现令人满意”，我们可以得到以下两条情感倾向性论述：

论述	持有者	对象	极性	强度	重要性
1	作者	XXX	贬	强	强
2	作者	XXX	褒	弱	弱

如果说句子是点，那么由句子构成的篇章是线，而由多篇文章组成的语料库就是面。在句子情感倾向分析的基础上，可以很方便地进行篇章的情感倾向分析，甚至可以得到海量信息的整体倾向性态势。

长期以来，客观性信息提取都是计算语言学的研究热点，但尚未研究透彻。近年来 ACE(自动内容提取会议)的评测结果也表明，命名实体识别和指代消解的性能尚可，但实体间关系的提取则显得很困难[4]，主观性信息的提取更是如此。这方面的研究即使在英文上也是少数，且集中在对句子情感倾向性的判断上 [Kim,2004] [Wiebe & Riloff,2005]。在此基础上，[Kim,2005]尝试识别情感倾向性论述的持有者。而关于系统地提取句子的情感倾向性信息的多个要素方面的研究，目前还少有报道。

对中文的研究也主要集中在句子情感倾向性论述的某个侧面。例如，[王波 2007]的主要工作是在情感倾向性论述定位评价对象。考察两个例句：

(a) 功能很全面，价格也很便宜。

(b) 我买电脑时最关心的是功能和价格。

功能和价格在例句 a 中是评价对象，但在例句 b 中并不是。他主要考察在只有规模很小的标注语料可用时，如何采用半监督自学习方法对评价对象进行迭代学习。

[王根 2007]则关注于句子情感倾向性的判断。他提出了一个分级模型，可以将句子的主客观性判别、褒贬分类和褒贬分级统一在一起：首先将句子分为主观句和客观句，对于主观句，分成赞扬和贬斥两类，每类再分成强烈和微弱两种强度；并提出了一种基于多重标记 CRF 的方法来加以解决。

[章剑峰 2007]所针对的具体任务是抽取评价词和目标对象之间的关联关系。这里的关联除了句法上的直接关联，也包括语义上的间接关联。目标对象被细分为直接评价对象和间接评价对象两种。如在例句“品牌 A 的造型很美观”中，评价词是“美观”，“造型”是“美观”直接评价的对象，而“品牌 A”是间接评价对象。他们把在同一句子中共现的评价词与评价对象作为候选集合，应用最大熵模型进行关系抽取。

篇章情感倾向性研究

篇章级情感倾向性分析，就是要从整体上判断某个文本的情感倾向性，即褒贬态度。有代表性的工作包括[Turney, 2002]和[Pang, 2002]对电影评论的分类。Turney 的方法是将文档中词和短语的倾向性进行平均，来判断文档的倾向性。这种方法基于情感倾向性词典，不需要人工标注了文本情感倾向性的训练语料；Pang 的任务是对电影评论的数据按照倾向性分成两类，他利用人工标注了文本倾向性的训练语料，基于 unigram 和 bigram 等特征，学习分类器。

将篇章作为一个整体，笼统地进行主观性分析存在很大局限性，其本质缺陷在于假设整个文本是针对同一个对象进行评论。而真实文本往往由包含多个对象，不同的对象所涉及到的观点、态度等主观性信息是有差异的。从另一方面看，篇章内的对象总数仍是有限的，不足以支撑对于整体倾向性的挖掘。因此，这两年根据情感倾向对篇章进行褒贬态度分类的研究有减少的趋势；更多的研究集中在篇章内进行情感倾向性论述的分析，以及在大规模数据集上进行整体倾向性分析。

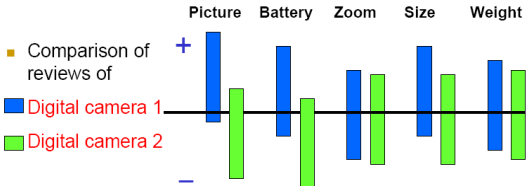
海量数据的整体倾向性预测

所谓整体倾向性预测，是针对海量数据而言的，其主要任务是：对从不同信息源抽取出的、针对某个话题的情感倾向性信息进行集成和分析，进而挖掘出态度的特点和走势。

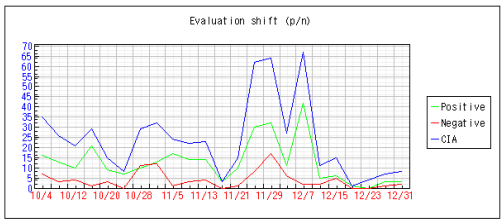
Durant 提出利用 Web log 来帮助对 blog

[4] http://www.nist.gov/speech/ tests/ace/index.htm

情感倾向的分类[Durant et al. 2006]。UIC 的 Liu 和 Hu 等人讨论了从评论中挖掘产品特性的方法，从而得到用户对于产品或者产品某个特性的整体倾向性[Hu & Liu, 2004][Liu et al. 2005]。例如，他们根据用户评论来比较两个款式的数码相机，并用如下图所示的可视化文摘来显示分析结果，每列代表相机的一个属性，水平线表示中立态度，彩条则反映了用户的褒贬度的主要取值范围。



日本富士通公司则开发出了从中文博客和论坛中提取对企业及其产品的评价信息的技术，根据从 web 上抓取的大量用户评论得到产品的整体信誉度，以图表的形式展现不同时间里企业和品牌的正面或负面等评价信息。例如，下图显示了某产品在 3 个月内用户评价的情况，其中绿、红两种颜色分别表示某一天持肯定、否定态度的评论数，而蓝线则表示评论总数。



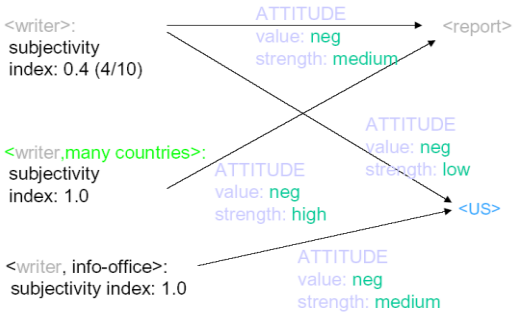
该系统受到了产业界的广泛关注，日经产业新闻、日刊工业新闻等报刊和网络媒体对此进行了广泛的报道。类似地，上海交通大学则开发了一个用于汉语汽车论坛的意见挖掘系统。其目的是在电子公告板、门户网站的各大论坛上挖掘并且概括顾客们对各种汽车品牌的不同性能指标的评论和意见，并且判断这些意见的褒贬性以及强度。然后，通过对文本处理的综合统计，给出可视化的结果[姚天昉 2006]。

标准语料库建设

由于文本情感倾向性分析即使在国外也是一个非常新颖的研究方向，一开始并没有一个文本情感倾向性分析的评测规范对该领域的研究任务进行清晰的定义。同时也没有一个被普遍接受的文本情感倾向性分析的标准语料库支撑关键技术的研究、评测和应用系统的开发。这个问题使得一段时间内该领域的研究显得比较混乱，不能进行客观的比较测试，严重地影响了该领域的研究水平和技术发展。

国外的研究人员已经意识到这个问题，并着手解决。影评数据集[5]是使用较多的一个语料库，它由电影评论组成，其中持肯定和否定态度的各 1000 篇。另外还有标注了整体褒贬极性的句子各 5000 句。影评库被广泛应用于词汇和篇章情感倾向研究，但由于未进行更细粒度的标注，它不适应句子情感倾向性分析的要求。

NRRC Summer Workshop 所开发的 MPQA(Multiple-Perspective QA)库是一个进行了深度标注的语料库，它标出了倾向性论述的持有者、对象、极性、强度等要素[MPQA][Wiebe,2005]。下图是用二分图表示的一组深层标注结果，原文的大意是美国发布的人权报告引起了许多国家的不满。图中左部表示情感倾向论述的持有者，右部为评价对象，箭头上的标记则显示了倾向性的极性和强度。



MPQA 语料库存在的主要问题是规模太小，只有 57 篇文章进行了深度标注，但它所建立的标注规范还是很有意义的。

关于语料标注，还有一个值得关注就问题，即极性和强度应该如何标注？通常把极性分为褒义、贬义和中性 3 类，而把强度分成强、中、弱 3 个级别。是否可以分得更细，

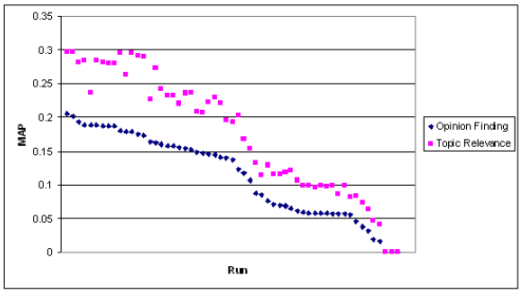
5 http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/

研究者们做了许多有意义的尝试，例如，[陈建美 2007]把情感细分为乐、好、怒、哀、惧、恶、惊等，共计 7 大类，20 小类。一般说来，较多的层次表现力更强，更能体现语言上的细微差别；但也会给人工标注带来困难，标注者之间更容易产生不一致，此外也会带来一定程度的数据稀疏问题。

系统评测

2006 年，文本检索会议（TREC）新增加了博客（blog）检索评测任务[6]，对于给定的查询，要求在博客数据集上（近 30GB，320 万篇）检索带有观点的文章。例如，给定查询对象“*Skype*”，要求检出的网页必须和 *Skype* 相关，且必须含有主观性信息，而不能是纯客观的叙述。除了观点检索任务之外，还有一个篇章态度分类的子任务，给定一批人工标注了正负极性的训练语料，用来测试基于监督学习的篇章态度分类系统。

如下图所示，在所提交的 50 多组检索结果中，观点检索的性能和常规的相关性检索都有很大的差距，前者约为后者的 2/3，说明主观性分析的准确率还有很大的提升空间[Ounis2006]。



NTCIR[7]的观点分析评测同样出现在 2006 年。不同于 TREC 所关注的观点检索，NTCIR 评测的主要任务是从新闻报道中提取主观性信息，并建立中、英、日 3 种语言的标准语料库。给定各个语种的句子，要求参加评测的系统判断句子是否和篇章的主题相关，并从句子中提取出观点持有者，评价词极性等信息。

下图是 NTCIR 观点分析的路线图，可以看出他们的目标是进行多语种、多信息源、多粒度、深层次的主观性信息提取。这一目标通过渐近方式实现，比如，即将开始的 NTCIR-7 评测已经转移到博客信息源，而对于情感倾向论述，也开始尝试提取被评价对象。

Genre	Subjectivity	Holder	Polarity	Strength
News	NTCIR-6	NTCIR-6	NTCIR-6	
Review	NTCIR-7	NTCIR-7	NTCIR-7	NTCIR-7
Blog	NTCIR-8	NTCIR-8	NTCIR-8	NTCIR-8

Stakeholder	Temporal	Language	Granuality	Application
		Chinese	single-sent	Summarization
NTCIR-7		English	clause	QA
NTCIR-8	NTCIR-8	Japanese	multi-sent	Opinion tracking
		CJE	document	Consistency checking
				Trend

结论

总结情感倾向性分析的研究现状，我们可以发现一下两个特点：

首先是开放性。情感倾向分析给现有的自然语言处理加入了许多新的研究内容。和文本检索相结合的产物，是观点检索；和信息提取相结合，即为主观性信息提取。这两个任务目前已经取得了 TREC 和 NTCIR 的关注。另外，如果和问题回答相结合，就是多视角的问题回答；和自动文摘特别是多文档自动文摘相结合，就是基于观点的文摘。这些都将成为很有意义的研究方向。

其次是和 Web2.0 技术的紧密结合。TREC 评测选择的是博客语料；NTCIR 首次评测是在新闻语料上进行的，从第二次起就转换到博客语料上。这是因为博客、论坛作为草根媒体，可以反映大众的真实情感和态度。其中，最为企业界关注的，是从 CGM(Consumer Generated Media)上获取产品评论信息，分析用户对于产品是持肯定还是否定的态度，并进行综合分析；而这也是最困难的任务。

和国外的研究相比，中文的情感倾向性分析有一定滞后，现有的主要工作集中在词语的情感倾向性分析。对情感倾向进行更细致的研究，特别是句子级的倾向性分析和海量信息的整体倾向预测，将是未来的主要研究趋势。与此同时，制订情感倾向性语料库标注规范，充分覆盖情感倾向性论述的要素；按照严格的程序进行人工标注和一致性

[6] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

[7] http://research.nii.ac.jp/ntcir/ntcir-ws6/opinion/index-en.html

检验，得到较大规模的细粒度标注语料库，并在此基础上对情感倾向分析方法进行客观公正的评测，必将是对中文情感倾向研究的重大贡献。

参考文献

[Stone,1966] Stone, Philip J., Dunphy, Dexter, Smith, Marshall, Ogilvie, Daniel, 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT

[王治敏2004] 王治敏，朱学锋，俞士汶，基于现代汉语语法信息词典的词语情感评价研究, Recent advancement in Chinese Lexical Semantics, Proceeding of 5th Chinese Lexical Semantics Workshop (CLSW-5), 2004, Singapore

[Hatzivassiloglou,1997] Hatzivassiloglou and McKeown, Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics, pages 174-181, Association for Computational Linguistics, Madrid, ES, 1997.

[Wilson2005] Theresa Wilson , Janyce Wiebe, and Paul Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, HLT-EMNLP-2005

[朱嫣岚等,2006] 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 年第 1 期

[Turney,2002] Turney Peter, 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. 417-424, 2002

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

[Wiebe,2001] J. Wiebe, J. M. A corpus study of evaluative and speculative language. In Proceedings of the 2nd ACL SIG on Dialogue Workshop on Discourse and Dialogue (Aalborg, Denmark).

[T'sou et al. 2005] T'sou Benjamin, Kwong Olivia, Wong Wei-Lung, Lai Tom. 2005, Sentiment and Content Analysis of Chinese News Coverage, International Journal of Computer Processing of Oriental Languages, 18: 171-183

[Kim,2004] Kim, S. & E. Hovy. 2004. Determining the Sentiment of Opinions. In: Proceedings of COLING-04: the 20th International Conference on Computational Linguistics, 2004

[Kim,2005] Soo-Min Kim and Eduard Hovy, Identifying Opinion Holders for Question Answering in Opinion Texts 2005, In: Proceedings of AAAI-05 Workshop on Question Answering Restricted Domains. 2005

[Wiebe & Riloff,2005] J. Wiebe and E. Riloff, Creating Subjective and Objective Sentence Classifiers from Unannotated Text. In: Proceedings of CICLING, 2005

[王根 2007] 王根，赵军，基于多重标记 CRF 的句子情感分析研究，全国第九届计算语言学学术会议，清华大学出版社，大连，2007

[王波 2007] 王波， 王厚峰，基于自学习策略的产品特征自动识别，全国第九届计算语言学学术会议，清华大学出版社，大连，2007

[章剑峰 2007] 章剑锋，张奇，黄萱菁，吴立德，中文评论挖掘中的主观性关系抽取，第三届全国信息检索与内容安全学术会议，苏州，2007

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

[Hu & Liu, 2004] Hu, Minqing, & Liu, Bing. 2004a. Mining and summarizing customer reviews. Pages 168-177 of: Proceedings of KDD '04

[Liu et al. 2005] Liu, Bing, Hu, Minqing, & Cheng, Junsheng. 2005. Opinion observer: analyzing and comparing

opinions on the Web. Pages 342-351 of: Proceedings of WWW '05

[Durant et al. 2006] Kathleen T. Durant & Michael D. Smith. 2006. Mining Sentiment Classification from Political Web Logs. Proceedings of WEBKDD '06

[姚天昉 2006] 姚天昉，聂青阳，李建超，一个用于汉语汽车评论的意见挖掘系统中国中文信息学会二十周年学术会议，2006 年，北京

[MPQA] Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, NRRC Summer Workshop on MPQA: Multi-Perspective Question Answering, Final Report, 2002

[Wiebe,2005] Wiebe, Janyce, Wilson, Theresa, and Cardie, Claire. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, 2005

[陈建美 2007] 陈建美，林鸿飞，杨志豪，基于贝叶斯模型的词汇情感消歧，全国第九届计算语言学学术会议，清华大学出版社，大连，2007

[Ounis2006] Iadh Ounis , Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff, Overview of the TREC-2006 Blog Track, Proceedings of TREC2006, Gaithersburg, USA