

观点句的多特征判别

方圆^{1,†} 冯帅² 陈锻生¹ 吴伊萍^{1,3}

1. 国立华侨大学计算机科学与技术学院, 厦门 361021; 2. 东北大学信息科学与工程学院, 沈阳 110001; 3. 泉州师范学院, 数学与计算机科学学院 泉州 362000

† 通信作者, E-mail: siegfang@qq.com

摘要 微博中的观点句包含着人们对事物的态度, 倾向以及其它丰富的信息。然而由于微博本身字数的限制和语言结构的自由, 从微博发现观点句面临着许多困难。通过采用句子中的词的词性、词的情感用类和词语之间的依存关系三类特征, 以句子中是否存在特征作为判别依据, 并使用 AdaBoost 算法将这三类特征的分类器加以整合成多特征分类器, 可以更好地发现观点句。NLP&CC 2012 的中文微博情感分析的评测表明, 此种多特征分类器的 F 值在微平均上达到 71.6%, 在宏平均上达到 71.3%

关键词 微博; 词性; 情感用词; 依存关系; Adaboost;

中图分类号

Distinguish Viewpoints Sentences by Multi-Feature

FANG Yuan^{1,†}, FENG Shuai², CHEN Duansheng¹, WU Yipin^{1,3}

1. School of Computer Science and Technology, National Huaqiao University, Xiamen 361021; 2. School of Information Science and Engineering, Northeastern University, Shenyang 110001; 3. School of Math and Computer Science, Quanzhou Normal University, Quanzhou 362000;

†Corresponding author, E-mail: siegfang@qq.com

Abstract The perspective sentences in microblogs have abundant information, including the opinions and tendencies of people to events. Since the limit of words and free grammar structure in microblogs, there have many problems to overcome. The basis for distinguish viewpoints sentences is the part-of-speech of word, the emotional dictionary and the dependency relations between words in a sentence. By use the existence of these three kinds of features to distinguish, the better classifier is built by AdaBoost algorithm than a single classifier with one feature. This year's NLP&CC evaluation of the analysis of emotion in Chinese microblog indicate that, this multi-feature classifier's F1-Measure can reach 71.6% in micro-average and 71.3% in macro-average.

Key words micorblogs; part-of-speech; emotional dictionary; dependency relation; Adaboost

随着微博的日趋流行, 新浪微博等网站已成为海量信息的发布体。这类多媒体迷你型博客, 允许用户通过 E-mail、web、手机、平板和电脑随时随地地发布文字、声音、图片和视频信息。相比于传统博客, 微博客最大的特点在于其“微”, 即单篇博客长短一般限制在 140 字之内。微博的发布形式多样而且灵活, 新闻的发布不再是报刊、杂志社、电视台和新闻网站的专利。只要用户身处新闻现场, 就可以使用手中的移动设备(如手机和平板)来发布新闻。一旦用户的微博发布, 收听此用户的其它用户也会看新闻信息, 他们可能会评论或者直接转发。这样新闻就以“口口相传”的方式迅速地为用户所知。微博的时效性可见一斑。微博中不包含着新闻, 还可能有着微博用户个人的所思所想, 或者关于某事, 某地和某人的所见所念。这引起了众多数据分析专业人员和自然语言处理学者的关注。

目前针对微博的自然语言处理研究中, 情感分析是一个研究热点和前沿课题。情感分析的目的是从文本中挖掘用户表达的观点以及情感极性。情感倾向的一种表达方式就是观点句。观点句是基于断言或评论并且带有个人情感和意向的抒发。关于观点句的分类可以追溯到意见挖掘中的主客观句子的分类。姚天昉等人^[1]针对主观性文本的特点提出了包括情感形容词在内的预选特征, 并通过实验选择稳定的特征。在这

些特征的基础上，林慧恩等人^[2]加入了HowNet的《情感分析用词语集》，并以两个连续的词性组合模式筛选主观句。这些方法并没有考虑到句子中的句法结构，而且评价对象一般考虑的是领域实体。一个观点的形成所需要的最基本的要素在于评价对象和评价词。评价对象和评价词的匹配对应关系称之为主观关系。樊娜等人^[3]将评价对象和评价词的句法路径信息与它们之间的距离应用到最大熵模型中以提取匹配关系。赵妍妍等人^[4]提出句法结构上的路径能够有效描述评价对象和评价词之间的匹配关系，并通过编辑距离进一步改进识别性能。

以上研究大多是在商品评论的数据基础上进行的。这些评论往往是围绕某种商品展开的。微博可以是围绕某个人、某件事、某件商品甚至某条微博展开。这使得微博在评价对象和评价词与商品评论有较大的差异。而且在微博上进行观点句判别的最大特点在于其字数的限制和语言结构的自由。因为字数的限制，其词语及其词性、依存关系的频率相对于普通文本大大减少，甚至为 0。因为语言结构的自由，句法结构上的分析进行得就相对困难。本文是对微博进行观点句判别的一次尝试。在词语上使用词性将形容词、动词、感叹词等包括进来；在词典上使用了多个词语集以判断词的情感用词类别；在句法结构上通过词与词的某种依存关系观察观点句。我们而不是以权重衡量或者以图结构判断，而是以句子是否包含这些特征作为分类的依据，通过AdaBoost^[5]将句子的多个特征整合成强分类器。此外，本文在特征的统计分析上使用准确率、召回率和F值（F1-Measure）作为统计指标分析特征对观点句分类的有效性。

1 句子的特征

1.1 词性与情感用词

组成句子的基本单位是词语。通常我们会将词按照词在造句中的不同作用分成实词和虚词两大类，其中实词可分为名词、形容词等 7 类，虚词可分为介词、连词、助词、叹词等 5 类。当人们感叹某事时，其抒写的语句往往包含叹词等虚词。而当人们在陈述某一事实时，名词和形容词又会大量出现。词的词性与词类相近，但是一般而言在实际应用中，词性的数量会比词类要多，要细。句子所包含的词的词性在一定程度上可以反映出句子的内容和作者或说话人的情感倾向。此外句子的长度要比段落、章节小得多，一个句子几乎不可能包含所有的词性，这使得以词性作为句子的特征成为可能。

此外，在表达情感和阐述观点时，人们还会使用特定的词语进行表达。比如表达感激之情时，会说“非常感谢”，“感激不尽”等；购物时遇到奢侈品时，会说“贵得要命”；遇到大晴天时，会发微博说“格外晴朗”。据此我们将用作同一情感表达目的的词语集合成一个情感用类，通过在语句中发现词语的情感用类，在一定程度上发现句子的含义。

1.2 依存关系

句子中的成分（如词语）是相互联系的。所谓依存关系是指一词语依存于另一词语的关系。最常见的一类关系比如定中关系，如“天真的小男孩”。这其中，“男孩”与“的”和“小”是定中关系。一般认为，现代依存语法的创立者为法国语言学家Tesnière^[6]。依存语法通过分析句子中语言成分之间的依存关系来揭示句子的句法结构（图 1）。

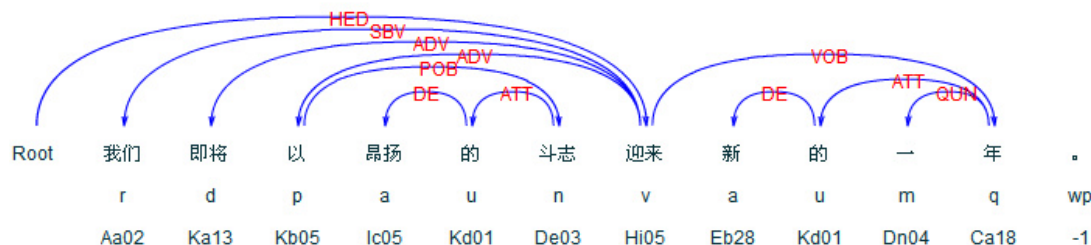


图 1 依存句法示例

Fig.1 an Example of Dependency Relations

如图 1 所示，“迎接”是核心动词，是支配其它成分的中心成分，而其本身却不受其它成分的支配。句子除“迎接”以外的成分都以某种依存关系从属于支配者。在 20 世纪 70 年代，Robinson^[7]提出依存语

法中关于依存关系的四条公理，中国学者针对处理中文信息提出了依存关系的第五条公理^[8]，如下：

- 1) 一个句子中只有一个成分是独立的；
- 2) 其它成分直接依存于某一成分；
- 3) 任何一个成分都不能依存与两个或两个以上的成分；
- 4) 如果 A 成分直接依存于 B 成分，而 C 成分在句中位于 A 和 B 之间，那么 C 或者直接依存于 B，或者直接依存于 A 和 B 之间的某一成分；
- 5) 中心成分左右两面的其它成分相互不发生关系。

句子中词语间的支配与被支配，依存与被依存的现象可以反映出各词语之间的语义修饰关系，它可以获得长距离的搭配信息，并与句子成分的物理位置无关。

2 多特征分类器

2.1 特征分类器

我们使用ICTPOS汉语词性标记集¹中的词性对句子中的词进行标记。ICTPOS汉语词性标记集有 23 种词性。我们为每个词性建立分类器，每个词性分类器用于匹配一种词性。例如名词分类器，当句子中包含名词时，名词分类器即将此句子判定为观点句。依存关系的标记采用Stanford依存关系集^[9]，共 57 种。同样的，我们也为每个依存关系建立分类器，用于匹配一种依存关系。此次实验共使用 6 个情感用词语集²，分别为主张词词语集、正面情感词语集、正面评价词语集、程度级别词语集、负面情感词语集、负面评价词语集。我们将同一个情感用词语集中的词视为同一个情感用类。一个情感用词分类器对应于一个情感用类，当句子中包含情感用词语集所包含的词语时，就将句子视为观点句。

如上所述，可以有 86 个特征分类器对句子中的观点句进行判别。通过机器学习的方式，我们可以将众多特征分类器加以整合，形成更强的分类器。

2.2 AdaBoost 训练

设训练集中的元素为由句子和其标志组成的二元组，设句子的集合为 X ，当句子为观点句时，将句子标记为 1，否则为-1。并设分类器的集合为 H ，则训练集可定义为

$$x_i \in X, |X| = m, \quad y_i \in Y = \{+1, -1\},$$
$$h_j \in H, |H| = n, \quad h_j(x_i) = \begin{cases} +1, & \text{如果 } h_j \text{ 判断 } x_i \text{ 为观点句} \\ -1, & \text{如果 } h_j \text{ 判断 } x_i \text{ 不为观点句} \end{cases}$$
$$(x_i, y_i) \in TrainSet, i = 1, \dots, m$$

在训练前，设每个句子的权重相等，且所有句子的权重相加为 1，即

$$D_1(i) = \frac{1}{m}, i = 1, \dots, m, \quad \text{其中，D 的下标 1 表示 AdaBoost 的第一次迭代。}$$

对于AdaBoost的第t次迭代，我们首先需要计算每个分类器 h_j 对样本 x_i 进行判别产生的误差，即

$$\varepsilon_j = \sum_{i=1}^m D_i(i) I(y_i \neq h_j(x_i)), \quad \text{其中 I 是指示函数, } I(y_i \neq h_j(x_i)) = \begin{cases} 1 & y_i \neq h_j(x_i) \\ 0 & y_i = h_j(x_i) \end{cases}$$

然后我们取误差与 0.5 的差的绝对值最大的那个分类器，记为 h_t ，并计算分类器 h_t 的权重 α_t

$$h_t = \arg \max_{h_t \in H} |0.5 - \varepsilon_t|, \quad \alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$$

接着我们需要判断 $|0.5 - \varepsilon_t| < \beta$ 是否成立以确定是否需要继续进行迭代。实验中我们将 β 预置为 0.05。

¹ ICTPOS 汉语词性标记集 http://ictclas.org/ictclas_download.aspx

² 知网情感分析用词语集 (beta 版): http://www.keenage.com/html/c_index.html

如果需要继续进行第t+1次迭代，训练集中样本的权重需要改变以适应已得到的分类器 h_t 。

$$D_{t+1}(i) = Z_t / \sum_{i=1}^m Z_t, Z_t = D_t(i) \exp(-\alpha_t y_i h_t(x_i))$$

迭代结束时即可得到强分类器 $H(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$ (其中 T 为迭代次数, sign 为符号函数)。当

$H(x)$ 的结果为+1时,所判别的句子即为观点句,否则为非观点句。

3 实验

本文先分别构造了基于词性、情感用词和依存关系的单特征弱分类器,通过剔除一些几乎无法判别观点句与非观点句的弱分类器,进而使用AdaBoost算法构造了更强的多特征分类器。NLP&CC2012的中文微博情感分析评测的评测结果¹表明,这种分类器在微平均上的F值达到了71.6%,在宏平均上的F值达到了71.3%。微平均以整个数据集为一个评价单元,计算整体的评价指标;宏平均以每个话题为一个评价单元,计算参评系统在该话题中的评价指标,最后计算所有话题上各指标的平均值。

实验在Java语言环境下使用中科院计算所的ICTCLAS²进行词性标记,并使用斯坦福大学的句法分析器stanford parser³进行依存句法分析,采用NLP&CC2012的中文微博情感分析评测的样例数据⁴进行统计和训练,通过AdaBoost将弱分类器联合为更强的分类器。训练数据中共有464个句子,其中观点句有271句,所占比例约为58.4%。

3.1 观点句上的特征统计

为了了解观点句上各种特征的分布情况,我们有必要对这些特征所能为观点句判别带来的潜力做一个估计,我们是通过计算使用每个特征进行判别得到的准确率、召回率和F值进行估计的。

$$\text{准确率} = \frac{\text{系统检索到的相关文件数}}{\text{系统返回的文件总数}}, \quad \text{召回率} = \frac{\text{系统检索到的相关文件数}}{\text{相关文件总数}}$$

$$F\text{值} = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}}$$

我们通过词性分类器、情感用词分类器和依存关系分类器计算词性、情感用词或依存关系的分类效果(准确率、召回率和F值)。部分结果如下所示:

表 1 词性分类器的分类效果
Table 1 Effect of POS Classifiers

词性	准确率(%)	召回率(%)	F1-Measure(%)
名词	65.06	84.50	73.52
人名	71.64	17.71	28.40
时间词	46.03	10.70	17.37
动词	58.16	84.13	68.78
名动词	57.78	19.19	28.81
形容词	73.66	5.57	10.36
名形词	81.48	8.12	14.76

¹ 中文情感分析&词汇语义关系抽取评测评测结果: http://tcci.ccf.org.cn/conference/2012/pages/page04_evars.html

² ICTCLAS 中文汉语分词系统: <http://ictclas.org/>

³ The Stanford Parser: A statistical parser: <http://nlp.stanford.edu/software/lex-parser.shtml>

⁴ 中文微博情感分析&词汇语义关系抽取评测通知: http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html

代词	71.55	64.95	68.09
数词	60.83	26.94	37.34
量词	74.16	24.35	36.67
副词	63.67	67.90	65.71
介词	55.29	25.09	34.52

表 2 情感用词分类器的分类效果
Table 2 Effect of Dictionary Classifiers

情感用词	准确率(%)	召回率(%)	F1-Measure(%)
主张词语	89.47	6.27	11.72
正面情感词语	46.97	11.44	18.40
正面评价词语	62.86	24.35	35.11
程度级别词语	84.29	21.77	34.60
负面情感词语	58.33	5.17	9.49
负面评价词语	92.31	31.00	46.41

表 3 依存关系分类器的分类效果
Table 3 Effect of Dependency Relations Classifiers

依存关系	准确率(%)	召回率(%)	F1-Measure(%)
root	60.00	94.10	73.28
dep	58.43	57.57	57.99
nsubj	64.02	62.36	63.18
nsubjpass	64.64	62.73	63.67
advmod	62.59	62.36	62.48
npadvmod	62.59	62.36	62.48
sdep	58.74	58.30	58.52

从表中我们可以观察到，有些分类器擅于判别一些特殊的观点句，如主张词语分类器，其召回率虽然只有 6.27%，但其准确率却达到了 89.47%。有些分类器可能就无法分辨观点句和非观点句，如依存关系 root 分类器，其召回率达 94%，但准确率却只有 60%。如前所述，观点句占所有句子的比例为 58.4%，，如果有一种分类器将所有的句子都判断为观点句的话，那么这种分类器的召回率就是 100%，准确率就只有 58.4%，所以依存关系 root 可能是一个所有句子几乎都有的“大众”特征。正是因为观点句上有如此之多参差不齐的特征，我们需要剔除了一些“大众”特征，通过 AdaBoost 将三种特征加以整合，组合成强分类器。

3.2 多特征分类器的组成

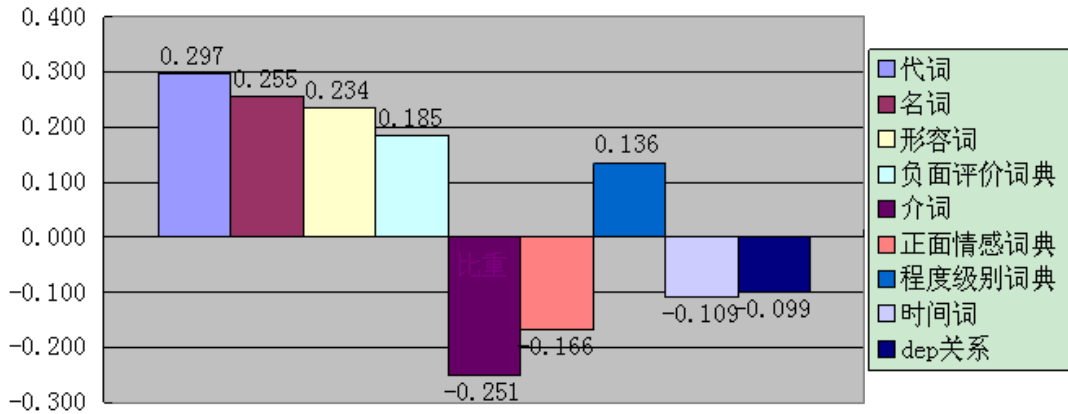


图 2 AdaBoost 强分类器中各分类器的比重

Fig.2 The Weight of Classifiers in Strong Classifier Produced by AdaBoost

如图所示即实验所得的多特征分类器中各特征分类器所占的比重，从左至右即为 AdaBoost 迭代选择的次序，其中依存关系 dep 权重最小，代词的权重最大，词性分类器占主体地位，但单单用词性分类器进行 AdaBoost 迭代产生的强分类器进行观点句判别，分类效果就不够好。如下图所示：

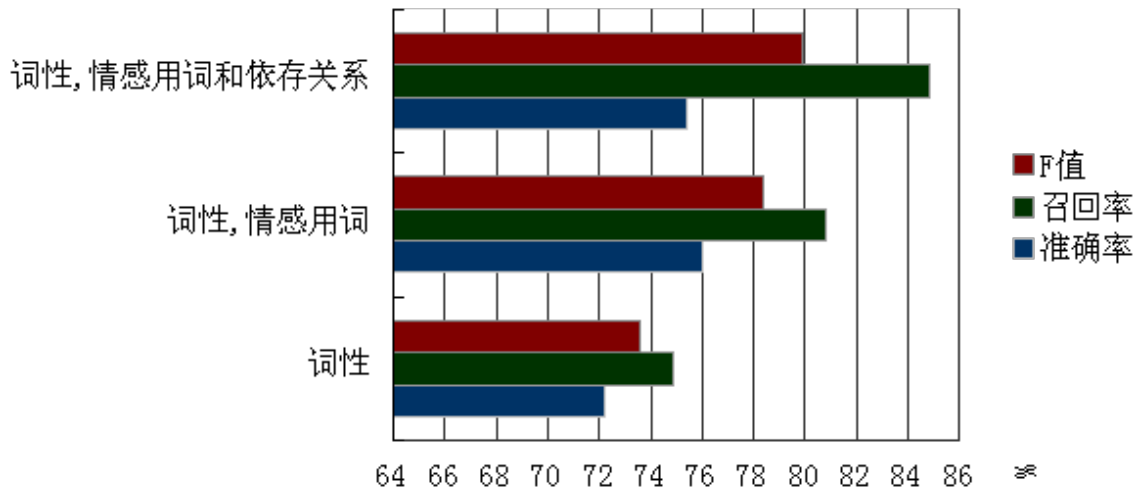


图 3 依次增加分类器种类得到的 AdaBoost 分类器的分类效果

Fig.3 The Effect of Classifiers Produced by Adding a Type of Classifier Each Time

从图中可以看出，情感用词分类器的加入起到了重要的辅助作用，将 F 值在 73.5% 的基础上提高了近 5%，随后依存关系的加入又将 F 值提高了 1%。可见词语的情感用类对于观点句的判别也是相当重要的。情感用类即为词语在用情感表达时所作的成分，包括程度用词、正负情感用词、主张用词、正负评价用词等。当我们表达对某事、某人、某商品的观点时，或多或少地都会带上自己的情感或者态度，而这些又是词性的短处。词语的情感用类在一定程度上补上这一短板。结合依存关系分类器的统计结果，可以看出依存关系对观点句分辨程度并不好，这集中体现在它们的准确率较低上。

4 结论

由于微博在评价对象和评价词上的多样性，本文使用词性，情感用词以及依存关系这些抽象特征来包容具体的词语及其之间的位置关系，并使用 AdaBoost 对特征分类器进行整合。针对微博字数少、语言结构自由的特点，特征均是以是否存在表征，而不以权重或者图结构体现。本文先统计了句子中单特征分类的准确率、召回率和 F 值，去掉“大众”特征，然后通过将所有单特征进行 AdaBoost 迭代训练得到多特征分类器，并通过这种多特征分类器进行观点句的判别。在多特征分类器中，词性和情感用词分类器是相互交叉，相辅相成的。然而依存关系在判别时只有较小的贡献，这一方面可能是使用 AdaBoost 进行训练

的原因,另一方面可能是因为依存关系的使用方法不对。我们下一步的研究重点在于考察依存关系与情感之间的关联,进一步探索词语的依存关系表达情感的方式。

参考文献

- [1] 姚天防,彭思崴.汉语主客观文本分类方法的研究[C].//第三届全国信息检索与内容安全学术会议论文集.2007:117-123.
- [2] 林慧恩,林世平.中文情感倾向分析中主观句子抽取方法的研究[C].//全国第 20 届计算机技术与应用(CACIS)学术会议论文集.2009:379-384.
- [3] 樊娜,蔡皖东,赵煜等.基于最大熵模型的观点句主观关系提取[J].计算机工程,2010,36(2):4-6.
- [4] 赵妍妍,秦兵,车万翔等.基于句法路径的情感评价单元识别[J].软件学报,2011,22(5):887-898.DOI:10.3724/SP.J.1001.2011.03767.
- [5] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119—139.
- [6] Tesnière L, Eléments de la Syntaxe Structurale. Paris, France: Klincksieck, 1959
- [7] Robinson J J, Dependency structures and transformational rules. Language, 1970, 46:259-285.
- [8] 郭艳华,周昌乐.一种汉语语句依存关系网协同生成方法研究[J].杭州电子工业学院学报. 2000, 20(4):24-32
- [9] Marneffe M D, Manning C D. Stanford typed dependencies manual[EB/OL], 2008(2011-09).
http://nlp.stanford.edu/software/dependencies_manual.pdf