计算机研究与发展
Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP
44(9): 1469~1475, 2007

基于语言建模的文本情感分类研究

胡 熠[1] 陆汝占[1] 李学宁[1,2] 段建勇[1] 陈玉泉[1]
[1](上海交通大学计算机科学与工程系 上海 200240)
[2](江南大学外语研究学院 无锡 214122)
(huyi@cs.sjtu.edu.cn)

Research on Language Modeling Based Sentiment Classification of Text

Hu Yi[1], Lu Ruzhan[1], Li Xuening[1,2], Duan Jianyong[1], and ChenYuquan[1]
[1](*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240)
[2](*School of Foreign Languages Study, Southern Yangtze University, Wuxi* 214122)

Abstract Presented in this paper is a language modeling approach to the sentiment classification of text. It provides the semantic information beyond topic in text summary when characterizing the semantic orientation of texts as "thumb up" or "thumb down". The motivation is simple: "thumb up" and "thumb down" language models are likely to be substantially different: they prefer to different language habits. This divergence is exploited in the language models to effectively classify test documents. Therefore, the method can be deployed in two stages: firstly, the two sentiment language models are estimated from training data; secondly, tests are done through comparing the Kullback-Leibler divergence between the language model estimated from test document and those two trained sentiment models. The unigrams and bigrams of words are employed as the model parameters, and correspondingly maximum likelihood estimation and smoothing techniques are used to estimate these parameters. Compared with two different classifiers, i.e. SVMs and Naïve Bayes, on movie review corpus when training data is limited, the language modeling approach performs better than SVMs and Naïve Bayes classifier, and on the other hand it shows its robustness in sentiment classification. Future works may focus on finding a good way to estimate better language models, especially the higher order n-gram models and more powerful smoothing methods.

Key words sentiment classification; language modeling; KL divergence; supervised learning; robustness

摘 要 提出了一种基于语言建模的文本情感分类的方法.将文本的情感倾向标记为"赞扬"或"批评",可以为文本提供主题之外的语义信息.为此提出了从训练数据中分别估计出代表"赞扬"和"批评"两种情感倾向的语言模型,然后通过比较测试文本自身的语言模型和这两种训练好的情感模型之间的Kullback-Leibler距离,分类测试文本的思路.各个模型的参数分别选用词形特征的unigram和bigram,而相应的参数估计也分别尝试了最大似然和平滑两种策略.当在电影评论语料上和代表不同分类模型的支持向量机及朴素贝叶斯分类器进行比较时,语言建模的方法表现出了较好的分类性能和鲁棒性.

关键词 情感分类;语言建模;KL距离;监督学习;鲁棒性

中图法分类号 TP18

传统上的文本分类往往关注于把文本映射到给定的主题,如体育、经济、政治等.然而,近年来对文本非主题分析的兴趣不断增加,其中就包括文本的情感分类.我们把文本所持的对自身主题"赞扬"

收稿日期:2006-09-20;**修回日期**:2007-06-11
基金项目:国家自然科学基金重大项目(60496326)

(thumb up)或"批评"(thumb down)的倾向性分析称为情感分类. 如一个主题是"影评"类的文本,对所评述的电影可能表示赞许也可能表示批评,这就是它的情感倾向. 给文本标注情感倾向可以为许多需要用到情感分析的智能系统提供帮助,如电影评论的信息检索等. 情感分类作为当前热门的研究方向已在不同的领域如电影评论、用户反馈等有所尝试[1-3].

一些有关情感分类的研究工作考虑的对象是词或短语,关心它们表达情感时的计算方法[4-5],其情感倾向被量化为一个实数值测度[5]. 在 Turney 等人的工作中,分别计算这些词或短语与"excellent"和"poor"之间的"逐点互信息(pointwise mutual information)"来区分它们的倾向[4]. 单个词或短语的情感倾向可以被用来进一步判断整个句子或篇章的情感倾向. 这种以特定词为基准来计算其他词情感倾向的思路也影响了国内的一些研究,如文献[6]中就是计算一般词语与知网(Hownet)中已标注褒贬的词语之间的相似度,然后选择倾向性明显的词语作为支持向量机的输入特征,分析文本的情感倾向,同时还采用语义否定规则提高分类效果.

还有一些研究使用机器学习方法把整个文本区分为"赞扬"和"批评". 文献[1]对一些机器学习算法做了对比,得出的结论是支持向量机(SVMs)相比其他的分类算法而言能获得较好的结果. 这些工作也表明,情感分类任务上的正确率通常要低于已知的主题分类的正确率. 其原因在于情感分类更多地涉及了人的心理世界,而用自然语言表达的情感往往微妙复杂,需要语言知识才能处理. 尽管以模式分类为出发点的机器学习算法取得了不错的效果,但还有进一步挖掘潜力的可能. 本文正是从模式分类的角度探讨了全文的情感分类问题.

本文提出了一种语言建模的方法,其重点在于从训练集中估计出表示"赞扬"和"批评"的语言模型,用于表示人在表达不同情感时的语言结构. 对一个同样用语言模型表示的测试文本,通过计算它和这两个情感模型的距离来评价它的情感倾向. 为了检验这个想法,我们把这种语言建模的方法和代表两种不同分类模型的典型分类器、支持向量机(区分模型)[7]和贝叶斯分类器(生成模型)[8]作了对比. 从本质上来讲我们的方法也是一种生成式的分类模型.

1 面向情感分类的语言建模的方法

本节提出了一种语言建模的方法来探测文本的情感倾向. 其主要假设就是"赞扬"对应的语言模型和"批评"对应的语言模型有可能是不一样的,因为它们可能倾向于不同的语言习惯,这样就可以通过探寻语言模型之间的差异把同样基于语言模型表示的测试文本区分为"赞扬"和"批评". 因此,语言建模的方法首先从训练数据中估计出这两种情感倾向的语言模型,然后用一个距离函数比较测试文本的语言模型和这两个情感语言模型之间的距离."赞扬"情感倾向的语言模型用 θ_P 表示,θ_P 是正例文本集中 *n-gram* 的概率分布. 相应的,θ_N 代表"批评"倾向的语言模型. 一个测试文本也生成一个语言模型 θ_d. 语言模型看做是语言单元的概率分布,表明观察到这些语言单元的可能性. 因此一个文本的语言模型可以和"赞扬"或"批评"的语言模型用某个分布上的距离来比较. 我们定义分类函数 φ 为

$$\varphi(d;\theta_P,\theta_N)=Dis(\theta_d,\theta_P)-Dis(\theta_d,\theta_N):\begin{cases}<0 & \text{"赞扬"}\\ >0 & \text{"批评"}\end{cases}, \tag{1}$$

这里 $Dis(p,q)$是两个分布 p 和 q 之间的距离. 式(1)表明这样一种分类思想:如果 $Dis(\theta_d,\theta_P)<Dis(\theta_d,\theta_N)$,测试文本 d 更接近"赞扬"的情感倾向;如果 $Dis(\theta_d,\theta_P)>Dis(\theta_d,\theta_N)$,$d$ 更接近于"批评"的情感倾向;如果 $\varphi(d;\theta_P,\theta_N)=0$,认为测试文本是"中立的",但本文不讨论这种情况. 我们利用 Kullback-Leibler Divergence(相对熵)作为语言模型之间的距离测度.

1.1 情感分类中的 Kullback-Leibler Divergence

两个概率分布 $p(x)$和 $q(x)$之间的 Kullback-Leibler Divergence,$D(p||q)$定义为

$$D(p\|q)=\sum_x Pr(x)\log\left(\frac{p(x)}{q(x)}\right), \tag{2}$$

很容易看到 $D(p\|q)$总是非负的,如果为 0,当且仅当 $p=q$. 尽管 KL-Divergence 不满足对称性和三角不等式,也就是说它并不是一个分布之间真正意义上的距离,但是把 KL-Divergence 当做"距离"来看待仍然很有用[9]. 考虑到我们的假设,我们设想通过 d 到 θ_P 和 θ_N 的距离来分类. θ_d 和 $\theta_P(\theta_N)$之间的 KL-Divergence 可以用(3)计算:

$$\begin{cases}D(\hat{\theta}_d\|\hat{\theta}_P)=\sum_{n\text{-}gram}Pr(n\text{-}gram\mid\hat{\theta}_d)\times\log\left(\frac{Pr(n\text{-}gram\mid\hat{\theta}_d)}{Pr(n\text{-}gram\mid\hat{\theta}_P)}\right),\\ D(\hat{\theta}_d\|\hat{\theta}_N)=\sum_{n\text{-}gram}Pr(n\text{-}gram\mid\hat{\theta}_d)\times\log\left(\frac{Pr(n\text{-}gram\mid\hat{\theta}_d)}{Pr(n\text{-}gram\mid\hat{\theta}_N)}\right),\end{cases} \tag{3}$$

$\hat{\theta}$代表真实模型θ的估计模型,$Pr(n\text{-}gram \mid \hat{\theta})$是给定估计模型$\hat{\theta}$时$n\text{-}gram$的概率. 把式(3)带入式(1)中,我们得到了一个情感分类函数:

$$\varphi(d;\hat{\theta}_P,\hat{\theta}_N) = D(\hat{\theta}_d \parallel \hat{\theta}_P) - D(\hat{\theta}_d \parallel \hat{\theta}_N) = \sum_{n\text{-}gram} Pr(n\text{-}gram \mid \hat{\theta}_d)\log\left\{\frac{Pr(n\text{-}gram \mid \hat{\theta}_N)}{Pr(n\text{-}gram \mid \hat{\theta}_P)}\right\}. \tag{4}$$

式(4)用$n\text{-}gram$表示一般意义下分类函数. 事实上,本文仅采用词语的 unigram 和 bigram 作为模型参数. 这是因为对更高阶的$n\text{-}gram$而言,现有的语料存在着严重的数据稀疏问题,所以我们没有讨论$n \geqslant 3$的情况. 但从理论上来讲,越高阶的语言模型越容易逼近语言现象的真实情况. 虽然我们只讨论了 unigram 和 bigram 的模型,但是可以用同样的范式(4)来操作较高阶的$n\text{-}gram$模型.

1.2 模型参数估计

为了调查语言建模方法的能力,我们使用两种方法估计 unigram 和 bigram 的分布:最大似然估计(MLE);对 unigram 和 bigram 语言模型的平滑估计.需要说明的是,我们的工作采用词形的 unigram 和 bigram 作为模型参数.

1) unigram 和 bigram 的最大似然估计

MLE 在模型估计中应用很广泛,所以我们不做更多的解释而是直接给出计算式(5),并简单说明.

$$\begin{cases} Pr_{\mathrm{MLE}}(w_i \mid s) = \dfrac{\#(w_i \text{ in } s)}{\#(* \text{ in } s)}, \\ s \in \{d,P,N\}, \text{ for unigram}, \\ Pr_{\mathrm{MLE}}(w_i \mid w_{i-1}, s) = \dfrac{\#(w_{i-1}w_i \text{ in } s)}{\#(w_{i-1} * \text{ in } s)}, \\ s \in \{d,P,N\}, \text{ for bigram}. \end{cases} \tag{5}$$

在式(5)中我们需要说明的是"s"它可以表示一个测试文本(d),"赞扬"文本集(P)或者"批评"文本集(N). $\#(n\text{-}gram)$指$n\text{-}gram$在对应文本集(用d,P或N表示)中出现的次数."$*$"代表任意一个词. 在本文的其余部分这些符号的含义一致.

当训练数据的规模相对于要估计的参数规模而言较小时,最大似然估计不是一个很好的选择:给一个没有观察到的$n\text{-}gram$直接分配一个 0 概率显然不够准确. 平滑描述的就是一种调整概率分配以期待获得更精确的模型的技术.

2) 对 unigram 的 Dirichlet Prior 平滑

Dirichlet Prior 平滑[10]是针对 0 概率问题常用的平滑方法,而且适用于 unigram 的平滑. 它是一种线性插值的方法,用于解决训练文本集相对较小时参数估计中的偏置问题:从可观察的$n\text{-}gram$中分配适当的折扣(非 0 的概率)给文本集中没有观察到的$n\text{-}gram$. 就我们用到的 unigram 语言模型,这种平滑估计定义为

$$Pr_{\mathrm{DP}}(w \mid s) = \begin{cases} Pr_\gamma(w \mid s), & \text{if word } w \text{ is seen}, \\ \alpha_s Pr_{\mathrm{MLE}}(w \mid C), & \text{otherwise}, \end{cases} \tag{6}$$

这里$Pr_\gamma(w \mid "s")$是用"s"指示的文本(集)中可观察的词w的平滑概率. $Pr_{\mathrm{MLE}}(w \mid C)$是在整个语料库$C$中$w$出现概率的最大似然估计. α_s是给没出现的词语分配概率值时的控制系数,使得所有的概率值之和等于 1. 一般来说,α_s依赖于所有的$Pr_\gamma(w \mid "s")$. 在我们的工作中探索了下面的平滑形式:

$$Pr_\gamma(w \mid s) = \frac{\#(w \text{ in } s) + \mu Pr_{\mathrm{MLE}}(w \mid C)}{\#(* \text{ in } s) + \mu}, \text{ to } s, \tag{7}$$

并且,

$$\alpha_s = \frac{\mu}{\mu + \mid C \mid}. \tag{8}$$

Dirichlet Prior 在很多 NLP 任务中很有效,但在本文用到的电影评论语料的情感分类上,它相对于简单的 MLE 并没有给出很大的提高(见实验部分).

3) bigram 的 Kenser-Ney 平滑

Kneser 和 Ney[11]介绍了一种基于绝对折扣思想的平滑想法,其中$n\text{-}gram$的低阶分布和高阶分布以一种新的方式组合起来. 基于 Kneser-Ney 的想法,Chen 和 Goodman[12]给出了他们的算法,通过选择低阶分布使得高阶分布的边际和训练语料的边际一致. 改进后的 Kneser-Ney 平滑在不同条件下相比于其他平滑技术而言都获得了最好的性能. 对于 bigram 模型,Chen 等人选择了一个平滑分布Pr_{KN},它满足下面对所有词w_i边缘分布的约束:

$$\sum_{w_{i-1}} Pr_{\mathrm{KN}}(w_{i-1}w_i) = \frac{\#(w_i)}{\sum_{w_i}\#(w_i)}. \tag{9}$$

可见,式(9)的左边是 bigram 平滑分布Pr_{KN}中w_i的 unigram 边缘概率,而右边则是训练数据中w_i的最大似然估计. 对我们的 bigram 平滑而言,我们定义 bigram 模型是等式(10)给出的形式:

$$Pr_{\mathrm{KN}}(w_i \mid w_{i-1}, s) = \frac{\max\{\#(w_{i-1}w_i) - D, 0\}}{\sum_{w_i}\#(w_{i-1}w_i)} + \frac{D}{\sum_{w_i}\#(w_{i-1}w_i)} N_{1+}(w_{i-1}\cdot) Pr_{\mathrm{KN}}(w_i), \text{ to } s, \tag{10}$$

D 是所有观察到的 bigram 上一个固定的概率折扣，Ney 等人建议 $D=\frac{n_1}{n_1+2n_2}$. 而且：

$$Pr_{KN}(w_i)=\frac{N_{1+}(\cdot\ w_i)}{N_{1+}(\cdot\ w_{i-1}\ \cdot)}, \quad (11)$$

符号 D, N_{1+}, n_1 和 n_2 的具体意义读者可以参考 Kneser 和 Chen 等人的论文[11-12].

2 实验结果

2.1 文本集和评测方法

Turney[3,13]发现电影评论是情感分类任务的若干领域中最难的一个. 他在 120 个文本集上处理电影评论的正确率是 65.83%(随机选择的正确率是 50%)，这是我们选择电影评论作为研究领域的原因. 我们的数据源是 Internet Movie Database (IMDB)文本集. 在本文我们对所有的 2000 篇文本预先作了词形还原和停用词的去除.

我们简单地采用了基于 3 倍交叉检验的平均正确率来评测算法的性能. 就分类结果而言，存在两种错误，包括把一个实际上是“赞扬”的文档归类成了“批评”，或者正好相反. 正确分类也有两种情况，即把一个实际上是“赞扬(批评)”的文档的确归类到了“赞扬(批评)”. 那么正确率可以定义如下：

$$Accuracy=\frac{A+D}{(A+D)+(B+C)}, \quad (12)$$

其中 A 和 D 代表了正确分类的两种情况，而 B 和 C 则代表了错误分类的两种情况. 平均正确率是基于 3 倍交叉验证的平均意义上的正确率.

2.2 实 验

我们设计了 3 个实验来研究我们的方法以及和支持向量机、贝叶斯分类器的比较. 第 1 个实验从 linear, polynomial, RBF 和 sigmoid 核中挑选一个适合情感分类任务的核函数. 第 2 个实验是比较了我们的方法和 SVMs、贝叶斯分类器的性能. 最后一个实验通过不断增加训练数据规模的方式调查基于语言模型方法的鲁棒性. 为保证实验结果不会因为不一致的特征选择而有所偏颇，这 3 个实验都选择了 2000 文本集上至少出现两次以上的 unigram 和 bigram 作为特征集，特征值是出现的频率.

就前两个实验而言，我们把这 2000 个电影评论分成 1200 个训练样本(600 个“赞扬”和 600 个“批评”)和 800 个测试样本(400 个“赞扬”400 个“批评”). 在最后一个鲁棒性实验中，我们使用的训练数据规模从 120 个文本(60/60)到 1200(600/600)，测试样本集是固定的 800 个文本(400/400)，形成了 10 个点的曲线，其中每个点都是 3 次实验的平均正确率，每次实验用到的是规模一样但是不同的训练数据.

1) SVMs 实验

SVMs 的实验对比了用 linear, polynomial, RBF 和 sigmoid 作为核函数的分类性能. 我们使用 Joachim 的 SVM^{light}工具包[14]来训练和测试，对不同核函数的其他参数都设置为系统默认值. 这个实验的目的是要看哪个核函数更适合情感分类任务.

表 1 给出了在 IDMB 语料上不同核函数的结果. 每个结果都是在不同特征集上基于 3 倍交叉验证得到的平均正确率.

Table 1 Comparison of Four Kernel Functions on the IDMB Corpus

表 1 IDMB 语料上 4 种不同核函数在分类性能上的比较

Features	# of Features	Linear	Polynomial	Radial Basis Function	Sigmoid
unigram	13693	78.21	59.59	50.09	49.25
bigram	18602	73.42	51.46	51.19	62.00

在两个特征集上最好的测试结果是线性函数得到的，所以我们将基于语言模型的方法和使用线性核的 SVMs 进行了比较.

2) 语言建模方法实验

在这个实验中我们评测了语言建模的方法，表 2 给出了在 IMDB 集上的实验结果：

Table 2 Comparison Between Language Modeling Approach and SVMs, NB on Word Unigram and Bigram

表 2 语言建模方法和 SVMs、贝叶斯分类器在词形 unigram 和 bigram 特征集上的比较

Features	# of Features	LM-MLE(%) (Uni-MLE/Bi-MLE)	LM-Smoothing(%) (Uni-DP/Bi-KN)	SVM^{light}(%)	Naïve Bayes(%)
unigram	13693	82.02	84.13	78.21	71.31
bigram	18602	61.62	73.80	73.42	68.76

表 2 中 Uni-DP(unigram 的平滑模型)在词形的 unigram 特征集上的性能是所有结果中最优的(84.13%)，相比 SVMs 最好的结果(78.21%)性能平均提高了 7.57%；相比朴素贝叶斯的最好结果(71.31%)性能平均提高了 17.98%. 简单的 Uni-MLE 也获得了相比 SVM 和 Naïve Bayes 较好的结

果.另一方面,在 bigram 特征集上语言建模的方法 Bi-KN 比 Naïve Bayes 的性能好很多,和 SVMs 相当,但是 Bi-MLE 的性能不佳.

就平滑技术而言:① Dirichlet Prior 在设置参数 $\mu=450$ 的基础上对 unigram 模型进行了平滑.平滑对 unigram 模型的性能有一些提高,而且在所有结果中是最好的.尽管基于最大似然估计的语言模型本质上无法很好地逼近真实模型,但是平滑是否一定能提高性能还不清楚,因为这里的提高很有限,只有 2.57%.原因可能是在目前语料规模不够大时,对 unigram 模型的平滑反而会带来起噪声作用的特征.这种现象让我们考虑能否把概率更多地分配到模型的某些关键参数(敏感概念)上来提高情感分类的性能.② Kneser-Ney 平滑对估计一个较好的 bigram 语言模型作了极大的贡献,相对 MLE 提高 19.77%. Kneser-Ney 平滑模型后获得了和 SVMs 可比的性能.原因也许可以阐述为:相对高阶的 bigram 模型理论上讲更容易逼近真实的语言事实,但由于实际用于参数估计的训练数据相比低阶的 uingram 模型稀疏得多,所以基于 MLE 的 bigram 模型不容易发挥作用,而一个强有力的平滑机制就可以产生明显的改善.

3) 鲁棒性实验

为了较全面地考察语言建模方法的性能,我们还通过逐步改变训练数据的方式,观察语言建模的鲁棒性(robustness),即训练数据不断增加时的表现.由于朴素贝叶斯的性能不是很好,再加上我们想和典型的区分模型作比较,所以这个实验主要和 SVMs 作了对比.

我们使用不同规模的训练数据来研究训练规模对语言建模方法和 SVMs 的影响.图 1 和图 2 表示了增长的训练集和固定的测试集(见前面的实验说明).

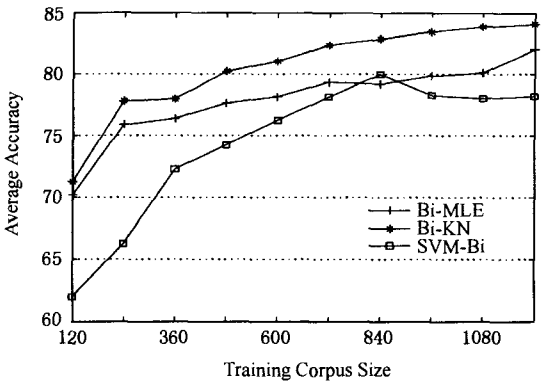

Fig. 1 Uni-MLE, Uni-DP and SVM-Uni on unigram.

图 1 在 unigram 上的 Uni-MLE, Uni-DP 和 SVM-Uni.

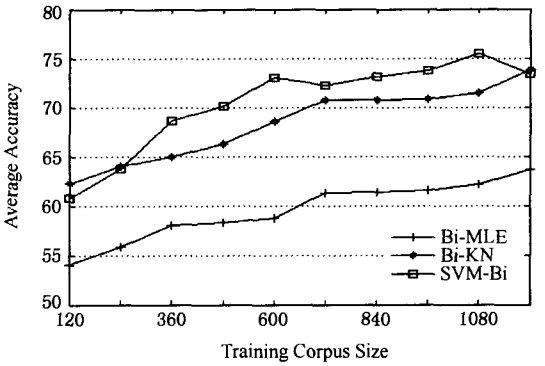

Fig. 2 Bi-MLE, Bi-KN and SVM-Bi on bigram.

图 2 在 bigram 上的 Bi-MLE, Bi-KN 和 SVM-Bi

图 1 中的结果表明,Uni-MLE 和 Uni-DP 都在小规模的训练样本上获得了较高的平均正确率.我们的方法能比 SVMs 挖掘更多的分类知识,即便是在训练语料规模很小时.而且在图 2 中,平滑后的 bigram 模型在小规模(120,240)的训练数据上也比 SVMs 性能好.

图 1 和图 2 还表现出了另一种现象:语言建模的方法随着训练语料的增加获得了越来越好的性能.尽管 SVMs 也体现了这种趋势,但是它的性能并不是随着训练数据增加而稳步增长,而是出现了振荡.在图 1 中,SVM-Uni 最好的结果(80.0%)出现在训练数据是 840 个训练样本时.在图 2 中,SVM-Bi 最好的结果(75.5%)出现在训练数据是 1080 个样本时.这两种现象都说明语言建模的方法鲁棒性更强一些.

3 结 论

在本文中我们提出了一个新的语言建模的方法来解决情感分类问题,从构成正例和反例的文本集中估计出两个语言模型,分别表示"赞扬"和"批评"情感倾向.当分类一个测试文本时,比较它的语言模型到这两个情感模型的距离来决定它的倾向.

从实验结果可以得出如下结论:当训练数据有限时语言建模的方法能得到比 SVMs、贝叶斯分类器更好的性能.另一方面,平滑技术对建立 bigram 模型的贡献很大,使得它获得和 SVMs 在相同特征集上可比较的结果,也高于贝叶斯分类器的正确率.主要和 SVMs 相比,语言建模的方法在两个方面做得更好:相对较好的分类结果和鲁棒性.这些似乎能表明语言建模的方法对情感分类问题而言是一种有希望的方法.需要说明的是,和同属于生成模型

的贝叶斯分类器相比,由于不必引入过多的假设,这种基于分布距离的想法或许更直接有效.

情感分类的困难也是显而易见的:"批评"的评论中可能包含明显的"赞扬"的 *n-gram*,即使这个评论有着强烈的"批评"论调;相反的情况也常出现.所有模式分类器都会面临这个困难.对语言建模方法而言,未来的研究目标也很明确,就是尽可能地估计较好的语言模型,特别是容易逼近真实语言现象的高阶语言模型,同时可以适当引入一些 *n-gram* 之间的语义连接来描述概念的关联,以充实语言模型在语义上的表述能力.

参 考 文 献

[1] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques [C]. EMNLP'02, Philadelphia, USA, 2002

[2] Michael Gamon. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis [C]. The 20th Int'l Conf on Computational Linguistics, Geneva, Switzerland, 2004

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

[4] Peter D Turney, Michael L Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus [R]. National Research Council Canada, Tech Rep: EGB-1094, 2002

[5] Vasileios Hatzivassiloglou, Kathleen McKeown. Predicting the semantic orientation of adjectives [C]. The 35th ACL/8th EACL, Madrid, Spain, 1997

[6] Xu Linhong, Lin Hongfei, Yang Zhihao. Text orientation identification based on semantic comprehension [J]. Journal of Chinese Information Processing, 2007, 21(1): 96-100 (in Chinese)
(徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100)

[7] C Burges. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167

[8] David D Lewis. Naive Bayes at forty: The independence assumption in information retrieval [C]. The 10th European Conf on Machine Learning, Chemnitz, Germany, 1998

[9] T M Cover, J A Thomas. Elements of Information Theory [M]. New York: Wiley, 1991

[10] C Zhai, J Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval [C]. SIGIR'2001, New Orleans, USA, 2001

[11] R Kneser, H Ney. Improved backing-off for m-gram language modeling [C]. The IEEE Int'l Conf on Acoustics, Speech and Signal Processing, Detroit, MI, USA, 1995

[12] S F Chen, J T Goodman. An empirical study of smoothing techniques for language modeling [R]. Harvard University, Tech Rep: TR-10-98, 1998

[13] Peter D Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]. ACL'02, Philadelphia, Pennsylvania, USA, 2002

[14] Thorsten Joachims. Making large-scale SVM learning practical [G]. In: Bernhard Scholkopf, Alexander Smola, eds. Advances in Kernel Methods—Support Vector Learning. Cambridge, MA: The MIT Press, 1999. 44-56

Hu Yi, born in 1978. Ph. D. candidate in Shanghai Jiao Tong University. His main research interests include natural language processing, intelligent information retrieval model and machine learning, *etc*.

胡 熠,1978 年生,博士研究生,主要研究方向为自然语言处理、智能信息检索模型和机器学习等.

Lu Ruzhan, born in 1940. Professor and Ph. D. supervisor in Shanghai Jiaotong University. His main research interests include Chinese corpus processing, Chinese intensional logic model and its applications, intelligent retrieval based on conceptual intensions (Internet, digital libararies), dialogue systems and semantic Web.

陆汝占,1940 年生,教授,博士生导师,主要研究方向为汉语语料库加工技术、汉语内涵逻辑模型及其应用、基于概念内涵的智能检索(互联网、图书情报)、对话理解系统、语义 Web.

Li Xuening, born in 1971. Ph. D. candidate in Shanghai Jiao Tong University. His main research interests include Chinese intensional logic model and dictionary extraction.

李学宁,1971 年生,博士研究生,主要研究方向为汉语内涵逻辑模型以及词典提取.

Duan Jianyong, born in 1978. Ph. D. candidate in Shanghai Jiao Tong University. His main research interests include natural language processing, information extraction, machine learning and bioinformatics, *etc*.

段建勇,1978 年生,博士研究生,主要研究方向为自然语言处理、信息抽取和生物信息学等.

Chen Yuquan, born in 1968. Associate professor in Shanghai Jiao Tong University. His main research interests include Chinese intensional logic model and its applications, intelligent retrieval based on concepts, Chinese corpus processing *etc*.

陈玉泉,1968年生,副教授,主要研究方向为汉语内涵逻辑模型及其应用、基于概念内涵的智能检索、汉语语料库加工等.

Research Background

Traditional wisdom of document categorization lies in mapping a document to given topics that are usually sport, finance, politics, etc. Whereas, in recent years there has been a growing interest in non-topical analysis, in which characterizations are sought by the opinions and feelings depicted in documents, instead of just their themes. This method of analysis is defined to classify a document as favorable (positive) or unfavorable (negative), which is called sentiment classification. Labeling documents by their semantic orientation provides succinct summaries to readers and will have a great impact on the field of intelligent information retrieval. In this paper, we present a new language modeling approach to sentiment classification. With respect to this generative model, we represent the "thumb up" and "thumb down" semantic orientation with their corresponding language models estimated from positive and negative collections. When classifying a test document, the distances of its language model from these two sentiment models are computed to determine its sentiment class. In terms of our experimental results, we conclude as follows: when training data is limited, the language modeling approach performs better than SVMs and Naïve Bayes classifier. On the other hand, the language modeling approach shows its robustness in sentiment classification. Thanks to the NSFC Major Research Program 60496326: Basic Theory and Core Techniques of Non Canonical Knowledge for supporting this study.