主客观识别中的上下文因素的研究[1]

蒙新泛　王厚峰

北京大学计算语言所教育部重点实验室，北京，100871

mxf@pku.edu.cn，wanghf@pku.edu.cn

摘 要：主客观判别是观点分析中的一个基本问题。在本文中，我们通过4组对比实验，分析了上下文信息对于主客观判别的影响。从实验中我们得出的结论是：引入上下文信息能够对主客观分类性能产生影响，但简单的信息引入方法反而会降低分类的准确度，只有在有针对性地选取特征以及分类方法时上下文信息才能起到提高性能的作用。我们的实验还表明了CRF模型是一种较有效的主客观识别方法。

关键词：情感分析，主客观识别，特征选择，机器学习算法

A Study on the Impact of Context Information in Subjectivity Detection

Xinfan Meng, Houfeng Wang

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, 100871

mxf@pku.edu.cn, wanghf@pku.edu.cn

Abstract: Subjectivity detection is a basic problem in sentiment analysis task. In this paper, we design 4 groups of experiments to analyze the impact of context information in subjectivity detection. Via these experiments, we draw the following conclusion: context information can influence the performance of the subjectivity detection. But a naïve way of incorporating the context information might lower the accuracy of subjectivity classification. To improve the performance, we have to carefully choose the features and classification methods. Our experiments also indicate that CRF model is a promising method to classify subjectivity text.

Keywords: sentiment analysis, subjectivity detection, feature selection, machine learning

1　引言

面向文本的观点（或情感）分析是近几年来自然语言处理研究的一个热点[1][6]。在国际会议和评测任务中，不乏相关的问题；颇具影响的国际评测会议TREC以及NTCIR等，都设置了倾向性分析的任务。

主客观分析是观点分析的基础，因为观点总是主观性的。所谓主客观分析，就是判断某个语言单位表达的是作者的主观观点还是作者陈述的客观事实。其中的语言单位可以有不同的粒度，如，篇章、段落、句子、短语或词。

根据目前的研究，主客观分析的方法大致可以划分为三类：（1）基于词典的方法。利用预先建立的词典（可以是人工标注也可以是机器自动获取的），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（例如SVM，最大熵），判断新数据应该划分为主观还是客观。（3）基于图的方法。利用求最小割的方法把文本在句子级别上切分为主客观两个部分[4]。

2　上下文信息

[1] 本研究受国家自然科学基金（No.60675035)和北京市自然科学基金（No.4072012)资助。

在寻找新的判别方法的同时，越来越多的信息被作为特征引入判断任务中。从直观上来说，一个句子与它的上下文环境有着密不可分的关系。我们认为这些上下文信息也有助于判别文本的主客观性，例如一个句子的前后句子都携带着主观信息，那么这个句子也极有可能是主观的。Pang 针对英文电影评论，讨论了上下文信息对于主客观判别的作用[4]。在本文中，我们针对中文，设计了多组对比实验，旨在通过实验讨论和分析上下文信息对于中文主客观判别的作用。

我们设计了 4 组实验，分别考察了（1）在引入和不引人上下文信息的情况下，（2）在使用不同算法的情况下，（3）在使用不同特征选取方法的情况下，主客观分类的准确度的变化情况。这 4 组实验分别是：

1. 没有引入上下文信息，使用 SVM，最大熵进行分类；
2. 简单地引入上下文信息，在使用不同的上下文窗口情况下利用 SVM，最大熵进行分类；
3. 将文本混合后再引入上下文信息，使用 SVM，最大熵进行分类；
4. 使用序列标注模型 CRF 进行分类。

在实验中使用的方法我们将在下一章中作详细的介绍。

3 实验

在本章中，我们将介绍采用的实验数据以及各组对照实验，而后针对实验结果给出分析。在实验中，主客观判别任务被当作一个二分类问题，分类的方法采用的是当前比较流行的机器学习方法。

3.1 实验数据

本文采用的实验数据是 NTCIR 评测中的简体中文和繁体中文的 Release 语料。其中，繁体中文语料来自从 1998 至 2001 年的《中国时报》，《中央日报》等多家报纸。简体中文语料来自从 1998 至 2001 年的《新华日报》以及《联合早报》。

简体中文和繁体中文语料下各自涵盖了 14 个主题的内容，主要是政治和财经报道，例如“亚洲金融危机”，“911 恐怖袭击”和“中国大陆洪水”等。两组语料下各自涵盖的主题有些是重叠的，但并不完全相同。在每一个主题下又划分为若干个文档，每个文档再划分为若干个句子。每个句子均由三个不同的标注者独立进行主客观标注，因而可能产生不一致，在我们的实验中，我们取多数标注者的意见作为标准答案，即一个句子，如果有两个人或三个人认为它是主观的，我们就认为它是主观的。这与 NTCIR 中的 Lenient 条件下的评价标准也是一致的[7]。

以下是我们对两组语料的一些统计数据：

表 1 两组语料的一些统计数据

语料	主题数	文档数	句子数	主观句子数	客观句子数
简体中文	14	252	4877	1869	3008
繁体中文	14	187	4655	2182	2483

由表 1，我们也可以确定在这两组语料上进行的主客观判别任务的准确度基线。

表 2 准确度基线

语料	基线
简体中文	0.617
繁体中文	0.532

3.2 实验 1

在本组实验中，我们将忽略上下文信息对主客观判别的影响，利用机器学习方法对每个句子

进行独立的分类，不考虑它前后句子所带来的影响。

在语料中，每个句子都已经是分割好的，因而我们只需要对句子进行分词和词性标注。在经过这些预处理步骤后，我们对句子进行特征提取。我们使用词袋子(bag of words)作为特征，句子中的每一个词语/词性对作为句子的一个特征，每个特征只有 1 和 0 两种取值。

在实验 1 中，我们采用了两种在自然语言处理领域中应用效果较好的机器学习方法作为主客观判别方法。一种是 SVM 方法，另一种是最大熵方法。

以下是对于两个数据集的十层交叉验证结果：

表 3 实验 1 结果

语料	SVM 准确度	最大熵准确度
简体中文	0.738	0.698
繁体中文	0.736	0.700

与表 2 中的基线相对比，可以看出这两种机器学习方法都能较好地分辨出句子的主客观类别。

3.3 实验 2

在本实验中，我们引入了句子的上下文信息来帮助对其主客观类别进行判别。

在实验中考虑了文本的顺序，对某个句子而言，在它之前的句子要比在它之后的句子对其影响更大。因而，我们只引入了句子之前的文本：对于一个句子，把它之前的 N 个句子的特征（特征的选取方法同上一节）作为特征加入这个句子的特征向量中。为了评价在不同的上下文窗口内准确度的变化情况。我们的实验针对多个 N（N = 2，3，4）进行了十层交叉验证。以下是我们的结果：

表 4 实验 2 结果

语料	N=2		N=3		N=4	
	SVM	最大熵	SVM	最大熵	SVM	最大熵
简体中文	0.688	0.671	0.663	0.659	0.670	0.647
繁体中文	0.643	0.643	0.643	0.609	0.633	0.610

从表中可以看到，当引入上下文信息后，准确度不仅没有上升，反而下降了。而且引入的上下文信息越多，准确率下降得越多。

3.4 实验 3

假如一个句子的主客观性与上下文没有关系，那么对语料进行随机混合打乱后再引入上下文信息能达到的精确度应该与实验 2 能达到的精确度保持一致。为此，我们进行了以下三个实验。

(1) 对不同主题下的句子进行随机混合

可以把不同主题的句子组合在一起。在混合后，一个句子的下一句可能来自于另一个主题。经过如此处理，我们按照实验 2 的方法引入 N=2 情况下的上下文信息。结果如表 5 所示：

表 5 不同主题混合情况下的实验结果

语料	SVM	最大熵
简体中文	0.589	0.638
繁体中文	0.565	0.614

(2) 对相同主题下的句子进行随机混合

对相同主题下的句子混合，重复上述实验，得到表 6 所示结果。

表 6 相同主题下混合的实验结果

语料	SVM	最大熵
简体中文	0.588	0.643
繁体中文	0.570	0.597

(3)对相同文档下的句子进行随机混合

同一个文档下的句子混合在一起，重复上述实验，得到表 7 所示结果。

表 7 相同文档下混合的实验结果

语料	SVM	最大熵
简体中文	0.629	0.655
繁体中文	0.602	0.633

从上述 3 组实验中可以看出，

1. 表 5 与表 6 对应的数据结果接近，但与表 7 的结果有着一定的差距；
2. 与实验 2 的表 4 相比，可以看出这组实验的准确度与实验 2 有较大的差距。

3.5 实验 4

CRF 方法是一个全局优化的序列标注方法，在对序列中一个元素进行分类的时候，不仅考虑到元素本身的数据，还考虑了它前后的元素所带来的影响。同时，CRF 还克服了隐马尔可夫模型以及最大熵马尔可夫模型的缺点，现在已经很好地被应用于词性标注，文本分段以及识别观点持有者等领域[5][8]。因而我们在这个实验中采用了这个模型。

在特征选择方面，出于性能方面的考虑，我们不能够把一个句子中的所有词语都作为 CRF 模型的特征。因此，我们有针对性地选择了以下三个特征：

1. 句子中是否包含了主张词。我们定义了一个主张词词典，其中包含了 69 个表达观点持有者发表观点的词语，例如：表示，透露，声明等等；
2. 句子中是否包含了情感词。所谓情感词是指带有主观色彩的词语，例如好，坏，美，丑等等。我们定义的情感词典中包含了 13653 个词语；
3. 句子中是否包含了连词。连词表达了前后句子之间的关系，例如并列，转折，递进等关系。我们也为此定义了一个连词词典，共包含了 20 个连词。

经过这样的特征提取步骤后，每一个句子被表示为了一个三维的向量。利用 CRF 模型进行十层交叉验证的结果在表 8 中给出：

表 8 CRF 模型实验结果

语料	准确度
简体中文	0.682
繁体中文	0.720

从表 8 中可以看出，在使用非常简单的特征的情况下，CRF 模型能够取得与实验 1 相近的准确度。

3.6 实验结果分析

为了对结果进行更好的分析，我们对两组语料中连续出现的主观或者客观的句子的个数做了统计。我们定义主客观连续长度为一段全是主观（或全是客观）的文本最多能包含的句子的个数。例如连续两个句子都是主观（客观），但是这两个句子的前后句子均为客观（主观），我们就认为该文本段的主客观连续长度为 2。对每一个主客观连续长度，我们对具有这个长度的文本段的个数进行了统计。结果如下（由于篇幅所限，只显示了主客观连续长度从 1 到 10 的文本段的个数）：

表 9 主客观连续长度-文本段个数统计

语料	主客观连续长度									
	1	2	3	4	5	6	7	8	9	10
简体	805	367	201	132	91	45	28	30	15	14
繁体	890	363	234	133	81	46	36	21	21	8

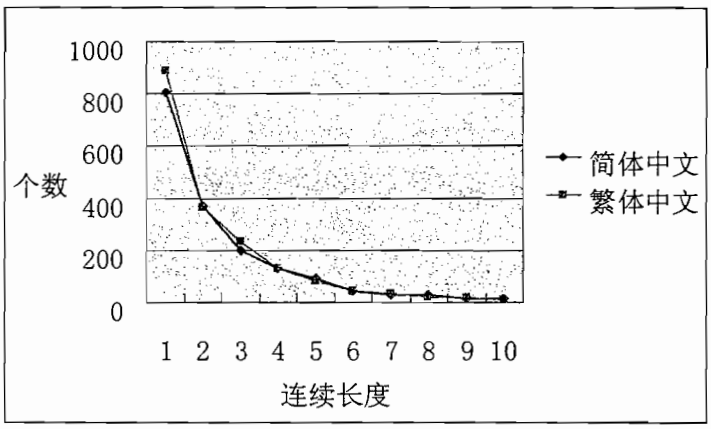

图 1 主客观连续长度—文本段个数（简体）

可以看到，主客观连续长度为 1 的文本段占了全部文本段个数的二分之一。这表明，在语料中，有二分之一的句子，它之前以及之后的句子与它的主客观标记是相反的。而且，连续长度越长的文本段的个数越少。这是因为我们使用的数据是新闻报导，新闻报导的措辞和行文方式比较简洁，每个句子表达的意思相对完整。

因而，简单地把一个句子上下文的特征加入到它的特征向量中，必然引入噪音，从而降低了分类的准确性。这在我们的实验 2 中得到了证实：在简单地引入上下文信息后准确度反而下降了。

但是，这并不能说明上下文信息对于我们进行主客观分类没有帮助，在实验 3 中，我们把语料按照不同的方法进行随机的混合。实验结果表明进行了混合之后，分类的准确度大大降低了。这表明上下文信息有助于判别句子的主客观性。

在实验 4 中，我们采用了序列标注模型 CRF 来进行分类，并且有针对性地选择了几个简单的特征。实验结果表明，在简单的特征下，CRF 模型已经能达到较高的准确率。在更好地选择特征的情况下，CRF 有望能达到更高的精确值。

通过对这 4 组实验的分析，我们得出了以下结论：

1. 上下文信息在文本主客观分类中能够起到一定作用；
2. 要有效地利用上下文信息，我们需要有针对性地采用合适的机器学习方法并设计合理的特征提取方法；
3. CRF 模型是一个能够有效利用上下文信息的模型。

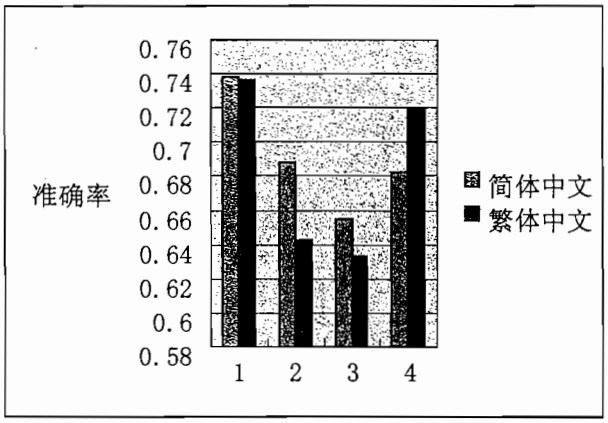

图 2 各组实验中取得的最高准确率

4 总结与展望

在本文中，我们设计了 4 组对照实验，验证了上下文信息因素在文本主客观分类中所起到的作用，并且讨论了使用哪些算法以及特征选择方法引入上下文信息能够更好地提高主客观分类的准确度。

在目前的工作中，我们选择的特征仍然比较简单，例如在 CRF 模型中，我们仅仅采用了三维向量表示一个句子，句子中的大量信息都丢失了。在进一步的工作中，我们将继续研究如何通过选择特征的方法来提高分类方法的准确度。

参 考 文 献

[1] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on EMNLP, pages 79-86. 2002.

[2] Ku, L.-W., Liang, Y.-T. and Chen, H.-H. Opinion extraction, summarization and tracking in news and blog Corpora. Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report. Pages 100-107. 2006.

[3] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In Proc. of the 2005 Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP2005), Vancouver, B. C., 2005

[4] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL), pages 271–278, 2004.

[5] J. Lafferty, A. K. McCallum & F. Pereira. 2001. Conditiona Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of 18th International Conference on Machine Learning.

[6] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[7] Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, Noriko Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. Proceedings of NTCIR-7 Workshop Meeting pages 185 -203, 2008.

[8] Xinfan Meng, Houfeng Wang. Detecting Opinionated Sentences by Extracting Context Information. Proceedings of NTCIR-7 Workshop Meeting pages 268 -271, 2008.