中文倾向性分析的研究

张猛，彭一凡，樊扬，李丹，林小俊，吴玺宏

北京大学言语听觉研究中心，北京，100871

E-mail: {zhangm, pengyf, fanyang, lidan, linxj, wxh}@cis.pku.edu.cn

摘　要：文本倾向性分析是自然语言处理中的一个热点问题。本文介绍了一套中文文本倾向性分析的方法，它包括词法分析和倾向性判别两个步骤。在词法分析中，基于条件随机场模型，对输入的文本进行分词和命名实体识别的一体化处理，从而有效地提高了分析性能。在倾向性判别中，从词汇、句子和篇章三个不同层次进行分析。其中在词汇层次上采用最大熵模型，根据上下文信息进行情感词识别和极性判别。在句子层次上根据构建的属性列表抽取评价对象，并通过修饰词判断其倾向性。在篇章层次上，以词汇判别结果为基础，采用支持向量机模型，融合多种信息对文本的主客观和极性进行判别。最后，本文在搜索引擎中加入文本倾向性分析功能，在检索到相关文档的同时，得到其褒贬倾向。

关键词：词法分析一体化，情感词，倾向性分析

Research on Chinese Orientation Analysis

Zhang Meng, Peng Yifan, Fan Yang, Li Dan, Lin Xiaojun, Wu Xihong

Speech and Hearing Research Center, Peking University, Beijing, 100871

E-mail: {zhangm, pengyf, fanyang, lidan, linxj, wxh}@cis.pku.edu.cn

Abstract: Orientation analysis is a hotspot in natural language processing. This paper mainly proposed some Chinese orientation analysis approaches, which included lexical analysis and orientation distinction. The lexical analysis integrated word segmentation and entity identification methods to improve analysis performance. Orientation distinction could be utilized on word, sentence and article levels. On the word level, it considered the context information to recognize the sentiment word and its polarity based on Maximum Entropy model. On the sentence level, it extracted evaluating objects according to pre-constructed property list, and predicted their orientation through modifying words. On the article level, it took the word level model as its fundamental, combined with multi-information to distinct articles' orientation by means of Support Vector Machine. Finally, it plugged orientation analysis function in the search engine to get corresponding documents with their orientation as well.

Keywords: integrated lexical analysis, sentiment word, orientation analysis

1　引言

随着计算机的普及和网络的发展，大量信息以电子文本的形式出现。面对信息爆炸带来的挑战，人们迫切需要更快更便捷的方法获取所需信息。倾向性分析就是在这样的背景下应运而生的。例如，在购买一款手机之前，我们往往会去一些网站或论坛，浏览其他用户的评价，这需要花费很多时间。然而在倾向性分析技术的帮助下，我们就可以快速地获得这款手机的综合评价。由此可以看出，倾向性分析有着广泛的应用前景。因此近几年，它已经成为自然语言处理中的一个热点问题。

本文提出了一套中文文本倾向性分析的方法，它包括词法分析和情感倾向性判别两部分。

在词法分析部分，该方法对输入文本进行分词、命名实体识别和词性标注。由于这三

者都可以视为序列标注任务，因此有很多序列标注模型可以应用，如条件随机场（Conditional Random Fields）[1]等。分词和命名实体识别之间存在着紧密联系，所以本文采用条件随机场模型，将分词和命名实体识别合并为一个序列标注任务，这样可以提高系统的整体性能。词性标注任务同样基于条件随机场模型，对分词结果标注词性信息。

倾向性判别部分可以在词汇、句子和篇章的不同粒度上进行。

在词汇层次上，我们首先基于 HowNet[2]构建候选情感词词典，然后利用英语情感词词典 Harvard&Lasswell[3]对候选情感词进行扩充，最后基于最大熵（Maximum Entropy）模型[4]，根据上下文信息，对这些候选情感词，进行情感词识别和极性判别。

在此基础之上，在句子层次对评价对象进行情感分析。我们首先利用属性列表挑选出评价对象，然后根据出现在其上下文中的修饰情感词的极性，给出评价的褒贬倾向。

在篇章层次上，利用情感词等多种特征，采用支持向量机（Support Vector Machine）模型[5]，对文本的主客观和极性进行判别。

最后，我们在搜索引擎中加入文档极性判别的功能，从而在检索到相关文档的同时，得到其褒贬倾向。

本文组织如下：第二部分介绍了一体化的词法分析方法；在此基础之上，第三部分详细阐述了各个任务所采用的方法；最后给出总结。

2 词法分析

词法分析是指自然语言处理系统对输入文本所进行的词汇级的处理过程。它是各种自然语言处理系统的首要分析步骤，是进行句法分析、语义分析的基础，也是倾向性分析的基础。分词、命名实体识别和词性标注是中文词法分析的主要任务。

2.1 分词

最近几年来，国际中文分词评测活动推进了自动分词技术的发展。基于字分类的策略已经成为汉语分词的主流方法。这种方法把分词问题视为字的分类问题，每个字根据其在具体词的不同位置，属于不同的类别，比如词首、词中、词尾、单字词等。例如：

词序列：　　汉语 的 分词 方法

字序列：　　汉　语　的　分　词　方　法

标记序列：　B　E　S　B　E　B　E

一个和字序列对应的标记序列，完全指明了一句话的切分方式。而分词的任务相应的转化为每个字在具体的上下文中的分类问题。这样，很多成熟的机器学习方法就可以应用到这个问题当中来。

2.2 命名实体识别

命名实体识别本质上是一项切割和标注的任务。因此可以看作是一个两阶段过程：首先确定命名实体的边界，进而确定命名实体的类别。

这样就有两种策略来处理该问题。

第一种策略是两阶段式。首先是切割阶段，可以采用类似于汉语分词的方法，即“B、I、O”标注方法，B 代表命名实体首，I 代表命名实体其他部分，O 代表非命名实体；然后，对于标为命名实体的词串进行分类；

第二种策略是将二者合并为一个序列标注任务，即“B-X、I-X、O”的方式，B-X 表示 X 类命名实体首，比如 B-PER 表示人名首。

一般认为第二种策略引入了更多的区分性信息，效果更显著。

2.3 分词和命名实体一体化方法

分词和命名实体识别之间存在着紧密关系。命名实体对应的专有名词往往是未登录词，而未登录词恰恰就是影响分词性能的最重要因素。同时，分词也会影响命名实体识别的性能。鉴于分词和命名实体识别之间的这种相互依赖关系，我们在序列标注的框架下，采用条件随机场模型[6]，将分词和命名实体识别合并为一个任务。实验表明，联合分析的性能优于单纯的分词和命名实体识别[7]。

针对本次评测，我们采用人民日报 2000 年 1 月份和 2 月份语料作为训练语料。分词采用 4 类标记：B、I、E、S，分别表示词首、词中、词尾和单字词。命名实体识别采用 4 类标记：PER、LOC、ORG、O，分别表示人名、地名、组织机构名和非命名实体。这样在一体化模型中就有 B-PER 等 16 种标记。采用的特征为基于 3 字窗长的 6 类特征模板，分别是：C-1、C0、C1、C-1C0、C0C1、C-1C1。其中 C-1、C0、C1 分别表示前一个字、本字、后一个字。

2.4 词性标注

我们将词性标注也视为序列标注问题，并采用条件随机场模型[6]进行建模。

我们采用人民日报 2000 年 1 月份和 2 月份语料作为训练语料。词性标记集采用《北京大学现代汉语语料库基本加工规范》。使用的特征除状态转移特征外，包括 W0、W-1、W1、L0、L-1、L1、B0、E0、W-1W0、W0W1、W-1W1，分别表示本词、前词、后词、本词长、前词长、后词长、词首字、词尾字、前词本词组合、本词后词组合、前词后词组合。

3 情感分析

3.1 情感词识别和极性判别

本次评测的任务 1 要求进行情感词的识别，任务 2 要求对识别出的情感词进行褒贬极性判别。这两个任务联系紧密，因此我们将它们作为一个任务来处理，即根据上下文环境直接挑选出情感词并判断褒贬极性，并分别采用有监督和无监督两种方法。

词的极性依赖于上下文，所以我们的极性判别策略是：首先利用 HowNet[2]情感词集合及英语情感词词典 Harvard&Lasswell[3]构建候选情感词词典[8]，然后从测试语料中挑选出能覆盖这些候选词的 1000 篇文章，进行手工标注。最后利用标注语料训练最大熵模型[9]，并根据上下文信息对候选情感词的极性进行判别。

例如“首先/c ，/w 我/r 向/p 全国/n 各族/r 人民/n 和/c 海外/s 侨胞/n ，/w 向/p 世界/n 各国/r 的/u 朋友/n 们/k ，/w 祝贺/v 新年/t 快乐/a ！/w”，假设当前词为“人民”，我们采用的特征模板及说明见表 1。

另外，我们还采用了无监督的方法来对候选情感词的极性进行判别，判别的具体步骤如下：

1) 从 HowNet[2]中挑选出情感词构建情感词词典；
2) 对于每一个候选情感词，首先查找情感词字典，若存在，则直接判断其极性；
3) 若没有出现在情感词典中，找出与其出现在同一个句子中的情感词，若没有则不对该候选情感词的极性进行判断；
4) 计算候选情感词与情感词之间的 SO-PMI（semantic orientation-pointwise mutual information）[8]，并设置阈值挑选情感词、判断其极性。

表 1 情感词判别模型的特征模板及说明

Tab.1 Table of feature templates and specifications used in model for sentiment word distinction

特征	说明	特征值
W_2	前两个词	全国
W_1	前一个词	各族
W0	当前词	人民
W1	后一个词	和
W2	后两个词	海外
P_1	前一个词的词性	r
P0	当前词的词性	n
P1	后一个词的词性	c

PMI 的计算公式如（1）：

$$PMI(word_1, word_2) = log(\frac{P(word_1 \ \& \ word_2)}{P(word_1)P(word_2)}) \quad (1)$$

其中 $P(word_1 \ \& \ word_2)$表示 $word_1$ 和 $word_2$ 同时出现的概率。在此基础之上，计算一个新的情感词与种子情感词的互信息概率 SO-PMI，公式如（2）：

$$SO\text{-}PMI(word) = \sum_{pword \in Pset} PMI(word, pword) - \sum_{nword \in Nset} PMI(word, nword) \quad (2)$$

其中，*Pset* 和 *Nset* 分别表示褒义和贬义种子情感词的集合。在评测中，我们仅针对形容词和动词，采用 SO-PMI 值进行判断。

3.2 评价对象抽取

任务 3 的语料涉及照相机、汽车、笔记本和手机四个领域。

我们的方法是：首先针对每个领域构造可能的商品属性列表，然后再根据上下文的情感词来对评价对象的极性进行判断。

对于商品属性列表的构建，我们首先从网上抓取了大量介绍这四类产品的网页，并从中抽取出商品属性，例如内存容量、内置摄像头等。

根据商品属性列表，我们抽取文本中的评价对象，并利用其上下文中出现的情感词，判断其极性。我们通过分析评测主办方提供的例子，发现在“优良的底盘技术”这样的描述中，凭借修饰词“优良”可以判断其对“底盘技术”的评价是褒义的。这种判别方法的关键是情感词词典的构建。情感词词典我们采用从 HowNet[2]中抽取出来的纯褒贬词（即仅作为褒义或贬义的词）。另外，我们发现单字情感词的极性判别难度较大，因此挑选出出现在商品属性词周围的单字情感词，进行手工标注。对于一个评价对象，当其上下文中的褒义情感词总数大于贬义情感词总数时，认为该评价是褒义的；否则，是贬义的。

运用这种方法，我们对评测主办方提供的样例进行分析判别，结果证明，该方法是有效的。

3.3 文本的主客观判别

客观性文本的句型、用词基本上都是规范的。而主观性文本不论在字和词的使用上，

还是在句型的使用上与客观性文本都有较大的差别。因此，普通的字频或词频统计对此类文本分类并不适用[10]。

经过对大量主客观文本的观察和分析，我们决定采用分类方法，并提出以下几类特征。这些特征反映了主客观文本的区别。

1) 情感词的个数是否大于一个阈值；
2) 文中叙述是否采用第一人称或者第二人称；
3) 是否包含感叹号；
4) 是否包含问号；
5) 是否包含主观性色彩强烈的词语，并且和第一人称或者第二人称在一个句子中共现。

我们的训练和测试语料采用手工标注的1000篇文档，分类器采用支持向量机[11]，但实验结果并不理想。通过分析，我们认为把主客观判别当作一个绝对的分类问题来处理是不合适的。由于任务4是要挑选出4000个主观性文本，所以实际上我们只需要挑选出主观性最明显的4000篇文档即可。

对于每篇文档，其主观程度由式（3）定义，

$$S(d) = P(d) + N(d) \tag{3}$$

其中，$S(d)$、$P(d)$、$N(d)$分别表示文档的主观程度、褒义程度和贬义程度。而 $P(d)$和 $N(d)$定义为，

$$P(d) = \sum_{w \in d} PW(w) \tag{4}$$

$$N(d) = \sum_{w \in d} PN(w) \tag{5}$$

其中 $PW(w)$和 $PN(w)$分别表示词 w 的褒义和贬义程度。对于式（4）和式（5）的计算，我们采用针对任务 1、2 得到的情感词极性判别模型进行判断，褒义取值为 1，贬义为-1，否则为 0；然后根据情感词前面程度修饰词的强弱和否定词得到最终的极性强度值。程度修饰词和否定词来自《现代汉语语法信息词典》[12]。

采用这种方法，当 $S(d)$大于某一阈值时，认为文档 d 为主观性文档。在手工标注的 1000 篇文档上，实验结果的准确率随阈值变化的曲线如图 1 所示。

由图 1 可以看出，这一比例最高为 87.68%，说明该方法是有效的。不同阈值下抽取文档在所有测试文档中所占比例的变化如图 2 所示。当阈值为 40 时，抽取主观文档所占比例为 17.26%，说明这种方法可以抽取出所需的文档。

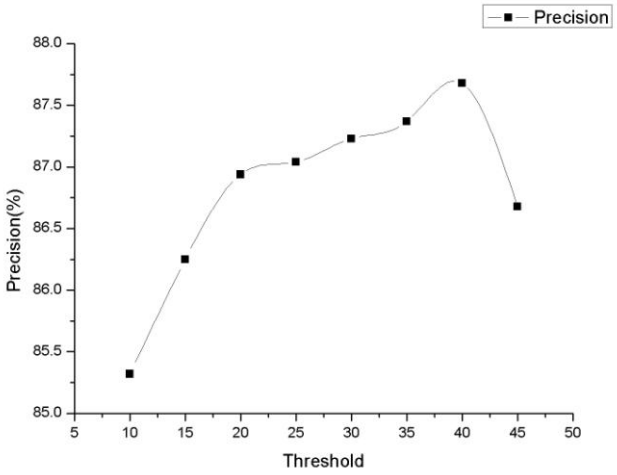

图 1 准确率随阈值变化的曲线

Fig.1 The curve of accuracy according to threshold

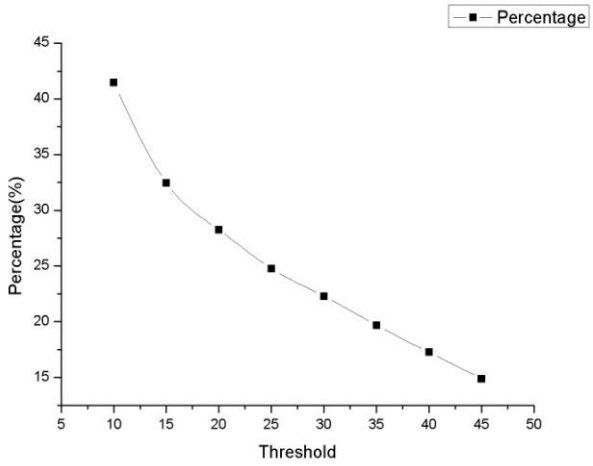

图 2 文档比例随阈值变化的曲线

Fig.2 The curve of document percentage according to threshold

3.4 文本褒贬极性判别

在这个任务中我们利用手工标注的 1000 篇文档，采用支持向量机模型[11]进行有监督的训练、对文档进行褒贬判别。其中抽取 800 篇作为训练语料，200 篇作为测试语料。选取的特征如下。实验结果的准确率为 72.92%。

1) 由文档中出现的词构成的特征向量

2) 褒义情感词统计量 *P(d)*

3) 贬义情感词统计量 *N(d)*

对于特征 1，没有采用词频，而是以是否出现作为特征[13]。

3.5 观点检索

任务 6 针对给定对象，要求找出包含关于该对象的倾向性观点的文章。

我们使用 Lucene[14]搭建检索平台。为避免分词结果与主题词出现不一致，首先对分词结果进行后处理，即，将测试中的主题词进行合并。主题词处理前后的检索情况对比如图 3 所示。

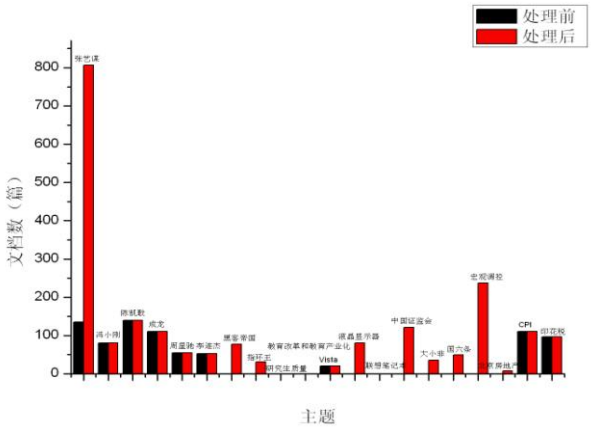

图 3 **主题词处理前后的检索情况对比**

Fig.3 The performance after post-processing

从图 3 可以看出，对于多个主题词，这样的后处理是必要的。另外，通过实验发现，有些主题词，例如“北京房地产”等，与它相关的文档既可能包含该词；也可能不包含该词，但同时包含了“北京”和“房地产”两个词。因此我们在对“北京房地产”检索后，又检索了同时出现“北京”和“房地产”的文档。主题词的切分方法是，利用从分词后的语料中提取的词典，对主题词进行切分。

对于检索到的文档的相关性得分，我们综合考虑了词条频率（term frequency）和倒排文档频率（inversed document frequency）的影响。对于文档的主客观和褒贬判别，采用任务 3.3、3.4 的方法进行判别。

4 总结

对于情感词的识别和极性判别，我们发现根据情感候选词的上下文信息进行极性判断是一种有效的方法。但是训练语料的标注容易受到主观因素的影响。因此，中文情感词标注语料库的建设是一个急需解决的问题，需要制定详尽的标注规范。

对于评价对象及其属性的抽取和评价问题，我们认为需要针对某个领域，尝试构建属性词典。在方法上，利用上下文的情感词信息可以解决一部分问题。

最后，在评测中我们发现，融合情感分析的信息检索可以给用户带来许多有价值的反

馈信息，但是这需要依赖词汇及文档的倾向性分析技术。

对于情感分析，我们认为仅从词汇、句法上进行分析并不能从本质上解决这个问题，而是应该建立在语义理解的基础上，利用本体知识、知识库，来更好地解决。

参 考 文 献

[1] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [A], In Proc. of ICML, 2001, pp. 282-289

[2] Dong Z, Dong Q. HowNet, http://www.keenage.com/zhiwang/e_zhiwang.html

[3] Stone P J, Dunphy D C, Smith M S, Ogilvie D M, The general inquirer: a computer approach to content analysis[M], MIT Press, 1966

[4] Berger, A., S. A. Della Pietra, and V. J. Della Pietra. Maximum Entropy Approach to Natural Language Processing [J], Computational Linguistics, 1996 , pp. 39-71

[5] Valdimir Vapnik. The Nature of Statistical Learning Theory [M]. Springer-Verlag, New York, 1995

[6] http://crfpp.sourceforge.net/

[7] 于佃海. 汉语词法分析的机器学习方法研究 [D]. 北京大学硕士研究生学位论文，2008 年

[8] 姚天昉，娄德成. 汉语情感词语义倾向判别的研究 [A]. 中文计算技术与语言问题研究-第七届中文信息处理国际会议论文集 [C]，2007

[9] http://maxent.sourceforge.net/

[10] 姚天昉，彭思崴. 汉语主客观文本分类方法的研究 [A]. 第三届全国信息检索与内容安全学术会议论文集 [C]，2007

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] 俞士汶，朱学锋，王慧等. 现代汉语语法信息词典详解 [M]，清华大学出版社，1998

[13] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques [A]. EMNLP，2002

[14] http://lucene.apache.org/