基于CRF和句法分析的中文微博情感分析[1]

陈豪，苏波，黄晨，刘功申[2]
上海交通大学信息内容分析技术国家工程实验室
中国上海

摘要　上海交通大学信息内容技术国家工程实验室参加了2012年CCF自然语言处理中文微博情感分析测评。在本次微博情感分析测评中，分别采用两种算法，提交了两组结果。第一种方法是采用条件随机场算法，对微博信息进行情感预测。第二种算法是，利用Standford Parser进行句法分析，然后，根据句子成分之间的修辞关系计算句子的情感。经过主办方的公开测试，两组算法的结果优异。

1 引言

情感分析[1-2]，又称为意见挖掘，是指通过自动分析来获得对于某件事物的褒贬意见，而随着近几年来微博使用者数量的急剧增加，这种让人方便发表自己态度以及看法的工具在整个互联网舆情中扮演越来越重要的角色。所以针对微博语料的情感分析变得非常有意义。

微博是一种限制文本长度的工具，大多数人在微博上发的内容一般都是以短文本的形式。但是其中也存在比较长，句子结构比较完整的句子。短文本就是内容较短的文本（一般长度不超过160字符），通常以新闻标题、微博、手机短信、电子邮件、购物评价等形式存在。有效的短文本情感倾向性分析技术可以帮助我们在海量信息中更准确地获取自己感兴趣的信息。

文章[3][4]提出了引入外部知识（如维基百科、搜索引擎返回的信息等）进行特征扩展的方法，弥补了短文本特征稀疏的缺点，提高了分类性能。但是这种依赖统计的特征引入方法容易受到噪声干扰，而且增加了算法的复杂度，不符合短文本处理快速高效的要求。文献[5]提出了采用CRFs对非常短的文本(10个字符以内的文本）进行字标注的方法进行分类，具有很好的分类效果，但不适用于一般的短文本。

本文针对这次测评发布语料的特殊性，采用两种方法对测评的微博语料进行测评标注，针对微博中短文本居多，并且字符数较少、特征稀疏等特点，使用一种CRFs的短文本情感倾向性分析算法，该算法采用文本处理中常用的序列标注算法，保存短文本词之间的联系。另外针对微博中也存在比较长的句子，所以另外一种算法采用了句法分析的方法，该算法使用了斯坦佛大学发布的句法分析器Standford Parser进行句法分析，通过依赖关系识别对每一句微博进行情感识别，得出正负结果。

2 利用CRFs进行短文本情感倾向性分析的实现

CRFs（Conditional Random Fields, 条件随机场)最早由John Lafferty等人于2001年提出的[6]。目前CRFs在数据分段、序列标注、命名实体识别、中文分词等自然语言处理任务中都有很好的表现。

CRFs是基于HMMs(隐式马尔可夫模型)与MEMs(最大熵模型)的基础上的改进。CRFs

[1] 国家自然科学基金项目支持（61272441，61171173）。
[2] 通信作者：刘功申（lgshen@sjtu.edu.cn）

使用条件特征，可以对特征进行全局归一化。它不是在给定当前状态的条件下定义下一个状态的分布，而是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，从而避免了 HMMs 的对数据进行不必要的独立性假设。而且 CRFs 很好的解决了 MEMS 的标注偏执问题。在现实的序列标注任务中，CRFs 性能往往都优于 HMMs 和 MEMs。

使用 CRFs 进行短文本情感倾向性分析的流程如图 1 所示。

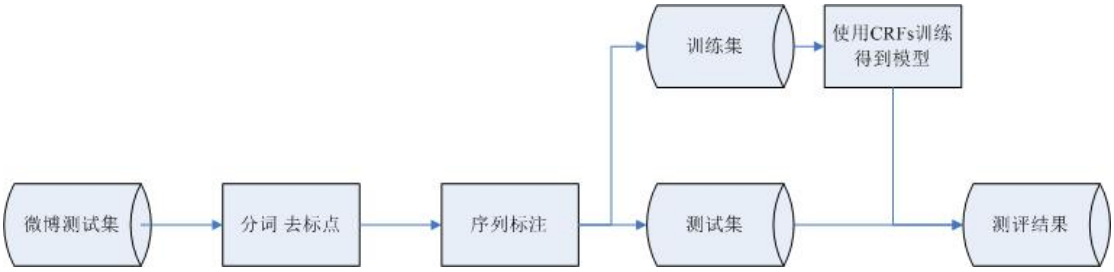

图 1. 基于 CRFs 的微博情感分析流程图

2.1 链式 CRFs 模型

CRFs是一个无向图上的指数概率模型，它采用了链式无向图结构计算给定观察值条件下输出状态的条件概率[9]。

令X=(x1,x2,...,xn)为可观测的输入序列(例如词性标注中的句子)，Y=(y1,y2,...,yn)为待预测的标记序列(例如词性标注中词性)，其中想xi表示X的第i个分量，yi是xi对应的标签。线性链CRFs定义标记序列Y的条件概率为：

$$p(y \mid x,\lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(y,x))$$

其中 Z(x)是归一化因子，是特征函数。特征函数分为两种，一种特征函数只与当前状态相关，另一种特征函数还与当前状态的前一个状态有关。对于离散型特征，函数的取值通常为 0 或 1。是对应特征函数的权重。

特征函数的权重可以使用最大似然估计法通过模型训练获得。对于序列标注，给定一个输入序列 X,模型用 Viterbi 算法求出以输入序列 X 为条件下具有最大条件概率的标记序列：

$$L(\lambda) = \sum_k \left[\log \frac{1}{Z(x^{(k)})} + \sum_j \lambda_j F_j(y^{(k)}, x^{(k)}) \right]$$

2.2 序列标注

图 2 是一个句子采用 CRFs 进行基本名词短语识别的例子。第一列是句子的单词，第二列是它们分别对应的词性，第三列是词的标注。其中第一列和第二列是已知数据，第三列在训练集中需给出，在测试集中未知。通过 CRFs 进行训练后可根据前面两列的信息预测出测试集中第三列的标注。

可以将 CRFs 预测序列标注的特点应用到短文本情感倾向性分析中。将短文本的每个词作为第一列，将短文本的情感倾向性作为标注作为第二列。如图 3 所示，每个词都标注为这个文本的类别房产，这样短文本就转化为一个标注后的序列，可以用于训练。测试的短文本只需给出每个词作为第一列，第二列文本类别为空，留待预测。

```
U.K.        JJ      B-NP
base        NN      I-NP
rates       NNS     I-NP
are         VBP     B-VP
at          IN      B-PP
their       PRP     B-NP
highest     JJS     I-NP
level       NN      I-NP
in          IN      B-PP
eight       CD      B-NP
years       NNS     I-NP
.           .       O
```

图 2

```
真的        正面
是          正面
非常        正面
好          正面
的          正面
想法        正面
值得        正面
进一步      正面
探讨        正面
```

图 3

2.3 特征模版

我们采用图 4 所示的特征模版来抽取特征：

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

# Bigram
B
```

图 4

```
...
真的        正面
是          正面
非常        正面>>current token
好          正面
的          正面
```

图 5

Unigram 是一元模版，定义的特征表示只与当前位置对应标签相关的特征。Bigram 是二元模版，定义了前一个位置和当前位置对应的标签相关的特征。

模板文件中(以#开头的注释和空格行除外)的每一行是一个模板。每个模板都是由%x[row,col]来指定输入数据中的一个 token。row 指定到当前 token 的行偏移，col 指定列位置。

如图 5 的一元模版中，当前 token 是“非常产”这个词时，%x[-2,0]就是 eight 的前两行，0 号列的元素（注意，列是从 0 号列开始的），即为“真的”。当前 token 为“正面”时，模版的每一行对应的扩展特征如表 1 所示。

表 1.

模版	扩展特征
%x[-2,0]	真的
%x[-1,0]	是

%x[0,0]	非常
%x[1,0]	好
%x[2,0]	的
%x[-1,0]/ %x[0,0]	是/非常
%x[0,0]/ %x[1,0]	非常/好

图 5 的二元模版中，‘B’表示模板自动产生当前 token 的标签和前一个 token 标签的合并。

当标签包正面、反面、中立这三种时，一元模版行"U02:%x[0,0]"对应的特征函数分别为：

func1 =

func2 =

func3 =

....

funcXX =

funcXY =

...

二元模版的特征函数为：

func1 =

func2 =

...

funcX =

...

用 L 表示输出标注的种类数目、N 表示模版生成的特征字符数目，则一元模版的特征函数有 L*N 种，二元函数的特征函数有 L*L*N 种。

链式 CRFs 的序列标注不仅保存了短文本的特征词，还保存了词之间的顺序关系，可以提高短文本情感倾向性分析的准确度，在这次微博情感倾向性测评中，得到了非常好的结果。

3 句法分析算法

针对微博语料的特殊性，本文对传统的句法分析算法做改进，下面给出流程图（图 6）。

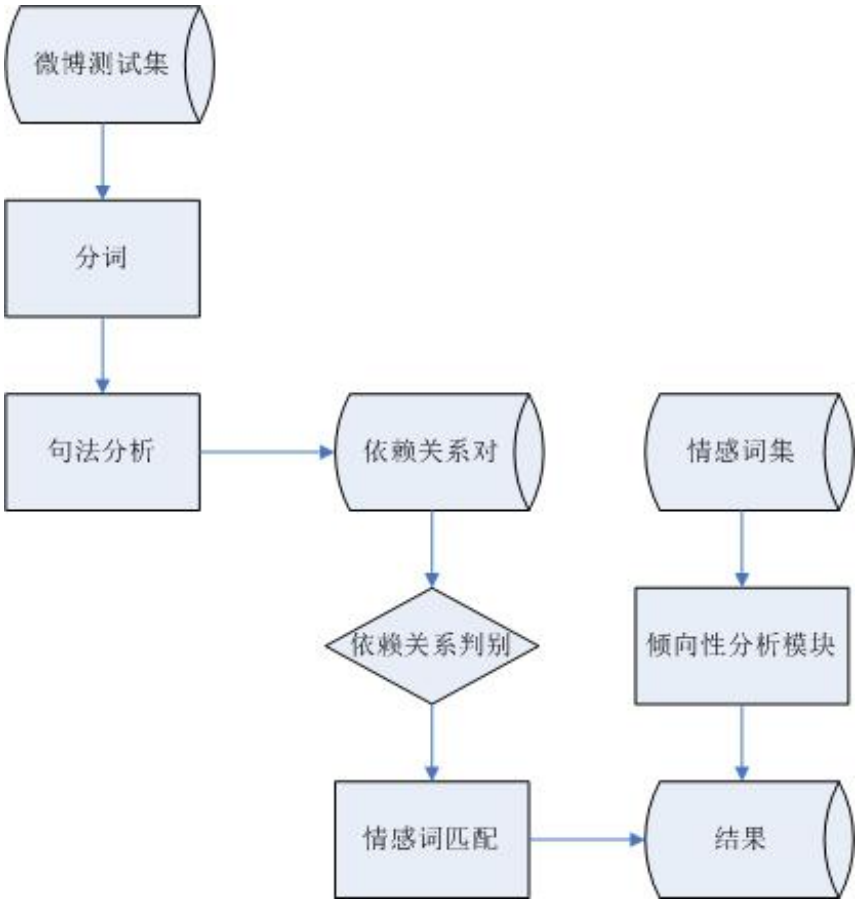

图 6.基于句法分析的微博情感分析流程图

首先对微博的测评预料进行分词，由于网络用语在微博中占了很大一部分，所以分词后还需要进行修正，将因为网络用语而分错的微博进行修正，其中，网络用语词表采用了新浪微博，腾讯微博的表情用语，并且整合了许多时下流行的网络用语。并且对这些网络表情，用语进行情感上的划分，区分褒贬，形成完善的情感词集。

句法分析器采用了 Standford Parser，从中提取每句微博的依赖关系对，考虑到微博预料特殊性。如果微博本身非常的短，根本不构成句法依赖关系，那么我们直接使用情感词匹配进行判别。

3.1 句法分析器

本文采用基于句法的情感倾向分析，使用了斯坦佛大学发布的 Standford Parser。Standford Parser 是由斯坦佛自然语言处理小组（Stanford NLP Group）研发的主要针对英语的句法分析器，但是逐渐在各种不同语言中完善。该句法分析器对分词后句子中的每一个词，进行词性标注，并且进行序号标记，遇到句号则从头开始标记，并且先对词进行词性标注，然后词与词之间产生依赖关系对。

对于中文的句法分析，Standford Parser采用了Pi-Chuan Chang ，Huihsin Tseng[7]等人针对中文句法分析定义的依赖关系，使用了基于类型的句法解析。下面列举一些本文使用到的依赖关系对。其中用nn，nsubj等符号表示两个括号中词之间的依赖关系。

表 2. 中文句法依赖关系

缩写	简短描述	中文例句	依赖类型
nn	复合名词	服务 中心	nn(中心，服务)
punct	标点	海关 统计 表明	punct(表明，,)
nsubj	名词词性主题	梅花 盛开	nsubj(盛开，梅花)
conj	和	设备 和 原材料	conj(原材料，设备)
dobj	直接对象	浦东 颁布 了 七十一 件 文件	dobj(颁布，文件)
advmod	状语修饰	部门 先 送上 文件	advmod(送上，先)
prep	介词短语修饰	在 实践 中 逐步 完善	prep(完善，在)
pobj	介词宾语	根据 有关 规定	pobj(根据，规定)
neg	负面修饰	以前 不 曾 遇到 过	neg(遇到，不)
comod	动词复合	颁布 实行	comod(颁布，实行)
amod	形容词修饰	跨世纪 工程	amod(工程,跨世纪)

表 2 中列举的是部分用到的依赖关系对，中文例句都来自参考文献。事实上，在很多情况下，Standford Parser 在进行句法分析后，得出的依赖关系往往不是如其简短描述中那么有局限性，比如：房间 还算 干净 整洁 。 这句话在处理后的依赖对为 nsubj(干净-3, 房间-1) advmod(干净-3, 还算-2)comod(干净-3, 整洁-4)。我们可以看到，当处理两个形容词并列关系时，该句法分析器也会将其归类为 comod 关系(动词复合关系)，可见 standford parser 对中文词性并没有很好的识别，但是另一方面也简化了关系的种类。本文在使用 Standford Parser 时，对其依赖关系对结果进行了总结归纳，针对句法分析器的分析结果特点进行了针对性的规则设置。用来识别评价语句的情感倾向。

3.2 基于句法分析的情感倾向性分析

一句完整的句子往往包含不同词性的词语，有名词，动词，形容词，副词等等，而且一句比较长的句子还会包含一些连接词，用来表达前后语句的关系，比如：但是，而且等等。比如：“距离/P 川沙/NR 公路/NN 较/AD 近/VA ,/PU 但是/CC 公交/NN 指示/NN 不对/VA ,/PU 如果/CS 是/VC “/PU 蔡陆线/NR ”/PU 的话/SP ,/PU 会/VV 非常/AD 麻烦/VA ./PU”

从这个句子中我们可以看到，其中“近”、“不对”、“麻烦”等词语具有明显的情感色彩，通过这些情感词语，我们可以很好判断这句评价语句的情感倾向性，其中“较”、“非常”等副词对他们对应的情感词“近”以及“麻烦”做了程度上的修饰，“但是”这个连接词也表达出了逗号区分开的两句短语句之间的转折关系。而这句句子中用到的依赖关系主要有这么几个：nsubj,nn,advmod,punct。

为了减少句法分析器错误依赖关系对的影响，我们定义窗口长度为 6，如果依赖关系对中的两个词距离大于 6，我们认为是没有效果的依赖对，不予考虑。

对于可以直接通过情感词以及依赖关系确定评价对象的情况：

1)：首先检索所有的依赖关系对中的情感词，如果情感词出现在 nsubj 关系对中，并且出现在关系对的左边，那么找到关系对右边的词语，这个词语必然是这个情感词语修饰的对象.

2)：找到这个修饰对象所在的关系对，是否存在 nn 的依赖关系，如果存在，那么 nn 依赖对中的两个词语一次合并成完整的修饰对象

3)：查找该情感词语的其他依赖关系对，如果存在 advmod 结构，并且是情感词语出现在关系对的左边，那么右边的词语就是修饰这个情感词语的副词，我们找到这个副词，并且做之后的副词程度匹配。

4)：继续寻找 advmod 依赖对，有时候往往会存在很多副词连续修饰一个情感词的情况，我们找到所有的修饰副词。

对于副词程度的匹配，本文定义了七种程度的副词，将“百分之百、倍加”等作为程度最强的副词，而将“轻度、相对”这种作为程度最弱的副词，而且设置了方面的副词，因为 Standford Parser 的 neg 依赖关系中会有不能匹配的方面副词以 advmod 的依赖关系出现。

5)：继续查找情感词语的依赖对，如果存在 neg 依赖关系，那么情感发生变化，生面的情感词语认为前面加了否定的副词修饰。

6)：检索是否存在 dobj 依赖对，如果存在，那么我们认为右边的词语是动宾结构的宾语，基于查找该宾语的依赖对中是否含有 nn 的结构，如果存在，那么 nn 依赖对中的词语将合并成为评价的对象，用于之后的匹配。

7)：检索是否存在比字结构，如果情感词出现在 prep 依赖对的左边，并且伴随着 pobj 依赖对的出现，我们认为这个情感词语修饰的是一个比较结构，pobj 依赖对的右边的词虽然不是直接形容评价对象的词语，但是，对于比较结构，一定是一个同类的评价对象，属于隐式的评价对象之一。

对于无法从以上依赖关系对中找到评价对象的情况，这里用两种方式来判别：

1)：对于短句中存在匹配到情感词语，但是无法用上面七条规则找到评价对象，我们向前寻找最近的标点符号，并且找到该标点的 punct 依赖对，并且有连接词出现在依赖对的左边，那么我们认为该情感词形容的评价对象是理它最近的一个之前的评价对象。

2)：对于短句中没有匹配到情感词汇，但是存在 punct 依赖对，并且连接词可以在递进，转折等连接词列表中匹配到，那么如果是转折关系，我们将之前最近的一个评价对象的倾向性做相反方向的处理，如果是递进关系则更加强调之前的结果。

4 测评结果

本文针对本次微博情感分析评测利用以上两种算法提交了两种结果，根据 2012 年CCF自然语言处理与中文计算会议发表的《中文微博情感分析测评结果》[8]。测评结果统计了正确率，召回率，以及F值。按照微平均和宏平均计算，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。CRFs算法提交的结果成绩优异，在正确率，召回率以及F值都获得了十分高的准确率。句法分析算法表现一般。

对于 CRFs 算法，本文对此次微博情感分析发布的 20 个主题的微博测评语料中一共选取了 1000 条语料作为训练集，并且对这 1000 条微博语料进行人工标注。经过公开评测后，其结果见表 3.

表 3. 基于 CRFs 算法的评测结果

微平均			宏平均		
正确率	召回率	F 值	正确率	召回率	F 值
0.853	0.743	0.794	0.854	0.745	0.794

对于句法分析算法，在微博情感测评的表现不理想。经过公开评测后，其结果见表4.

表4. 基于句法分析的评测结果

微平均			宏平均		
正确率	召回率	F 值	正确率	召回率	F 值
0.597	0.532	0.563	0.585	0.525	0.552

5 总结

从这次的测评结果看，CRFs 算法在短文本情感分析测评中效果显著，由于微博属于短文本，并且大部分的微博语句表达网络化，直白化，有时候往往一句表达观点的微博只有一个表情，或者几个直接的词。对于这种情况，句法分析算法表现得力不从心，大部分的微博无法形成有效的依赖关系。而本文针对这种情况使用词表直接匹配的方式并不能达到理想的效果。句法分析算法还是比较实用在句子结构比较完整，比较书面的中长语句中。

参考文献

[1] PangB, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification usingmachine learning techniques[C] / /Proceedings of the 2002 Conference on Emp iricalMethods in Natural Language Processing (EMNLP) , Morristown, NJ , USA: Association for Computational L inguistics, 2002: 79 - 86.

[2] Turney P.Thumbs up or Thumbs down? Semantic orientation app lied to unsupervised classification of reviews[C] / /Proceedings of the 40 th AnnualMeeting of the Association for ComputationalL inguistics (ACL) , Morristown, NJ , USA: Association for Computational L inguistics, 2002:417-424.

[3] X. Phan, L. Nguyen, S. Horiguchi. Learning to Classify Short and Sparse Text and Web with Hidden Topics from Large-scale Data Collections [C]//The International World Wide Web Conference Committee (IW3C2) (2008)

[4] Mehran Sahami, Timothy D. Heilman. A Webbased kernel function for measuring the similarity of short text snippets [C] //Proceedings of the 15th international conference on World Wide Web (2006)

[5] 滕少华, 基于 CRFs 的中文分词和短文本分类技术 [D] // 北京：清华大学 (2009)

[6] John Lafferty, Andrew McCallum, Fernando C.N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C] //Proceedings of the 18th International Conference on Machine Learning (2001)

[7] Pi-Chuan Changa, Huihsin Tsengb, Discriminative Reordering with Chinese Grammatical Relations Features / /aComputer Science Department, Stanford University, Stanford, CA 94305.bYahoo! Inc., Santa Clara, CA 95054

[8] 中国计算机学会中文信息技术专业委员会(CCF TCCI),中文微博情感分析测评结果[C]// 自然语言处理与中文计算会议(2012)

[9] Hanna M. Wallach. Conditional Random Fields: An Introduction [R] // Technical Reports, Department of Computer & Information Science, University of Pennsylvania (2004)