文章编号: 1006-2475(2016) 08-0027-05

基于词向量与句法树的中文句子情感分析

相若晨,孙美凤

(扬州大学信息工程学院,江苏 扬州 225009)

摘要: 随着互联网的快速发展,网络中充斥着海量主观性文本,如何对这些主观性语句进行情感倾向性判断是文本情感分析的关键。本文提出一种基于词向量和句法树的中文句子情感分析方法。针对目前大量网络新词的使用所带来的问题,以已有标注的情感词典为基础,采用词向量的方法判断词语之间的语义相似度,从而得到未知词语的情感极性。针对情感极性转移现象,定义相应的情感判断规则。在此基础上,利用句子的句法树结构,对句子进行情感倾向性分析。实验证明,该方法在一定程度上解决了网络新词的问题,有效提高了句子情感分析的准确率和召回率,且具有领域适用性。

关键词: 情感词典; 词向量; 句法树; 情感倾向性分析

中图分类号: TP391. 1 文献标识码: A doi: 10. 3969/j. issn. 1006-2475. 2016. 08. 006

Sentiment Analysis of Chinese Sentences Based on Word Embedding and Syntax Tree

XIANG Ruo-chen, SUN Mei-feng

(College of Information Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract: With the rapid development of Internet, the network is filled with a lot of subjective texts. How to judge the emotional polarity of these subjective statements is the key of the text sentiment analysis. In this paper, a method of sentiment analysis of Chinese sentences based on the word embedding and syntax tree structure is proposed. In view of the large number of network words, word embeddings are used to compute the semantic similarity between words, and the emotional polarity of the target word is gained. Some sentiment rules are defined for the phenomenon of emotional polarity transfer. Then, we judge the sentiment of sentences according to the syntax tree structure of the sentences. Experiments show that this method can solve the problem of the network words. Simultaneously, the precision and recall rate of the method are improved, and it also can be used widely in different domains.

Key words: sentimental lexicon; word embedding; syntax tree; sentiment analysis

0 引 言

情感分析又称为观点挖掘,旨在分析文本中所蕴含的情感信息,判断人们对某事物或事件的观点和态度。随着互联网的飞速发展和大规模普及,网络上的信息量每天都以指数级的速度增长,通过文本情感分析技术可以对网络中的海量信息进行有效利用,应用到网络舆情分析、企业市场分析、信息预测、实现情感机器人等多个领域。

情感分析自 2002 年由 Pang Bo [1] 提出之后,获得了很大程度的研究。目前的研究方法主要集中于基于情感词典的方法和机器学习的方法。

机器学习的方法把句子情感倾向性分析当作一个分类问题处理,文献 [2-5] 通过提取文本情感特征,采用支持向量机、朴素贝叶斯等机器学习的方法构造情感分类器,进而对待测文本进行分析。但这类方法的领域通用性普遍较差,且有监督的方法通常需要大规模的标注语料,人工标注需要耗费大量的时间和精力。

基于情感词典的方法将句子看成词语的集合,根据情感词典抽取文本中的情感词,由正负情感词的数量判断情感倾向,这种方法完全依赖情感词典,且没有考虑词序、句法和语义信息,导致情感分析的准确率较低。学者们在此基础上对基于情感词典的方法不断改进,引入句法分析的方法和基于规则的方法 [6-9] ,在很大程度上提高了情感判断的准确率。文献 [10] 提出了将句法分析与情感词典相结合的分析

收稿日期: 2016-01-26

作者简介: 相若晨(1990-) ,女,江苏南京人,扬州大学信息工程学院硕士研究生,研究方向: 文本情感分析; 孙美凤(1970-) ,女,江苏泰州人,副教授,硕士生导师,博士,研究方向: 对等网络应用,流量识别。

方法,利用情感词在句子中的成份、情感指数权重以及与其他情感词之间的组合共现关系计算出综合的情感指数,提高了情感分析的正确率,并且具有适用性。文献[11]提出了基于词典和规则集的中文微博情感分析方法,在不同的语言层次上定义了规则,结合情感词典对微博文本进行了从词语到句子的多粒度情感计算,并且验证了该方法的有效性。但这些方法同样依赖于情感词典,情感词典的不完善极大地限制了文本情感分析的性能。随着网络词语的流行和使用,依靠人工完善和扩充情感词典的方法显然是不可行的。

基于此,本文在情感词典的基础上,提出一种基于词向量和句法树的中文句子情感倾向性分析方法。首先利用神经网络训练的词向量来判断未知情感词与基准词之间的语义相似度,从语料自身挖掘词语之间的语义关系,摆脱对外部资源的依赖,能够很好地解决新词的情感判断问题。在此基础上,将句法分析与规则的方法相结合,根据句子的句法树结构和规则集,由词语的情感倾向性自底向上不断迭代,得到整个句子的情感倾向性,充分考虑了句子的句法和语法信息,提高了句子情感判断的准确性。

1 基于词向量的词典扩充

1.1 基础词典的构建

图 1 是本文构建的基础词典。正负情感词由 HowNet[12] 提供的情感词语集筛选后得到,并在此基础上进行了网络词语的扩充,对这些词语的情感极性进行了人工标注。程度词由 HowNet 提供的程度级别词筛选后得到。词典中还包括 13 个常用否定词和 17 个转折关系词。

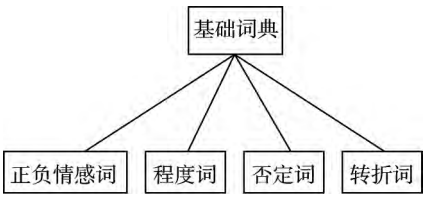

图 1 基础词典组成

1.2 基于词向量的词语倾向性判断

词向量最早是由 Hinton 等人在 1986 年的论文“Learning Representations by Back-propagating Errors”[13] 中提出的,但一直到 2000 年之后才逐渐开始被人重视。其基本思想是通过训练,将每个词语映射成一个实数向量,词向量之间的余弦距离表示词语之间的语义相似度。

词向量是语言模型训练的产物,本文使用的 Word2vec[14-16] 是一个将词表征成实数向量的高效工具,包括 CBOW 模型和 Skip-gram 模型。选用 Skip-gram 模型作为本文的向量训练模型。

Skip-gram 很好地解决了 n 元模型受窗口大小限制的问题,它允许跳过一些词,窗口内的词都会两两计算概率。同时,模型中的每个词向量表征了上下文的分布。图 2 是 Skip-gram 模型的示意图。

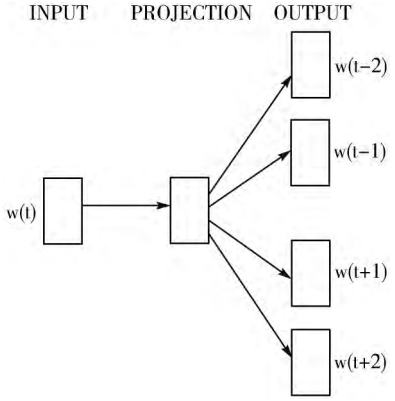

图 2 Skip-gram 模型示意图

Skip-gram 模型采用当前词 w(t) 训练上下文的词向量。假设存在一个词组序列 $w_1, w_2, w_3, \cdots, w_t$,Skip-gram 的目标是使公式(1)的值最大化:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j}\mid w_t)$$

其中,c 是窗口大小,T 是训练文本的大小。

本文利用词向量的距离判断句子中出现的新词的情感倾向性。将 1.1 节中构建的基础词典作为基准词,并获取其向量表示。根据公式(2)计算句子中出现的新词与基准词的语义相似度:

$$\text{similarity}(x_i, y) = \cos(x_i, y) = \frac{x_i \cdot y}{\|x_i\| \times \|y\|} \quad (2)$$

其中 x_i 为基准词的词向量,y 为目标词的词向量。

未知词语倾向性判断的具体步骤如下:

1) 获取基准词的向量表示 x_i。

2) 对待测句子进行分词后,将句子中每个词语与基础词典进行匹配,当出现新词时,获取其词向量 y。

3) 遍历所有基准词,计算目标词与基准词的 similarity 相似度。如果存在相似度大于 0.7,则将与目标词相似度最高的基准词的倾向性赋予目标词,并将目标词加入扩充词典;若不存在与目标词相似度大于 0.7 的词,则将目标词视为中性词。

2 基于句法树的句子情感倾向性分析

本文的句子情感倾向性分析模型是基于句子的句法二叉树结构的,利用词语之间的依赖关系和规则的方法,将句子情感判断转化为基于树的符号计算。

2.1　句法树的构造

本文采用目前开源中文句法分析器中比较具有代表性的 Stanford Parser 对句子进行句法分析。Stanford Parser 是由斯坦福大学自然语言处理小组开发的,一个高度优化的概率上下文无关文法和词汇化依存分析器。通过对句子进行句法分析,能够得到词语之间的依赖关系,借助所得到的二元依存关系,采用自底向上的方法逐层构建句子的二叉树结构。

句法树构造算法如下:

1) 输入第一个(i = 1) 词语 V,压入栈中保存。

2) 输入下一个词语(i ++) ,压入栈中。读取该句子的依存关系。

3) 若栈顶相邻的 2 个词语是依存对,则将它们生成父结点 F,父结点的序号 i 更新为支配词的序号。2 个子结点退栈,F 压入栈中保存。若此时栈中有词语,转步骤 3,没有转步骤 2。

4) 若栈顶相邻的 2 个词语不是依存对,则转步骤 2。

5) 当栈中压入最后一个父结点,则将句法树输出。

如句子“这家酒店环境还不错,但是服务特别差。”,图 3 是句子的分词结果,表 1 是由句法分析得到的依存关系。图 4 是句子的句法树结构。

这	家	酒店	环境	还	不错	，	但是	服务	特别	差	。
1	2	3	4	5	6	7	8	9	10	11	12

图 3　例句的分词结果

表 1　例句中词语的依存关系

支配词—从属词
这 - 1—家 - 2
酒店 - 3—这 - 1
环境 - 4—酒店 - 3
不错 - 6—环境 - 4
不错 - 6—还 - 5
差 - 11—但是 - 8
差 - 11—服务 - 9
差 - 11—特别 - 10
不错 - 6—差 - 11

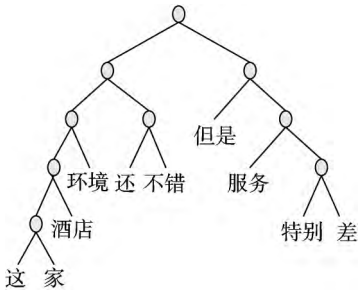

图 4　例句的句法二叉树结构

2.2　句子情感分析模型

在本文的模型中,句子中的每个词语对应为句法树的叶子结点,由 1.2 节的方法可以得到每个叶子结点的倾向性。根据 2.1 节的句法树结构,将相应的子结点组合成父亲结点,结合规则的方法判断父亲结点的倾向,并自底向上不断迭代,直至得到根结点的情感,即为整个句子的情感倾向性。图 5 是本文模型的示例图。

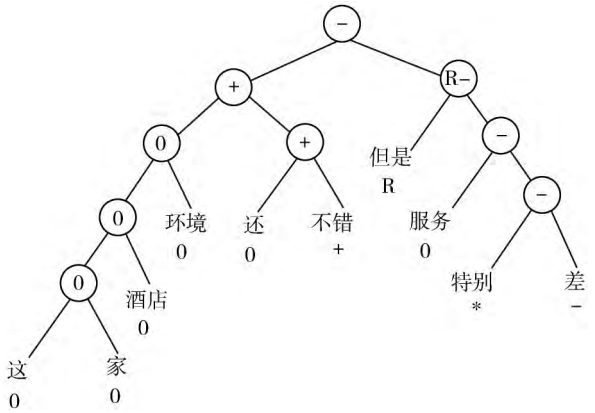

图 5　基于句法树的句子情感分析模型

在计算父节点情感标签时,本文定义了针对情感极性转移问题的处理规则。

情感极性转移是指由于一些特殊的语言结构,句子的整体情感与其中包含的词汇的情感极性不一致的现象。情感极性转移现象是影响句子情感分析准确性的重要原因。如句子“他并不是个好人”,“好人”是褒义情感色彩,但句子整体却是贬义情感。又如上述复句“这家酒店环境还不错,但是服务特别差。”,前半句是褒义情感色彩,后半句是贬义情感色彩,而在复句中,更强调转折词后的情感色彩,所以该句子所表达的对酒店的整体评价是贬义的。

因此,本文定义了如下规则集。为了便于本文方法的表达和实现,将词典中不同类型的词语用不同的符号进行标注。

1) 单句。

情感词　正面情感词标注为“+”,负面情感词标注为“-”。

中性词　中性词不影响句子情感,标注符号为“0”。

规则 1　如果其中一个子结点符号为 0 时,那么父结点符号即为另一子结点的符号。例　[0 +] ⇒ +。

程度词　程度词在句中起加强或减弱情感的作用,只能说明情感表达的程度,并不改变句子情感极性。程度词的符号为“* ”。

规则 2　如果其中一个子结点符号为* ,那么父结点符号即为另一子结点的符号。例　[* +] ⇒ +。

否定词　否定词是影响情感极性转移现象的主要因素之一，当否定词出现在情感词前时，会直接改变原来的情感极性。否定词符号为“!”。

规则 3　如果情感词为褒义倾向，则整体为贬义倾向；如果情感词为贬义倾向，则整体为褒义倾向。即：[+] ⇒ -；[-] ⇒ +。

规则 4　单句中，可能会出现左右结点分别为极性相反的符号，例如“我讨厌喜欢这本书的人”，根据句子的主谓结构，定义计算规则为：父结点的符号取左结点的情感符号，即：[+ -] ⇒ +，[- +] ⇒ -。

2) 复句。

转折词　在复句中，句子采用转折关系，一般是为了突出强调转折连词后的内容，即转折词后面才是说话人所要表达的真正情感。“虽然、尽管”等转折前接词符号为“L”，“但是、可是”等转折后接词符号为“R”。

规则 5　当遇到转折后接词时，保留符号至根结点处理，取转折符号“R”后面的情感符号。例：[- R +] ⇒ +。

规则 6　若句子中只有单一的转折前接词，则根结点情感极性与转折前接词后的情感相反。例：[L - +] ⇒ +。

3 实　验

3.1 基于词向量方法的验证实验

为了验证本文提出的基于词向量的词典扩充方法的有效性，人工标注了 200 个正负情感词作为基准词，获取与基准词相似度最高的 3 个词加入扩充词典，将这 3 个词的情感极性标注为该基准词的情感极性。对扩充词典中词语的情感极性进行人工标注，与根据词向量相似度得到的情感标注进行对比。

训练语料来源于 2012 版全网新闻数据(Sogou-CA)[17]，语料大小为 711 MB。使用 Stanford 分词器对文本进行分词，以空格隔开，利用分好词的训练语料，分别对 CBOW 模型和 Skip-gram 模型进行实验。训练维度为 50 ~ 200。实验结果如图 6 所示。

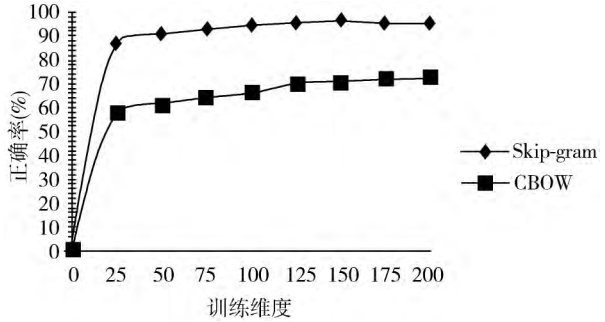

图 6　实验结果对比

从实验结果可以看出：1) Skip-gram 模型的准确率明显高于 CBOW 模型。CBOW 模型不考虑词语之间的先后顺序，因此在处理中文时语义准确率相对于 Skip-gram 较低。2) 使用 Skip-gram 模型，扩充词典中词语的情感极性准确率基本维持在 95%，训练维度为 150 时准确率最高。由此证明，利用词向量的相似度判断词语的情感倾向以及扩充词典的方法是有效的。

3.2 参数设置

本实验用来设置 1.2 节中未知词语倾向判断方法中的相似度的阈值。

随机选取 100 个词语，并对其进行人工标注。计算每个词语与基础词典中词语的相似度并排序，对 1.2 节的方法设置不同的阈值，将实验结果与人工标注结果对比，比较结果如表 2 所示。

表 2　不同阈值下词语倾向判断正确率

阈值	正确率
0.6	0.89
0.7	0.96
0.8	0.92

由实验结果可以发现，阈值设置为 0.7 时，正确率最高。

3.3 基于句法树的句子情感计算实验

3.3.1 实验数据

语料来自谭松波整理的中文情感挖掘语料——ChnSentiCorp[18]，包括酒店、电脑(笔记本)和书籍这 3 个领域的评论文本，共包含 12000 篇平衡语料，每个领域中正面和负面评价类各 2000 篇。筛选后选取 9000 个句子作为本文的测试语料，其中每个领域正面和负面评价各 1500 句。

3.3.2 评价指标

本文方法的评价指标为：准确率 P(Precision)、召回率 R(Recall)和 F 值(F-measure)。

$$P = \frac{\text{判断正确的该类句子个数}}{\text{返回结果为该类的句子个数}} \quad (3)$$

$$R = \frac{\text{判断正确的该类句子个数}}{\text{标准结果中该类句子的个数}} \quad (4)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (5)$$

3.3.3 实验结果及分析

为了验证基于词向量词典扩充的作用和本文句子情感倾向性分析模型的有效性，分别将本文方法与传统的情感词典方法以及目前最先进的方法进行对比，采用以下 4 种方法在相同的测试集上进行实验：

A：传统的情感词典方法；

B：加入基于词向量的情感词典方法；

C: Socher 提出的向量空间语义组合模型[19];

D: 本文基于词向量与句法树的方法。

实验结果如表 3 所示。

表 3　实验结果

	PosP	PosR	PosF	NegP	NegR	NegF
A	0.632	0.712	0.669	0.686	0.608	0.645
B	0.723	0.649	0.684	0.721	0.789	0.753
C	0.775	0.807	0.790	0.810	0.791	0.800
D	0.844	0.812	0.828	0.827	0.858	0.842

实验结果分析:

1) 相较于传统的情感词典的方法,结合基于词向量的词典扩充的方法后,准确率和召回率有所提高,说明基于词向量判断未知新词情感倾向性的方法有利于提高句子情感倾向性分析的准确率。

2) 本文方法的准确率、召回率和 F 值均高于其他 3 种方法,因为情感极性转移现象在中文句子中很常见,而传统的情感词典方法和 Socher 的模型都没有考虑该因素。

3) 本文方法的准确率和召回率都在 80% 以上,说明本文方法在句子情感倾向性分析任务中有较好的表现。

4) 测试集是在 3 个领域的语料中随机抽取的,说明本文提出的情感倾向性分析方法不受领域限制。

4　结束语

本文提出了基于词向量与句法树的句子情感倾向性分析方法,该方法在情感词典的基础上,利用词向量判断词语之间的语义相似性,解决了未知新词的情感倾向性判断问题。同时利用句法树结构结合语义规则判断句子的情感倾向性,并通过实验验证本文方法具有较高的准确率和召回率。但本文的方法只针对句子的褒贬倾向性分析,下一步的工作还需要在情感倾向性分析任务的基础上进行情感强度分析以及情感多分类研究。

参考文献:

[1] Pang Bo, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques [C] // Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002,10: 79-86.

[2] 王素格,杨安娜,李德玉,等. 基于支持向量机的文本倾向性分类研究 [J]. 中北大学学报(自然科学版), 2008,29(5): 421-425.

[3] Ye Qiang, Zhang Ziqiong, Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches [J]. Expert Systems with Applications, 2009,36(3-Part 2): 6527-6535.

[4] Eirinaki M, Pisal S, Singh J. Feature-based opinion mining and ranking [J]. Journal of Computer and System Sciences, 2012,78(4): 1175-1184.

[5] 刘鲁,刘志明. 基于机器学习的中文微博情感分类实证研究 [J]. 计算机工程与应用, 2012,48(1): 1-4.

[6] 党蕾,张蕾. 一种基于知网的中文句子情感倾向判别方法 [J]. 计算机应用研究, 2010,27(4): 1370-1372.

[7] 吴江,唐常杰,李太勇,等. 基于语义规则的 Web 金融文本情感分析 [J]. 计算机应用, 2014,34(2): 481-485.

[8] 冯时,付永陈,阳锋,等. 基于依存句法的博文情感倾向分析研究 [J]. 计算机研究与发展, 2012,49(11): 2395-2406.

[9] Gao Kai, Xu Hua, Wang Jiushuo. A rule-based approach to emotion cause detection for Chinese micro-blogs [J]. Expert Systems with Applications, 2015,42(9): 4517-4528.

[10] 肖红,许少华. 基于句法分析和情感词典的网络舆情倾向性分析研究 [J]. 小型微型计算机系统, 2014,35(4): 811-813.

[11] 王志涛,於志文,郭斌,等. 基于词典和规则集的中文微博情感分析 [J]. 计算机工程与应用, 2015,51(8): 218-225.

[12] 知网. HowNet [DB/OL]. http://www.keenage.com/, 2015-12-19.

[13] Rumelhart D E, Hintont G E, Williams R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-536.

[14] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space [C] // Proceedings of the 2013 International Conference on Learning Representations Workshop Track. 2013.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C] // Advances in Neural Information Processing Systems. 2013: 3111-3119.

[16] 熊富林,邓怡豪,唐晓晟. Word2vec 的核心架构及其应用 [J]. 南京师范大学学报(工程技术版), 2015,15(1): 43-48.

[17] 搜狗网. 全网新闻数据(SogouCA) [DB/OL]. http://www.sogou.com/labs/dl/ca.html, 2012-08-10.

[18] 谭松波. 中文情感挖掘语料—ChnSentiCrop [DB/OL]. http://www.nlpir.org/?action-viewnews-itemid-77, 2011-04-19.

[19] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces [C] // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1201-1211.