基于词汇语义和句法依存的情感关键句识别

冯 冲,廖 纯,刘至润,黄河燕
(北京理工大学计算机学院,北京 100081)

摘 要: 门户网站、博客和论坛中的新闻性文章往往都带有自己的情感倾向性,而情感关键句的识别对判断文章的情感倾向、了解社会动态和舆情状况有着非常重要的作用. 传统方法主要基于词汇特征,未能充分利用潜在的句法和语义信息. 本文提出了一种基于词汇语义和句法依存的情感关键句识别方法. 该方法首先通过构建情感词典和关键词词典获取词汇语义信息,然后利用一种新颖的面向情感关键句提取算法获取句法依存信息,最后把情感关键句的识别问题看成一个是否为情感关键句的二分类问题加以解决. 在 COAE2014 公开评测数据集上进行的实验表明本文方法的准确率和召回率均显著优于其他方法.

关键词: 情感关键句; 词汇语义; 句法依存; 支持向量机
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2016)10-2471-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.10.027

Sentiment Key Sentence Identification Based on Lexical Semantics and Syntactic Dependency

FENG Chong, LIAO Chun, LIU Zhi-run, HUANG He-yan
(*Beijing Institute of Technology, Beijing* 100081, *China*)

Abstract: A lot of news articles in the portal, blog and forums always have their own emotional orientations and sentiment key sentence identification plays an important role in distinguishing emotional orientation of one article, supervising social trends and public sentiment state. The traditional lexicon-based methods totally depended on lexical semantics and did not excavate the implied syntactic structure. So a hybrid method of sentiment key sentence identification based on lexical semantics and syntactic dependency is proposed in this paper. This approach first gets lexical semantics knowledge from emotion lexicon expansion and keywords lexicon construction, and then this paper proposes a novel dependency templates extraction algorithm for syntactic dependency information to build a dependency knowledge base, finally we regard sentiment key sentence identification as a classification task and perform identification through different groups of features. Experimental results on COAE2014 dataset show that this approach notably outperforms other baselines of sentiment key sentence identification on precision and recall.

Key words: sentiment key sentence identification; lexical semantics; syntactic dependency; support vector machine

1 引言

网络作为一种新型媒体不但成为各种社会思潮、利益诉求和意识形态较量的场所,而且也成为民众评议时政、谈论是非、交流观点的集散地. 抽取出一篇文章的情感关键句,对网络舆情监测和分析有着重要的作用. 情感关键句识别技术的研究目的在于自动从海量信息中抽取出与主题相关的情感关键句,富有学术意义和实用价值.

情感关键句,又常被称作主题情感句,主要包含两个要素:主题关键词和情感关键词. 主题关键词用来概括篇章的主题;情感关键词用来概括情感倾向. 目前,关于情感关键句识别方面的研究并不多. 林政、谭松波[1]等提出了一种情感关键句抽取算法,算法考虑句子的 3 类属性:情感属性、位置属性和关键词属性,并将抽取出的情感关键句分别用于有监督和半监督的情感分类,取得了不错的效果;2014 年,在中文信息学会主办的第六届中文倾向性评 COAE(Chinese Opinion Analysis Evaluation)的任务一中提出了面向新闻的情感关键句抽取与判定任务,要求在给定新闻集合(每篇文章已切成

收稿日期:2015-02-03;修回日期:2015-07-20;责任编辑:马兰英
基金项目:国家重点基础研究发展计划(No. 2013CB329605, No. 2013CB329303);国家自然科学基金重点项目(No. 61132009, No. 61201351);国家高技术研究发展计划 863 项目(No. 2015AA015404)

句子)中,判别每篇文章的情感关键句. 本文研究工作采纳此评测任务所提出的任务定义,并利用相同的公开数据集进行了实验和对比分析.

总体来看,情感关键句识别的研究尚不成熟,目前还处于起步阶段. 而中文语言的灵活性及表达的多样性,也使情感关键句识别的研究相对更加困难. 目前情感关键句识别的方法大多仅基于规则或仅基于统计,两者结合的尚不充分. 而且在抽取和分析过程中大都只利用到浅层分析,并未研究如何挖掘句子的深层信息. 本文将情感关键句的识别问题转化为情感关键句二元分类问题. 首先采用点间互信息(PMI)对情感词典进行扩充,从而得到领域性极强、召回率高的情感词典,并采用 LDA 和 TextRank 相结合的方法构建关键词词典. 然后,对文章中的句子进行过滤,保留含有情感词或关键词的句子,再对保留下来的句子进行依存分析,进一步挖掘句子的深层语义信息,并根据本文提出的依存模板提取算法构建依存模板知识库. 最后,将提取出的情感词的出现概率、关键词的 TextRank 得分、依存模板的出现概率、和位置特征按照一定的规则抽象成向量,利用 SVM 进行分类. 实验结果表明,本文方法在准确率和召回率上均大幅度超越了 COAE2014 评测公布的最佳成绩.

2 面向情感关键句的词汇语义分析

由于情感词和主题词是情感关键句的两个重要组成成分,因此我们通过情感词典扩充和关键词词典构建来获取词汇语义信息.

2.1 情感词典扩充

构建一个覆盖面大、精确率高的情感词典在近些年受到人们的普遍关注[2,3]. 目前,文本情感分析研究领域还没有一部完整且通用的情感词典. 本文采用知网 hownet① 和简体中文的 NTUSD② 构成基础情感词典,并对基础情感词典进行扩充. 扩展情感词典的方法主要有基于语义相似度[4,5]和基于同义词的方法[6,7]. 本文采用点间互信息(PMI),通过计算词语间的语义相似程度,构建出一个领域相关的情感词典 DEL(Domain-related Emotion Lexicon),其计算公式如下:

$$\mathrm{PMI}(w_1,w_2)=\log\left(\frac{P(w_1\&\ w_2)}{P(w_1)P(w_2)}\right) \tag{1}$$

式中 $P(w_1\&\ w_2)$ 表示 w_1 和 w_2 在同一个句子中共同出现的概率,$P(w_1)$ 和 $P(w_2)$ 分别表示两个词语单独出现的概率.

基于点间互信息 PMI 算法过程如下:

(1)对语料进行预处理,并按词性筛选出名词、动词和形容词作为候选词.

(2)分别计算上文的基础情感词典中每个词与这些候选词之间的点间互信息.

(3)对于基础情感词典中的每个词,选取前五个互信息高的词语,与其出现频率一起加入基础情感词典,生成最终的领域相关的情感词典 DEL.

2.2 关键词词典构建

所谓关键词词典 KL(Keywords Lexicon)构建,就是从一篇给定的文本中自动抽取出若干有意义的词语或词组,抽取方法既可以通过训练语料[8,9]构建模型实现,也可以借助于词语之间的关系直接从文本本身抽取. 关于无监督关键词抽取方法的研究,主流方法可归纳为三种:基于 TF-IDF 统计特征、基于主题模型[10,11]和基于词图模型[12~14]的关键词抽取方法. 本文首先提出了一种新的加权方法 PCFO,然后采用 LDA 和 TextRank[12]相结合的词图模型进行关键词抽取,构建关键词词典.

(1)PCFO:一种图模型混合加权方法

对于图中的任一结点 v 来说,其重要性得分由其相邻结点的贡献组成,而其本身的得分也将被转移到相邻结点. 通过观察发现,一个结点对相邻结点集的影响力主要可以分解为四个组成部分:位置重要性的影响力(position)、覆盖重要性的影响力(coverage)、频度重要性的影响力(frequency)和共现重要性(co-occurrence)的影响力. 因此,本文提出了一种新的图模型混合加权方法 PCFO.

令 w_{ij} 表示结点 v_i 和 v_j 的整体影响力权重,$\alpha,\beta,\gamma,\delta$ 分别表示这四类不同的影响力所占的比重,则两节点之间的权值可以设为以下形式:

$$w_{ij}=\alpha w_{\mathrm{pos}}(v_i,v_j)+\beta w_{\mathrm{cov}}(v_i,v_j)+\gamma w_{\mathrm{freq}}(v_i,v_j)+\delta w_{\text{co-occur}}(v_i,v_j) \tag{2}$$

其中 $\alpha+\beta+\gamma+\delta=1$.

(a)$w_{\mathrm{pos}}(v_i,v_j)$ 表示节点 v_i 的位置影响力传递到 v_j 的权重,计算公式如下:

$$w_{\mathrm{pos}}(v_i,v_j)=\frac{P(v_j)}{\sum_{v_t\in \mathrm{Out}(v_i)}P(v_t)} \tag{3}$$

其中,$P(v_j)$ 表示节点 v_j 的位置重要性得分,具体赋值方式如下:

$$P(v)=\begin{cases}\lambda, & v\text{ 所对应的词语在标题中出现}\\ 1, & v\text{ 所对应的词语在标题中不出现}\end{cases}$$

其中,λ 是一个比 1 大的数字,实验中,经过验证选择 $\lambda=1.5$.

(b)$w_{\mathrm{cov}}(v_i,v_j)$ 表示节点 v_i 覆盖影响力传递到 v_j 的权重,计算公式如下:

$$w_{\mathrm{cov}}(v_i,v_j)=\frac{1}{|\mathrm{Out}(v_i)|} \tag{4}$$

① http://www.keenage.com/html/c_index.html

② http://www.datatang.com/data/11837

其中,$|\mathrm{Out}(v_i)|$表示节点 v_i 的出度.

(c)$w_{\mathrm{freq}}(v_i,v_j)$表示节点 v_i 的频度影响力传递到 v_j 的权重,计算公式如下:

$$w_{\mathrm{freq}}(v_i,v_j)=\frac{f(v_j)}{\sum_{v_t\in \mathrm{Out}(v_i)} f(v_t)} \tag{5}$$

其中,$f(v_j)$表示节点 v_j 所代表的词语在文章中出现的次数.

(d)$w_{\text{co-occur}}(v_i,v_j)$表示节点 v_i 的共现影响力传递到 v_j 的权重,计算公式如下:

$$w_{\text{co-occur}}(v_i,v_j)=\frac{Co(v_i,v_j)}{\sum_{v_t\in \mathrm{Out}(v_i)} Co(v_i,v_t)} \tag{6}$$

其中,$Co(v_i,v_j)$表示节点 v_i,v_j 所代表的词语在一定窗口内共现的次数.

(2)考虑主题分布的词图模型构建

TextRank 的思想来源于 PageRank,通过把文本分割成若干组成单元并建立图模型,利用投票机制对文本中的重要成分进行排序. 但传统的词图模型中每个节点是以相等的概率随机跳转的,这种方法容易产生局部最优的情况. 因此,在随机游走的过程中考虑文章的主题分布,把每一个词属于特定主题的概率 $P(z|w)$ 作为该主题下词的随机跳转概率,即 $P(z_t|v_i)=P(z|w)$,$P(z|w)$由 LDA 模型求得. 因此,在考虑主题分布的词图模型中,根据 TextRank 的打分策略,按照式(7)给每一个词赋予一个不同主题下的得分:

$$R_{Z_t}(v_i)=\lambda\sum_{j:v_j\to v_i}\frac{w(v_j,v_i)}{\mathrm{Out}(v_j)}R_{Z_t}(v_j)+(1-\lambda)P(z_t|v_i) \tag{7}$$

其中,λ 是一个阻尼因子,表示每个节点都有 $1-\lambda$ 的概率随机跳转到图中的其他节点. $w(v_j,v_i)$表示节点 v_j 到 v_i 的边的权值(由上文 PCFO 方法求得),$\mathrm{Out}(v_j)$表示由 v_j 出发的所有边的权值之和,$P(z_t|v_i)$表示 v_i 节点所代表的词属于当前主题的概率. $R_{z_t}(v_i)$表示节点 v_i 在主题 z_t 下的得分,迭代上述式子,直到收敛.

最后按照式(8)对所有主题下的得分加权求和得到一个最终的得分,排序取排名较高的节点作为最终的关键词提取结果.

$$R(v_i)=\sum_{t=1}^{k} R_{z_t}(v_i)\times P(z_t|d) \tag{8}$$

其中,$R_{z_t}(v_i)$表示节点 v_i 在主题 z_t 下的得分,$P(z_t|d)$表示该篇文档属于主题 z_t 的概率,$R(v_i)$表示节点 v_i 的最终得分.

因此,关键词提取算法如下:

算法 1　关键词提取算法

```
输入:语料集 corpus
输出:每篇文档对应的关键词词典 KL
for doc in corpus:
  for sen in doc:
      分词,词性标注,去除停用词;
  for topic in doc:
      构建图模型 G=(V,E);
      按式(7)迭代计算每一个节点在特定主题下的得分;
  按照式(8)计算每一个节点的最终得分;
  按照最终得分对节点排序;
  KL=节点代表的词+最终得分
  return KL
```

3　面向情感关键句的依存句法分析

依存句法分析[15,16]主要通过语言单位的各个组成部分来体现句子中的结构信息. 依存关系文法中将每个句子的谓语动词作为一句话的中心,认为它可以支配其他成分而它本身是不受其他任何成分的制约的,其他所有被支配的成分都附属于其支配者并存在某种依存关系. 依存关系反映的是中心词和与其相互依存的附属词之间的语义依赖关系[17]. 例如句子“笔者认为这必将受到严厉惩罚.”,使用 LTP[18]进行依存分析结果如图 1 所示,该句中心词为“认为”,与中心词相依存的依存关系为“SBV”、“VOB”和“WP”关系.

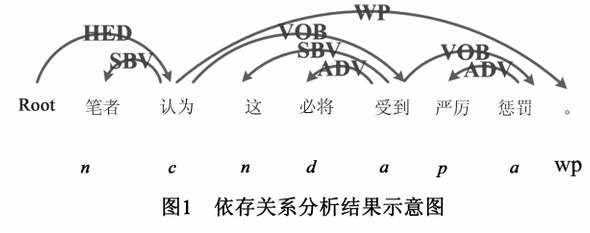

图1　依存关系分析结果示意图

面向情感关键句的依存模板提取算法如下:

算法 2　面向情感关键句的依存模板提取算法

```
输入:经过预处理得到的语料 T(情感关键句与非情感关键句),领域相关的情感词典 DEL,依存分析结果 DP
输出:依存知识库 DKB
for word in sentence of T:
  if word in DEL or Hownet advocating words:
      CoreWord += word;
  if word.relate == 'HED' in DP:
      CoreWord += word;
  for word in sentence:
      if word.parent in CoreWord and word.relation != WP:
          dpWords += word +word.relation
//笔者(SBV)认为(HED)受到(VOB)
//ForeRelations = SBV and BackRelations = VOB.
  for word in dpWords:
      if word.ip < CoreWord.ip:
          ForeRelations += relation
```

```
else:
    BackRelations += relation
// SBV + 认为 + VOB
        for forerelation in ForeRelations:
            for backrelation in BackRelations:
                template += forerelation + CoreWord + backrelation
//选取最终模板
for template in 情感关键句 and 非情感关键句:
    计算模板出现概率
    if frequency in 情感关键句 > 非情感关键句:
        Final_templates += template
DKB += Final_templates + frequencies
return DKB
```

4 利用 SVM 进行情感关键句识别

本文提出四种 SVM 的候选特征:情感词特征,关键词特征,依存模板特征和位置特征. 针对情感词、关键词和依存模板特征,分别选取领域相关的情感词典 DEL、关键词词典 KL 和依存知识库 DKB 中排名较高的前 n 位的得分,与该类特征的维数一起作为相对应部分的特征. 此外,由于中文文章的文章结构不外乎“总-分-总”、“分-总”、“总-分”、“分-分-分”,而上述第四种形式是非常少见的,因此有关作者主观情感及看法的句子,即情感关键句,一般都出现在文章的开头或结尾. 因此,针对位置特征[1],实验选择两种打分函数进行实验,第一种采用改进后的正态分布 Normal 形式,如下:

$$\text{score}_{\text{sen}}(\text{pos}(\text{sen})) = \frac{1}{\sqrt{2\pi\sigma}}\left(1 - e^{-\frac{(\text{pos}(\text{sen})-\mu)^2}{2\sigma^2}}\right) \quad (9)$$

其中 $\mu = \frac{n}{2}\text{pos}(\text{sen})$ 表示句子在文章中的位置.

第二种采用抛物线的形式,打分函数如下:

$$\text{score}_{\text{sen}}(\text{pos}(\text{sen})) = a \times \text{pos}(\text{sen})^2 + b \times \text{pos}(\text{sen}) + c \quad (10)$$

其中,$-\frac{b}{2a} = \frac{n}{2}$,$a > 0$,$b < 0$,$\text{pos}(\text{sen})$ 表示句子在文章中的位置.

5 实验与分析

5.1 实验系统和数据集

本文情感关键句识别的主要流程如图 2 所示:

(1)预处理:分词词性标注①、去除停用词.

(2)分别对句子进行词汇语义和句法依存分析:扩展情感词典、构建关键词词典,构建依存知识库.

(3)根据扩展后的情感词典和构建的关键词词典,按规则的方法对句子进行过滤,获取含有情感词和关键词的候选情感关键句.

(4)生成候选情感关键句的 4 种特征:情感词、关键词、依存模板和位置特征.

(5)使用 SVM 进行分类,判别一个句子是否是情感关键句.

本文数据集采用 COAE2014 公开的评测数据集,共包含 1994 篇文档. 使用 SVM 进行分类时,使用 4047 句情感关键句,以及非情感关键句 5000 句作为训练集;972 句情感关键句,以及 7325 句非情感关键句进行测试,并采用传统的准确率、召回率和 F 值对提取结果进行评价.

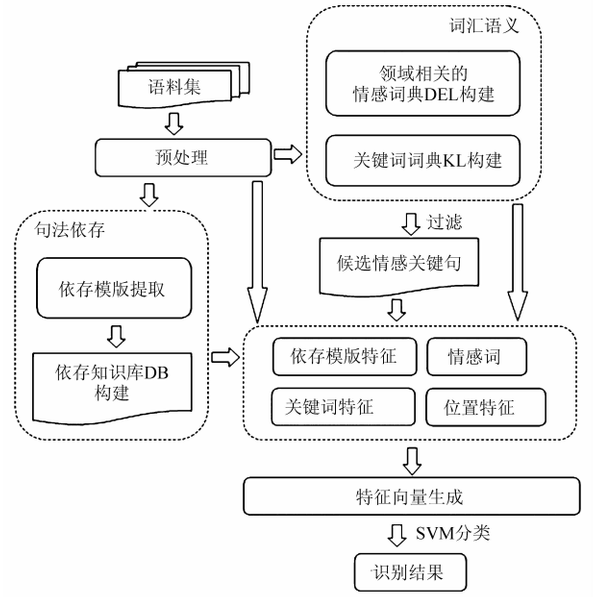

图2 情感关键句识别的方法流程

5.2 情感词典

情感词典作为情感关键句的重要特征,实验采用情感词典扩充前后在情感关键句中的覆盖率,验证其完整性和适应性. 通过实验发现,在 5119 句情感关键句中,对于扩展前的情感词典,出现情感词的情感关键句有 3721 句,覆盖率为 72%;而对于扩展之后的情感词典,出现情感词的情感关键句有 5019 句,覆盖率高达 98%,由此可见,情感词典的扩展在一定程度上大大提高了情感词的覆盖率,在一定程度上弥补了基础情感词典与领域不相关的不足. 但仅仅依赖情感词匹配的方法远远不能达到目的,情感词典要和其他方法相互配合才能达到更好抽取情感关键句的目的.

5.3 不同关键词词典构建方法的比较

关键词信息是情感关键句的一个重要元素,因此关键词提取效果将直接影响情感关键句抽取的准确率. 实验主要采用了四种关键词提取方法,分别采用三种不同的加权方法:距离的倒数、共现次数、PCFO 方法分别与传统的 Tf - idf 方法进行情感关键句识别. 结果

① http://www.ltp-cloud.com/

如表 1 所示.

表 1　不同关键词词典构建方法的比较

Methods	P(%)	R(%)	F(%)
Tf-idf	31.49	48.63	38.22
1/distance	32.51	49.57	39.26
Co-occurrence	33.84	50.10	40.39
PCFO	**36.29**	**52.35**	**42.87**

实验结果表明,本文提出的 PCFO 方法大大提升了情感关键句提取的效果. 这主要是因为本文采用 LDA 与 TextRank 相结合的方法,克服了传统图模型中随机游走的缺点,并采用 PCFO 方法综合考虑位置、覆盖、频度、共现四种影响力对图模型的权值进行修正. 为使 PCFO 方法达到最优,实验研究了 $\alpha,\beta,\gamma,\delta$ 四个参数对实验结果的影响. 实验采用 5 种不同的($\alpha,\beta,\gamma,\delta$)的组合,结果如图 3 所示,其中 1、2、3、4、5 分别代表(0,1,0,0)、(0.5,0.5,0,0)、(0.3,0.4,0.3,0)、(0.2,0.3,0.2,0.3)与(0.25,0.25,0.25,0.25)五种组合.

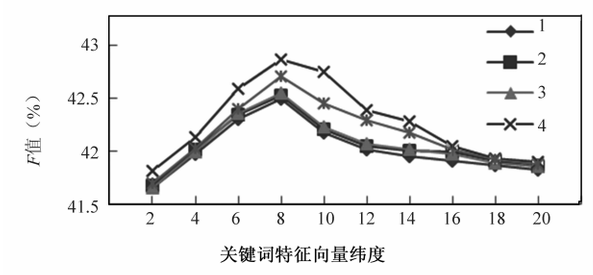

图3　不同($\alpha,\beta,\gamma,\delta$)组合与关键词向量维度下的情感关键句抽取的$F$值的比较

从图中可以看到,当选择 8 维作为 SVM 关键词向量维度,并使用第 4 种组合(0.2,0.3,0.2,0.3)时,实验效果最好. 在关键词向量维度选择上,过大的维度反而会降低分类能力;而在四种影响力的组合上,综合考虑位置、覆盖、频度、共现四种影响力的组合远比只考虑一部分的效果要好.

5.4　不同特征组合的比较

实验采用 4 种候选特征的不同组合加入 SVM 进行实验:情感词(Sentiment)与关键词(Keyword);情感词、关键词与依存模板信息(dp);情感词、关键词、依存模板信息与采用改进的高斯分布($P1$)进行打分的位置信息;情感词、关键词、依存模板信息与采用抛物线($P2$)进行打分的位置信息. 实验结果如表 2 所示.

表 2　不同 SVM 特征组合的比较

Methods	P(%)	R(%)	F(%)
Sentiment + Keyword	23.04	50.02	31.54
Sentiment + Keyword + dp	33.24	50.79	40.49
Sentiment + Keyword + dp + $P1$	35.13	51.76	41.85
Sentiment + Keyword + dp + $P2$	**36.29**	**52.35**	**42.87**

实验结果表明,依存分析大大提升了情感关键句识别的效果. 同时,位置特征部分使用抛物线形式优于改进的正态分布形式. 这主要是因为正态分布曲线在篇章首尾部分过于平滑,对篇章首尾部分的句子打分函数值变化不是很大,不能很好地体现出篇章首尾句子的重要性.

5.5　不同情感关键句识别方法的比较

本节比较了本文融合词汇语义和句法依存的方法(Lexicon + Syntax(Rules + Statistics))与其他四种基本方法:COAE2014 任务 1 的最好结果(COAE)、基于词汇(Lexicon[1])、人工标注 500 条数据作为训练集(COAE-500labelled)和去掉本文情感关键句识别流程中第三步,即不预先过滤掉一部分句子的方法(Lexicon + Syntax(Statistics)),实验结果如表 3 所示.

表 3　不同情感关键句识别方法的比较

Methods	P(%)	R(%)	F(%)
COAE	10.41	38.88	16.42
Lexicon[1]	12.18	29.13	17.19
COAE-500labelled	16.74	39.09	23.44
Lexicon + Syntax(Statistics)	30.70	50.79	38.27
Lexicon + Syntax(Rules + Statistics)	**36.29**	**52.35**	**42.87**

实验结果表明,融合了词汇语义和句法依存信息的方法大大提升了情感关键句识别效果. 而且,即使仅选择 500 条人工标注的句子进行实验,仍然取得了比 COAE 和基于词汇方法更高的效果. 另外,当使用情感词典、关键词词典对语料进行规则过滤的时候,相当于一个降噪的过程,以保证达到更高的准确率 P、召回率 R 和 F 值.

6　总结和未来工作

本文提出了情感关键句识别的新思路,将其看作一个二元分类问题,通过情感词典的扩充与关键词词典的创建,首先对所有文章中的句子进行规则过滤,然后选择情感词、关键词、依存模板和位置特征,利用 SVM 分类器完成识别. 实验结果表明,该方法显著优于前人方法.

然而,有些问题还有待更深入的研究,下一步工作中将重点探究如下问题:(1)考虑对句子进行短语结构分析,并将其与依存分析相结合,共同为情感关键句的识别服务;(2)对现有的依存模板进行同义词扩展,改进依存关系提取算法,尝试提出更具普遍意义的依存关系提取算法;(3)在汉语情感关键句语料库的建设上,统计构建一个更大规模的情感关键句集合势在必行.

参考文献

[1] 林政,等. 基于情感关键句抽取的情感分类研究[J]. 计

算机研究与发展,2012,49(11):2376 – 2382.
Zheng Lin,et al. Sentiment classification analysis based on extraction of sentiment key sentence[J]. Journal of Computer Research and Development,2012,49(11):2376 – 2382. (in Chinese)

[2] E Riloff,et al. A corpus-based approach for building semantic lexicons[A]. In Proceedings of the second conference on empirical methods in natural language processing [C]. eprint arXiv:cmp-lg/9706013,1997. 117 – 124.

[3] V Hatzivassiloglou,et al. Predicting the semantic orientation of adjectives[A]. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics[C]. Spain:Universidad Nacional de Educaci6n a Distancia,1997. 174 – 181.

[4] Turney,P D,et al. Measuring praise and criticism:Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS),2003,21(4),315 – 346.

[5] 朱嫣岚,等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报,2006,20(1),14 – 20.
Zhu Yanlan,et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese information processing,2006,20(1):14 – 20. (in Chinese)

[6] 田久乐,等. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版),2010,28(6):602 – 608.
Tian Jiule,et al. Words similarity algorithm based on tongyiciCilin in semantic web adaptive learning system[J]. Journal of Jilin University(Information Science Edition),2010,28(6):602 – 608. (in Chinese)

[7] 王素格,等. 基于同义词的词汇情感倾向性判别方法[J]. 中文信息学报,2009,23(5):68 – 74.
Wang Suge,et al. A synonyms based word sentiment orientation discriminating[J]. Journal of Chinese information processing,2009,23(5):68 – 74. (in Chinese)

[8] Frank,E.,et al. Domain – specific keyphrase extraction [A]. In Proceedings of 16th International Joint Conference on Artificial Intelligence[C]. New York:Association for Computing Machinery,1999. 668 – 673.

[9] Turney,P D. Learning algorithms for keyphrase extraction [J]. Information Retrieval,2000,2(4):303 – 336.

[10] Blei,D M,et al. Latent dirichlet allocation[J]. The Journal of machine Learning research,2003,3:993 – 1022.

[11] Pasquier,C. Task 5:Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation [A]. In Proceedings of the 5th international workshop on semantic evaluation[C]. USA:Association for Computational Linguistics,2010. 154 – 157.

[12] Liu,Z,et al. Automatic keyphrase extraction via topic decomposition[A]. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing [C]. USA:Association for Computational Linguistics,2010. 366 – 376.

[13] Page,L,et al. The PageRank citation ranking:Bringing order to the web[J]. Stanford Infolab,1999,9(1):1 – 14.

[14] Mihalcea,et al. TextRank:Bringing order into texts[A]. In Proceedings of Empirical Methods in Natural Language Processing[C]. Stroudsburg:Association for Computational Linguistics,2004. 404 – 411.

[15] Hermjakob,U. Parsing and question classification for question answering[A]. In Proceedings of the workshop on Open-domain question answering-Volume 12[C]. Stroudsburg:Association for Computational Linguistics,2001. 1 – 6.

[16] Baoshun Hu,et al. An answer extraction algorithm based on syntax structure feature parsing and classification[J]. In Chinese journal of computers,2008,31(4):662 – 676.

[17] Li xin,et al. Learning question classifiers[A]. In Proceedings of the 19th international conference on Computational linguistics-Volume 1[C]. Stroudsburg:Association for Computational Linguistics,2002. 1 – 7.

[18] Che Wanxiang,et al. Ltp:A chinese language technology platform[A]. In Proceedings of the 23rd International Conference on Computational Linguistics:Demonstrations [C]. Stroudsburg:Association for Computational Linguistics,2010. 13 – 16.

作者简介

冯 冲(通信作者) 男,1977 年生于河南驻马店,北京市海量语言信息处理和云计算应用工程研究中心副研究员,主要研究方向为社会媒体处理、机器翻译、信息抽取等自然语言处理相关领域.
E-mail:fengchong@ bit. edu. cn

廖 纯 女,1990 年生于河南驻马店,北京理工大学计算机学院硕士研究生,主要研究方向为社交网络,评价对象评价词抽取,情感倾向性分析.
E-mail:cliao@ bit. edu. cn