DOI: 10.11992/tis.201806006
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180716.1113.002.html

半监督自训练的方面提取

曲昭伟[1], 吴春叶[1], 王晓茹[2]
(1. 北京邮电大学 网络技术研究院, 北京 100876; 2. 北京邮电大学 计算机学院, 北京 100876)

摘 要: 方面提取是观点挖掘和情感分析任务中的关键一步, 随着社交网络的发展, 用户越来越倾向于根据评论信息来帮助进行决策, 并且用户也更加关注评论的细粒度的信息, 因此, 从海量的网络评论数据中快速挖掘方面信息对于用户快速决策具有重要意义。大部分基于主题模型和聚类的方法在方面提取的一致性上效果并不好, 传统的监督学习的方法效果虽然表现很好, 但是需要大量的标注文本作为训练数据, 标注文本需要消耗大量的人力成本。基于以上问题, 本文提出一种基于半监督自训练的方面提取方法, 充分利用现存的大量未标签的数据价值, 在未标签数据集上通过词向量模型寻找方面种子词的相似词, 对每个方面建立与数据集最相关的方面表示词集合, 本文方法避免了大量的文本标注, 充分利用未标签数据的价值, 并且本文方法在中文和英文数据集上都表现出了理想的效果。

关键词: 方面提取; 词向量; 半监督; 自训练; 未标签数据; 观点挖掘; 种子词; 相似词
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2019)04−0635−07

中文引用格式: 曲昭伟, 吴春叶, 王晓茹. 半监督自训练的方面提取 [J]. 智能系统学报, 2019, 14(4): 635–641.
英文引用格式: QU Zhaowei, WU Chunye, WANG Xiaoru. Aspects extraction based on semi-supervised self-training[J]. CAAI transactions on intelligent systems, 2019, 14(4): 635–641.

Aspects extraction based on semi-supervised self-training

QU Zhaowei[1], WU Chunye[1], WANG Xiaoru[2]
(1. Institute of Network Technology, Beijing University of Posts and Telecommunication, Beijing 100876, China; 2. College of Computer Science, Beijing University of Posts and Telecommunication, Beijing 100876, China)

Abstract: Aspect extraction is a key step in opinion mining and sentiment analysis. With the development of social networks, users are increasingly inclined to make decisions based on review information and pay more attention to the fine-grained information of comments. Therefore, it is important to help users to make these decisions by quickly mining information from massive comments. Most topic-based models and clustering methods do not work well in terms of consistency in aspect extraction. The traditional supervised learning method works well, but it requires a large amount of annotation text as training data, and labeling text requires a lot of labor costs. Based on the above issues, a method for aspects extraction based on semi-supervised self-training (AESS) is proposed in this paper. The method takes full advantage of the large amount of unlabeled data that exist in the web. Words similar to seed words on the unlabeled datasets using a word vector model are found, and multiple aspects word sets that are most related to the data set are constructed. Our approach avoids a large number of text annotations and makes full use of the value of unlabeled data, and our method has made good performance in both Chinese and English datasets.

Keywords: aspect extraction; word vector; semi-supervised; self-training; unlabeled data; opinion mining; seed words; similar words

随着互联网的发展, 用户逐渐借助网络平台来发表自己对于产品和服务的意见。这些评论往往由句子组成的短文本的形式出现, 涉及产品的

收稿日期: 2018−06−02. **网络出版日期**: 2018−07−17.
基金项目: 国家自然科学基金项目 (61672108).
通信作者: 曲昭伟. E-mail: zwqu@bupt.edu.cn.

一个或者多个方面意见。因此句子级别的观点挖掘[1]任务一直是研究的热点。观点由 4 个元素组成：方面、持有者、观点内容及情感。这四者之间所存在的联系为:观点的持有者针对某一方面发表了具有情感的观点内容。方面提取[2]是观点挖掘任务的子任务之一。简短专注于观点句子中的实体部分提取，例如："口味很好，服务周到，值得推荐"。这里的"口味"和"服务"就是方面术语，在方面提取中又涉及两个子任务：1) 提取评论句子的黄金方面也叫做主体方面，是评论中各个方面表示的总称，例如美食评论包括黄金方面"食物""服务"等，黄金方面"食物"包含多种多个方面表示词如"味道、口味"等；2) 从评论语料库中学习所涉及的方面表示词。

近年来，潜在狄利克雷分布 (LDA)[3]及其变种[1,4-5]已成为用于方面提取的主导无监督方法。LDA 将语料库建模为主题 (方面) 的混合，并将主题作为词类的分布。尽管基于 LDA 的模型发现的各个方面可能会很好地描述一个语料库，但发现提取出的各个方面质量差，通常由无关或松散相关的概念组成。因为基于 LDA 的模型需要估计每个文档的主题分布，但是评论语句通常是由句子组成的短文本，对于主体分布的评估造成困难，所以效果不好。

监督学习的方法是近年来流行的研究方法，深度学习中卷积神经网络 (convolutional neural network, CNN)[6-7]被应用于方面提取任务,并展现出卓越的效果。文献 [8] 提出一种基于 7 层深度卷积网络的模型来对句子进行标记训练,从而对评论进行方面提取，而且效果比较理想。然而，监督学习需要大量的标注文本作为训练数据,人工标签成本太高，而且主观性太强。

本文利用提出了一种新的半监督自训练的方法确定黄金方面[9]后利用少量标注的方面种子词，在未标签数据集上提取方面表示相似词，建立多个方面表示词集合，解决方面表示词集问题，利用丰富的方面词集合来识别文本的，能够避免大量的人工标注，并且本文方法在实际中文数据集和英文数据集上都产生了理想的效果。对实际的数据集进行了如下 3 个方面的研究：

1) 根据研究评论数据，发现评论的针对性很强，基本是针对某项产品或者服务给出自己的体验和建议。而且数据结构具有鲜明的产品特色，句子语言简短观点明确，会经常使用到明显的方面表示词来发表意见。

2) 评论往往涉及一个或者多个方面，以下一个简单的例子来自美团网 (http://bj.meituan.com/meishi/) 美食评论数据来说明研究的意义。例如："口味清淡，服务员态度很好，就是价格有点贵"。这句评论涉及了对餐厅食物的"口味""服务"以及"价格"3 个方面的评价，而且对于不同的方面给出了不同的意见。采用方面表示向量来对涉及的方面进行向量表示。方面提取作为观点挖掘的第一步，来确定评论涉及的多个方面。

3) 考虑到评论中涉及的含蓄表达，例如："还挺好吃的，排队等了半小时，不过还是很好吃"。句子中并没有明确的方面表示词，但是根据关键词"好吃"可以确定是针对食物的方面意见。针对这种没有明确方面表示名词的提取方面形容词来识别方面。

基于以上的研究，本文提出的半监督自训练方法能够确定方面表示词并且自动对评论进行方面识别。首先通过计算数据集中每个词的 TF-IDF 值，确定数据集的黄金方面，进一步从部分标签数据中获取方面表示种子词，利用词向量模型在实际未标签的数据集中寻找相似词，获得的与种子词相似的方面表示词，补充到对应的黄金方面词典里，扩充方面词典。并对目标文本进行方面识别，得到文本的方面表示向量。

1　相关工作

方面提取[10]是观点挖掘任务的基础性工作，在过去的十几年间，许多学者已经在方面提取上做了很多研究工作。主要专注于两个方向的研究，无监督和有监督方法方面提取过程可看作一个文本序列标注问题,因此可利用带监督的统计模型对序列标注过程进行训练从而提取句子的方面表示。适用此问题的典型带监督学习的方法有隐马尔可夫模型 (hidden Markov model, HMM)[10]、条件随机场模型 (conditional random rield, CRF)[11]等，文献 [10, 12-14] 采用一种编入词汇的 HMM 模型来提取显式方面。 最近，提出了不同的神经模型[15-16]，以自动学习基于 CRF 的方面提取的特征。但是监督学习需要大量的标签数据作为训练集，数据标注需要耗费大量的人力成本。

无监督的学习方法可以避免标签依赖问题。潜在狄利克雷分布 (LDA)[3]已成为方面提取的主导无监督方法。LDA 将语料库建模为主题 (方面) 的混合，并将主题作为词类的分布。虽然基于 LDA 的模型挖掘到的各个方面得到一个很好的描述，但提取出的各个方面质量不好，通常由无关或松散相关的概念组成。因为基于 LDA 的

模型需要估计每个文档的主题分布, 但是评论语句通常是由句子组成的短文本, 对于主体分布的评估造成困难。文献 [17] 以 Apriori 算法为基础采用关联规则挖掘方法找出频繁出现的名词并名词短语作为候选方面, 然后, 将错误的词语通过剪枝算法进行过滤, 最终形成方面集合。大多数方法是基于 LDA, 文献 [18] 提出了一个生成共现词对的主题模型 (BTM)。

半监督模型在方面提取中既避免了大量的文本标注, 也可以利用数据的内部大量的信息来进行方面提取。文献 [19] 提出了两个半监督模型: SAS(seeded aspect and sentiment model) 和 ME-SAS(maximum entropy-SAS model)。SAS 是个混合主题模型在提取方面后提取观点内容,后者将方面与内容联合提取。同时, 模型中加入相应种子词汇, 但是可移植性较差。

本文提出的基于半监督自训练的方法进行方面提取, 不仅避免监督学习中的大量的标签数据依赖问题; 而且, 解决了无监督主题模型中存在的短文本的方面提取结果不稳定的问题, 并且在中文和英文数据集上都产生很好的性能。

2　半监督自训练模型的构建

2.1　半监督自训练模型

自训练的过程中从未标签的训练数据集上学习到贴近数据集合的方面表示词。首先计算数据集单词的 TF-IDF, 并对结果进行排序以便确定黄金方面。对于确定的黄金方面结果, 随机选择少量的数据进行人工方面标注, 从标注结果中选取黄金方面表示词作为方面表示种子词。基于方面种子词, 利用词向量模型进行方面表示词学习, 扩充方面表示词集合。利用新生成的集合对部分标签数据进行方面识别验证, 并生成对应的方面向量。直到交叉验证的结果的正确率不再上升, 则得到了最终的方面表示词集合。自训练模型架构如图 1 所示。

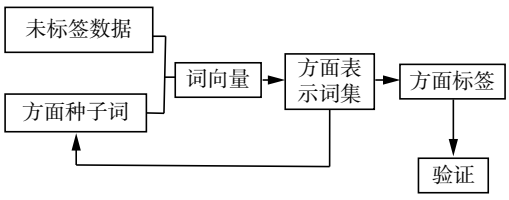

图 1　自训练模型架构

Fig. 1　Self-training model architectural overview

2.2　黄金方面确定方法

为了获得数据集上的黄金方面, 计算了数据集中单词的 TF-IDF 值。用以评估一个单词对于一个语料库中的其中一份文档的重要程度。单词的重要性随着它在该文档中出现的次数增加, 但同时会随着它在语料库中出现的频率下降。其中词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的次数, 逆向文件频率 (inverse document frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的 IDF, 可以由总文件数目除以包含该词语之文件的数目, 再将得到的商取对数得到。对于在某一特定文件里的词语 t_i 来说, 它的重要性可表示为

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

式中: $n_{i,j}$ 是该词 t_i 在文件 d_j 中出现的次数, 而分母是在文件 d_j 中所有单词出现次数之和。逆向文件频率是一个词语普遍重要性的度量。某一特定词语的 IDF, 可以由总文件数目除以包含该词语之文件的数目, 再将得到的商取对数得到:

$$idf_i=\log\frac{|D|}{\left|\{j:t_i\in d_j\}\right|+1} \tag{2}$$

式中: $|\boldsymbol{D}|$表示预料库的总文件数, $\left|\{j:t_i\in d_j\}\right|$ 表示包含词语 t_i 的文件数目, 分母加上 1 避免分母为 0。因此, 单词 t_i 在文档 d_j 中的 TF_IDF 值可以表示为

$$\text{TF\_IDF}_{i,j}=tf_{i,j}\times idf_i \tag{3}$$

2.3　方面表示词集的建立

为了获得方面表示词集合, 引入词向量模型, 利用模型学习到与数据集相关的丰富准确的方面表示词集。在每次方面词的学习过程中计算与已经确定的词集合的相似性, 保留每个方面不重复的前 10 个词, 扩充方面词典, 学习的过程在交叉验证结果的正确率下降前停止。这是一个自训练的过程, 来确定方面表示词集。因此, 构建词向量[20] 模型是非常重要的。

为了构造单词向量, 引入函数 $\boldsymbol{f}(w_{t-n+1},w_{t-n+2},\cdots,w_t)$来拟合单词序列的条件概率 $p(w_t|w_1,w_2,\cdots,w_{t-1})$。通过引入连续的单词向量和平滑概率模型, 可以在连续空间中获得序列概率, 从而减轻稀疏性和维数灾难。该模型如图 2 所示, 其中 $1\leqslant i\leqslant n$,$1\leqslant j\leqslant n$(n 表示黄金方面数)。线性嵌入层帮助 $N-1$ 个单词向量通过共享的 $\boldsymbol{D}\times\boldsymbol{V}=\boldsymbol{C}$ 映射到具有分布向量的 $N-1$ 个向量, $\boldsymbol{V}$ 是字典的大小, $\boldsymbol{D}$ 是嵌入向量的维数。需要学习的单词向量存储在矩阵 $\boldsymbol{C}$ 中。前向反馈神经网络 g 由 tanh 隐藏层和 softmax 输出层组成。在网络上, 由嵌入层产生的 $N-1$ 个单词向量被映射到长度为 $\boldsymbol{V}$ 的概率分布向

量。基于上下文的字典中词的条件概率可以估计为

$$p(w_i|w_1,w_2,\cdots,w_{t-1}) \approx f(w_i,w_{t-1},\cdots,w_{t-n+1}) = g(w_i,c(w_{t-n+1}),\cdots,c(w_{t-1})) \tag{4}$$

模型参数可以通过最小化交叉熵规则化损失函数来拟合：

$$L(\theta) = \frac{1}{T}\sum \log f(w_t,w_{t-1},\cdots,w_{t-n+1}) + R(\theta) \tag{5}$$

模型参数 θ 包括嵌入层矩阵 $\boldsymbol{C}$ 的元素和反向传播神经网络模型 g 中的权重。这是一个巨大的参数空间。

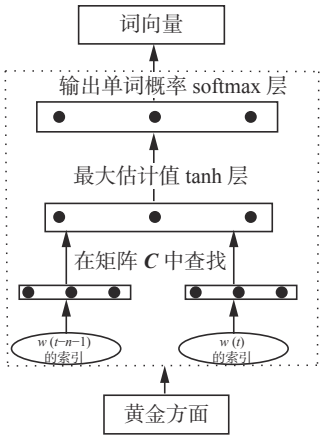

图 2　词向量生成模型

Fig. 2　Word embedding generation model

为了从上下文中预测目标词的过程中学习词向量。Skip-gram[21-22] 模型的正向计算在数学上构造为

$$p(w_o|w_i) = \frac{e^{U_o\cdot V_i}}{\sum_j e^{U_j\cdot V_i}} \tag{6}$$

式中：$\boldsymbol{V}_i$ 是嵌入层矩阵中的列向量，也称为 $\boldsymbol{w}_i$ 的输入向量；$\boldsymbol{U}_j$ 是 softmax 层矩阵的行向量，也称为 $\boldsymbol{w}_i$ 的输出向量。因此，Skip-gram 模型是计算输入词的输入向量与目标词的输出向量之间的余弦相似度，最后对 softmax 进行归一化。学习的模型参数是这些词向量。但是，计算词汇的相似性并将其直接标准化是一项耗时的任务。

考虑到上下文，目标单词属于一个子集的概率服从以下逻辑回归函数分布：

$$p(w_i \in D_1|\text{context}) = \frac{1}{1+e^{-U_{D_{\text{root}}}\cdot V_{w_t}}} \tag{7}$$

式中 $U_{D_{\text{root}}}$ 和 V_{w_t} 都是模型的参数。

划分子集并重复上面的过程，使用二叉树可以获得 $\log V$ 的深度。叶节点逐一对应原始字典词，非叶节点对应于类似类集的词。从根节点到任何一个叶节点只有一条路。同时，从根节点到叶节点的方式是随机游走过程。因此，可以基于这个二叉树来计算叶节点出现的概率。例如，如果样本及其二叉树中的目标词编码为 $\{1,0,1,\cdots,1\}$，则似然函数为

$$\begin{aligned} &p(w_t|\text{context}) = p(D_1 = 1|\text{context}) \\ &p(D_2 = 0|D_1 = 1)\cdots p(w_k|D_k = 1) \end{aligned} \tag{8}$$

每个项目是方程中的逻辑回归函数，并且可以通过最大化似然函数来获得非叶节点向量。一个词条件概率的变化会影响其路径中所有非叶节点的概率变化，间接影响其他词汇出现在不同程度的条件概率。因此，为了准确地获得方面词的向量，n(黄金方面数) 个 skim-gram 模型构建并输入已经由模型上一次迭代产生的数据。该模型的 n 个部分分别基于 n 个语料库构建词向量，避免了词向量的交互。单词嵌入可以很容易地获得类似的单词。

3　实验

本节描述实验的建立过程，以及本文实验在实际的中文和英文数据集的效果，并且与已有的经典方法进行对比，最后对实验结果进行了分析评估。

3.1　实验的建立

数据集：采用一个中文数据集和一个英文数据集来评估本文方法。对于中文数据集，爬虫从美团网获取的 71 万条美食评论。英文的公共数据集 Citysearch corpus 是一个餐馆评论语料库，以前的研究文献 [5, 23-24] 也广泛使用，其中包含超过 5 万个来自 Citysearch New York 的餐厅评论。文献 [23] 还提供了一个从语料库中手工标记的 3 400 个句子的子集。这些有标签的句子用于评估方面识别。有 6 个手动定义的方面标签：Food、Staff、Ambience、Price、Anecdotes、Miscellaneous，数据集分布见表 1。

表 1　数据集描述

Table 1　Data set description

数据集	训练集	测试集	总数
Citysearch corpus	279 859	1 490	281 349
Chinese	700 000	10 000	710 000

数据预处理：为了获得中文方面表示词的集合，将随机选择的 1 500 份美食评论平均分成 5 组，5 位评估者被要求按照涉及的方面进行手动标注评论。所有的中文评论都被分词工具 jieba 分割。并且去除标点符号和停用词。英文数据集只选取了 Food、Staff、Ambience 三个方面黄金方面的数据，去除停用词和标点符号，并且把单词的变形转换成最原始的形态。2 个数据集的单词分布结果见表 2，黄金方面和部分方面表示词的示例见表 3。

基准方法：为了评估本文模型，选择了两个基准方法。

表 2　数据单词集描述

Table 2　Data word set description

数据集	单词集	食物	价格	服务	环境
Chinese	Seed	17	20	12	14
	Final	115	142	76	90
Citysearch corpus	Seed	13	—	13	23
	Final	100	—	100	75

LocLDA[23]: 该方法使用了 LDA 的标准实现。为了防止全局主题的提取并将模型引向可评价方面, 将每条评论作为一个单独的文档处理。模型的输出是对数据中每条评论的方面分布。

SAS[19]: 该方法是一个混合主题模型, 在用户感兴趣的类别上给定一些种子词, 自动地提取类别方面术语。这个模型在已知的主题模型上, 对于方面提取具有很强的竞争性。

表 3　黄金方面和部分方面表示词

Table 3　Gold aspects and representative words

黄金方面	方面表示词(美团美食)	方面表示词(Restaurant)
Flavor/Food	味道、口味、菜味、饭菜、菜色、风味、菜系、菜品、口感、肉质	food、cuisine、meal、quality、healthy、fusion、describe、desert、dinner
Price	价格、价钱、菜价、价位、单价、经济、票价、物价、消费水平、价格公道	Charge、paid、bill、dollar、expensive
Service/Staff	服务质量、素质、效率、态度、敬业、客气、热心、热情、服务员、服务生	server、waiter、host、personnel、waitress、hostess、manner、overbearing、server、manager
Ambience	环境、效果、卫生、视听、音响效果、整洁、整齐、清洗、音效、视觉、音质、景色、设施、室内	atmosphere、environment、surroundings、classy、lively、sexy、relax、chill、hang、enjoy、quiet、enjoying

评估方法: 把每个方面识别的过程看作一个二分类的过程, 因此方面提取的效果通过 precision、recall 率和 $F_1$3 个指标来衡量, $\text{precision}=\frac{\text{TP}}{\text{TP+FP}}$, $\text{recall}=\frac{\text{TP}}{\text{TP+FN}}$, $F_1-\text{score}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$, 对于每个二分类过程, 存在 4 种可能的情况, 正类被预测成正类 (TP), 负类被预测成正类 (FP), 负类被预测成负类 (TN), 正类被预测成负类 (FN)。

3.2　评估和结果

对于英文餐厅的数据集, 评估“Food”“Staff ”和“Ambience”3 个主要方面, 因为其他方面的数据在词语使用或写作风格上都没有表现出明确的模式, 这使得这些方面甚至很难被人类识别。中文数据集评估了“食物”“价格”“服务”和“环境”4 个方面, 根据计算数据集的各单词的 TF-IDF 来确定数据集的黄金方面。本文模型在英文数据集和中文数据集上的结果如表 4、表 5 和图 3、图 4 所示。

通过图表的对比可以观察到: 1) 在英文数据集上, 本文方法 (AESS) 在 3 个方面的召回率都高于其他方法, 本文方法在员工和环境方面识别的 F_1 分数高于其他 2 种方法。AESS 食物的 F_1 比 SAS 差, 但其召回率非常高。分析了原因, 发现大多数句子没有提及到味道或者食物的名词。例如, “挺好吃的”这个句子的真实标签就是食物。

2) 在中文数据集上, 本文方法 (AESS) 在食物、价格和环境方面识别的 F_1 分数高于其他方法, 4 个方面的召回率都高于其他方法。本文方法在中文数据集上明显优于其他 2 种方法, 可能的原因有中文数据集是具有特色的美食评论数据, 中文在语法表达上和英文不同, 语句简短甚至没有固定的语法, 对于主题提取比较困难, 基于数据集创建词典, 避免这类问题, 因此效果比较好。

表 4　3 种方法在相同的英文数据集上的 3 个黄金方面确定的结果对比

Table 4　Comparison of results determined by the three methods on the three gold aspects of the same English data set

方面	方法	准确率	召回率	F_1分数
Food	**AESS**	0.712	**0.902**	0.796
	Loc LDA	**0.898**	0.648	0.753
	SAS	0.867	0.772	**0.817**
Staff	**AESS**	**0.892**	**0.740**	**0.809**
	LocLDA	0.804	0.585	0.677
	SAS	0.774	0.556	0.647
Ambience	**AESS**	0.595	**0.852**	**0.700**
	LocLDA	0.603	0.677	0.638
	SAS	**0.780**	0.542	0.640

表 5　3 种方法在相同的中文数据集上的 4 个黄金方面确定的结果对比

Table 5　Comparison of results determined by the three methods on the four gold aspects of the same Chinese data set

方面	方法	准确率	召回率	F_1分数
Flavor	**AESS**	0.909	**0.985**	**0.946**
	Loc LDA	**0.928**	0.557	0.696
	SAS	0.753	0.655	0.701
Price	**AESS**	**0.715**	**0.770**	**0.742**
	LocLDA	0.421	0.675	0.519
	SAS	0.525	0.532	0.528
Service	**AESS**	0.907	**0.710**	0.796
	LocLDA	**0.942**	0.653	0.771
	SAS	0.935	0.705	**0.804**
Ambience	**AESS**	**0.773**	**0.892**	**0.828**
	LocLDA	0.575	0.675	0.621
	SAS	0.625	0.685	0.654

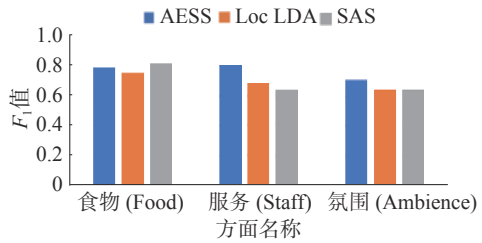

图 3　3 种方法在相同的英文数据集上的 3 个黄金方面确定 F_1 结果对比

Fig. 3　The F_1 results that three methods determine the three gold aspects on same English data set

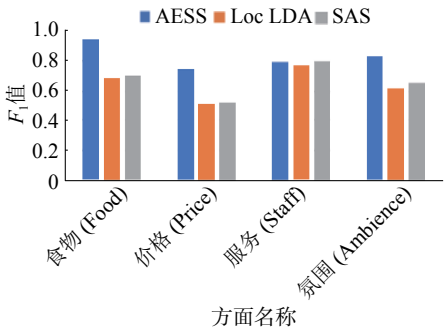

图 4　3 种方法在相同的中文数据集上的 4 个黄金方面确定结果对比

Fig. 4　The F_1 results that three methods determine the four gold aspects on same Chinese data set

4　结束语

本文提出一种基于半监督自训练的方面提取方法，避免了监督学习的标签数据依赖问题，并且在方面提取的结果中解决了以往无监督模型的方面聚类效果不一致的问题。本文有 3 个方面的贡献：1) 人工标注少量的方面，作为方面表示的种子词，利用词向量获得与语料相关的丰富的方面表示词典集合，确定方面表示单词集合，解决方面表示单词确定的困难；2) 通过计算数据集单词的 TF-IDF 值来确定数据集黄金方面，对每个句子进行多个方面的识别，并采用方面向量表示文本的包含的方面；3) 本文方法同时应用到中文美食评论和英文公开评论数据集，并对比了两种经典的方面提取方法。但是本模型对单词种子词比较敏感，未来可以进一步在方面提取的基础上基于方面对内容进行挖掘，将具有更重要的意义。

参考文献：

[1] LIU Bing. Sentiment analysis and opinion mining[C]//Proceedings of the Synthesis Lectures on Human Language Technologies. Toronto, Canada, 2012: 152–153.

[2] 刘倩. 观点挖掘中评价对象抽取方法的研究 [D]. 南京: 东南大学, 2016.
LIU Qian. Research on approaches to opinion target extraction in opinion mining[D]. Nanjing: Southeast University, 2016.

[3] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993–1022.

[4] TITOV I, MCDONALD R. Modeling online reviews with multi-grain topic models[C]//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 111–120.

[5] BRODY S, ELHADAD N. An unsupervised aspect-sentiment model for online reviews[C]//Proceedings of the Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 804–812.

[6] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493–2537.

[7] PORIA S, CAMBRIA E, GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2539–2544.

[8] PORIA S, CAMBRIA E, GELBUKH A. Aspect extraction for opinion mining with a deep convolutional neural network[J]. Knowledge-Based Systems, 2016, 108: 42–49.

[9] HE Ruidan, LEE W S, NG H T, et al. An unsupervised

neural attention model for aspect extraction[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 388–397.

[10] 韩忠明, 李梦琪, 刘雯, 等. 网络评论方面级观点挖掘方法研究综述 [J]. 软件学报, 2018, 29(2): 417–441.
HAN Zhongming, LI Mengqi, LIU Wen, et al. Survey of studies on aspect-based opinion mining of internet[J]. Journal of Software, 2018, 29(2): 417–441.

[11] JIN Wei, HO H H. A novel lexicalized HMM-based learning framework for web opinion mining[C]//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 465–472.

[12] LI Fangtao, HAN Chao, HUANG Minle, et al. Structure-aware review mining and summarization[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 653–661.

[13] JIN Wei, HO H H, SRIHARI R K. OpinionMiner: a novel machine learning system for web opinion mining and extraction[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 1195–1204.

[14] WANG Wenya, PAN S J, DAHLMEIER D, et al. Recursive neural conditional random fields for aspect-based sentiment analysis[J]. arXiv preprint arXiv:1603.06679, 2016.

[15] CHEN Huimin, SUN Maosong, TU Cunchao, et al. Neural sentiment classification with user and product attention[C]//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1650–1659.

[16] CHINSHA T C, JOSEPH S. A syntactic approach for aspect based opinion mining[C]//Proceedings of 2015 IEEE International Conference on Semantic Computing. Anaheim, USA, 2015: 24–31.

[17] YAN Xiaohui, GUO Jiafeng, LAN Yanyan, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1445–1456.

[18] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 142–150.

[19] WANG Linlin, LIU Kang, CAO Zhu, et al. Sentiment-aspect extraction based on restricted boltzmann machines[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015.

[20] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[21] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 3111–3119.

[22] GANU G, ELHADAD N, MARIAN A. Beyond the stars: improving rating predictions using review text content[C]//Proceedings of the 12th International Workshop on the Web and Databases. Rhode Island, USA, 2009.

[23] ZHAO W X, JIANG Jing, YAN Hongfei, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid[C]//Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts, USA, 2010: 56–65.

[24] MUKHERJEE A, LIU Bing. Aspect extraction through semi-supervised modeling[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Jeju Island, Korea, 2012: 339–348.

作者简介:

曲昭伟, 男, 1970 年生, 教授, 主要研究方向为数据挖掘、人工智能、无线传感器网络。承担多项横向课题。发表学术论文 50 余篇。

吴春叶, 女, 1992 年生, 硕士研究生, 主要研究方向为数据挖掘、Web 挖掘、机器学习和 Web 搜索引擎。

王小茹, 女, 1980 年生, 副教授, 主要研究方向为人工智能、计算机视觉、图像理解、精准搜索与大数据数据挖掘。获得国家发明专利 3 项。发表学术论文 36 篇, 出版学术著作 6 部, 译著 2 部。