文章编号：2095-2163(2021)04-0025-05 中图分类号：TP391 文献标志码：A

基于词嵌入和自注意力机制的方面提取算法

吴杭鑫，张云华
（浙江理工大学 信息学院，杭州 310018）

摘 要：方面提取是情感分析中的关键步骤，随着互联网的快速发展，短文本数据迅猛增加，对短文本数据加以整理和利用极为重要。本文针对短文本的特殊性，提出了短文本模型 WESM。与现有模型不同的是，本文引入了词汇共现网络，丰富了词汇的上下文信息，针对中文数据，引入了 cw2vec 模型，能够充分利用中文词语的语义信息；为了提高短文本的上下文语义缺失，引入了自注意力机制，能够丰富模型的上下文语义信息，提高方面词汇权重，在词汇聚类过程中，降低了非方面词汇的影响。相较于传统方面提取算法性能有着显著的提升。

关键词：方面提取；词嵌入；自注意力机制

Aspect extraction algorithm based on word embedding and self-attention mechanism

WU Hangxin, ZHANG Yunhua
(School of Information, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Aspect extraction is a key step in sentiment analysis tasks. With the rapid development of the Internet, the data of short has increased rapidly, and it is important to organize and make use of those. The main work of this paper is as follows: For the particularity of short text, this paper proposes a short text model WESM. Different from the existing models, this paper introduces a vocabulary co-occurrence network to enrich the context information of the vocabulary. As for Chinese data, the cw2vec model has been introduced, which will make full use of the context semantic information; in order to improve the lack of contextual semantics of short texts, this paper introduces a self-attention mechanism, which can enrich the contextual semantic information of the model and increase the weight of the terms. In the process of clustering, the influence of non-aspect words is reduced. Compared with the traditional extraction algorithm, the performance has been significantly improved.

【Key words】 Aspect extraction; Word embedding; Self-attention mechanism

0 引 言

方面信息提取[1]是从给定原始文本中提取出表征实体、实体属性或反映实体某一侧面的信息。方面信息是方面情感的直接受体，一般为一个词语或者短语。例如，在句子“今天的晚餐既美味又实惠”中，“美味”和“实惠”分别评价了晚餐的两个不同侧面，且赋予了正向的情感极性，所以可作为方面信息提取出来。

方面提取任务是方面级别情感分类任务的前提和基础。近年来随着互联网的发展，越来越受到业界的关注。早期的研究人员主要采用基于语义特征的方法来训练模型[2-4]。但此类模型的性能受人为定义特征的影响较大，相对费时、费力，且对于研究人员的操作能力与资源质量有着较强的依赖性。近期，性能表现较好的方面提取算法，主要以基于词共现网络和基于图的方法为主[5-7]。

受上述方法的启发，本文提出基于词嵌入和自注意力机制的方面提取算法（World Embedding and Self - attention Model for Aspect Extraction，简称 WESM），主要工作如下：

（1）利用基于词汇共现网络的来进行方面提取，相较于传统的主题模型，能够有效克服短文本存在稀疏性等特点，可以发现一些不常见的主题。

（2）引入自注意力机制，解决由于长距离依赖问题而造成的上下文信息忽略问题，能够充分捕捉词的上下文语义信息。

（3）应用细粒度的汽车评论数据集及来自购物网站的抓取数据集，与当前主流相关算法进行了比较。实验结果表明，所提出的 WESM 模型的性能优于相关工作，适合于方面提取任务。

1 相关工作

方面提取是观点挖掘领域中的细分任务，在过去的数十年中，大量学者在方面提取上做了大量研

作者简介：吴杭鑫（1994-），男，硕士研究生，主要研究方向：智能信息处理；张云华（1965-），男，博士，教授，硕士生导师，主要研究方向：软件架构、软件工厂、智能信息处理。

收稿日期：2020-12-04

究工作。如,在文献[8-9]中提出了一个词汇 HMM 模型来提取文本的显示方面;文献[10]提出基于监督的条件随机场模型来提取显示方面。但是监督学习需要大量的标签数据,耗费大量的人力。

无监督的学习方法可以省去大量的数据标注工作。以 pLAS(probabilistic latent semantic analysis)[11] 和 LDA(Latent Dirichlet allocation)[12] 模型为主的方法,通过在文档与单词间搭建"主题"这一桥梁,来进行方面提取,已经被许多研究者应用于方面提取的任务中。然而,这类主题模型基本都是针对长文本方面提取,对于短文本任务无法取得良好效果。针对短文本特性,文献[13]提出了 BTM(Biterm Topic Model, 简称 BTM)模型,它与 LDA 模型不同的是使用了 biterm 进行建模,能够更好的发掘文章的隐藏主题;文献[14]提出了词汇网络共现主题模型(Word NetWork Topic Model,简称 WNTM),通过词汇共现网络中语义紧凑的潜在词群,发现不常见的主题,取得了良好效果。上述方法在针对短文本这一特定方向时,综合表现较好,其共性在于挖掘上下文隐含的语义关系解决短文本存在的稀疏性等特点。

综上所述,短文本数据相对于长文本主要存在的问题是文本稀疏性大、语义信息不足以及主题不平衡等。针对这些问题,本文提出了 WESM 模型,通过词汇共现网络,解决了难以发现罕见主题的问题;通过引入针对中文的词嵌入模型,能够更好的发掘出丰富中文词汇的语义信息;通过引入自注意力机制,缓解上下文语义缺失的问题,提高了算法的性能。

2 模型描述

本文提出了一种基于词嵌入和自注意机制的方面提取算法(WESM)。该模型基于词汇共现网络,在整个语料库上构建伪文档,相较于传统的 LDA 模型,词汇共现网络有着明显的优势,能够充分利用整个语料库的语义信息。其次,其节点之间的边权值表示两个词汇在上下文中共现的次数。通过针对中文的词嵌入模型 cw2vec 来训练词汇,丰富词汇的潜在语义信息,得到词汇的向量表示。然后输入到自注意力机制模块中,其特点在于可以无视词汇之间距离,捕获长距离的依赖关系。算法模型架构如图 1 所示。

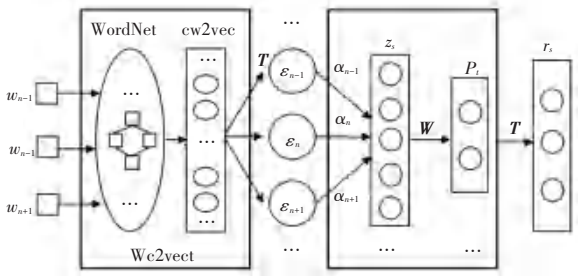


图 1 WESM 模型图

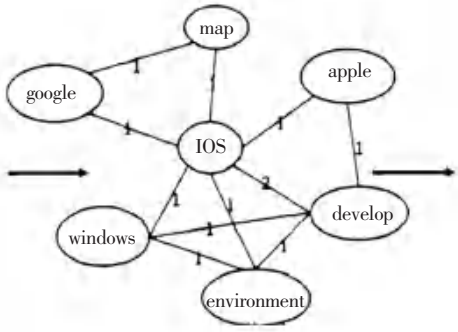
Fig. 1 Model of WESM

其中,WordNet 是词汇共现网络,w_{n-1}、w_n、w_{n+1} 是输入量,分别表示语料库中的词汇、网络中节点是词汇、节点之间的边权重表示两个节点词汇共现的次数。cw2vec 是中文单词向量模型,经过该模型的训练可以得到词汇的向量表示,即 ε_{n-1}、ε_n、ε_{n+1}。经过词嵌入模型后,进入自注意力机制模块,该模块主要是为了得到词汇的上下语义信息,Z_s 表示相应句子的嵌入表示,W 为过滤矩阵,T 代表高维空间向量矩阵。

WordNet 是词汇共现网络(WNTM)模型,从关键词之间的共现关系角度来建立网络。考虑到语义的联系是相互的,所以该网络是一个无向有权图。其中节点表示关键词,权值表示两个词汇共同出现的次数。显然,若节点之间的边权值越大,则它们之间的关系越紧密。

WNTM 模型由网络图、邻近表、伪文档三部分组成,如图 2 所示。其中,伪文档是和邻近表由词汇与相邻节点生成的,描述了节点之间可能存在的关系。

Doc 1	Google Map for IOS
Doc 2	IOS developed by Apple
Doc 3	IOS developed environment for windows



word	pseudo-document
map	google IOS
google	map IOS
apple	google IOS
IOS	google map apple windows of nvironment devolop devetop
winknes	environment OIS develop
develop	apple IOS IOS womdpws envronment
environment	OIS windows develop

图 2 WNTM 模型图

Fig. 2 Model of WNTM

词汇贡献网络生成相应的伪文档，步骤如下：

（1）根据词汇 w_i 的邻近词汇表 L_i，与潜在词群 z_i，进行“主题-单词”概率分布采样，得出相应关系表达式：$\Theta_i \sim Dir(\alpha)$；

（2）对潜在词群 z，进行“伪文档-主题”概率分布采样，得到表达式：$\varphi_z \sim Dir(\beta)$；

（3）对于邻近词汇表 L_i 中的每个词汇 w_j：

①根据“伪文档-主题”概率分布，采样主题 $z_j \sim Dir(\Theta_i)$；

②根据“主题-单词”概率分布，采样单词 $w_j \sim Dir(\varphi_{z_j})$。

其中，Θ、φ 分别表示邻近表中潜在词群出现的概率分布、词汇属于潜在词群的概率分布。

由于 WNTM 模型包含了词汇的上下文语义信息，因此将词汇 w_i 的邻近词表 Θ_i 的主题比例作为词汇 w_i 的主题比例，其计算公式如下：

$$P(z \mid d) = \sum_{w_i} P(z \mid w_i) P(w_i \mid d), \tag{1}$$

其中，Θ_i 可以表示为 $P(z \mid w_i)$，$P(w_i \mid d)$ 可以看成词汇的经验分布，计算公式如下：

$$P(w_i \mid d) = \frac{n_d(w_i)}{Len(d)}. \tag{2}$$

式中，$n_d(w_i)$ 表示词汇 w_i 在文档 d 中的词频，$Len(d)$ 表示文档 d 的长度。由于短文本数据的特点，使得长文本主题相关方法对其处理效果欠佳。而基于 WNTM 模型构建的伪文档中包含了所有的主题信息，学习伪文档上的主题分布，能够解决短文本数据稀疏性问题。

TC2vec 的另一个部分是 cw2vec 模型，该模型以中文笔画信息作为特征，捕捉汉字词语的语义和结构层面信息，获得分布式向量词并以负采样进行计算。

cw2vec 模型使用一种基于 n 元笔画的损失函数，公式如下：

$$L = \sum_{w \in D} \sum_{c \in T(w)} \log sigmod(sim(w,c)) + \lambda E_{w' \sim P}[\log sigmod(-sim(w,w'))], \tag{3}$$

其中，w 和 D 分别表示词语和词语归属的训练语料；c 和 T(w) 是词语的上下文和词语上下文窗口内的所有词语集合；λ 是负采样的数量，由总数乘以负采样比例得到；$E_{w' \sim P}[\sim]$ 是期望，并且选择的负采样 w' 服从部分 P。因此，语料中出现次数越多的词语越容易被采样，公式如下：

$$sim(w,w') = \sum_{q \in S(w)} \vec{q} \cdot \vec{w'}. \tag{4}$$

式中，$\vec{q}$ 和 S(w) 分别表示 n 元笔画向量及当前词语 w 所对应的 n 元笔画的集合。该方法是将当前词语拆解为对应的 n 元笔画。如“人”字，可拆解为一撇一捺。

在词嵌入层后增加自注意力层，通过自注意力机制获取文本的上下文语义信息，重构句子嵌入表示，转化为 r_s 的形式。

通过加权和的方式将方面信息纳入到重现后的句子中，计算公式如下：

$$z_s = \sum_{i=1}^{n} a_i e_{w_i}. \tag{5}$$

式中，e_{w_i} 表示第 i 个词汇的向量，词汇的嵌入表示和上下文环境将共同决定注意力机制的权值，即 a_i 的数值。计算公式如下：

$$y_s = \frac{1}{n} \sum_{i=1}^{n} e_{w_i}. \tag{6}$$

$$d_i = e_{w_i}^T \cdot M \cdot y_s. \tag{7}$$

$$a_i = \frac{\exp(d_i)}{\sum_{j=1}^{n} \exp(d_j)}. \tag{8}$$

其中，y_s 由组成句子词汇的向量和求均值得到，是句子向量的嵌入。通过模型训练获得矩阵 $M(M \in R^{d \times d})$，并在句子向量和词汇向量之间进行映射，以获得词汇和方面相关信息。a_i 表示注意机制的权重，公式如下：

$$r_s = T^T \cdot p_t, \tag{9}$$

$$p_t = softmax(W \cdot z_s + b). \tag{10}$$

其中，p_t 表示方面嵌入权重，将 z_s 从 d 维降到 k 维，然后通过 softmax 函数标准化得到。W 和 b 从训练模型中获得的。

若直接进行后续训练，将会产生较大的重构误差。因此，本文采用最大边界相关函数（Contrastive Max-margin Objective Function），其公式如下：

$$J(\theta) = \sum_{s \in D} \sum_{i=1}^{m} \max(0, 1 - r_s z_s + r_s n_i). \tag{11}$$

其中，D 代表语料库，n_i 代表负样本。训练使得 r_s 与 z_s 大体相似，并且与 n_i 最大限度不同。$\{E、T、M、W、b\}$ 为训练得到的模型参数。

3　实验及分析

3.1　数据集和评价指标

为验证模型的有效性，采用公开数据集（细粒度汽车评论标注语料数据集）通过网络爬取相关评论信息（某购物网站关于手机的相关数据）进行测

试,并且都详细标注了用户评论中的评价对象和评价特征。数据集详细信息见表 1。

表 1 数据集和训练集

Tab. 1 Data sets and training sets

		句子	方面	包含方面的句子
数据 1	训练集	3 261	2 781	1 852
	测试集	900	741	579
数据 2	训练集	2 616	1 860	1 312
	测试集	750	612	436

本文选取精确率(*Precision*)、召回率(*Recall*)和 $F-score$ 值来评估模型的整体性能。精确率计算公式如下:

$$precision = \frac{TP}{TP + FP}. \tag{12}$$

召回率也称为查全率(*Recall*),计算公式如下:

$$recall = \frac{TP}{TP + FN}. \tag{13}$$

实验引入了 $F-score$ 值,用来调节查准率和查全率,公式如下:

$$F-score = (1-\beta^2)\frac{precision \times recall}{\beta^2(precision + recall)}, \tag{14}$$

其中,β 为 $F-score$ 中权重参数。

3.2 实验结果及分析

本文使用主题聚合度(Topic Coherence)来评价得到的方面聚类之间的相似性,它与方面词相似度之间呈正相关,其计算公式如下:

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D(v_m^{(t)}, v_l^{(t)}) + \varepsilon}{D(v_l^{(t)})}. \tag{15}$$

其中,t 代表某个方面;M 表示在方面集合选取词汇数量;$V^{(t)}$ 代表方面 t 中 n 个方面词汇;$v_m^{(t)}$ 和 $v_l^{(t)}$ 分别表示方面 t 中的两个方面词汇;$D(\sim)$ 表示参数(方面词汇)的共现次数。当只有一个参数时,表示该词汇的出现次数。为了防止 log 函数的值为 0 造成计算错误,本文设置 ε 为 1。公式如下:

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} log\frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}. \tag{16}$$

本文选取 LDA、BTM、WNTM、WESM 算法在数据集上进行对比实验,模型表现如图 3、图 4 所示。其中主题数统一选 10,先验参数 $\alpha = \frac{50}{K} = 5, \beta = 0.05$ 学习率设置为 0.025。

从图 3 和图 4 中可以看出,WNTM 模型在主题句聚合度上的实验结果优于 LDA 和 BTM 模型,而 WESM 算法又优于 WNTM 模型。由于数据都是短文本数据,用于长文本的 LDA 模型在实验结果上略逊色于 WNTM 模型和 BTM 模型,而 WESM 算法是在 WNTM 的基础上增加了词向量模型和自注意力机制,能够更细粒度的利用词汇语义信息。主题聚合度得分越高,则模型所得到的主题质量更好,证明了引入自注意机制能够有效丰富语境语义,提高主题质量。

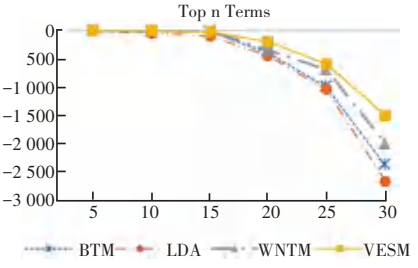


图 3 数据集 *D*1 主题聚合度

Fig. 3 Data set *D*1 topic aggregation degree

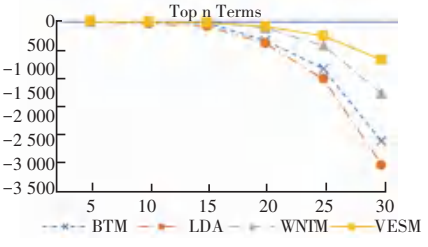


图 4 数据集 *D*2 主题聚合度

Fig. 4 Data set *D*2 topic aggregation degree

图 3 和图 4 通过各主题聚合度得分表现,说明了 WESM 模型主题聚合效果上表现更出色。下面将通过查准率、召回率以及 $F1$ 值对模型的其它方面做进一步验证。其中本文选取前 n(其中 n = 10,20,30,40)个词汇计算各项指标,结果如图 5、图 6 所示。

从图 5 和图 6 可以看出,WESM 模型的平均查准率比其它三个模型更好。对图表进一步观察对比发现,针对短文本提出的 WNTM 和 BTM 模型在查准率上的表现优于 LDA 模型;通过 WNTM 和 WESM 的对比发现,引入词嵌入和自注意力机制确实有利于查询率的提高,验证了其对于方面提取性能的提升是有效果的。

通过图 7 和图 8 可以看出,随着词汇数量的增加,各模型的 $F1$ 值都呈现下降的趋势。但 WESM 模型在实验中的表现还是优于其他模型。验证了词向量和自注意力机制能够丰富词汇的上下文语义特征,从而提高方面提取的性能。

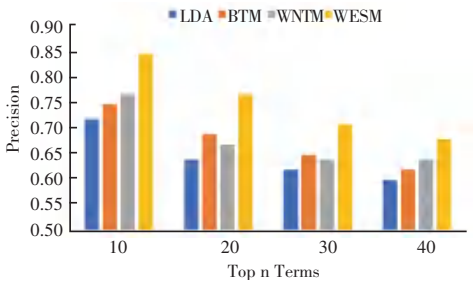


图 5　数据集 *D*1 平均查准率

Fig. 5　Average precision of data set *D*1

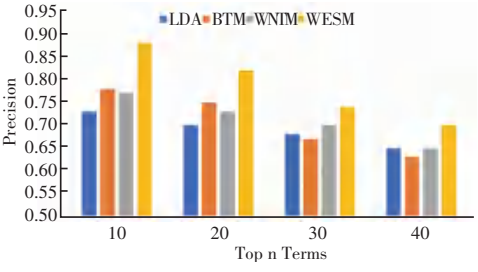


图 6　数据集 *D*2 平均查准率

Fig. 6　Average precision of data set *D*2

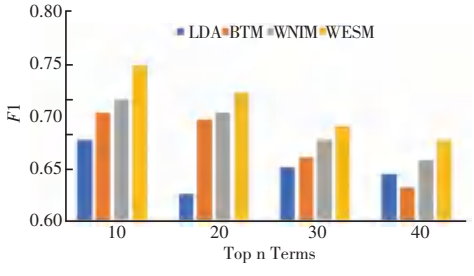


图 7　数据集 *D*1 平均 *F*1 值

Fig. 7　Average *F*1 value of data set *D*1

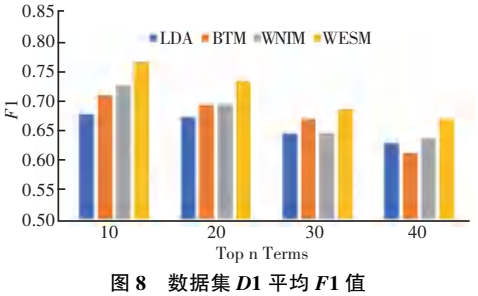


图 8　数据集 *D*1 平均 *F*1 值

Fig. 8　Average *F*1 value of data set *D*2

4　结束语

本文在词汇共现网络基础上引入了词向量模型和自注意力机制，提出了方面提取算法 WESM。实验结果表明，基于本文提出的方面提取算法及自注意力机制的引入，能够丰富词汇的语义信息，得到文本的上下文语义信息。通过应用两个数据集实验，对比了相关的方面提取方法，证明了本文算法的优势所在。但是，模型仍旧存在一些不足。如，方面词汇聚类簇数要人为进行设置、参数的设定直接影响模型的性能等问题。因此，在后续研究中可以考虑是否能够将这些步骤都通过算法训练得到，减少人为对算法的影响。

参考文献

[1] 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.

[2] LIU H Y，ZHAO Y Y，QIN B，et al. Comment target extraction and sentiment classification [J]. Chinese Information Processing，2010，24(1)：84-88.

[3] YANG X，SU J. Coreference resolution using semantic relatedness information from automatically discovered patterns [C]// the 45th Annual Meeting of the Association for Computational Linguistics，2007：528-535.

[4] LANG J，XIN Z，QIN B，et al. Coreference resolution with integrated multiple background semantic knowledge [J]. Chinese Information Processing，2009，23(3)：3-9.

[5] MATSUO Y，ISHIZUKA M. Keyword extraction from a single document using word cooccurrence statistical information. Int´I Journal on Artificial Intelligence Tools，2004，13(1)：157-169.

[6] ZHANG Y，ZHU W. Extracting implicit features in online customer reviews for opinion mining [C]// the 22nd Int'l Conf. on World Wide Web Companion，2013：103-104.

[7] XIA L，WANG Z，CHEN C，et al. Research on feature-based opinion mining using topic maps [J]. The Electronic Library，2016，34(3)：435-456.

[8] JIN W，HO H H. A novel lexicalized HMM-based learning framework for Web opinion mining [C]// the 26th Annual Int'l Conf. on Machine Learning，2009：465-472.

[9] JIN W，HO H H，SRIHARI R K. Opinion miner：A novel machine learning system for Web opinion mining and extraction [C]// the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining，2009：1195-1204.

[10] XU B，ZHAO T J，WANG S Y，et al. Extraction of opinion targets based on shallow parsing features [J]. Acta Automatica Sinica，2011，37(10)：1241-1247.

[11] HOFMANN T. Probabilistic latent semantic indexing [C]// the 22nd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval，1999：50-57.

[12] Blei DM，Ng AY，Jordan MI. Latent dirichlet allocation [J]. Machine Learning Research，2003，3：993-1022.

[13] YAN X，GUO J，LAN Y，et al. A biterm topic model for short text [C]// the 22nd international conference on World Wide Web. 2013：1445-1456.

[14] ZUO Y，ZHAO J，XU K，Word network topic model：a simple but general solution for short and imbalanced text [J]. Knowledge and Information Systems，2016：379-398.

[15] ARORA S，GE R，HALPERN Y，et al. A practical algorithm for topic modeling with provable guarantees [C]// International Conference on Machine Learning. 2013：280-288.