

基于依赖联系分析的观点词对协同抽取



赵威^{1,2} 林煜明¹ 王超强¹ 蔡国永¹

1 桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004

2 华东师范大学数据科学与工程学院 上海 200062

(331205121@qq.com)

摘要 同一类商品下,观点词对中包含的观点目标和观点词通常有着很强的观点依赖联系,因此可以通过对评论句子中单词间的观点依赖联系进行分析来提取观点词对。首先,构建评论句子的依赖联系分析模型来获取评论句子中每个单词之间的依赖联系信息,文中选择的基本模型是 LSTM 神经网络;然后,假设评论句子中所包含的观点词对中的一项是已知的,并将该已知项作为模型的注意力信息,使得模型能够从评论句子中有重点地提取出与该已知项具有强观点依赖联系的单词或词组,并将其作为观点词对中的另一未知项;最后,将观点依赖联系得分最高的词对作为观点词对并输出。文中进一步设计了一种复合模型,通过结合两种包含不同已知项信息的上述模型,来实现在不需要提前知道已知项的情况下观点词对的挖掘。

关键词: 观点词对;观点依赖联系分析;注意力机制;神经网络

中图法分类号 TP391

Opinion Word-pairs Collaborative Extraction Based on Dependency Relation Analysis

ZHAO Wei^{1,2}, LIN Yu-ming¹, WANG Chao-qiang¹ and CAI Guo-yong¹

1 Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China

2 School of Data Science & Engineering, East China Normal University, Shanghai 200062, China

Abstract In the same category of commodities, opinion word-pairs usually have strong opinion dependence relation to the opinion targets and the opinion words contained in them. Therefore, in the extraction process of opinion word-pairs, they can be extracted by analyzing the opinion dependence relations among the words in the review sentences. Firstly, a dependency relation analysis model is constructed to obtain the dependency relation information of each word in a review sentence, and the basic model is defined as LSTM neural network. Secondly, it is assumed that one of the item that opinion word-pairs contained in review sentence is known, and the known item is used as the model's attention information, so that the model can focus on extracting the words of phrases associated with the known item with strong opinion dependence from the review sentence as another unknown item in the opinion word-pairs. Finally, the word-pairs with the highest score of the opinion dependence relation are output as the opinion word-pairs. Then a compound model is designed to realize the mining of opinion word pairs without knowing the known items in advance by combining the two models which contain the information of different known items in the opinion word-pairs.

Keywords Opinion pair, Opinion dependency relation analysis, Attention mechanism, Neural network

1 引言

在信息时代飞速发展的背景下,各种网络平台已经成为大多数人的主要生活方式,尤其在电子商务领域,用户通过在线购物平台进行了大量的消费和评论,因此产生了海量的用

户信息,获取这些信息不仅能够为商家提供有价值的反馈信息,从而进一步提高商品质量或对用户的服务质量,其还能够作为用户的参考,使其筛选出性价比更高的产品。因此,研究如何从海量评论信息中挖掘出用户的观点信息,具有重要的应用价值和现实意义。

到稿日期:2019-06-26 返修日期:2019-09-03 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:广西自然科学基金(2018GXNSFDA281049);国家自然科学基金(61662015, U1711263);广西创新驱动发展专项资金项目(桂科AA19046004);桂林电子科技大学研究生教育创新计划资助项目(2018YJXC48);广西可信软件重点实验室研究课题(kx201916)

This work was supported by the Guangxi Natural Science Foundation (2018GXNSFDA281049), National Natural Science Foundation of China (61662015, U1711263), Science and Technology Major Project of Guangxi Province (AA19046004) and Innovation Project of Guet Graduate Education (2018YJXC48) and Project of Guangxi Key Laboratory of Trusted Software(kx201916).

通信作者:林煜明(ymlin@guet.edu.cn)

随着大数据时代的来临,用户观点信息挖掘出现了新的机遇和挑战,例如,不断完善的神经网络算法和深度学习思想已经在情感极性判断的研究领域取得了不小的成果,但是传统的神经网络算法并不能很好地应用到观点词对的提取上。这是因为单纯地通过神经元之间的信息交互,很难从评论句子中单独学习到观点词对中存在的依赖关系,即使能够独立地提取出观点目标和观点词,也很难将其进行组合形成能够表达用户观点信息的观点词对。而注意力机制的提出则有效缓解了该问题,带有注意力机制的神经网络算法能够有重点地进行学习。本文的研究内容之一就是,围绕神经网络算法和深度学习的思想对英文评论信息中的观点词对信息进行提取。

本文的主要贡献如下:

(1)提出了一种观点依赖联系分析方法,对评论句子中存在的观点依赖联系进行分析,从而提取出评论句子中包含的用户观点词对方法,简称基于观点依赖联系的观点词对自动提取方法。

(2)假设观点词对中的一项是已知的,利用 LSTM 对评论句子的所有单词进行依赖联系分析,并根据注意力机制的思想找出与已知项具有强观点依赖联系的单词或词组作为该观点词对的另一未知项。

(3)基于上述观点词对挖掘思路设计了一种复合模型,通过结合两种基于不同已知项的上述观点词对抽取模型,实现观点词对的协同抽取。

(4)在不同数据集上的实验结果表明,本文提出的观点词对提取算法可以有效地提取出评论信息中存在的观点词对。

2 相关工作

目前,根据挖掘过程的不同,可以将观点词对的主要研究工作分为两大类,即管道式的观点词对挖掘方式和传播式的观点词对挖掘方式。

管道式的观点词对挖掘存在的挑战是,一次性的观点挖掘可能不能将评论信息中存在的所有观点词对挖掘出来,而接下来的提纯操作并不能挖掘出新的观点词对。在该方面已有很多研究工作,例如,Ding 等^[7]提出通过构建全局词典依赖的方法来进行观点挖掘的工作,但构建全局词典的成本较高且很难建立完整的全局词典。Wu 等^[2]提出的方法是通过评论信息中的单词进行语法依赖分析和分组,来构建评论信息的语法依赖树,最终通过观点目标和观点词之间存在的关系提取出观点词对。Xu 等^[3]提出了一种情感图游走算法,首先利用句法分析构建情感图,然后通过随机游走方法评估候选集的置信度,从而提取出高置信度的观点词对,但其只考虑了观点词对中的语义关系。Wu 等^[4]提出了一种无监督的学习方法,主要是通过句法分析和定义语法规则来确定观点词对候选集,然后将这些候选集作为训练集对 GRU 模型进行训练,省去了通过人工标注获取大量带标签训练集的成本。Danilo 等^[5]同样在预处理方面做了研究,通过对比不同的预处理方法对不同算法结果的影响,找到了一种能够提高模型挖掘效果的预处理方法。Rodrigo 等^[6]提出了一种语言独立

的观点目标挖掘方法,其主要思想是把观点词对的挖掘看成序列数据的识别标注。Ding 等^[7]提出虽然目前观点挖掘已经成为了一个非常热门的研究工作,但是很少有人提出通过分析观点目标和观点词的共指消解来进行观点词对的挖掘。Mariana 等^[8]和 Nguyen 等^[9]设计了一种很巧妙的方法来进行观点词对的挖掘,具体是通过两种语言之间语义的相似性,交叉进行观点挖掘工作。Tu 等^[10]提出了一种新的思想,他们认为对于评论信息中的观点挖掘不能只是通过判断文本上的情感极性来进行,还要考虑第三维度的时序方向。

与管道式的观点词对挖掘过程不同,传播式的观点词对挖掘技术将观点词对挖掘定义为一种迭代过程。这种思路存在的挑战是,如果在某次迭代过程中产生的挖掘结果存在问题,则会导致接下来的错误率越来越高。在该方面同样存在很多研究工作。例如,Zhang 等^[11]设计了几种观点目标和观点词间的语法规则,根据这些语法规则依次迭代筛选出观点目标和观点词,并通过商品的主题模型引导模型对得到的观点词对进行过滤。Hai 等^[12]则是在文献[11]方法迭代过程开始之前,就通过似然函数对观点目标和观点词进行了过滤。Qiu 等^[13]通过构建一个句法依赖树和一个初始观点词字典的方式,交叉迭代地进行观点目标和观点词的提取。Yang 等^[14]利用先验知识定义挖掘规则,并将观点挖掘的工作看作在评论信息中找寻最优解的过程。马尔可夫模型是一种能够很好地进行传播式观点词对挖掘的基础模型,Luis 等^[15]最先把马尔可夫的概念引入到观点挖掘的领域,设计了细粒化的挖掘模型,从而构建出一种实用的词典。Jiang 等^[16-17]提出了一种基于字对齐模型进行观点词对挖掘的方法,之后通过整合监督学习和非监督学习的优缺点,提出了一种基于主动学习的半监督学习方法,从而解决了依靠人工标注造成的时间消耗和低准确度的问题。

目前,通过带有注意力机制的深度神经网络进行情感分析的研究工作已经取得了很大成功,但是这些方法都没有很好地利用语言方面的先验知识。Lei 等^[18]在利用带有注意力机制的 LSTM 模型进行评论句子的情感分析时,将情感词典加入到注意力的生成过程。Abhishek 等^[19]把注意力机制加入到 LSTM-CRF 模型中来提高该模型对于观点词对的挖掘效果,具体来讲就是先利用双向 LSTM 对评论句子进行分析,根据获取到的结果进行同一评论句子下不同观点目标的观点表达的注意力求解,之后利用 CRF 对得到的每个单词的注意力进行依赖关系分析。

3 问题描述与动机

一条评论句子中用户的观点信息通常是由一组或多组观点词对组成,一组观点词对中包含了用户所关注商品的观点目标和其所持有的观点态度用词。因此,观点词对挖掘工作存在的主要挑战是,如何能够成对地从评论信息中提取出这些“观点目标”和“观点词”。根据观点词对中存在的观点依赖联系这一特性,模型设计的基本思想是首先对评论信息中每个单词之间的依赖联系进行学习,然后通过带有注意力权重信息的神经网络模型,有重点地学习隐藏在评论句子中的具

有强观点依赖联系的“观点目标”和“观点词”。

为了能够更好地对评论数据进行分析 and 提取,我们以句子为基本单位对每条评论进行划分。评论句子集合表示为 $S = \{s_1, s_2, \dots, s_n\}$, 每条评论句子中包含的观点词对表示为 $WP_i = \{wp_{i1}, wp_{i2}, \dots, wp_{im}\}$, 其中 $wp_{ij} = \langle ot_{ij}, ow_{ij} \rangle$ 。 ot_{ij} 和 ow_{ij} 分别表示“观点目标”和“观点词”。模型的最终目标是: 在对输入的评论句子 s_i 进行观点依赖联系分析后, 能够以最小误差输出评论句子 $s_i \in S$ 中存在的观点词对集合 WP_i 。由于句子中每个单词之间都普遍存在依赖联系, 想要从中找出观点依赖信息并不容易。最直接的一种方法是, 假设观点词对中观点目标或者观点词是已知的, 并记录为观点词对中的已知项 K , 之后模型只需从包含已知项 K 的依赖联系中, 找到与其具有强观点依赖联系的单词或词组, 作为观点词对中的另一未知项 uK , 如式(1)所示:

$$uK = M(s_i; K, \theta) \quad (1)$$

其中, M 表示观点词对挖掘模型, 能够根据输入的评论句子和已知项, 输出与已知项相关的未知项; θ 表示评论句子的观点依赖联系参数。但通常情况下, 观点词对中的观点目标和观点词都是未知的, 因此需要找到一种能够从评论句子中确定已知项的方法 $K = f(s_i)$ 。综上所述, 最终可以确定模型的目标函数为:

$$\begin{cases} K = f(s_i) \\ M^* = \arg \min_{\theta \in R} \sum_{s_i \in S} \text{loss}(\langle K, M(s_i; K, \theta) \rangle, l_i) \end{cases} \quad (2)$$

其中, R 表示参数的样本空间; l_i 表示评论句子 i 中真实包含的观点词对; loss 用来评估模型输出结果与真实的观点词对之间的差异, loss 值越小表示模型提取结果的误差越小, 模型的效果就越好。

常用的损失函数有平方损失 (Square loss) 和交叉熵 (cross entropy)。现在令 l_i^p 表示通过模型得到评论句子 i 中的观点词对, 则平方差估计的公式可以表示为:

$$\text{loss}_{sq}(l^p, l) = \sum_{wp_j \in l_i^p} (p(wp_j | l_i) - p(wp_j | l_i^p))^2 \quad (3)$$

交叉熵的公式可以表示为:

$$\text{loss}_{ce}(l^p, l) = \sum_{wp_j \in l_i^p} p(wp_j | l_i) * \log \frac{1}{p(wp_j | l_i^p)} \quad (4)$$

4 基于观点依赖联系分析的观点词对挖掘

本节提出的模型为基于观点依赖联系分析的观点挖掘模型 (Opinion Mining based on Opinion Dependency analysis, ODA), 如图 1 所示。

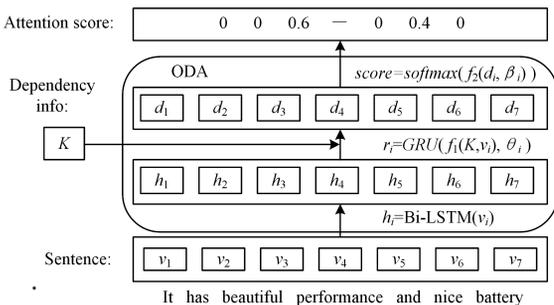


图 1 基于观点依赖联系分析的观点挖掘模型

Fig. 1 Opinion mining based on opinion dependency analysis

该模型主要包括两部分内容:

(1) 对输入的评论句子进行分析, 利用双向 LSTM 获取到能够包含评论句子中每个单词之间依赖联系的中间向量表示。

(2) 将观点词对中的已知项作为匹配项, 加入模型的注意力生成过程中, 最终生成与已知项具有强观点依赖联系的未知项, 并组合成观点词对进行输出。

4.1 评论句子上下文依赖信息的获取

ODA 模型的核心思想是依据观点词对中的已知项 K , 结合对评论句子的分析, 获取到与该已知项 K 具有强观点依赖联系的单词或词组, 并将其作为观点词对中的另一未知项, 因此需要先对整条评论句子进行扫描, 获取每个单词所包含的依赖信息。

具体思路如图 2 所示, 设计一个双向 LSTM 神经元, 该神经元接收来自两个方向 LSTM 的依赖特征信息传递, 即从评论句子最左端开始向评论句子最右端扫描的正向 LSTM 神经网络信息传递 fw , 以及从评论句子最右端开始向评论句子最左端扫描的反向 LSTM 神经网络信息传递 bw 。Bi-Lstm 神经元的主要作用是接收并综合考虑评论句子中每个单词 V_i 来自两个方向的 LSTM 信息传入, 从而得到评论句子在单词 V_i 位置处的依赖信息特征表示。具体计算式如式(5)所示:

$$C_i = b_i * fw_i + t_i * bw_i \quad (5)$$

其中, C_i 表示当前神经元的细胞状态, 包含了当前单词所持有的依赖信息特征; $t_i = \sigma(W_t * [x_i, bw_{i+1}] + bias_t)$ 和 $b_i = \sigma(W_b * [fw_{i-1}, x_i] + bias_b)$, 分别表示两个方向上依赖信息的输入门, 用来控制加入当前结点后两个方向上的依赖信息特征输入, 从而更新当前单词结点的细胞状态。

然后, 通过 tanh 函数将每个 Bi-LSTM 神经元状态信息转换为非线性关系, 将信息控制门作为权重来控制信息的输出, 最终输出的结果为当前结点位置评论句子上下文特征信息的中间信息表示。

$$o_i = \sigma(W_o * [fw_{i-1}, x_i, bw_{i+1}] + bias_o) \quad (6)$$

$$h_i = o_i * \tanh(C_i) \quad (7)$$

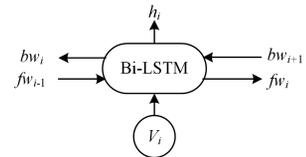


图 2 双向 LSTM 神经元的设计结构

Fig. 2 Design structure of Bi-LSTM neurons

4.2 基于已知项的注意力运算策略

由于观点词对中存在观点依赖联系, 已知项 K 可以引导模型找到与其具有强观点依赖联系的未知项。给定已知项 K 与评论句子 $s_i = \{v_1^{(i)}, \dots, v_n^{(i)}\}$, ODA 模型首先通过对评论句子 s_i 的扫描得到带有评论句子上下文依赖信息的中间信息表示 $H = \{h_1^{(i)}, \dots, h_n^{(i)}\}$, 然后获取已知项 K 与评论句子中每个单词 $v_j^{(i)}$ 的观点依赖联系信息 $d_j^{(i)}$ 和其注意力得分 $e_j^{(i)}$ 。

观点依赖联系信息的挖掘过程中, ODA 模型会通过一个张量计算器 f 来对已知项 K 与当前的单词中间文本标识进

行观点依赖信息的编码操作,计算出一个包含二者的观点依赖联系信息的复合信息 $\beta_j^{(i)}$:

$$\beta_j^{(i)} = f(h_j^{(i)}, \mathbf{K}) = \tanh(h_j^{(i)} \mathbf{W}_\beta \mathbf{K}) \quad (8)$$

其中, \mathbf{W}_β 是一个二维张量,可以将已知项 \mathbf{K} 与 $h_j^{(i)}$ 进行连接,形成已知项与该结点单词的观点依赖联系复合向量表示,之后添加 \tanh 的目的是将二者转换成非线性的函数联系,以获取更加多样的依赖联系特征。在对当前结点与已知项的观点依赖联系信息复合向量 $\beta_j^{(i)}$ 进行分析时,会考虑上个结点传来的上文观点依赖信息 $d_j^{(i-1)}$,如果上个节点与已知项的观点依赖信息并没有当前节点的强,则需要对上文的观点依赖信息进行遗忘,并将当前结点的观点依赖信息保存在 $d_j^{(i)}$ 中,接着传入到下个结点的观点依赖联系分析过程中。采用 GRU 神经元能够非常容易地对已知项的观点依赖联系进行分析和传递,具体方法如下:

$$d_j^{(i)} = (1 - z_j^{(i)}) * d_j^{(i-1)} + z_j^{(i)} * \tilde{d}_j^{(i)} \quad (9)$$

其中, $\tilde{d}_j^{(i)} = \tanh(\mathbf{W}_d^{(i)}(r_j^{(i)} * d_j^{(i-1)} + \mathbf{U}_d^{(i)} \beta_j^{(i)}))$, $z_j^{(i)} = \sigma(\mathbf{W}_z^{(i)}(d_j^{(i-1)} + \mathbf{U}_z^{(i)} \beta_j^{(i)}))$, $r_j^{(i)} = \sigma(\mathbf{W}_r^{(i)}(d_j^{(i-1)} + \mathbf{U}_r^{(i)} \beta_j^{(i)}))$ 。这里, $z_j^{(i)}$ 表示信息更新门,用来时刻保留与已知项的强观点依赖联系,并同时弱观点依赖联系进行遗忘; $r_j^{(i)}$ 表示重置门,可以结合上文信息判断当前结点单词与已知项观点依赖联系的强弱,从而控制 $d_j^{(i-1)}$ 信息的传输。

但经过分析发现,如果只从一个方向对评论句子进行扫描,每个结点只能考虑上文的观点依赖信息,而无法对下文的观点依赖信息进行分析,因此模型所获取的观点依赖联系信息是不可靠的。于是,考虑从两个方向上对评论句子进行扫描,扫描结束后会得到两组不同方向上的观点依赖联系信息,即正向扫描结果 $D_F^{(i)} = \{d_{j_1}^{(i)} \cdots d_{j_n}^{(i)} \cdots d_{j_m}^{(i)}\}$ 和反向扫描结果 $D_B^{(i)} = \{d_{b_1}^{(i)} \cdots d_{b_n}^{(i)} \cdots d_{b_m}^{(i)}\}$,然后利用式(10)将同一结点两个方向的扫描结果进行整合:

$$d_j^{(i)} = r_j^{(i)} * d_{j_1}^{(i)} + (1 - r_j^{(i)}) * d_{j_2}^{(i)} \quad (10)$$

其中, $r_j^{(i)}$ 表示信息控制门,能够保留在两个方向上获取到的与已知项具有更强观点依赖联系的信息,并遗忘掉较弱的观点依赖联系信息。

由于模型在对评论句子进行扫描的过程中,始终保留了与已知项具有强观点依赖联系的信息,因此可以通过式(11)分析 $h_j^{(i)}$ 与 $d_j^{(i)}$ 的关联度,从而判断该结点处的单词与已知项的观点依赖联系的强弱,即每个单词的注意力权重得分 $e_j^{(i)}$ 为:

$$e_j^{(i)} = \text{softmax}(\mathbf{W}_j^{(i)}[d_j^{(i)}, h_j^{(i)}] \mathbf{C}^{(i)}) \quad (11)$$

其中, $\mathbf{W}_j^{(i)}$ 表示一个观点依赖联系分析矩阵能够分析出 $d_j^{(i)}$ 与 $h_j^{(i)}$ 观点依赖的强弱, $\mathbf{C}^{(i)}$ 表示一个二分类参数,能够将分析结果转化为一个二分类问题,类别包括 $\{UK, O\}$,其中 UK 表示该结点为观点词对中的未知项, O 则表示非未知项。取二者得分最大的一项对该结点进行标注,最终可以得到两种标注结果,一类是标注为 UK 的结果概率,另一类是标注为 O 的结果概率。 Softmax 函数的作用是将得到的提取结果进行归一化操作,方便模型的训练。

5 基于多候选池的 Coupled-ODA 模型

ODA 模型的运用条件比较苛刻,需要在观点词对中一项已知的前提下提取另一未知项,但在实际的观点词对挖掘工作中,观点词对的两项一般都是未知的。因此,本文在 ODA 模型的基础上提出了一种 CODA(Coupled-ODA)模型,该模型的基本思想是利用训练集分别以观点词对中观点目标和观点词为已知项学习到两种 ODA 模型,即 TO-ODA 和 OT-ODA,然后将这两个 ODA 模型相结合,通过将各自的输出分别作为另一个模型的已知项来达到协同提取观点词对的目的,而且该模型只需将待提取观点词对的评论句子输入即可。图 3 给出了 CODA 模型的结构。

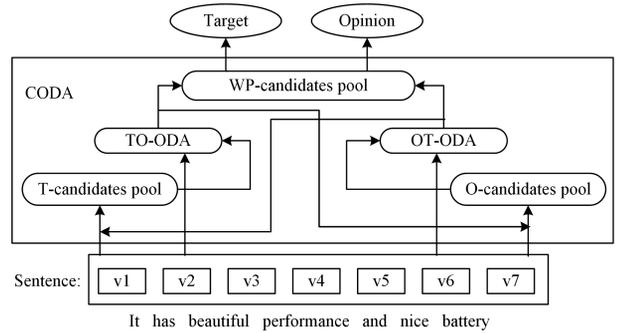


图 3 双-CODA 模型的结构

Fig. 3 Structure of Coupled-ODA model

在 CODA 模型中需要通过设置候选池的方式来收集两个 ODA 模型的输出结果,并将候选池中得到的候选结果作为下一次 ODA 模型进行观点词对提取的已知项。

5.1 候选池的初始化与更新策略

CODA 中包含 3 个候选池,分别是用来收集观点目标的观点目标候选池 T-candidates pool、用来收集观点词的观点词候选池 O-candidates pool,以及用来收集观点词对的观点词对候选池 WP-candidate pool。候选池的作用是对两个 ODA 模型的提取结果进行收集,其中 T-candidates pool 和 O-candidates pool 还要为 ODA 模型下一轮的观点词对提取提供已知项。3 个候选池中包含的数据信息有筛选出的未知项信息或者观点词对信息,以及其对应的得分和所在的评论句子 ID。保存得分的目的是在更新候选池时始终保存得分最高的候选项;保存评论句子 ID 的目的是方便模型找到已知项或观点词对对应的评论句子,并进行相关提取操作。因此,在模型开始运行时,T-candidates pool 和 O-candidates pool 需要进行初始化的操作。候选池的初始化是指通过一些简单的手段对评论句子进行过滤,达到将观点目标和观点词被粗略压缩到一定范围的目的,允许初始化结果只囊括部分观点目标和观点词。通过这种手段可以使得模型的提取结果很快收敛。候选池初始化的方法有很多,如:

(1) 直接将整条评论句子中的所有单词作为候选池的初始化结果。这种方式可能会导致模型提取速度和收敛速度变慢,影响提取的效率和准确率。

(2) 利用单词词性进行筛选,如选择名词生成观点目标候选池,选择形容词生成观点词候选池。

(3)通过设计两种简单的 LSTM 模型,来分别进行观点目标和观点词的粗略提取。

经初始化得到的 T-candidates pool 和 O-candidates pool 中每个候选项得分默认为 1。每次更新过程都会对候选池中的每个 candidate 得分进行重新评估,具体的更新算法如算法 1 所示。

算法 1 T-candidates pool 和 O-candidates pool 候选池更新算法

输入:old_pool,new_candidates

输出:new_pool

1. for 对于新生成候选集 new_candidates 中每个候选项 candidate:
2. if 候选池 old_pool 中已经包含了 candidate:
3. 取 new_candidates 中保存的 candidate 得分,记为 score₁
4. 取 old_pool 中保存的 candidate 得分,记为 score₂
5. 得到 candidate 新的得分 $score = \frac{score_1 + score_2}{2}$
6. 用 score 覆盖 old_pool 中 candidate 之前的得分
7. end if
8. else:
9. 将 candidate 连同其得分和所在评论句子的 ID 保存到 old_pool 中
10. end else
11. end for
12. 对 old_pool 中存在的 candidates 按照得分进行降序排序
13. 保存 old_pool 中得分最高的最多前 δ 个 candidate 到 new_pool

算法 1 中的输入包含了一个 new_candidates 候选集,该候选集是由 CODA 模型中的 WP candidate pool 产生的。根据不同的需求,只保留观点词对中观点目标信息或者观点词信息到 new_candidates 候选集中。WP-candidate pool 的更新方式与 T-candidates pool 和 O-candidates pool 略有不同,具体更新方法将在 5.2 节介绍。

候选池会随着模型的提取过程不断更新,直至 WP-candidate pool 中包含的候选项不再发生变化,或者模型迭代次数达到上限,最终模型会输出观点词对候选池 WP-candidate pool 中存在的观点词对。

5.2 基于不确定候选池的观点词对生成策略

由于候选池中存在的观点目标和观点词是随着模型提取过程动态变化的,因此 CODA 模型在每次迭代过程开始时,都会分别提取出 T-candidates pool 和 O-candidates pool 候选池中各自包含的 n 个潜在候选项作为已知项进行未知项的提取,并将每个已知项通过模型筛选得到的观点词对与 WP-candidate pool 候选池中包含的候选项进行对比。最好的结果是提取结果与候选池中候选项完全相同,则模型停止迭代并将 WP-candidate pool 中的候选结果输出;最坏的结果是提取结果与 WP-candidate pool 候选池中的候选项完全不重复,因此会得到最多 $2n$ 个潜在的提取结果,因此可以根据式(12)重新计算每个提取结果的得分。

$$(l_i^{wp})^* = \frac{l^{ot} + l^{lo} + l^{wp}}{3} \quad (12)$$

其中, l^{ot} 表示在观点词已知的情况下得到的观点词对 wp_i 得分, l^{lo} 表示在观点目标已知的情况下得到的观点词对 wp_i 得分, l^{wp} 表示 WP-candidate pool 候选池中包含的观点词对 wp_i

得分。在 WP-candidate pool 候选池中保存得分最高的前 n 项提取结果,并分别用其中的观点目标和观点词替换 T-candidates pool 和 O-candidates pool 中的候选项。

根据 CODA 的定义,由于已知项不是提前确定的,因此 CODA 筛选出的已知项可能并非评论句子中真正观点词对的已知项。因此,对于 l^{ot} , l^{lo} , l^{wp} 的计算都需要设置一个惩罚系数 $\xi \in (0, 1]$ 来指导模型抛弃不可靠已知项的筛选结果, ξ 越大说明已知项越可靠。在 CODA 模型中, ξ 定义为 l^{ot} , l^{lo} 各自的已知项分别在 T-candidates pool 和 O-candidates pool 中的得分,而对于 l^{wp} 来说,由于我们默认在 WP-candidate pool 中存在的观点词对是目前获取到的最可靠的观点词对,因此将 l^{wp} 的惩罚系数默认设置成 1,于是可以将式(12)更改为:

$$(l_i^{wp})^* = \frac{\xi^{ot} l^{ot} + \xi^{lo} l^{lo} + l^{wp}}{3} \quad (13)$$

另外,通过分析评论数据可以发现,观点词对中的观点目标和观点词有可能是一个单词或者一个词组,因此在对已知项进行定义时,需要通过适当的方法把可能是词组的已知项转换成复合向量来替代,具体方法如式(14)所示:

$$\mathbf{K} = \frac{\sum_i^{N_k} \alpha_i \mathbf{V}_i}{N_k} \quad (14)$$

其中, α_i 表示权重信息, \mathbf{V}_i 表示已知项中每个单词的词向量, N_k 表示已知项中包含的单词个数。通过将已知项中所有单词向量进行叠加,来最大可能地保存所有单词的依赖信息。

6 实验与结果

6.1 度量标准

本文采用精确度(P)、召回率(R)和 F 值 3 个指标来对实验结果进行评估。

$$P = \frac{|A_r|}{|A|}, R = \frac{|A_r|}{|W|}, F = \frac{2PR}{P+R} \quad (15)$$

其中,模型的提取结果集合为 A ,正确的观点词对集合为 W ,正确的提取结果集合为 A_r 。

6.2 数据集描述

本节选用的实验数据是已经被广泛应用于观点词对研究的样本信息数据集 Customer Review Dataset,该数据集来源于 Amazon 和 C|net,是观点词对提取研究工作常用的公用数据集,具体如表 1 所列。

表 1 Customer Review Dataset 的相关信息

Table 1 Relevant information of Customer Review Dataset

| Dataset | Training | Test | Total |
|----------|----------|------|-------|
| Apex | 666 | 74 | 740 |
| Canon | 538 | 59 | 597 |
| Nokia | 492 | 54 | 546 |
| Nikon | 311 | 35 | 346 |
| Creative | 1544 | 172 | 1716 |

该数据集中包含 5 种商品的评论信息,包括两类照相机(Canon 和 Nikon)、一类手机(Nokia)、一类 MP3 音乐播放器(Creative)和一类 DVD(Apex)。每类商品包含了多个用户的评价信息,评论信息按照单个句子进行存储,并在每条评论句子最前端标注出了该评论句子中包含的观点目标信息。表 1

列出了每类商品包含的句子个数,每个商品的评论数据被分成了两份,分别是训练数据和测试数据。由于该数据集中只包含了每条评论句子中的观点目标,因此本文按照已有的观点目标通过人工的方式标注了观点词,最终获取了每条评论句子中的观点词对信息并存储在每条评论句子的最前端。

6.3 实验和结果

表 2 列出了不同方法在多个数据集上的实验结果。

表 2 不同数据下不同方法的实验结果

Table 2 Experimental results of different methods

| 相机评论数据集 | | FASTR | Prop-dep | CR_WP | CODA |
|----------|----------------|-------|----------|-------------|-------------|
| Nikon | P | 0.68 | 0.87 | 0.86 | 0.85 |
| | R | 0.79 | 0.82 | 0.84 | 0.75 |
| | F ₁ | 0.73 | 0.84 | 0.85 | 0.79 |
| Canon | P | 0.71 | 0.83 | 0.84 | 0.85 |
| | R | 0.81 | 0.80 | 0.82 | 0.84 |
| | F ₁ | 0.76 | 0.81 | 0.83 | 0.84 |
| Nokia | P | 0.69 | 0.85 | 0.89 | 0.84 |
| | R | 0.77 | 0.86 | 0.89 | 0.75 |
| | F ₁ | 0.73 | 0.85 | 0.89 | 0.80 |
| Creative | P | 0.67 | 0.79 | 0.81 | 0.86 |
| | R | 0.79 | 0.81 | 0.80 | 0.81 |
| | F ₁ | 0.73 | 0.80 | 0.81 | 0.84 |
| Apex | P | 0.70 | 0.89 | 0.92 | 0.87 |
| | R | 0.78 | 0.82 | 0.85 | 0.89 |
| | F ₁ | 0.74 | 0.85 | 0.88 | 0.88 |

从实验结果中可以看出,本文提出的方法在 Canon, Creative 和 Apex 这 3 个评论数据集中的实验效果优于其他 4 组的实验结果,但在 Nikon 和 Nokia 两组评论数据集中表现不佳,从表 1 的评论数据信息中可以发现这两种商品所包含的评论数据量是最少的,因此可能存在模型没有得到充分训练的情况,导致提取效果受到了一定的影响。

(1)与 FASTR 模型相比,本文提出的模型的挖掘效果得到了明显提高。由于 FASTR 模型对观点目标和观点词的词性进行了限制,因此忽略了其他词性中可能包含的观点词对。此外,该模型采用最近邻算法进行观点词对的挖掘,其局限性在于窗口大小会有一定的限制,不能挖掘出窗口之外的观点目标或观点词,观点词对的提取过程没有摆脱距离上的约束。本文提出的模型是对整条评论句子进行分析的,因此不存在距离上的限制。

(2)与 Prop-dep 模型相比,本文提出的模型效果更好。由于 Prop-dep 模型需要人为定义观点词对中存在的语法关系,因此不可避免地会遗漏很多不常见的语法关系,从而导致不能有效提取出所有的观点词对。而本文提出的模型能够有效学习到待挖掘评论信息中存在的所有语法关系,并将这种语法关系转换为观点联系特征来高效地提取观点词对。

(3)与 CR_WP 模型相比,本文提出的模型在部分数据集上的效果更好。由于 CR_WP 模型采用随机游走的方法来更新包含观点词对的二部图信息,因此对于得到的新观点词对结点置信度和边的权重信息,无法判断信息更新的有效性。而本文提出的模型利用观点依赖特征对观点词对进行挖掘,

从而能够保证有重点地学习到真实的观点词对。

本文提出了两种 ODA 模型的生成方法,第一种是将观点词对中的观点目标作为已知项进行未知项提取的 TTO-ODA 方法;另一种是将观点词对中的观点词作为已知项进行未知项提取的 OTT-ODA 方法。两种模型的实验对比结果如图 4 所示。

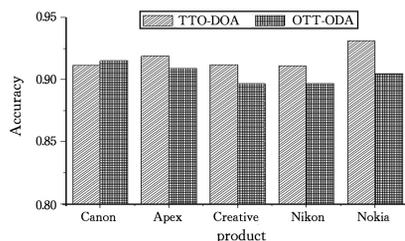


图 4 不同结构下 ODA 模型的实验效果

Fig. 4 Experimental results of ODA models with different structures

从实验结果可以看到, TTO-ODA 模型的实现效果普遍优于 OTT-ODA 模型,其原因可能是同一观点词可以修饰几种观点目标,但是同一观点目标大多数情况下只会被一种观点词修饰,因此相对于观点目标,观点词的提取会更加稳定。

结束语 本文提出了一种利用评论句子中存在的观点依赖联系进行观点词对的提取方法,主要结合 LSTM 能够长期保存信息的特点来提取评论句子中包含的具有强观点依赖联系的观点词对。本文首先介绍了基于观点依赖联系分析的观点挖掘模型(ODA),并对 ODA 模型具体的设计原理进行了详细的介绍;然后给出了将 ODA 模型扩展为 CODA 模型的过程;最后在有效的数据集上进行了验证性实验,证明了所提方法的有效性。

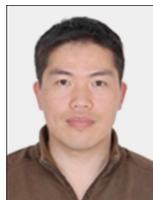
参考文献

- [1] DING X, LIU B, YU P S. A holistic lexicon-based approach to opinion mining[C]// International Conference on Web Search & Data Mining, 2008:231-240.
- [2] WU Y, ZHANG Q, HUANG X, et al. Phrase Dependency Parsing for Opinion Mining[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009:1533-1541.
- [3] XU L, LIU K, LAI S, et al. Mining Opinion Words and Opinion Targets in a Two-Stage Framework[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013:1764-1773.
- [4] WU C, WU F, WU S, et al. A Hybrid Unsupervised Method for Aspect Term and Opinion Target Extraction[J]. Knowledge-Based Systems, 2018;148:66-73.
- [5] DANILO E, DENILSON G, IVES P, et al. Analysis of Document Pre-Processing Effects in Text and Opinion Mining[J]. Information, 2018,9(4):100.
- [6] AGERRI R, RIGAU G. Language Independent Sequence Labeling for Opinion Target Extraction[J]. Artificial Intelligence, 2019,268:85-95.

- [7] DING X, LIU B. Resolving Object and Attribute Coreference in Opinion Mining [C] // The 23rd International Conference on Computational Linguistics Proceedings of the Main Conference. 2010;268-276.
- [8] ALMEIDA M S C, PINTO C, et al. Martins. Aligning Opinions: Cross-Lingual Opinion Mining with Dependencies[C]// Meeting of the Association for Computational Linguistics & International Joint Conference on Natural Language Processing. 2015: 408-418.
- [9] NGUYEN Thi Thanh Thuy, NGO Xuan Bach, et al. Cross-Language Aspect Extraction for Opinion Mining[C]//10th International Conference on Knowledge and Systems Engineering. 2018;67-72.
- [10] TU W, CHEUNG D, MAMOULIS N. Time-sensitive opinion mining for prediction[J]. Association for the Advancement of Artificial Intelligence, 2015:4214-4215.
- [11] ZHANG L, LIM S H, LIU B. Extracting and Ranking Product Features in Opinion Documents[C]// International Conference on Computational Linguistics. 2010;1462-1470.
- [12] HAI Z, CHANG K, CONG G. One seed to find them all: mining opinion features via association[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012;255-264.
- [13] QIU G, LIU B, BU J, et al. Opinion Word Expansion and Target Extraction through Double Propagation[J]. Computational Linguistics, 2011, 37(1):9-27.
- [14] BISHAN YANG C C. Joint Inference for Fine-grained Opinion Extraction[C]// Meeting of the Association for Computational Linguistics. 2013;1640-1649.
- [15] MOJICA L G, NG V. Fine-Grained Opinion Extraction with Markov Logic Networks[C] // 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). 2015;271-276.
- [16] JIANG X, LIN Y, LI Y, et al. Collective Extraction for Opinion Targets and Opinion Words from Online Reviews[C]// 7th International Conference on Cloud Computing and Big Data. 2016: 367-373.
- [17] LIN Y, JIANG X, et al. Semi-supervised collective extraction of opinion target and opinion word from online reviews based on active labeling[J]. Journal of Intelligent and Fuzzy Systems, 2017, 33(6):3949-3958.
- [18] WANG H, ZHANG C, YIN H, et al. A Unified Framework for Fine-Grained Opinion Mining from Online Reviews[C]// 49th Hawaii International Conference on System Sciences (HICSS). 2016;1134-1143.
- [19] LADDHA A, MUKHERJEE A. Aspect Specific Opinion Expression Extraction using Attention based LSTM-CRF Network [J]. CoRR abs. 2019;1902.02709.



ZHAO Wei, born in 1995, postgraduate. His main research interests include opinion mining and so on.



LIN Yu-ming, born in 1978, Ph.D, professor. His main research interests include opinion mining, knowledge graph, and massive data management.