

文章编号: 2095-2163(2019)03-0259-04

中图分类号: TP391.1

文献标志码: A

# 从餐馆评论中提取方面术语

雷 叶

(西安石油大学 计算机学院, 西安 710065)

**摘要:** 方面术语提取是基于方面的情感分析中的一个关键的任务,其目的是从在线用户评论中提取关键的方面术语。本文通过在现实数据上使用了一个解决方面术语提取任务的新框架。该框架通过挖掘2个有用的线索,即意见摘要和方面预测的历史。意见摘要是从整个输入语句中提取出来的关键词,以每个当前标记为条件进行方面预测,因此确定的摘要可以帮助对该标记进行方面预测。另一条线索是方面预测的历史,是从以前的方面预测中提取出来的关键词,以便利用坐标结构和标注模式约束来更好地做出方面预测。用此模型分析餐馆的用户评论,最终的实验结果则展示了良好的提取结果。

**关键词:** 方面提取;长短期记忆网络;注意力机制;展望

## Aspect term extraction for restaurant reviews

LEI Ye

(School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China)

**[Abstract]** Aspect term extraction is a key task in aspect-based sentiment analysis, whose purpose is to extract key aspect terms from online user reviews. This paper uses a new framework to address the aspect term extraction task by using it on real data. The framework digs through two useful threads, namely the opinion summary and the aspect forecast history. The opinion summary is a keyword extracted from the entire input statement, and the aspect prediction is based on each current tag, so the determined summary can help the aspect prediction of the tag. Another clue is the history of aspect prediction, which is the extraction of key words from previous aspect prediction in order to make better aspect prediction with coordinate structure and annotation pattern constraints. This model is used to analyze the restaurant user reviews, and the final experimental results show that the research achieves a good extraction result.

**[Key words]** aspect term extraction; Long Short-Term Memory Networks; attention mechanism; expectation

## 0 引言

随着互联网的快速发展,网上购物、网上点餐等方便快捷的生活方式日益深入人们的生活。与之相适应的是,人们在这些平台上发表的评论信息也正在呈指数级的方式增长。这些信息数量庞大,在一定程度上有着重要的研究价值。对这些评论信息进行分析,不仅能引导消费者的消费行为,而且有利于商家掌握消费者需求,从而有针对性地做出产品改进。快速有效地提取、利用网络信息已成为学界关注的焦点,从文本中提取方面术语也是其中的一个关键研究课题。现今,许多研究人员将方面术语提取表示为序列标记问题或标记级分类问题。传统的序列模型如条件随机域(CRFs)<sup>[1-3]</sup>、Long Short-Term Memory Networks(LSTMs)<sup>[4]</sup>以及SVM等分类模型<sup>[5]</sup>都被应用于方面术语提取的任务处理,性能较为合理。但是上述方法也有不足,也就是根据任务的定义,方面术语应该与表示意见的词相一致。因此,这些方法对非主观性句子中经常使用的方面

词的提取上就容易出现偏差,如“The restaurant installation is very comfortable, does not affect the dining”中的“restaurant”一词,不应该提取出来,因为这个句子中并没有表达任何观点。

有一些文献在进行方面术语提取研究时考虑了意见术语。文献[6]提出递归神经条件随机域(Recursive Neural Conditional Random Fields, RNCRF),在一个框架内开展方面和观点提取。通过联合抽取和基于依赖的表示学习,建立了方面意见关系的模型。RNCRF的一个假设是依赖关系解析将捕获同一句话中方面术语和观点词之间的关系,从而使联合提取受益。这种假设通常适用于简单的句子,但对于一些复杂的结构,如分句和括号,这种假设是很脆弱的。此外,由于RNCRF的网络结构依赖于输入的依赖树,因此存在依赖解析错误。CMLA<sup>[7]</sup>在不使用句法的情况下对方面意见关系进行建模。这是使2个任务能够通过注意机制共享信息。例如,方法中利用全局意见信息,直接计算方面原型和个人意见隐藏表示之间的关联得分,而后执

**作者简介:** 雷 叶(1996-),女,硕士研究生,主要研究方向:自然语言处理。

**收稿日期:** 2019-03-05

行加权聚合。然而,这种聚集可能会带来噪声。分析可知,这一缺陷多半在一定程度上是由注意机制继承而来的。

为了更好地利用意见信息辅助方面项的提取,研究将整个输入语句的意见信息提取作为意见摘要,这种提取是以特定的当前令牌为条件进行方面预测的。继而,将意见摘要作为当前方面预测的特征之一。以“the hotel is nice but not very quiet”这一句为例,当本文的模型对“hotel”这个词进行预测时,将运行生成一个以“hotel”为条件的对整个句子的意见总结。由于“hotel”与“quiet”(一个点评词)之间存在着很强的相关性,因此点评总结会传达更多关于“quiet”的信息,从而帮助预测“hotel”作为一个高概率的方面。需要注意的是,意见摘要是在辅助意见检测任务的初始意见特征基础上的,这些初始意见特征在意见词区分上已经显出一定成效。此外,研究还提出了一种新的转换网络,可以帮助增强“hotel”和“quiet”之间的良好关联,从而使生成的意见总结包含更少的噪音。

这里,对方面预测的研究进行梳理后可知:一方面,在序列标注中,之前时间步长的预测是减少当前预测误差空间的有用线索。例如,在  $B-I-O$  标注中(参见 2.1 节),如果之前的预测为  $O$ ,则当前的预测不能为  $I$ ;另一方面,研究中还观察到,有些句子包含多个方面的术语。例如,苹果在产品质量、美学、工艺技术等方面都是无与伦比的,而客户服务在各方面都有协调的结构。在此结构下,利用之前预测的常用方面术语(如产品质量),可以帮助模型找到不经常出现的方面术语(如工艺)。为了捕捉上述线索,本文研发的模型提取了之前方面检测的信息,以便更好地预测当前状态。

总而言之,本文将通过一个该框架在网站评论的应用,进行方面术语的提取。研发步骤如下,旨在利用意见摘要和方面检测历史来更准确地提取方面术语。过程中,使用了 2 个标准的长短时记忆网络(LSTMs)来构建初始方面和记录顺序信息的意见表示。为了在每个时间步将历史信息编码到初始方面表示中,截断历史注意,从最近的方面预测中提取有用的特性,并生成具有历史意识的方面表示。研究还设计了一个选择性的转换网络,在每一步都能得到意见摘要。也就是,文中应用方面信息来转换最初的意见表示,并将注意力投放在转换后的表示上以生成意见摘要。实验结果表明,本文用该框架提取了网站评论的关键词。

## 1 相关工作

研究最初,有学者基于意见词通常对于方面周围的观察,开发了一个用于处理方面术语提取的自引导框架。文献[8-9]中,基于复杂的句法模式对 aspect terms 和 opinion words 进行共提取。然而,在处理非正式的在线评论时,依赖语法模式会出现解析错误。针对这一缺陷,文献[10-11]采用基于单词的翻译模型。具体来说,这些模型将方面术语提取任务表示为单语单词对齐过程,而方面-意见关系由对齐链接捕获,而不是单词依赖关系。方面术语提取任务也可以表示为一个令牌级序列标记问题。最近,基于神经网络的模型,如基于 LSTM<sup>[4]</sup> 和 CNN<sup>[12]</sup> 方法成为主流。在此基础上,提出了联合提取方面和观点的神经模型。文献[6]在单个基于树的递归神经网络中完成这 2 项任务。其中的网络结构依赖于依存关系语法解析,但是用于非正式的评论很容易出错。CMLA<sup>[7]</sup> 由标准 GRUs 之上的多个注意层组成,提取方面和意见词。同样,MIN<sup>[13]</sup> 在多任务学习框架下,使用多个 LSTMs 交互地进行方面术语的提取和意见单词的提取。本文的框架与前述工作的不同之处在于:

(1)能够过滤意见摘要,在原始意见表示中加入每一步的方面特征。

(2)利用方面检测的历史信息,捕捉坐标结构和以前的方面特征。

研究至此,最终的实验结果表明,本文中使用的模型<sup>[14]</sup>有着出色的表现。

## 2 模型

### 2.1 方面术语

给定一个包含了  $T$  个单词的序列  $X = \{x_1, \dots, x_T\}$ , 方面术语提取可以表示为记号/单词级序列标记问题,用于预测方面标记序列  $y = \{y_1, \dots, y_T\}$ , 其中每个  $y_i$  来自一个有限的标签集  $Y = \{B, I, O\}$ , 由其描述了可能的方面标签。方面术语示例见表 1。

表 1 方面术语示例

Tab. 1 Sample of aspect terms

$X$	I	love	the	operation	system	and	preloaded	Software
$Y$	$O$	$O$	$O$	$B$	$I$	$O$	$B$	$I$

表 1 中,  $B$ 、 $I$  和  $O$  分别表示方面跨越空间的开始、内部(inside)和外面(outside)。

### 2.2 模型描述

该模型包含 2 个关键部分,即:截断历史记录

注意力和选择转换网络,用于捕获历史信息的方面检测和意见摘要。该模型建立在2个LSTMs上,分别为方面术语提取和辅助意见检测任务生成关键字表示。方法中将面向方面检测历史信息集成到当前的面向特征中,生成一个新的面向历史的感知表示。选择转换网络时,首先根据当前方面候选对象计算一个新的意见表示。然后,利用一个双线注意网络,根据新的意见表示与当前方面表示的关联,计算出意见摘要作为新意见表示的加权和。最后,将历史感知方面表示和意见摘要连接起来作为当前时间步长方面预测的特征。方面术语和意见总结的预测公式可表示为:

$$P(y_i^T | x_i) = \text{Softmax}(W_f^T y_i^T + b_f^T) \quad T = \{1, 2\}, \quad (1)$$

当  $T = 1$  时,表示方面术语网络中使用的公式;当  $T = 2$  时,表示意见总结中作为全连接层的预测。 $W_f^T$  和  $b_f^T$  表示全连接层的参数。

本次研究中,损失函数可写作如下数学形式:

$$L_T = -\frac{1}{T} \sum_{i=1}^T P(y_i^{T,g} | x_i^T) \odot \log P[(y_i^T | x_i)] \quad T = \{1, 2\}, \quad (2)$$

最终的损失函数表示为:

$$J = L_1 + L_2. \quad (3)$$

### 3 实验

#### 3.1 模型对比

本数据集选自网上截取的数据案例进行分析,由此研究得到的模型比较数据结果,详见表2。

表2 数据集和训练集

Tab. 2 Data sets and training sets

		句子	方面	包含方面的句子
数据1	训练集	3 045	2 358	1 484
	测试集	800	654	422
数据2	训练集	1 315	1 192	832
	测试集	685	542	401

同时,关于本次研究选用的对比模型可概述如下:

(1)LSTM:普通的双向LSTM,带有预先训练好的单词嵌入。

(2)CRF-1:带有基本特征模板的条件随机字段。

(3)CRF-2:带有基本功能模板和单词嵌入的条件随机字段。

(4)CMLA<sup>[6]</sup>:CMLA是一个多层架构,每层由2个耦合的GRUs组成,用来建模方面术语和意见词之间的关系。

本文的框架致力于提取以意见信息为辅助的方面术语,后两种模型则是联合提取方面和意见,研究中拟使用现有的意见词典来提供较弱的意见监督。对比结果见表3。

表3 各种方法对比结果

Tab. 3 The comparison of various methods results %

Models	数据1	数据2
CRF-1	72.77	62.67
CRF-2	74.01	67.54
LSTM	75.71	68.26
CMLA	77.80	70.73
方面术语提取	79.52	71.46

由对比结果可以知道,CMLA利用注意机制提取意见信息来帮助方面提取,本文提出的框架始终比其它框架表现得更好。对其性能优势可表述如下:

(1)在本文的模型中,意见摘要是在对当前方面特征进行选择转换后加以利用的,因此,意见摘要在一定程度上可以避免由于直接应用常规注意而产生的噪声。

(2)通过历史的关注,本文的模型可以在坐标结构中一些常用方面的指导下发现一些不寻常的方面。

带有基本特征模板的CRF在运行性能上仍有待完善,因此研究中特别添加了CRF-2作为另一个基线。由表3可知,在所有数据集上,使用word嵌入的CRF-2比CRF-1获得更好的结果。因此,上述对比表明,word嵌入是有用的,包含结构信息的嵌入可以进一步提高性能。

#### 3.2 方面提取结果

本文从美国的大众点评网站(www.yelp.com)爬取到香港地区的餐馆的评论内容,利用这些评论数据来组织仿真测试,对爬取的数据进行了分句处理后,则转入注意力机制的模型中进行实验,实验结果见表4。

表4 实验结果

Tab. 4 Experimental result

用户评论的句子	提取出的方面关键词	权重
Super filling and big portions of beef chunks.	super	0.056
	filling	0.275
	big	0.038
	portion	0.067
	beef	0.281
	chunk	0.282

实验结果表明,研究可以从句子中提取出相应

的关键词,为现在对句子的情感分析提供了帮助。

## 4 结束语

时下,餐馆评论是在自然语言处理方面受到多方重视与关注的研究内容。为了更准确地提取方面术语,研究探索了2种重要类型的信息,即历史信息的方面检测和意见摘要。设计了截断历史注意和选择性转换网络两部分。实验结果表明,该模型提取出的关键词都有助于后续对于句子做文本分析。本次研究对于深度学习知识的挖掘就是为了给人们的现实生活带来更多便利,做句子的情感分析也是大数据分析领域的一个热门研究方向,可以预期在未来的相关工作中也将会收获到更大的研究进展和更多的应用成果。

## 参考文献

- [1] CHERNYSHEVICH M. Ihs r&d belarus: Cross-domain extraction of product features using CRF [C]//Proceedings of the 8<sup>th</sup> International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistic, 2014: 309-313.
- [2] TOH Z, WANG Wenting. Dlirec: Aspect term extraction and term polarity classification system [C]// Proceedings of the 8<sup>th</sup> International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistic, 2014: 235-240.
- [3] TOH Z, SU Jian. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features [C]// Proceedings of the 10<sup>th</sup> International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California: Association for Computational Linguistic, 2016: 282-288.
- [4] LIU Pengfei, JOTY S, MENG H. Fine-grained opinion mining with recurrent neural networks and word embeddings [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistic, 2015: 1433-1443.
- [5] MANEK A S, SHENOY P D, MOHAN M C, et al. Aspect term

extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier[J]. World Wide Web, 2017, 20(2): 135-154.

- [6] WANG Wenya, PAN Jialin, DAHLMIEIER D, et al. Recursive neural conditional random fields for aspect-based sentiment analysis[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistic, 2016: 616-626.
- [7] WANG Wenya, PAN Jialin, DAHLMIEIER D, et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). San Francisco, California, USA: AAAI, 2017: 3316-3322.
- [8] POPESCU A M, ETZIONI O. Extracting product features and opinions from reviews [C]// HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: ACM, 2005: 339-346.
- [9] QIU Guang, LIU Bing, BU Jiajun, et al. Opinion word expansion and target extraction through double propagation [J]. Computational Linguistics, 2011, 37(1): 9-27.
- [10] LIU Kang, XU Liheng, ZHAO Jun. Opinion target extraction using word-based translation model [C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: ACM, 2012: 1346-1356.
- [11] LIU Kang, XU Hengli, LIU Yang, et al. Opinion target extraction using partially-supervised word alignment model [C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, China: ACM, 2013: 2134-2140.
- [12] 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.
- [13] LI Xin, LAM W. Deep multi-task learning for aspect term extraction with memory interaction [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistic, 2017: 2886-2892.
- [14] LI Xin, BING Lidong, LI Piji, et al. Aspect term extraction with history attention and selective transformation [J]. arXiv preprint arXiv:1805.00760, 2018.

(上接第258页)

比如地域、运营商等进行详细分析,同时还详细研究了手机号对应的手机设备上一次插卡的电信运营商为海外运营商的特性,将欺诈的画像做出更加精细的描绘,依据结果和高风险维度结论,可以建立黑名单,可以形成一些防范欺诈的手机号规则引擎或者反欺诈模型,对互联网金融的反欺诈应用和优化具有很大的实际价值。

## 参考文献

- [1] 薛又轩. 大数据应用在互联网金融借贷行业的实践与探索[J].

银行家, 2016(5): 119-121.

- [2] 刘扬, 姬建华. 大数据在传统贷款风控中的应用 [J]. 科技传播, 2018, 1(3): 119-120, 168.
- [3] 丁濛濛. 基于规则引擎的互联网金融反欺诈研究 [J]. 电脑知识与技术, 2018, 14(1): 1-3.
- [4] 李赞妮. 神经网络模型在银行互联网金融反欺诈中的应用探索 [J]. 金融科技时代, 2018, 1(8): 24-28.
- [5] 窦路路, 石秀金. 基于深度学习的银行卡交易反欺诈技术研究 [J]. 智能计算机与应用, 2018, 8(4): 85-87, 91.
- [6] 仵伟强, 后其林. 基于机器学习模型的消费金融反欺诈模型与方法 [J]. 现代管理科学, 2018, 1(10): 51-54.