硕士学位论文

结合观点评价词的属性级情感分析研究

RESEARCH ON ASPECT-BASED SENTIMENT ANALYSIS WITH OPINION TERMS

张义策

哈尔滨工业大学

2020 年 6 月

国内图书分类号：TP391.1　　学校代码：10213
国际图书分类号：004.9　　密级：公开

工学硕士学位论文

结合观点评价词的属性级情感分析研究

硕士研究生：张义策

导　　师：刘远超副教授

申请学位：工学硕士

学　　科：计算机科学与技术

所在单位：计算机科学与技术学院

答辩日期：2020 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.1
U.D.C: 004.9

Dissertation for the Master's Degree in Engineering

RESEARCH ON ASPECT-BASED SENTIMENT ANALYSIS WITH OPINION TERMS

Candidate:	Zhang Yice
Supervisor:	Associate Prof. Liu Yuanchao
Academic Degree Applied for:	Master of Engineering
Specialty:	Computer Science and Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June, 2020
Degree-Conferring-Institution:	Harbin Institute of Technology

摘 要

属性级情感分析关注的是文本中评价对象及其情感倾向。如句子“屏幕不行，电池可以”中，评价对象为“屏幕”和“电池”，相应的情感倾向为负面和正面。观点评价词，也称情感词，是指文本中表达主观情感的单词或者短语，如上面例子中的“不行”和“可以”。直观来看，观点评价词对情感倾向的确定非常重要。但目前的大多数属性级情感分析工作对观点评价词的关注比较少。因此，本文重点关注了属性级情感分类中的观点评价词。

本文的主要研究工作如下：

1. 面向评价对象的评价词抽取。以往的工作大多将观点评价词抽取视为一个辅助任务，而没有将其视为属性级情感分类中一类重要特征。为了实现此类特征的提取，本文关注了面向评价对象的评价词抽取任务，并基于 BERT 提出了一个抽取模型。对于评价词抽取任务，除了使用序列标注的方法，本文还尝试了抽取式问答中常用的边界预测的方法。实验结果表明，本文所提出的模型在 f1 值上远优于现有的方法（7%-8%），边界预测的方法在 f1 值上略优于序列标注的方法，在训练和解码速度上远超序列标注的方法。

2. 结合观点评价词的属性级情感分类。得到观点评价词之后，本文研究了观点评价词在属性级情感分类中的作用机制。首先，设计了两个指标对观点评价词在属性级情感分类中的贡献度进行量化。在两个量化指标上的实验表明，随着训练的进行，模型将更加关注文本的观点评价词。为了更直接地学习到这种关注，本文对 BERT 模型的自注意力层上进行了修改。实验表明，使用本文所提出的模型来指示出文本的观点评价词后，相比于不指示观点评价词，属性级情感分类的性能得到了提高，准确率最高提高了 1.45%。

3. 结合观点评价词的端到端的属性级情感分析。以往的端到端的属性级情感分析工作都只输出评价对象及其情感倾向，而没有在输出结果中结合观点评价词。为了结果的完整性，本文关注了观点三元组抽取任务。观点三元组，由评价对象、相应的观点评价词及情感倾向构成。现有的工作将观点三元组抽取建模为两个阶段，首先分别抽取评价对象和观点评价词，然后两两组对送入分类器中判断二者的对应关系。本文则提出了一个端到端的模型去进行观点三元组的抽取，在抽取评价对象和观点评价词的同时，抽取它们的关系。实验表明，本文提出的模型在多个数据集的多个指标上优于现有的方法，在观点三元组的 f1 值上最高提高了 13%。

关键词：属性级情感分析；观点评价词；端到端的模型

Abstract

Aspect-based sentiment analysis(ABSA) aims to determine the sentiment polarity of a review on a given aspect term (also named opinion target). For example, in the sentence "*the screen is not good while the battery is OK*", the aspect term are "*screen*" and "*battery*", and the corresponding sentiment polarity are negative and positive. the Opinion term, also called sentiment words, refer to words or phrases that express subjective emotions in the text, such as "*not good*" and "*OK*" in the above example. Intuitively, opinion terms are very important for determining sentiment polarity. However, most of works on ABSA pay less attention to opinion terms. Therefore, this dissertation focuses on opinion terms of ABSA.

The main work of this dissertation is as follows:

1. Target-oriented opinion words extraction(TOWE). Most of the previous work take the opinion term(word) extraction as an auxiliary task, rather than regarding the opinion term as an important feature in sentiment classification. In order to extraction opinion terms for a specific target, we focuses on TOWE task, and proposes an extraction model based on BERT. For opinion term extraction, in addition to sequence labeling, we also try the boundary prediction method, which is commonly used in extractive question answering. The experimental results show that the proposed model is far superior to the existing method (7%-8%) in f1 measure, and the boundary prediction method is slightly better than the sequence labeling in f1 measure, and have a faster training and decoding speed.

2. aspect-level sentiment classification with opinion terms(ASCO). After obtaining opinion terms, we studies the mechanism of opinion terms in aspect-level sentiment classification(ASC). Two indicators are designed to quantify the contribution of opinion terms in ASC. Experiments on the two quantitative indicators show that as the training progresses, the model will pay more attention to opinion terms. In order to learn this attention more directly, we modify the self-attention layer of BERT model. Experiments show that with our modified BERT to indicate the opinon terms of the textm, the performance of ASC is improved and the impovement of the accuracy score is up to 1.45%.

3. End-to-end ABSA with opinion terms. Previous end-to-end ABSA model only output the targets and the corresponding sentiment polarity, without opinion terms. For

the sake of completeness of ABSA results, we focuses on the task of aspect sentiment triplet extraction(ASTE), in which a triplet is composed of a target, the corresponding opinion terms and the sentiment polarity. In the existing work, ASTE is modeled as two stages. In stage one, targets and opinion terms are extracted. In stage two, they are coupled to form candidate pairs pool and then sent to a classifier to identify the legitimate ones. This paper proposes an end-to-end model for ASTE, extracting targets and opinion terms among with their relationship. Experiments show that the proposed model is superior to the existing methods on multiple indicators in multiple datasets, and the f1 measure on triplets is increased by up to 13%.

Keywords: aspect, sentiment analysis, opinion term, end-to-end model

目　录

第 1 章 绪论

1.1 课题背景及研究意义

文本情感分析，也称意见挖掘，旨在对带有情感色彩的主观文本进行研究分析。这些主观文本表达了人们对某一对象的观点、意见、态度、评价、立场等，常带有较强的情感倾向，毫无疑问具有非常广泛的应用价值。借助文本情感分析，互联网上存积的和不断产生的大量主观文本可以得到量化且直观的分析，为政府和企业提供决策依据。依据分析的粒度，文本情感分析可分为文档级情感分析、句子级情感分析和属性级情感分析。其中，现阶段的研究工作主要关注的是属性级情感分析。

作为一个细粒度的自然语言处理任务，属性级情感分析 (aspect-based sentiment analysis, ABSA) 关注的是文本在特定评价对象 (也称属性词，aspect term/opinion target) 上的的情感倾向 (通常包括正向、负向和中性)。如在句子“电池不错，屏幕太小”中，其中的评价对象为“电池”、“屏幕”，相应的情感倾向分别为正面、负面。

属性级情感分析通常被拆分为两个子任务：(1) 给定文本，识别文本中的评价对象，称为评价对象抽取 (aspect term extraction)；(2) 给定文本和文本中的评价对象，识别该文本在此评价对象上的情感倾向，称为属性级情感分类 (aspect-level sentiment classification)。现阶段的属性级情感分析研究的任务定义和实验数据大都来源于 SemEval 系列国际语义评测[1–3]。有研究者[4–6] 为提高评价对象和情感倾向预测结果的可解释性及性能，在 SemEval 评测数据上标注了句中的观点评价词，并引入了评价词抽取任务 (opinion term extraction)，即抽取文本中表达主观情感的单词或短语，如句子“电池不错，屏幕太小”中的“不错”、“太小”。

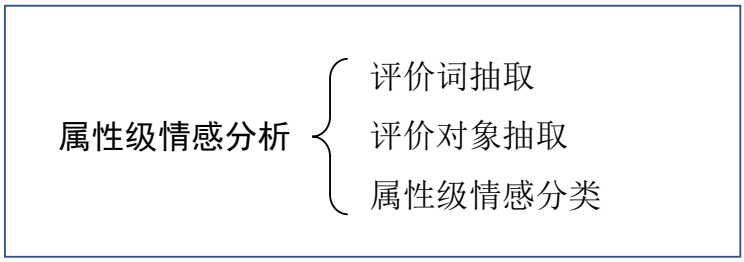


图 1-1 属性级情感分析中的三个子任务

1.2 国内外研究现状

由于 SemEval 系列评测在属性级情感分析领域发挥非常重要的作用，因此在介绍具体的模型和方法之前，先对这些评测和数据进行了简单的介绍。然后，分别介绍属性级情感分类、评价对象抽取以及两个任务的联合学习的研究现状。

1.2.1 SemEval 评测中的属性级情感分析任务

现阶段的属性级情感分析研究的任务定义和实验数据大都来源于 SemEval 系列国际语义评测。在 SemEval2014 的任务 4 定义了 4 个子任务：

(1) 评价对象抽取 (aspect term extraction)。给定一个句子，该任务要求返回句子中存在的所有评价对象，这些评价对象是特定实体的不同属性。如输入句子“我喜欢这里的员工和服务，而不是食物”，返回 {员工, 服务, 食物}。

(2) 预测特定评价对象的情感倾向 (aspect term polarity)。给定一个句子及句子中出现的评价对象，输出相应的情感倾向，包括四类：正向、负向、中性和冲突。如输入句子和评价对象“我讨厌他们的烤肉，但是他们的沙拉不错”，对“烤肉”输出负面，对“沙拉”输出正面。

(3) 属性类识别 (aspect category detection)。给定一个句子，要求输出句子中所讨论的属性类。与子任务 (1) 中的评价对象不同，这里的属性类定义地更加粗糙，而且属性类本身不需要出现在给定的句子中。在发布的餐饮数据集中，定义了 5 个属性类，分别是食物、服务、价格、环境和其他。如对于句子“这个餐厅太贵了，不过菜单还是棒棒哒”，输出价格, 食物。

(4) 预测在属性类上的情感倾向 (aspect category polarity)。给定一个句子以及相应的属性类，判断句子在属性类上表达的情感倾向。如对于句子“这个餐厅太贵了，不过菜单还是棒棒哒”以及属性类集合价格, 食物，对于属性类“价格”输出负向，对属性类“食物”输出正向。

考虑到预测评价对象的情感倾向和预测属性类上的情感倾向两个子任务的相似性，在本文中将他们统称为属性级情感分类 (aspect-level sentiment classification)。并且，与现有的大多数研究一样，本文主要关注文本中的评价对象及其情感倾向。

随后的 SemEval2015 和 SemEval2016，在 SemEval2014 任务定义的基础上继续举办了属性级情感分析的评测，并发布了更多的数据集。属性级情感分析中的大多数研究都在这些数据集上进行的。如图1-2所示，这些数据集标注了文本中的评价对象和相应的情感倾向。此外，一些研究者[5,6] 在 SemEval 数据集的基础上标注了文本中观点评价词；还有一些研究者[4,7] 在此基础上标注了观点评价词和评价对象的对应关系。

```
<sentence id="813">
   <text>All the appetizers and salads were fabulous, the steak was mouth
   watering and the pasta was delicious!!!</text>
   <aspectTerms>
      <aspectTerm term="appetizers" polarity="positive" from="8" to="18"/>
      <aspectTerm term="salads" polarity="positive" from="23" to="29"/>
      <aspectTerm term="steak" polarity="positive" from="49" to="54"/>
      <aspectTerm term="pasta" polarity="positive" from="82" to="87"/>
   </aspectTerms>
</sentence>
```

图 1-2 SemEval2014 属性级情感分析任务的数据样例

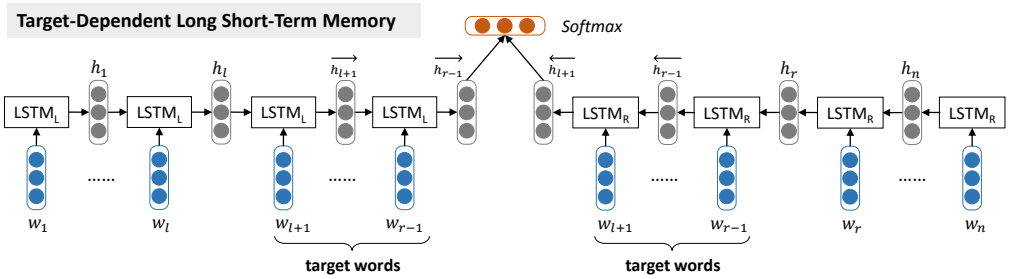
1.2.2 属性级情感分类

文档级情感分类可以视做一个文本分类问题，直接使用 LSTM、textCNN、RCNN 等文本分类模型即可解决。而属性级情感分类的输入除了句子本身外，还有句子中的评价对象。因此，如何将评价对象融合到已有的文本分类模型中是一个关键问题。对于句子“我讨厌他们的烤肉，但是他们的沙拉不错”，在决定在评价对象“烤肉”上的情感倾向时，模型需要捕捉的关键特征是“讨厌”。即需要通过一个机制，使用评价对象对句子中的信息进行选择，然后进行分类。依据信息选择（或者称为信息融合）的方式，早期①的属性级情感分类模型可以大致分为三类。

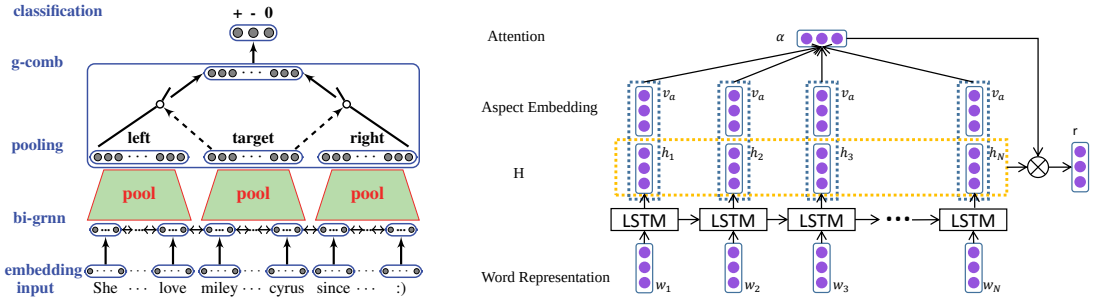
(1) **基于位置信息的方法**。由于评价对象在句子中出现，最为直接的方式便是利用评价对象的位置信息。如在 TD-LSTM[8] 中，首先通过双向 LSTM 后得到句子的表示，接着使用评价对象对应位置上的表示进行后续的分类。此类方法中比较有代表性的还有 Duy-Tin Vo 等人 (2015)[9] 和 Meishan Zhang 等人 (2016)[10] 的工作，它们在 CNN 上结合了评价对象的位置信息。然而，只使用位置信息无法充分地完成对所需信息的选择。如在句子“续航真的可以，屏幕是可惜了”，评价对象“屏幕”，与“可以”和“可惜”的距离相等。仅仅使用位置信息可能会造成信息的遗漏和混淆，需要其他更加强大的信息选择方法。

(2) **基于注意力机制的方法**。相比基于位置的方法，注意力机制则显得更加灵活。因此这类方法中涌现了大量工作，比较有代表性的是 AT-LSTM[11]。它在使用双向 LSTM 获得句子的表示后，不再只使用特定位置的表示，而是将评价对象作为查询，将句子中每个词的表示作为键值，通过注意力机制实现信息的选择，获得了比 TD-LSTM 更好的性能。类似的研究还有 Sentic LSTM[12]、IAM[13]、BILSTM-ATT-G[14]、HEAT[15]、He Ruidan 等人的工作[16,17] 和 Tnet[18]，其中 Sentic LSTM[12]、HEAT[15] 和 He Ruidan 等人的工作[17] 使用了外部知识。此外，也有研

① 这里说指的早期是指 2014 年-2018 年这段时期。



a) TD-LSTM 的模型结构



b) Meishan Zhang 等人 (2016)[10] 的工作中提出的模型

c) AT-LSTM 的模型结构

图 1-3 早期的属性级情感分类模型

究者使用记忆网络进行属性级情感分类，相关的研究工作有 MemNet[19]、RAM[20]、Conv-Memnet[21]。

(3) **其他方法**。除了注意力机制，门控机制也是一种优秀的信息选择方法，在运行速度上优于注意力机制。但由于其不够灵活，所以相关工作比较少。典型的工作是 GCAE[22]，提出了 GTRU(gated tanh-relu units) 模块来代替原本的卷积核。首先，句子通过词向量层得到 $X = [v_1, v_2, \cdots, v_L]$，随后 X 被输入到在 GTRU 模块中。在 GTRU 模块中，一方面如式1-2所示，使用窗口为 k 的卷积核在 X 上生成句子的特征表示；另一方面如式1-1所示，综合 X 和评价对象的表示 v_a 生成多个门，用来对句子的特征表示进行选择。

$$a_i = \mathrm{ReLU}(X_{i:i+k} * W_a + V_a v_a + b_a) \tag{1-1}$$

$$s_i = \mathrm{Tanh}(X_{i:i+k} * W_s + b_a) \tag{1-2}$$

$$c_i = s_i \times a_i \tag{1-3}$$

AF-LSTM[23] 则是提出了 association layer 来进行评价对象和句子的融合。PhraseRNN[24] 建立了一个基于评价对象的二叉短语依存树 (Target Dependent Binary Phrase Dependency Tree)，然后使用递归神经网络 (recursive neural network) 预测评价对象上的情感倾向。

当然，这三种方法也不是相互独立的，可以在基于注意力机制的方法和基于门控机制的方法中使用位置信息，也可在模型设计中将注意力机制和门控机制结合。在 2018 年，随着 BERT 为代表的预训练语言模型的兴起，在预训练语言模型的基础上进行属性级情感分类成为了趋势。比较有代表性的是 Sun Chi 等人 (2019)[25] 的工作，通过将评价对象构造为一个辅助句子，将属性级情感分类转化为一个句对分类的任务，然后使用 BERT 模型求解。比如对于评价对象“电池”，构造辅助句子“电池上的情感倾向是正面/负面/中性”，然后分别输出正确或错误。此外，Hu Xu 等人 (2019) 的工作[26] 探究了后训练 (post train) 对模型性能的影响。

1.2.3 评价对象抽取

大体上，评价对象抽取中的方法可以分为有监督的方法和无/半监督的方法。由于句子中评价对象的标注相对困难，2014 年以前，出现了非常多的无监督和半监督的工作，它们的主要思想是将一类评价对象视为主题模型中的一个主题，或者根据单词之间的联系从一个种子集合进行扩展。本文主要关注有监督的评价对象抽取，因此不对这些方法进行过多介绍。而有监督的方法可以依据是否结合观点评价词分为两类，下面分别进行介绍。

有监督的评价对象抽取可以建模为一个序列标注问题，使用流行的 LSTM-CRF[27] 模型即可解决。也有研究者使用 CNN 进行评价对象的抽取，提出了 DE-CNN[28]。在 DE-CNN 中，领域词向量和通用词向量被结合到一起，实验表明这对模型性能有较大的提高。除此之外，还有一些应用句法信息的工作[29,30]。而 19 年的一个工作[31] 则较为新颖地将评价对象抽取任务建模为一个 Seq-to-Seq 的任务。

句子	这	手	机	电	池	不	错	，	就	是	屏	幕	太	小	。
字标签	O	O	O	BA	IA	BO	IO	O	O	O	BA	IA	BO	IO	O

图 1-4 评价对象和观点评价词的联合抽取任务中的联合字标签举例。其中 BA/BO 表示评价对象/观点评价词的开始字符，IA/IO 表示评价对象/观点评价词的中间字符，O 表示其他字符。

2016 年，Wang Wenya 等人[5] 率先在 SemEval 属性级情感分析数据集上标注了文本中的观点评价词，并进行了评价对象和观点评价词的联合抽取。如图1-4所示，可以通过建立一套联合的字标签融合为将评价对象和观点评价词抽取建模成一个序列标注任务。考虑到评价对象和观点评价词往往是相互对应的，如图1-4中的“电池”和“不错”，这与命名实体识别任务 (named entity recognition, NER) 有所不同，因为 NER 任务中不同类型的实体之间通常没有很强的对应关系。因此，Wang Wenya 等人的工作提出了 RNCRF[5] 的模型，使用了递归神经网络来建模评

价对象和观点评价词之间的关系。

随后，在 2017 年，Wang Wenya 等人又提出了 CMLA[32](coupled multi-layer attentions) 模型，即使用多层的注意力机制实现二者的交互。其中，每一层执行两次注意力，分别用于抽取评价对象和观点评价词。同年，Li Xin 等人则通过记忆机制完成二者的交互，提出了 MIN[6](memory interaction network) 模型。在 MIN 模型中，句子首先经过两个 LSTM 分别得到评价对象抽取和观点评价词抽取两个任务的隐含表示 H^A 和 H^O，然后计算相应的摘要表示 m^A 和 m^O，随后将 H^A 和 m^O 拼接起来用于后续的评价对象抽取，将 H^O 和 m^A 拼接起来用于后续的观点评价词抽取。CMLA 和 MIN 中评价对象和观点评价词的交互机制可以简化为图1-5。据我所知，没有发现文献直接比较二者的性能，但二者均与 RNCRF 的性能进行对比，根据比较结果推测二者的性能相差无几。

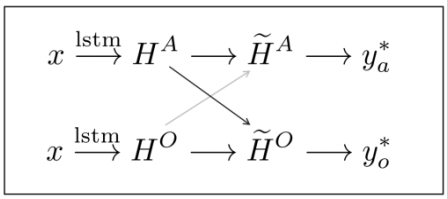


图 1-5 评价对象抽取和观点评价词抽取两个任务交互的示意图。图中的斜线在 CMLA 中指的是注意力机制，在 MIN 中指的是记忆机制。y_a^* 指的是评价对象字标签的预测值，y_o^* 指的是观点评价词字标签的预测值。

除此之外，18 年提出的 HAST[33] 以及 19 年提出的 DOER[34]，使用了情感词典来协助评价对象的抽取。

1.2.4 评价对象抽取和属性级情感分类的联合学习

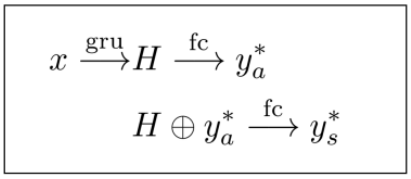
评价对象抽取是一个序列标注任务，而属性级情感分类则是给定评价对象的分类任务。为实现二者的联合学习，如图1-6所示，研究者们常通过将属性情感分类也建模为一个序列标注任务，然后就可以使用一个联合的字标签进行求解。这里的联合字标签可以理解为类型为某种情感倾向的实体。

句子	这	手	机	电	池	不	错	，	就	是	屏	幕	太	小	。
ATE	O	O	O	B	I	O	O	O	O	O	B	I	O	O	O
ASC	O	O	O	P	P	O	O	O	O	O	N	N	O	O	O
联合	O	O	O	BP	IP	O	O	O	O	O	BN	IN	O	O	O

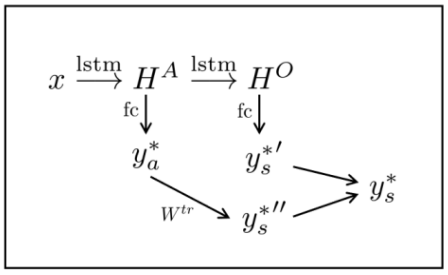
图 1-6 评价对象抽取和观点评价词抽取的联合字标签举例。其中，B 和 I 分别表示评价对象的开始字符和中间字符，P 和 N 分别表示正向情感和负向情感，O 表示其他字符

2018 年，Ma Dehong 等人首次提出评价对象抽取和属性级情感分类的联合学

习任务，并提出了 HMBi-GRU[35](hierarchical multi-layer bidirectional gru) 模型。如图1-7 a)所示，评价对象上字标签的预测值参与了联合的字标签预测中。但在实际的应用，联合字标签的方法也许会出现两种不一致问题：(1) 边界不一致，即模型输出的评价对象上的字标签和联合字标签中评价对象的边界不相同；(2) 情感不一致，即在模型输出的联合字标签中同一个评价对象内情感极性不一致。



a) HMBi-GRU 模型的简易示意图。其中，x 表示句子，gru 表示多层双向 GRU，fc 表示全连接层，$\oplus$ 表示矩阵拼接。图中，y_a^* 指的是评价对象字标签的预测值，y_s^* 指的是联合字标签的预测值。



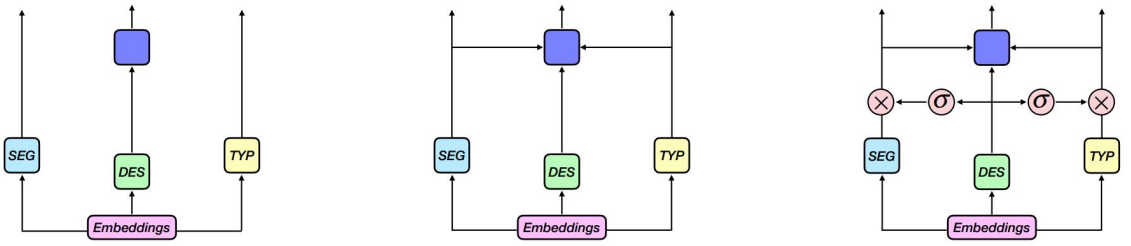
b) Li Xin 等人[36] 所提出模型的简易示意图。其中，x 表示句子，fc 表示全连接层，W^{tr} 为评价对象字标签到联合字标签的转移矩阵。

图 1-7 两个评价对象抽取和属性级情感分类的联合学习模型的简易示意图

因此，2019 年，Li Xin 等人[36] 相应地提出了两种机制来缓解这两种不一致。这里主要介绍其中的 BG(Boundary Guidance) 机制，即在得到评价对象字标签的预测值之后，如图1-7 b)所示，使用一个转移矩阵将其映射为联合字标签上的一个预测值，该预测值参与到最终预测值的生成过程中。实验结果表明，BG 机制可以有效地提高模型整体上的性能。同年，Hu Minghao 等人的工作[37] 为了解决情感不一致性的问题，放弃了联合字标签的方法，使用了 extract-then-classify 的模式。在该模式中，先进行评价对象的抽取，然后在评价对象的表示上进行属性级情感分类。实验表明在 BERT 模型上，extract-then-classify 模式优于联合字标签的方法。

Zhang Xiao 等人 (2019)[38] 则从 modular learning 的角度考虑，首先将联合字标签拆解为两个部分字标签（即评价对象的字标签和情感分类的字标签），然后使用部分字标签的预测值辅助联合字标签的预测。如图1-8所示，文章设计了三种辅助模型，其中带门控机制的信息流动模型 LSTM-CRF-TI(G) 取得了最好的成绩，比直接进行联合标签的序列标注在 f1 提高了约 2%。Luo Huaishao 等人 (2019)[34] 则使用了一个交叉共享单元 (cross-shared unit) 完成评价对象抽取和情感分类的交互。He Ruidan 等人 (2019)[39] 将评价对象和观点评价词的联合抽取、属性级情感分类、文档级情感分类、文档级领域分类这个四个任务置于一个框架下，使用共享编码

器以及消息传递机制 (message-passing mechanism) 的方式进行联合学习。



(a) LSTM-CRF-T (b) LSTM-CRF-TI (c) LSTM-CRF-TI (G)

图 1-8 Zhang Xiao 等人 (2019)[38] 中提出的三种模型。其中，SEG 指评价对象抽取，TYP 指情感分类，DES 指联合学习。

在上述的工作中，Li Xin 等人 (2019)[36]、Luo Huaishao 等人 (2019)[34]、He Ruidan 等人 (2019)[39] 在模型中使用了观点评价词或情感词典。但 Li Xin 等人 (2019)[36] 和 Luo Huaishao 等人[34](2019) 的工作中只是将观点评价词的抽取简单的视为一个辅助任务，而没有过多的考虑其与情感分类之间的联系。而 Ruidan 等人 (2019)[39] 提出了一个 opinion transmission 机制，通过在自注意力机制中增加观点评价词注意力权重的方式来增强文本的表示。

在 2020 年，Wan Hai 等人[40] 将属性类情感分类和属性词情感分类结合在一起，定义了一个新的任务：Target-Aspect-Sentiment Detection(TASD)。即给定一个句子，输出句子在某个属性类上的评价对象和情感倾向。同年，Peng Haiyun 等人 (2020)[7] 提出了 ASTE(aspect sentiment triplet extraction) 任务，即给定一个句子，输出句子中的观点三元组，包括评价对象、观点评价词和情感倾向。

1.2.5 研究现状的总结及现有工作的不足

本节叙述了属性级情感分析任务上的研究现状。属性级情感分析任务可以分为两个子任务：评价对象抽取和属性级情感分类。早期的研究工作主要集中在属性级情感分类任务上。在属性级情感分类中，为了进行句子和评价对象信息的有效融合，最早人们提出了基于位置的方法，随后基于注意力机制的方法开始大放异彩，大量的相关模型被提出。2018 年后，研究重点慢慢变为两个子任务的联合学习（即端到端的属性级情感分析）。同时，以预训练语言模型为基础的研究也慢慢增多。

本人认为这些研究工作有以下不足：

(1) 对属性级情感分析中的观点评价词的研究较少。观点评价词，也称情感词，是指文本中表达主观情感的单词或短语。直观来看，观点评价词对于文本情感极性的预测至关重要，一方面可以提高情感倾向预测结果的可解释性，另一方面可

以大大提高情感倾向预测的准确率。如果给观点评价词赋予一个分数来表示其情感倾向和强度，一系列的观点评价词被称为情感词典 (sentiment lexicon)。在文档级情感分类中，关于情感词典的相关研究工作比较多。这些工作主要关注的是，如何挖掘和扩展情感词典，以及如何将情感词典应用在情感分类中。然而，在属性级情分析中，与观点评价词和情感词典的相关研究不多。关于观点评价词，大部分现有的研究工作都只是将观点评价词的抽取简单地视作一个辅助任务，来提高评价对象抽取和情感分类的性能。本文认为这些研究工作对观点评价词的研究不够深入，如没有挖掘观点评价词与评价对象的对应关系，没有将观点评价词作为情感分类中的一类重要特征进行应用等。因此，这些研究工作对属性级情感分析中对观点评价词的利用是不够充分的。

(2) 大部分提出的模型和方法可扩展性不够强。早期的研究工作大多将评价对象抽取和属性级情感分类视为独立的子问题进行独立的研究。如此得到的模型可以通过流水线的方式联合起来，即先应用评价对象抽取模型，得到句子中的评价对象；然后对所抽取到的评价对象进行情感极性分类。流水线模式不仅会带来错误累计问题，而且无法利用到任务间的关联性。因此，近些年出现了一些同时抽取评价对象以及情感倾向的研究工作。然而，这些流水线的模型和端到端的模型，大多依赖复杂的模型结构和机制进行不同对象间的交互及任务间的交互，其可扩展性不强。这导致这些模型通用性不够强。随着以 BERT[41] 为代表的预训练语言模型的流行，以往研究工作中的模型和方法难以得到有效的应用。

1.3 本文主要研究内容

本文重点关注了属性级情感分析中的观点评价词。如图1-9所示，在本文的研究中，一条完整的用户观点包括三部分：评价对象、观点评价词以及相应的情感倾向。本文的最终目标是构建一个可以从大规模评论文本中挖掘用户观点的端到端的模型。为了实现这一目标，本文分三步开展了研究工作。首先，研究了如何对文本中特定的评价对象对应的观点评价词进行抽取；随后，研究了如何将所得到的观点评价词应用到属性级情感分类任务中；最后，研究了一个结合观点评价词的端到端的属性级情感分析模型，以实现从评论中抽取观点三元组的目的。因此，本文的主要研究内容可以分为三个部分：

(1) 第二章对面向评价对象的评价词抽取工作进行了研究。考虑到目前的大部分工作没有挖掘评价对象和评价词的对应关系，研究了如何抽取评价对象对应的观点评价词。这部分重点探讨了如何进行评价对象和句子的融合，以及除序列标注之外的评价词抽取方式。

(2) 第三章对属性级情感分类中的观点评价词进行了研究。为了研究观点评价词在情感分类中的重要性和作用机制，在自注意力机制上提出了两个指标，来量化观点评价词在情感分类中的贡献度。在所提出指标上的实验结果表明，在属性级情感分类中，模型会对观点评价词予以更大的关注。随后，提出了一个结合观点评价词的属性级情感分类模型，通过修改 BERT 模型中的自注意力机制来指示出文本中的观点评价词。

(3) 第四章对结合观点词的端到端属性级情感分析（即观点三元组抽取）进行了研究。不同于以往的研究工作，在这一部分一个完整观点不仅包含评价对象及其情感倾向，还包含相应的观点评价词。观点三元组抽取的难点在于如何在抽取评价对象和观点评价词的同时，确定它们之间的对应关系。为此，首先抽取句子中的评价对象，然后为评价对象中的每个词抽取相应的观点评价词。并相应地设计了一个解码算法来解码评价对象对应的观点评价词。

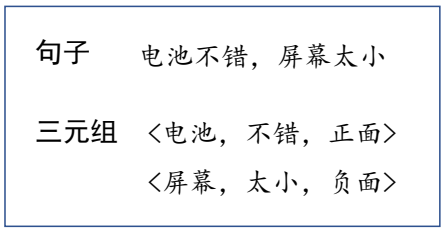


图 1-9 用户观点举例

第 2 章 面向评价对象的评价词抽取

2.1 引言

2016 年，Wang Wenya 等人[5] 在 SemEval 属性级情感分析数据集上标注观点评价词并进行了评价词抽取任务。随后的许多工作[6,32,36,39] 都在属性级情感分析中引入了评价词抽取任务。然而，它们大多数都是将评价词抽取和评价对象抽取两个任务置于多任务学习的框架内进行二者的联合学习，主要的目的是希望通过两个任务间的交互，来提高评价对象抽取任务的性能。还有一些工作[33,34] 引入了情感词预测任务，即预测句子中的词是否为情感词，来增强属性级情感分析模型中编码器的编码能力。

本章认为这些工作中存在两个问题：(1) 仅仅将评价词抽取视作一个辅助任务，而没有将观点评价词作为情感分类中的一类重要特征去进行研究和应用，因此对评价词抽取任务缺乏足够的重视；(2) 仅仅将评价对象抽取和评价词抽取视为两个序列标注问题，而没有研究评价对象和评价词之间的显式联系，如二者的对应关系。在句子“我讨厌他们的烤肉，但是他们的沙拉不错”中，评价对象为“烤肉”和“沙拉”，观点评价词为“喜欢”和“不错”。显然，“烤肉”和“讨厌”对应，而“沙拉”和“不错”对应。本章认为挖掘这种对应关系是重要的，因为当成对抽取评价对象和观点评价词后，可以直接将观点评价词作为一种重要特征，用于后续的属性级情感分类。这既能增加分类的性能，又可以增加模型的可解释性。

因此，在 2019 年，Fan Zhifang 等人[4] 提出了面向评价对象的观点评价此抽取 (target-oriented opinion words extraction, TOWE) 任务。给定一个句子和句子中的评价对象，TOWE 任务旨在抽取句子中描述和评价该评价对象的观点评价词。如此，便可得到观点评价词和评价对象的对应关系。Fan Zhifang 等人[4] 在 SemEval 数据集上标注了观点评价词以及观点评价词与评价对象的对应关系，构造了四个可以用于 TOWE 任务的数据集；并针对 TOWE 任务，在 LSTM 的基础上设计了一个称为 IOG-LSTM 的基线模型。

考虑到预训练语言模型在许多任务上表现出了卓越的性能，且在自然语言处理中的应用越来越广泛，而 Fan Zhifang 等人所设计的 IOG-LSTM 由于扩展性不强，无法直接与之结合。因此，本章旨在设计一个可以应用 BERT 模型的面向评价对象的评价词抽取模型。此外，本章还探讨了除序列标注外的评价词抽取方法。

2.2 任务定义

给定一个句子（或者称为单词序列）$\text{seq} = \{w_1, w_2, \cdots, w_{|\text{seq}|}\}$，以及序列中的一个评价对象 $\text{target} = \{w_{t1}, w_{t2}, \cdots, w_{tn}\}$，TOWE 任务的目的是抽取句子中在该评价对象对应的观点评价词。

比如对于句子 {我, 不禁, 感叹, 你, 是, 怎么, 用, 这么, 少的, 钱, 做出, 这么, 美味的, 食物}，TOWE 任务需要对评价对象“钱”输出“少的”，对评价对象“食物”输出“美味的”。

2.3 提出的模型

在以往的工作中，评价词抽取被建模为一个序列标注任务，首先进行句子的编码，然后使用经典的条件随机场 (conditional random field, CRF) 进行求解。而面向评价对象的评价词抽取任务则需要将评价对象融合到句子中，因此在这里将评价对象视作一个句子，和原句拼为一个句对输入到 BERT 模型中。即如式2-1和式2-2所示，构造 BERT 输入中的词序列 seq^{pair} 和分割向量 (segment_ids)seq^{seg}。

$$\text{seq}^{\text{pair}} = \{[\text{CLS}], w_1, \cdots, w_{|\text{seq}|}, [\text{SEP}], w_{t1}, \cdots, w_{tn}, [\text{SEP}]\}, \tag{2-1}$$

$$\text{seq}^{\text{seg}} = \{\underbrace{0, 0, \cdots, 0, 0}_{|\text{seq}|+2}, \underbrace{1, \cdots, 1, 1}_{|\text{target}|+1}\}. \tag{2-2}$$

得到 seq^{pair} 和 seq^{seg} 之后，将其输入到 BERT 中，得到句子的表示 $H^{\text{seq}} \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$ 和评价对象的表示 $H^{\text{target}} \in \mathbb{R}^{d_{\text{model}} \times |\text{target}|}$。注意，在接下来的评价对象抽取中只有 H^{seq} 被使用，而 H^{target} 则被舍弃。

$$H^{\text{seq}}, H^{\text{target}} = \text{BERT}(\text{seq}^{\text{pair}}, \text{seq}^{\text{seg}}). \tag{2-3}$$

为了识别句子中的观点评价词，本章使用了两种识别方法。第一种方法将评价词抽取视为一个序列标注任务，使用流行的 CRF 去解决；第二种方法则是仿照抽取式问答 (extractive question answering) 任务，通过预测观点评价词开始位置和结束位置来进行抽取，本文称之为基于 span 的方法。下面对两种方法分别进行介绍。

条件随机场

相比于直接基于发射分数进行预测，CRF 考虑到字标签之间的相关性，在整体的字标签序列上计算分数。

$$p(\mathbf{y}|H^{\text{seq}}) = \frac{\exp(s(H^{\text{seq}}, \mathbf{y}))}{\sum_{\mathbf{y} \in Y} \exp(s(H^{\text{seq}}, \mathbf{y})}, \tag{2-4}$$

其中，Y 是所有可能的字标签序列，$s(H^{\text{seq}}, \mathbf{y}) = \sum_i A_{y_{i-1}, y_i} + P_{i, y_i}$ 是评分函数。

A 为字标签上转移矩阵，A_{y_{i-1},y_i} 为从 y_{i-1} 到 y_i 的转移概率，$P = \text{Linear}_{\text{emission}}(H^{\text{seq}})$ 为发射概率。接下来，便使用负对数似然函数计算损失：

$$\mathcal{L}^{|\text{seq}|,\text{target}} = -\log p(\mathbf{y}|H^{\text{seq}}). \tag{2-5}$$

解码阶段则使用维特比 (viterbi) 算法。

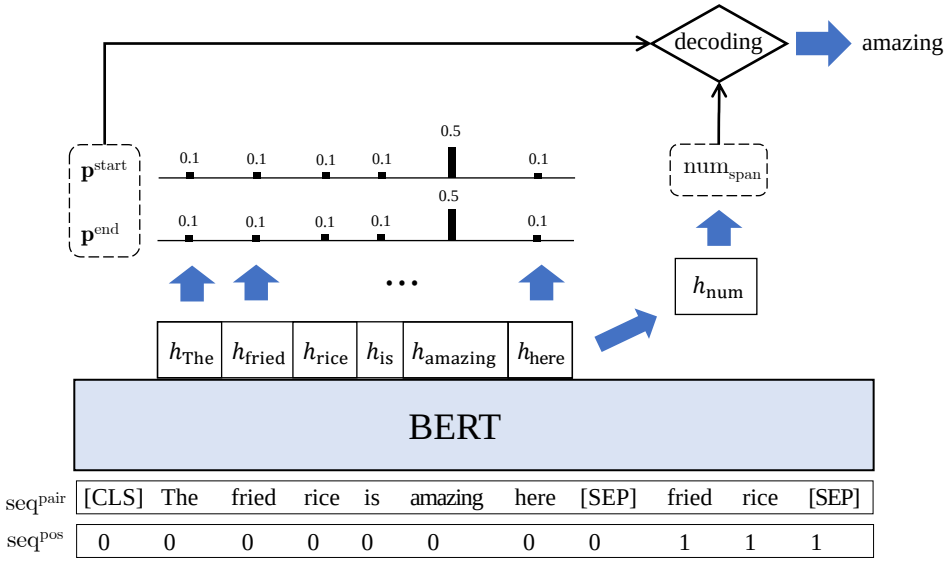


图 2-1 基于 span 的方法的模型示意图。

基于 Span 的方法

虽然 CRF 是目前序列标注中最为流行的方法，但 CRF 存在搜索空间大，运行速度慢的特点。不同于 CRF，基于 span 的方法采用另一种思路识别句子中的观点评价词。如同在抽取式问答中，基于 span 的方法通过预测实体的边界，即开始位置和结束位置，来抽取观点评价词。

首先，通过两个全连接层和 softmax 层得到在开始和结束位置上的概率分布：

$$\mathbf{g}^{\text{start}} = \text{Linear}_{\text{start}}(H^{\text{seq}}), \tag{2-6}$$

$$\mathbf{p}^{\text{start}} = \text{softmax}(\mathbf{g}^{\text{start}}), \tag{2-7}$$

$$\mathbf{g}^{\text{end}} = \text{Linear}_{\text{end}}(H^{\text{seq}}), \tag{2-8}$$

$$\mathbf{p}^{\text{end}} = \text{softmax}(\mathbf{g}^{\text{end}}). \tag{2-9}$$

在训练阶段，通过构造向量 $\mathbf{y}^{\text{start}} \in \mathbb{R}^{|seq|}$ 和 $\mathbf{y}^{\text{end}} \in \mathbb{R}^{|seq|}$ 来向模型传递实体边界的真实值。注意，由于一个评价对象有可能对应多个观点评价词，所以上述两个向量并不一定是一个 one-hot 向量。接下来，使用交叉熵损失函数在边界的预测值和真实值上计算损失：

$$\mathcal{L}^{(|\text{seq}|,\text{target})} = -\sum_i \mathbf{y}_i^{\text{start}} \log(\mathbf{p}_i^{\text{start}}) - \sum_j \mathbf{y}_j^{\text{end}} \log(\mathbf{p}_j^{\text{end}}). \tag{2-10}$$

在解码阶段，由于一个评价对象有可能对应多个观点评价词，传统的解码方法就不适用了。因此，Hu Minghao 等人[37](2019) 提出了一种多 span 解码算法，算法流程如图2-2所示。这是一个启发式的贪心剪枝算法，下面对算法流程进行简要的描述。

首先，从 $\mathbf{g}^{\text{start}}$ 和 $\mathbf{g}^{\text{end}}$ 中根据值大小分别选择前 M 个值，将相应的索引存入 $\mathbf{S}$ 和 $\mathbf{E}$ 中。然后，根据 $\mathbf{S}$ 和 $\mathbf{E}$ 生成所有候选的 span 集合 $\mathbf{R}$，并对每个 span 计算一个分数 u_l（计算方法如图中第 6 行所示）。这里所选择的候选 span 需要满足两个条件：(1) 开始位置不大于结束位置，(2) 开始位置的分数加上结束位置的分数不小于一个给定的阈值 γ。接下来对候选 span 集合 $\mathbf{R}$ 进行剪枝。首先选择分数最大的 span，将其从候选集合 $\mathbf{R}$ 中移动到输出集合 $\mathbf{O}$ 中，然后移除 $\mathbf{R}$ 和此 span 重叠的 span(图中第 13 行)。重复这个过程直到 $\mathbf{R}$ 为空或者 $\mathbf{O}$ 内的元素数目超过 K 个。最后，$\mathbf{O}$ 即为解码得到的 span 集合。

Algorithm 1 Heuristic multi-span decoding

Input: $\mathbf{g}^s, \mathbf{g}^e, \gamma, K$
$\mathbf{g}^s$ denotes the score of start positions
$\mathbf{g}^e$ denotes the score of end positions
γ is a minimum score threshold
K is the maximum number of proposed targets

1: Initialize $\mathbf{R}, \mathbf{U}, \mathbf{O} = \{\}, \{\}, \{\}$
2: Get top-M indices $\mathbf{S}$, $\mathbf{E}$ from $\mathbf{g}^s$, $\mathbf{g}^e$
3: **for** $\mathbf{s}_i$ in $\mathbf{S}$ **do**
4: **for** $\mathbf{e}_j$ in $\mathbf{E}$ **do**
5: **if** $\mathbf{s}_i \le \mathbf{e}_j$ and $\mathbf{g}^s_{\mathbf{s}_i} + \mathbf{g}^e_{\mathbf{e}_j} \ge \gamma$ **then**
6: $\mathbf{u}_l = \mathbf{g}^s_{\mathbf{s}_i} + \mathbf{g}^e_{\mathbf{e}_j} - (\mathbf{e}_j - \mathbf{s}_i + 1)$
7: $\mathbf{r}_l = (\mathbf{s}_i, \mathbf{e}_j)$
8: $\mathbf{R} = \mathbf{R} \cup \{\mathbf{r}_l\}, \mathbf{U} = \mathbf{U} \cup \{\mathbf{u}_l\}$
9: **while** $\mathbf{R} \neq \{\}$ and size$(\mathbf{O}) < K$ **do**
10: $l = \arg\max \mathbf{U}$
11: $\mathbf{O} = \mathbf{O} \cup \{\mathbf{r}_l\}; \mathbf{R} = \mathbf{R} - \{\mathbf{r}_l\}; \mathbf{U} = \mathbf{U} - \{\mathbf{u}_l\}$
12: **for** $\mathbf{r}_k$ in $\mathbf{R}$ **do**
13: **if** f1$(\mathbf{r}_l, \mathbf{r}_k) \neq 0$ **then**
14: $\mathbf{R} = \mathbf{R} - \{\mathbf{r}_k\}; \mathbf{U} = \mathbf{U} - \{\mathbf{u}_k\}$
15: **return** $\mathbf{O}$

图 2-2 Hu Minghao 等人[37](2019) 提出的启发式的多 span 解码算法。图中，$\mathbf{g}^s$ 即本章中的 $\mathbf{g}^{\text{start}}$，$\mathbf{g}^e$ 即本章中的 $\mathbf{g}^{\text{end}}$。

本文认为上述的启发式解码算法存在较大的问题。在该解码算法中，M, γ, K 为算法中的超参数。其中，模型对 γ, K 的值比较敏感。当忽略 span 之间的重叠时，可以发现在上述的解码算法中，γ, K 起到的作用是控制算法返回 span 集合的大小。当返回的 span 集合越大，则模型整体的召回率越高；当返回的 span 集合越

小，则模型整体的精确率越高。因此，对于不同的数据集需要仔细调节这两个超参数，以使得模型达到最佳的性能，这无疑为增加了工作量。而且，对于同一数据集中的不同样例，本文认为它们所对应的最优超参数的值也应不同。

因此，为了消除解码算法中的参数 γ, K，本章采取的思路是通过预测句子中 span 的数目，来控制解码算法返回的 span 集合的大小。如下式所示，span 数目预测被视为一个分类问题：

$$h_{\text{pool}} = \text{MaxPooling}(H^{\text{seq}}), \tag{2-11}$$

$$h_{\text{start}} = \sum_{t} \text{ReLU}(g_t^{\text{start}}), \tag{2-12}$$

$$h_{\text{end}} = \sum_{t} \text{ReLU}(g_t^{\text{end}}), \tag{2-13}$$

$$h_{\text{num}} = \text{ReLU}(\text{Linear}_{\text{num1}}(h_{\text{pool}}; h_{\text{start}} + h_{\text{end}})), \tag{2-14}$$

$$g^{\text{num}} = \text{Linear}_{\text{num2}}(h_{\text{num}}), \tag{2-15}$$

$$p^{\text{num}} = \text{softmax}((h^{\text{num}}). \tag{2-16}$$

当得到 span 数目的预测值之后，便可以将其作为解码算法中的 K 带入。

2.4 数据及实验设置

表 2-1 Fan Zhifang 等人[4](2019) 中使用的数据集的统计信息。

	Datasets	Train		Test	
		sentence	target	sentence	target
D1	Restaurant14	1627	2643	500	865
D2	Laptop14	1158	1634	343	482
D3	Restaurant15	754	1076	325	436
D4	Restaurant16	1079	1512	329	457

所使用的属性级情感分类数据集同 Fan Zhifang 等人[4]，它的统计信息如表2-1所示。这些数据来自 SemEval2014、SemEval2015 和 SemEval2016。原始的数据只标注了评价对象和相应的情感倾向，没有标注评价对象对应的观点评价词。因此，Fan Zhifang 等人在这些数据上为每个评价对象标注了相应的观点评价词。

本节所使用的 BERT 模型为 `bert-base-uncased`。为了减少训练时间，使用了混合精度的训练方式①，并且设置 opt_level 为 01。优化器则使用了学习率为

① https://github.com/NVIDIA/apex

3e-5 [①] 的 Adam 优化器，并且以批大小为 16 在训练数据上运行了 5 个迭代。在启发式的解码算法中，设置了 M 为 10，K 为 5，而阈值 γ 在每个数据集上调节。

由于关于 TOWE 任务的工作不多，所以可以比较的模型有限。**TC-LSTM** 是指通过拼接的方式将评价对象的信息融合到句子中。首先，在评价对象的词向量上平均池化得到评价对象的向量。然后将句子中的每个词的词向量和评价对象向量拼接在一起，送入双向 LSTM 中进行序列标注。**IOG-LSTM** 是 Fan Zhifang 等人[4] 提出的，句子的表示由三部分构成：评价对象左侧的上下文，评价对象右侧的上下文以及全局上下文。**Bert-CRF, Bert-span** 是本章提出的模型，分别表示使用 CRF 和 span 抽取的方法进行评价对象的抽取。

2.5 实验结果及分析

本章使用 F1 值来衡量评价词抽取的性能。如表2-2所示，得益于 BERT 强大的编码能力，使用了 BERT 后评价词抽取的性能大大提高，在四个数据集上分别超过 IOG-LSTM 7.1%、8.6%、8.0% 和 8.2%。此外，本文所使用的 span 抽取方法在性能优于基于 CRF 的序列标注方法，在 f1 值上平均提高了 1.6%。值得注意的是，Bert-span 上的精确率明显地高于召回率。这或许是因为解码算法中的诸多限制导致解码得到的 span 数目较少。这或许是解码算法可以改进的一个方向。考虑到 span 抽取性能的优越，并且具有更小的搜索空间和更快的解码速度，本文认为在序列标注的任务上可以尝试使用 span 抽取的方法代替 CRF。

表 2-2 在 TOWE 任务上模型性能的比较。

	指标	TC-LSTM	IOG-LSTM	Bert-CRF	Bert-span
D1	P	0.6765	0.8285	0.8654	0.8980
	R	0.6767	0.7738	0.8738	0.8466
	F1	0.6761	0.8002	0.8696	**0.8716**
D2	P	0.6245	0.7324	0.7670	0.8693
	R	0.6014	0.6963	0.7778	0.7390
	F1	0.6121	0.7135	0.7723	**0.7989**
D3	P	0.6606	0.7606	0.7789	0.8384
	R	0.6016	0.7071	0.7931	0.7890
	F1	0.6294	0.7325	0.7859	**0.8130**
D4	P	0.7346	0.8525	0.8883	0.9185
	R	0.7288	0.7851	0.8933	0.8800
	F1	0.7310	0.8169	0.8908	**0.8988**

① 使用了 warmup 比率为 0.1 的 `cosine_with_hard_restarts_schedule_with_warmup`，随着训练的进行自动地调整学习率。

本章还就本文所改进的解码算法进行了实验。如表2-3所示，在 f1 值上，本文所提出的解码方法优于 Hu Minghao 等人 (2019)[37] 提出的解码算法（以下称为原始的解码算法）。对比表2-2和表2-3还可以发现，使用了原始的解码算法后 span 抽取方法在整体性能上不如 CRF 的方法，这也印证了 Hu Minghao 等人 (2019) 工作中所得的实验结论。为了分析本文提出的解码方法中 span 数目的影响，本文还在解码时给定 span 数目的真实值，来查看 f1 值的变化。从表2-3中可以看出，给定 span 数目的真实值后，f1 值大大提高。这表明 span 数目预测的性能对评价词抽取上的 f1 值有很大的影响。进一步提高 span 数目预测的性能或许是未来工作的重点。

表 2-3 基于 span 的方法中解码方式的对比。decode1 是指 Hu Minghao 等人[37](2019) 提出的解码算法，其中超参数 γ 的值经过了多次调整后取 f1 最优时的值。decode2 是指本文提出的解码方法。decode3 是指在解码时，给出 span 数目的真实值。$\mathrm{acc}_{\mathrm{span}}$ 是指模型在 span 数目预测上的准确率。

数据集	解码方法	P	R	F1	$\mathrm{acc}_{\mathrm{num}}$	γ
D1	decode1	0.8993	0. 8233	0.8596	-	7
	decode2	0.8980	0.8466	**0.8716**	0.9051	-
	decode3	0.9150	0.8680	0.8909	1.0000	-
D2	decode1	0.8621	0.7055	0.7759	-	6
	decode2	0.8693	0.7390	**0.7989**	0.8693	-
	decode3	0.8779	0.8113	0.8433	1.0000	-
D3	decode1	0.8069	0.7627	0.7842	-	4
	decode2	0.8384	0.7890	**0.8130**	0.9197	-
	decode3	0.8581	0.8093	0.8330	1.0000	-
D4	decode1	0.8893	0.9029	0.8960	-	4
	decode2	0.9185	0.8800	**0.8988**	0.9256	-
	decode3	0.9295	0.9048	0.9170	1.0000	-

2.6 本章小结

考虑到目前大部分的评价词抽取工作，都没有挖掘观点评价词与评价对象的对应关系，本章关注了面向评价对象的评价词抽取 (TOWE) 任务。TOWE 任务旨在抽取文本中某个评价对象相对应的观点评价词，而非文本中的所有观点评价词。这要求在进行句子编码的时候，需要进行句子与评价对象之间的信息融合。针对现有模型可扩展性不强的问题，本章使用了一个基于 BERT 的抽取模型，尝试了使用边界预测的方法代替传统的以 CRF 为代表的序列标注方法。实验结果表明所提出的模型大大优于现有的模型，所使用的边界预测的方法优于基于 CRF 的方法。

第 3 章 结合观点评价词的属性级情感分类

3.1 引言

直观来看，观点评价词对情感倾向的预测至关重要。如对于句子“这手机电池不错，就是屏幕太小”，其中的观点评价词为“不错”和“太小”；通过这些观点评价词，模型可以清晰地判断句子在评价对象“电池”和“屏幕”上的情感倾向。有理由相信，捕捉这些观点评价词，正是模型进行情感分类的第一步。虽然深度学习模型对句子具有较强的编码能力，但在现有的研究工作中，观点评价词抽取的性能仍然不够高，如在 He Ruidan 等人 (2019)[39] 的工作中观点评价词抽取的 f1 值在 SemEval2016 数据集上仅为 0.72。这表明让模型捕获观点评价词仍比较困难，特别当训练数据较少的情况时。这意味寄希望于模型通过在通用预料上的预训练过程或者在特定任务上微调的过程，来具备捕捉观点评价词的能力是不现实的。因此显示地指示出文本中的观点评价词对属性级情感分类是有必要的。

然而，结合观点评价词的属性级别情感分类 (aspect-level sentiment classification with opinion terms, ASCO) 的研究工作却非常稀少。如果给观点评价词加上一个分数来表示其情感倾向和强度，一系列这样的观点评价词可以称为情感词典。在早期的文档级情感分类中，存在许多关于情感词典的研究工作，它们主要关注如何从情感文本中挖掘和扩展情感词典，如何将得到的情感词典应用到情感分类中。但是在属性级情感分析中，大部分引入观点评价词的工作，都是为了通过观点评价词和评价对象的交互，提高评价对象抽取任务的性能。仅有的几个在属性级情感分类中应用观点评价词的工作，也大多是让观点评价词抽取和属性级情感分类共享一个文本编码器的方式，将观点评价词抽取作为一个辅助任务；同时缺乏对属性级情感分类中观点评价词的作用机制进行探讨和分析。

因此，本章关注了属性级情感分类中的观点评价词。首先，探究了观点评价词在属性级情感分类的重要性和作用机制。此处的分析是在 BERT 模型上进行的，为了对观点评价词的贡献度进行量化，本章在自注意力机制的基础上提出了两个定量指标，并在这些指标上进行了计算和分析。（增加）随后，针对 ASCO 任务，通过对 BERT 的自注意力层进行修改，加强了编码阶段对观点评价词的关注，实现了观点评价词与评价对象的交互。

3.2 探究观点评价词对属性级情感分类的重要性

本节首先简要介绍了一个基于 BERT 的属性级情感分类模型，并介绍了在此模型上设计的两个量化指标。随后通过在这些指标上的实验分析，探究观点评价词语属性级情感分类上的重要程度和作用机制。

3.2.1 任务定义

给定一个句子（或称为单词序列）$\text{seq} = \{w_1, w_2, \cdots, w_{|\text{seq}|}\}$，以及序列中的一个评价对象 $\text{target} = \{w_{t1}, w_{t2}, \cdots, w_{tn_t}\}$，属性级情感分类任务的目的是确定句子在该评价对象上的情感倾向 $y_{\text{target}} \in \{\text{POS}, \text{NEG}, \text{NEU}\}$。

例如，对于单词序列 {我, 不禁, 感叹, 你, 是, 怎么, 用, 这么, 少的, 钱, 做出, 这么, 美味的, 食物}，属性级情感分类任务需要对评价对象“食物”输出正面，对评价对象“钱”也输出正面。

3.2.2 用于分析的属性级情感分类模型

如图3-1所示，给定一个句子，首先将 BERT 作为编码器来获得句子的表示。然后，在评价对象对应的表示上执行最大池化得到 h_{target}。基于 h_{target}，通过一个全连接层预测句子在该评价对象上的情感倾向。

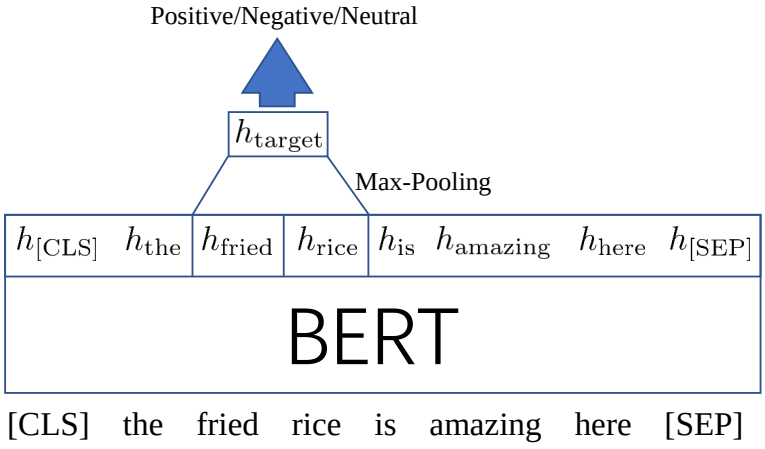


图 3-1 一个简单的属性级情感分析模型

3.2.3 量化观点评价词的贡献度

不妨记输入的句子（或者称为词序列）为 $\text{seq} = \{w_1, w_2, \cdots, w_{|\text{seq}|}\}$。其中的词可以被划分为三类：观点评价词 $\mathcal{O}$、评价对象词 $\mathcal{T}$ 和其他词 $\mathcal{N}$。在这里提出一个假设，如果观点评价词对属性级情感分类是重要的，那么在 BERT 的自注意力层中 $\mathcal{O}$ 和 $\mathcal{T}$ 之间的注意力得分就会相对较高，或者比微调前的高。我们的量化指标都是以这个假设为前提。因此，为了量化这种注意力得分上的提高，本节提出

了两个指标：(1) 平均注意力得分和 (2) 最终贡献度。

平均注意力得分

不妨记在 BERT 模型的第 l 层的第 h 个注意力头中，从词 w_i 到词 w_j 的注意力得分为 $s_{h,l}^{w_i \to w_j}$。于是，从 w_i 到 w_j 的平均注意力得分为：

$$\mu_s^{w_i \to w_j} = \frac{\sum_{h,l} s_{h,l}^{w_i \to w_j}}{n_{\text{head}} * n_{\text{layer}}}, \tag{3-1}$$

其中，n_{head} 为注意力头的个数，n_{layer} 为 BERT 中的 transformer 层数。接下来，进一步计算不同类型单词间注意力得分上的均值。例如，从观点评价词到评价对象词的平均得分为：

$$\mu_s^{\mathcal{O} \to \mathcal{T}} = \frac{\sum_{w_i \in \mathcal{O}, w_j \in \mathcal{T}} \mu_s^{w_i \to w_j}}{|\mathcal{O}| * |\mathcal{T}|}. \tag{3-2}$$

最终贡献度

在此指标中，将自注意力机制简化为信息的流动，将注意力权重视作不同词之间信息流动的比例。在信息流动的视角下，可以计算一个词语对其他词语隐含表示的贡献度。记 BERT 模型的第 l 层中从词 w_i 到 w_j 的注意力权重为 $\alpha_l^{w_i \to w_j}$。需要注意的是，这里的 α_l 是在第 l 层上所有注意力头上注意力权重的均值，有 $\sum_j \alpha^{w_i \to w_j} = 1$。那么在第 1 层中 w_i 对 w_j 的隐含表示的贡献度为 $\alpha_1^{w_i \to w_j}$。在考虑残差连接和归一化之后，这个贡献度为

$$c_1^{w_i \to w_j} = \frac{\alpha_1^{w_i \to w_j} + 1\{i = j\}}{2}. \tag{3-3}$$

其中，$1\{\cdot\}$ 是一个函数，当大括号中的表达式成立返回 1，不成立返回 0。接下来，基于前一层的贡献度，计算下一层的贡献度。这个迭代过程如式3-4所示。与式3-2相似，进一步计算不同类型单词间的贡献度。如观点评价词到评价对象的贡献度的计算如式3-5所示。由于情感分类是在评价对象在最后一层上的隐含表示的基础上进行的，因此称 $c_{n_{\text{layer}}}^{* \to \mathcal{T}}$ 为最终贡献度。

$$c_{l+1}^{w_i \to w_j} = \sum_k c_l^{w_i \to w_k} \frac{\alpha_{l+1}^{w_k \to w_j} + 1\{k = j\}}{2}, \tag{3-4}$$

$$c_{n_{\text{layer}}}^{\mathcal{O} \to \mathcal{T}} = \frac{\sum_{w_i \in \mathcal{O}, w_j \in \mathcal{T}} c_{n_{\text{layer}}}^{\mathcal{O} \to \mathcal{T}}}{|\mathcal{O}| * |\mathcal{T}|}. \tag{3-5}$$

3.2.4 数据及实验设置

所使用的属性级情感分类数据集同 He Ruidan 等人 (2019)[39]，它的统计信息如表3.3.3所示。这些数据来自 SemEval2014 和 SemEval2015。其中情感倾向分为

四类：正向 (positive)、负向 (negative)、中性 (neutral)、冲突 (conflict)。在实际实验中，同大多数工作一样忽略了冲突类。数据集中的观点评价词由 Wang Wenya 等人[5] 标注。

本节所使用的 BERT 模型为 `bert-base-uncased`，它共有 12 层，每一层包含 12 个注意头。为了减少训练时间，使用了混合精度的训练方式，并且设置 opt_level 为 01。优化器则使用了学习率为 3e-5 ① 的 Adam 优化器，并且以批大小为 16 在训练数据上运行了 4 个迭代。

表 3-1 He Ruidan 等人 (2019)[39] 中使用的数据集的统计信息。

	Datasets	Train		Test	
		aspect	opinion	aspect	opinion
D1	Restaurant14	3699	3484	1134	1008
D2	Laptop14	2373	2504	654	674
D3	Restaurant15	1199	1210	542	510

3.2.5 实验结果及分析

如表3-2所示，所使用的属性级情感分类模型的准确率为 0.8 左右，这表明微调后的 BERT 模型对属性级情感分类任务具有一定的理解能力。因此，猜想微调后的 BERT 在注意力的计算上与微调前有较大变化。为了验证这种猜想，首先了在微调前后的 BERT 上计算所提出的平均注意力得分指标，并着重对微调后的注意力得分的提升进行了分析。随后，计算了所提出了最终贡献度的指标，并以案例分析的形式展示了词语在每一层上的注意力权重。

表 3-2 3.2.2节中使用的属性级情感分类模型在三个数据集上的准确率。

Datasets	准确率
D1	0.8491
D2	0.7821
D3	0.8362

如表3-3所示，我们可以发现 BERT 微调前后，不同类型的词之间的注意力得分有了明显的变化。不妨将这种变化记为 BOAT(**b**ias **o**n **a**ttention scores between different **t**ypes of words)。特别是从观点评价词到评价对象的平均注意力得分有了明显的提升，同时观点评价词到其他词的平均注意力得分也有一定的提升。这表明在微调后的 BERT 模型中，观点评价词得到了更大的关注，即观点评价词对属

① 使用了 warmup 比率为 0.1 的 `cosine_with_hard_restarts_schedule_with_warmup`，随着训练的进行自动地调整学习率。

表 3-3 Restaurant14 数据集上 BERT 微调前后，不同类型单词间的平均注意力得分。表中，BERT 是指微调前的 BERT 模型，ASC Model 是指在属性级情感分类任务上微调后的模型。在计算相同类型单词间的注意力时，排除了同一单词间的注意力分数；为了便于展示和比较，每个模型中的平均注意力分数都减去了 $\mu_s^{\mathcal{N}\to\mathcal{N}}$。

Direction	BERT	ASC Model	Improvement
$\mathcal{N}\to\mathcal{N}$	0.	0.	0
$\mathcal{N}\to\mathcal{O}$	-0.0689	-0.0631	0.0058
$\mathcal{N}\to\mathcal{T}$	-0.1161	-0.0698	0.0463
$\mathcal{O}\to\mathcal{N}$	-0.1613	-0.0340	**0.1273**
$\mathcal{O}\to\mathcal{O}$	0.5873	1.0561	**0.4688**
$\mathcal{O}\to\mathcal{T}$	-0.3135	-0.0888	**0.2247**
$\mathcal{T}\to\mathcal{N}$	-0.1435	-0.2223	-0.0788
$\mathcal{T}\to\mathcal{O}$	-0.2549	0.2641	-0.0092
$\mathcal{T}\to\mathcal{T}$	1.0457	1.5766	**0.5309**

性级情感分类具有十分重要的作用。

除了发现存在从观点评价词到评价对象的注意力得分的提升外，本节还发现这种提升呈现一定的规律。如图3-2所示，我们可以发现随着层数的增加，这种提升呈增加的趋势，并在第 11 层达到最大值。这表明随着层数的增加，BERT 对单词的理解层能力逐渐增强，对观点评价词识别能力逐渐增强，boat 也逐渐增大。此外，从图3-3中可以看出，在所有的注意力头上均存在注意力得分的提升，但是不同注意力头上的提升差异不大，提升最大的是第 11 个注意力头和第 6 个注意力头。

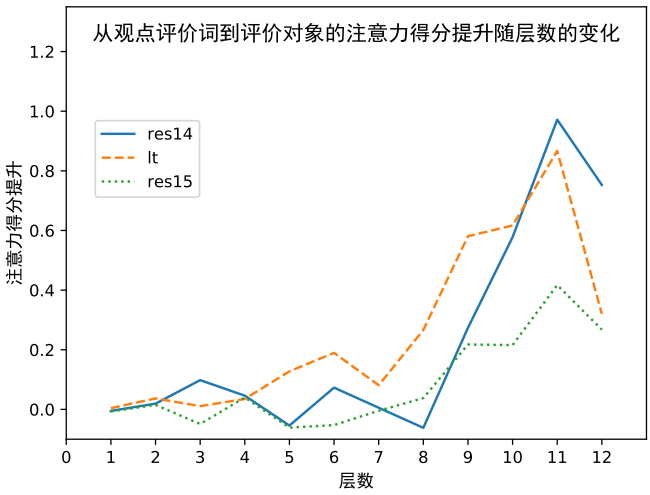


图 3-2 BERT 微调后，从观点评价词到评价对象的注意力得分提升随层数的变化。

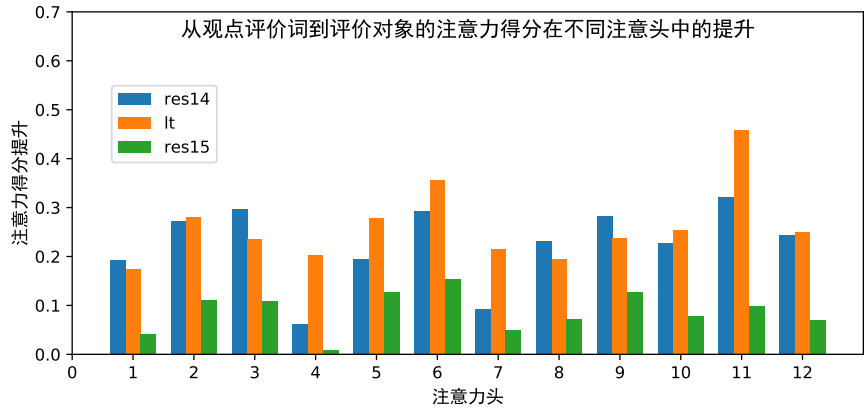


图 3-3 BERT 微调后，从观点评价词到评价对象的注意力得分在不同注意力头上的提升。

不同于平均注意力得分，最终贡献度可以更加简要地展现某类单词对于情感分类的影响。如表3-4所示，在本节所提出的最终贡献度的指标上，观点评价词的最终贡献度明显高于其他类型的单词，甚至高于评价对象词。这再次印证了观点评价词对属性级情感分类十分重要。

表 3-4 BERT 微调后，不同类型单词的最终贡献度。

Datasets	$\mathcal{N}$	$\mathcal{O}$	$\mathcal{T}$
D1	0.0245	**0.0326**	0.0250
D2	0.0208	**0.0369**	0.0282
D3	0.0192	**0.0310**	0.0258

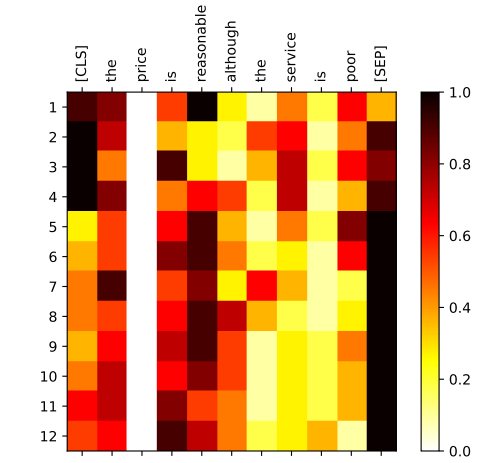
在图3-4中最终贡献度被逐层展开，可以从图中看出，对评价对象而言，来自 [CLS] 和 [SEP] 上的注意力权重非常高，这是 BERT 本身的特性造成的；除此之外，来自观点评价词的注意力相较其他单词十分显著。如在图3-4 a) 中，reasonable 为句中的观点评价词，其对评价对象 price 在第 1、5、6、77、8、9、10 和 12 层的注意力权重均比较高。

3.3 结合观点评价词的属性级别情感分类模型

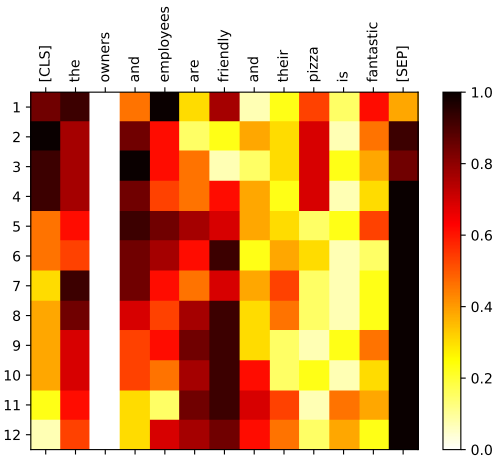
3.3.1 任务定义

给定一个句子（或称为单词序列）$\text{seq} = \{w_1, w_2, \cdots, w_{|\text{seq}|}\}$，以及序列中的一个评价对象 $\text{target} = \{w_{t1}, w_{t2}, \cdots, w_{tn_t}\}$ 和所有的观点评价词 $\mathcal{O} = \{w_{o1}, w_{o2}, \cdots, w_{on_o}\}$，ASCO 任务的目的是确定句子在该评价对象上的情感倾向 $y_{\text{target}} \in \{\text{POS}, \text{NEG}, \text{NEU}\}$。

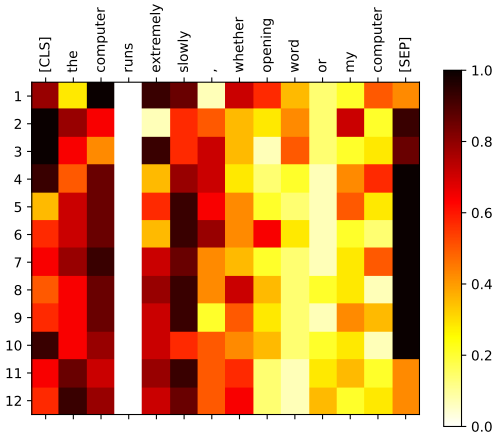
例如，对于单词序列 {我, 不禁, 感叹, 你, 是, 怎么, 用, 这么, 少的, 钱, 做出, 这



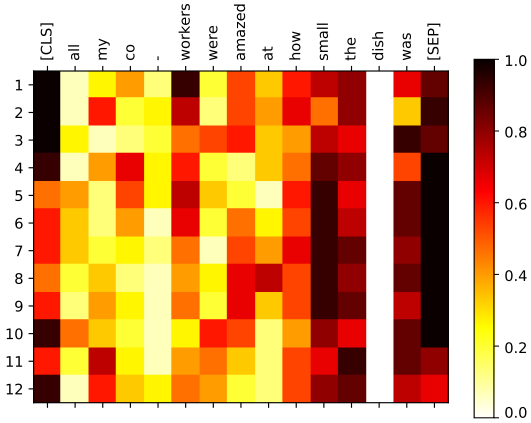
a) the price is reasonable althought the service is poor



b) the owner and employees are friendly and their pizza is fantastic



c) the computer runs extremely slowly, whether opening word or my computer



d) all my co-workers were amazed at how small the dish was

图 3-4 BERT 微调后在每一层上对评价对象的注意力示例。为了方便显示，在每一层上对注意力进行排序，然后从高到低依次涂色，故图中对应位置的值不是实际的注意力权重，而是权重大小的相对顺序。

么，美味的，食物}，句子中的观点评价词 $\mathcal{O}=\{$少的，美味的$\}$，ASCO 任务需要对评价对象“食物”输出正面，对评价对象“钱”也输出正面。注意到在这里给出的观点评价词是所有出现在句子中的观点评价词，而不是针对某个评价对象的观点评价词。

3.3.2 所提出的模型

在上一节中，我们可以发现从观点评价词到评价对象上的平均注意力得分是相对较高的，在微调过后注意力得分出现了 BOAT。这种 BOAT 是 BERT 在属性级情感分析任务上微调之后得到的，这意味着我们可以通过一个更加直接的方式

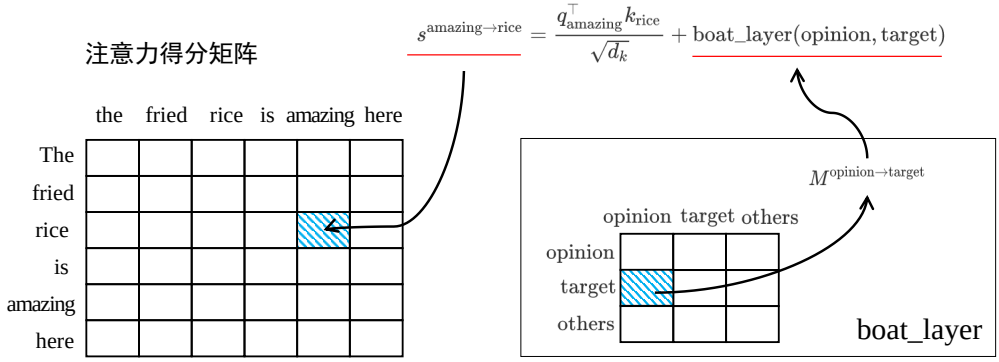


图 3-5 boat 层示意图。

来学习到这种 BOAT。基于这种出发点，本节在 BERT 的自注意力层的基础上，设计了一个 BOAT 层来根据词的类型更加直接地学习 BOAT。在本小节中将首先介绍 BOAT 层和在 BERT 的注意力机制上的修改，接着介绍如何将修改后的 BERT 应用到 ASCO 任务上。

偏置矩阵

如图3-5所示，在计算单词间的注意力时，引入一个额外的偏置 (bias)。因此修改后的自注意力机制变为

$$\text{Attention}(Q, K, V, B) = \text{softmax}(\frac{QK^{\top}}{\sqrt{d_k}} + B)V, \tag{3-6}$$

其中，$Q, K \in \mathbb{R}^{|\text{seq}|\times d_k}$，$V \in \mathbb{R}^{|\text{seq}|\times d_v}$，偏置矩阵 $B \in \mathbb{R}^{|seq|\times|seq|}$。

偏置矩阵由 BOAT 层计算。不妨令类型向量 $\text{seq}^{\text{type}} = \{e_1, e_2, \cdots, e_{|\text{seq}|}\}$，其中 $e_i \in \{\text{opi}, \text{tar}, \text{oth}\}$ 表示单词 w_i 的类型（类型向量的构造会在本节后面介绍）。那么偏置矩阵 B 的计算过程如下：

$$B_{ij} = M^{e_j\to e_i}, \tag{3-7}$$

其中，$M \in \mathbb{R}^{3\times3}$ 是 BOAT 层中一个可学习的参数。直观上，其他单词到其他单词的偏置应该为 0，因此设置 $M^{\text{oth}\to\text{oth}} = 0$。

修改后的多头注意力

为了将偏置矩阵应用到 BERT 的多头注意力上，首先将参数 M 扩展为一个三维的张量 $\mathbf{M} \in \mathbb{R}^{3\times3\times d_{\text{boat}}}$。此时，偏置张量 $\mathbf{B} \in \mathbb{R}^{|\text{seq}|\times|\text{seq}|\times d_{\text{boat}}}$ 的计算过程如下：

$$\mathbf{B}_{ij} = \mathbf{M}^{e_j\to e_i}. \tag{3-8}$$

因此，BOAT 层和修改后的多头注意力机制变为

$$\text{MultiHead}(Q, K, V, \text{seq}^{\mathcal{O}}, \text{seq}^{\mathcal{T}}) = \text{Concat}(\text{head}_1, \cdots, \text{head}_{n_h})W^O \tag{3-9}$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V, \mathbf{B}W_i^B)$$

$$\mathbf{B} = \text{BOATLayer}(\text{seq}^{\mathcal{O}}, \text{seq}^{\mathcal{T}}; \mathbf{M})$$

其中，$W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, $W_i^B \in \mathbb{R}^{d_{\text{highlight}} \times 1}$ 以及 $W^O \in \mathbb{R}^{n_{\text{head}} d_v \times d_{\text{model}}}$ 都是模型参数。注意到 BOAT 层不是共享的，每一多头自注意力层都对应着一个 BOAT 层。

结合观点评价词的属性级情感分类模型

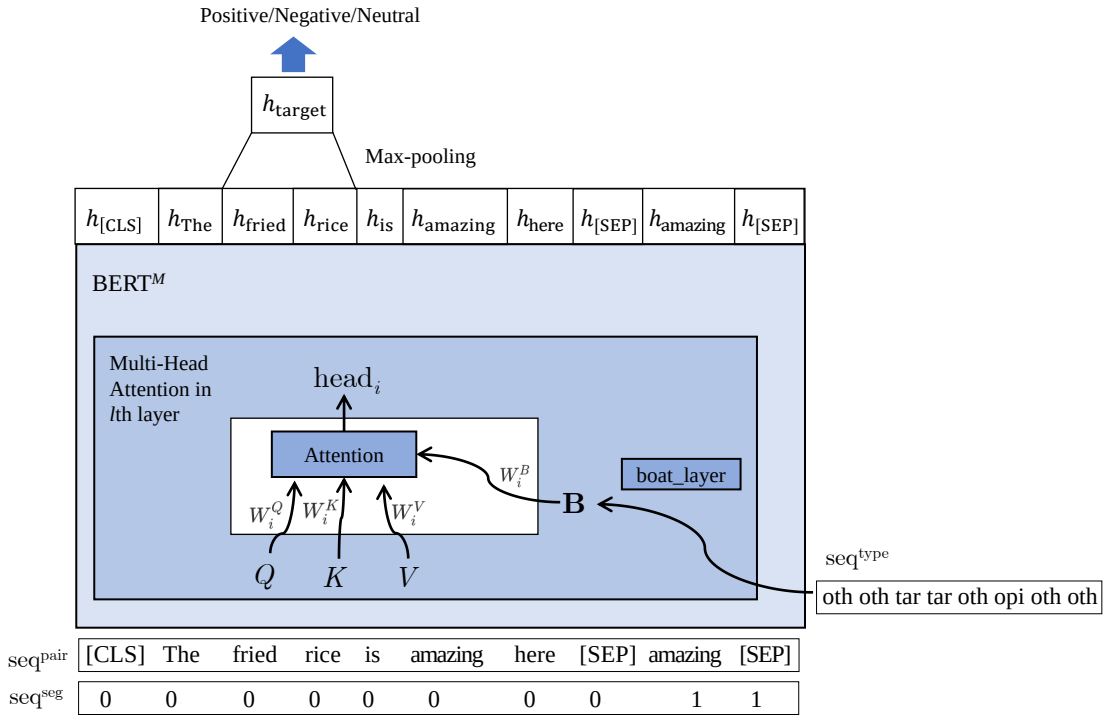


图 3-6 结合观点评价词的属性级情感分类模型。

上面描述了如何使用 BOAT 层生成偏置矩阵，如何将偏置矩阵引入到注意力机制中，不妨记修改后的 BERT 为 BERT^M。在这里，将描述如何使用 BERT^M 进行属性级情感分类，大体框架同图3-5。

给定句子 seq 和其中的观点评价词 $\mathcal{O}$ 以及评价对象 target。首先如式3-10和式式3-11所示，通过将观点评价词视作一个句子，构造 BERT^M 输入中的词序列 seq^{pair} 和分割向量 (segment_ids)seq^{seg}。

$$\text{seq}^{\text{pair}} = \{[\text{CLS}], w_1, \cdots, w_{|\text{seq}|}, [\text{SEP}], w_{o1}, \cdots, w_{on_o}, [\text{SEP}]\}, \tag{3-10}$$

$$\text{seq}^{\text{seg}} = \{\underbrace{0, 0, \cdots, 0, 0}_{|\text{seq}|+2}, \underbrace{1, \cdots, 1, 1}_{|\mathcal{O}|+1}\}. \tag{3-11}$$

然后依据观点评价词 $\mathcal{O}$ 和评价对象 target 构造一个类型向量 seq^{type}，用来作为 BERT^M 中 BOAT 层的输入，构造过程如式3-12所示。注意到，由于采用的是句

对分类的方式，为第二个句子中的观点评价此增加一种类型，记作 opi′。

$$\text{seq}^{\text{type}} = \{\text{oth}, e_1, \cdots, e_{|\text{seq}|}, \text{oth}, \underbrace{\text{opi}', \cdots, \text{opi}'}_{|\mathcal{O}|}, \text{oth}\} \tag{3-12}$$

$$\text{where } e_i = \begin{cases} \text{tar} & w_i \in \text{target}, \\ \text{opi} & w_i \in \mathcal{O}, \\ \text{oth} & \text{其他情况}. \end{cases}$$

得到 seq^{pair}、seq^{seg} 和 seq^{type} 后，首先将三个向量输入到 BERT^M 中，得到句子的表示 $H^{|\text{seq}|} \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$。为了引入评价对象的信息，在评价对象对应的表示上执行最大池化得到评价对象的表示 $h_{\text{target}} \in \mathbb{R}^{d_{\text{model}} \times 1}$。有了评价对象的表示，即可通过一个全连接层预测句子在该评价对象上的情感倾向。形式化的计算过程如下：

$$H^{|\text{seq}|}, H^{|\mathcal{O}|} = \text{BERT}^M(\text{seq}^{\text{pair}}, \text{seq}^{\text{seg}}, \text{seq}^{\text{type}}), \tag{3-13}$$

$$h_{\text{target}} = \text{MaxPooling}(H_{t1}^{|\text{seq}|}, H_{t2}^{|\text{seq}|}, \cdots, H_{tn_t}^{|\text{seq}|}), \tag{3-14}$$

$$\hat{y}_{\text{target}} = \text{softmax}(W h_{\text{target}} + b), \tag{3-15}$$

其中，$W \in \mathbb{R}^{3 \times d_{model}}$，$b \in \mathbb{R}^{3 \times 1}$ 是模型参数。得到情感倾向的预测值 $\hat{y}_{\text{target}}$ 之后，使用交叉熵损失函数计算损失。

3.3.3 数据及实验设置

所使用的实验数据同3.2.4节，其统计信息如表。本节所使用的 BERT 模型为本节所使用的 BERT 模型为 `bert-base-uncased`。为了减少训练时间，使用了混合精度的训练方式，并且设置 opt_level 为 01。优化器则使用了学习率为 3e-5 [①] 的 Adam 优化器，设置批大小为 16。由于三个数据集的训练数据数目差异较大，因此根据数据量设置了不同的迭代次数，训练数据最多的数据集 D1 为 5 个迭代，训练数据规模中等的数据集 D2 为 6 个迭代，训练数据最少的数据集 D3 为 8 个迭代。

实验中使用准确率来衡量情感分类的性能，分别记使用观点评价词之前和之后的准确率为 acc1 和 acc2，通过二者的差异来展现观点评价词的作用。实验中对每个模型都随机初始化了 5 次，然后将运行所得的结果平均起来作为该模型的性能。

由于在属性级情感分类中使用观点评价词的研究工作比较少，可以用来对比

① 使用了 warmup 比率为 0.1 的 `cosine_with_hard_restarts_schedule_with_warmup`，随着训练的进行自动地调整学习率。

的模型也比较少。**IMN**[39] 将观点评价词和评价对象抽取以及属性级情感分类置于一个多任务学习的框架内，并设计了一个名为 opinion transmission 的机制实现二者的交互。同时，IMN 中也使用了大量（超过 3 万条）文档级情感分类的数据辅助模型进行句子的编码。**BERT-pair** 是指将观点评价词作为一个辅助句子，将情感分类转变为句对分类任务，而没有使用 BERT^M 模型，即在式3-13中使用 BERT 代替 BERT^M。

3.3.4 实验结果及分析

如表3-5所示，我们可以发现使用观点评价词之后，情感分类的准确率有了较大幅度的提高，特别是在数据集 D1 上达到了 1.45%。其次，在大多数情况下本节所提出的 BERT^M 优于 IMN 和 BERT-pair，可以更加充分的利用到观点评价词中的信息。但是在数据集 D3 上，使用 BERT^M 带来的准确率提高低于 IMN 和 BERT-pair，这或许是因为数据集 D3 上的训练数据太少导致 BOAT 层难以训练。这一点也可以通过 IMN 在 D3 的准确率上得以体现。由于在 IMN 的训练过程中，大量的文档级情感分类文本被使用，所以 D3 上训练数据少的问题得到了缓解。

表 3-5 在 ASCO 任务上的模型比较。图中准确率为运行 5 次后的平均结果。

	acc	IMN	BERT-pair	BERT^M
D1	acc1	0.8305	0.8486	0.8486
	acc2	0.8389	0.8546	0.8631
D2	acc1	0.7512	0.7875	0.7875
	acc2	0.7536	0.7893	0.7922
D3	acc1	0.8453	0.8472	0.8472
	acc2	0.8564	0.8525	0.8514

3.4 本章小结

本章对属性级情感分类中的观点评价词进行重点研究。首先研究了观点评价词在属性级情感分类的作用机制，通过设计两个指标对观点评价词在属性级情感分类中的贡献度进行量化。在两个量化指标上的实验表明，随着训练的进行，模型将更加关注文本中的观点评价词，观点评价词对属性级情感分类具有非常重要的作用。因此，在属性级情感分类任务中指示出文本中的观点评价词是有必要的。为了在属性级情感分类中指示观点评价词，本章设计了一个 BOAT 层，通过句子中词语的类型来让模型更直接学习到对某类词语的关注，实现了让模型更加“注意”句子中的观点评价词。并在公开数据集上超过基线方法达到最优。

第 4 章 结合观点评价词的端到端的属性级情感分析

4.1 引言

在本章中一个完整的观点包含三个要素：评价对象、观点评价词、情感倾向。如句子“我讨厌他们的烤肉，但是他们的沙拉不错”中包含两个观点三元组：< 烤肉, 讨厌, 负面 >、< 沙拉, 不错, 正面 >。本章的标题“结合观点评价词的端到端的属性级情感分析”便是指抽取[①] 文本中所有的观点三元组，因此也称观点三元组抽取 (aspect sentiment triplet extraction, ASTE) 任务。

以往的研究工作中，端到端的属性级情感分析大都只关注句子中的评价对象和相应的情感倾向，而对文本中的观点评价词缺乏足够的关注。然而，句子中的观点评价词不仅可以提高情感倾向预测结果的可解释性，还可以大大提高情感倾向预测的准确率。此外，与值为正面、负面或中性的情感倾向相比，观点评价词往往蕴含了更加丰富的含义。如在句子“这屏幕太小”和句子“这屏幕太暗”中，虽然观点评价词表达的情感倾向都是负面的，但是在含义上有所差别，同时也可以作为表达情感倾向的原因。再如在句子“这沙拉真棒”和句子“这沙拉还不错吧”中，虽然观点评价词表达都是正面的情感，但是表达的情感强度不同。

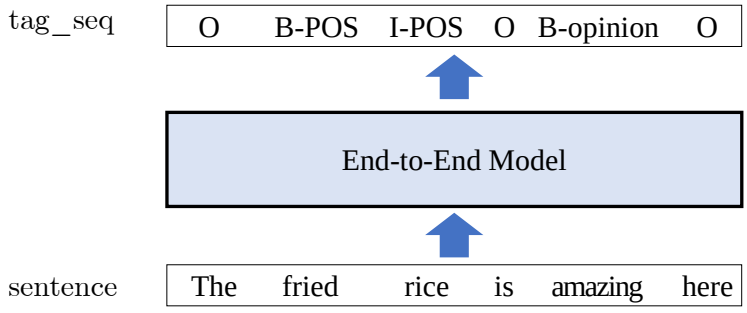


图 4-1 以往的端到端的属性级情感分析任务的示意图。

如图4-1，以往的端到端的属性级情感分析工作可以简化为一个序列标注问题。其中，评价对象和情感倾向分别作为字标签的边界和类型，观点评价词也作为另一类实体。本章认为这种任务模式存在两个问题：(1) 未编码评价对象和观点评价词之间的对应关系，这导致了结果的不完整；(2) 未在预测情感倾向时，综合考虑评价对象和观点评价词的信息。因此，为了属性级情感分析任务的丰富性和完整性以及情感预测性能的进一步提升，本文认为有必要在抽取评价对象和预测情感

① 由于评价对象和观点评价词都是在文本中出现的，所以对二者应用抽取描述没有问题。而情感倾向没有在文本中出现，应该使用的是分类，但为了表达的简洁，本文统一称为抽取。

倾向的同时，抽取观点评价词及评价对象和观点评价词的对应关系。

与本章较为相关的工作是 Fan Zhifang 等人在 2019 年的工作[4]，他们认为应该成对抽取评价对象和观点评价词，并提出了 TOWE 任务。该任务假设已经事先获取了句子中的评价对象，旨在与预测评价对象相对应的观点评价词。TOWE 任务可以视作 ASTE 任务中的一个子任务。此外，在 2020 年，Peng Haiyun 等人[7] 首次正式地提出了 ASTE 任务，并在 SemEval 属性情感分析数据集上进行了标注，提供了 4 个可用于 ASTE 任务的数据集。他们同时设计了一个流水线模式的基线模型。本章的实验将在这些数据集上开展。

4.2 任务定义

给定一个句子（或称为单词序列）$\text{seq} = \{w_1, w_2, \cdots, w_{|seq|}\}$，ASTE 任务的目的是输出句子中所有的观点三元组 T，观点三元组由评价对象、观点评价词和情感倾向构成，即有 $T = \{(t_1, O^{t_1}, p_1), \cdots, (t_{n_t}, O^{t_{n_t}}, p_{n_t})\}$，其中 t_i 为评价对象，O^{t_i} 为评价对象 t_i 对应的观点评价词的集合，p_i 为评价对象的情感倾向。注意 O^{t_i} 可以为空集。

例如，对于句子 {我, 不禁, 感叹, 你, 是, 怎么, 用, 这么, 少的, 钱, 做出, 这么, 美味的, 食物}，ASTE 任务需要输出 {< 钱,{少的}, 正面 >,< 食物,{美味的}, 正面 >}。对于句子 {这里的, 环境, 和, 可口的, 食物, 是, 我们, 常来, 的, 原因}，ASTE 任务需要输出 {< 环境,{}, 正面 >,< 食物,{“可口的}, 正面 >}。

4.3 提出的模型

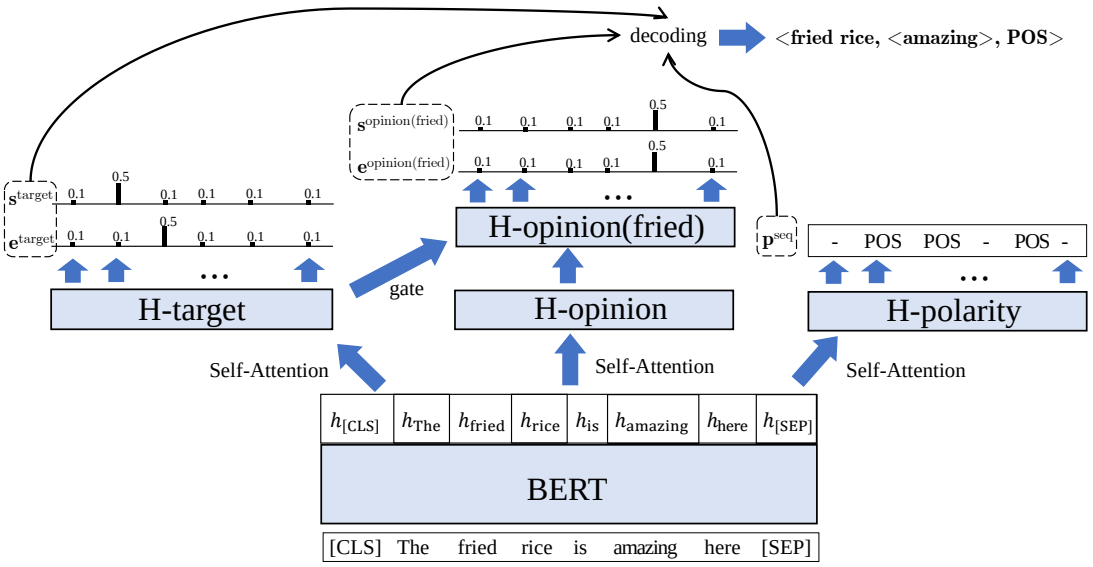


图 4-2 对于 ASTE 任务，所提出的模型的整体框架。

模型的整体框架

如图4-2所示，首先通过 BERT 模型进行句子的编码。然后通过自注意力机制获得句子在评价对象上的表示 H^{target}，句子在观点评价词上的表示 H^{opinion} 以及句子在情感倾向上的表示 H^{polarity}。在所得的句子表示 H^{target} 上，通过边界预测的方法（以下称为 span 抽取）进行评价对象的抽取；在所得句子表示 H^{polarity} 上，进行情感标签序列的预测。接着通过门控机制，综合 H^{target} 和 H^{opinion} 得到特定词语上的句子表示 $H^{\text{opinion}(w_i)}$。在 $H^{\text{opinion}(w_i)}$ 上进行词 w_i 对应的观点评价词的抽取。最后，根据以上的结果进行解码，得到文本中的评价对象和观点评价词，以及对应的关系和情感倾向，以此获取文本中的观点三元组。

获得特定任务上的句子表示

给定句子 seq，将其输入到到 BERT 模型中得到句子表示 $H \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$。然后通过注意力机制，根据 H 得到不同任务上的句子表示。

$$H = \text{BERT}(\text{seq}), \tag{4-1}$$

$$H^{\text{target}} = \text{SelfAttentionLayer}(H^{\text{bert}}), \tag{4-2}$$

$$H^{\text{target}} = \text{SelfAttentionLayer}(H^{\text{bert}}), \tag{4-3}$$

$$H^{\text{target}} = \text{SelfAttentionLayer}(H^{\text{bert}}). \tag{4-4}$$

其中，SelfAttentionLayer 为自注意力层，源于 transformer[42] 模型，包含多头自注意力机制和一个残差连接。

$$Q = \text{Linear}_q(H^{\text{input}}), \tag{4-5}$$

$$K = \text{Linear}_k(H^{\text{input}}), \tag{4-6}$$

$$V = \text{Linear}_v(H^{\text{input}}), \tag{4-7}$$

$$H^{\text{att}_1} = \text{MultiHead}(Q, K, V), \tag{4-8}$$

$$H^{\text{att}_2} = \text{Linear}_{\text{output}}(H^{\text{att}_1}), \tag{4-9}$$

$$H^{\text{output}} = \text{LayerNorm}(H^{\text{input}} + H^{\text{att}_2}), \tag{4-10}$$

其中，$H^{\text{input}}, H^{\text{output}} \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$。

获得特定词语上的句子表示

得到 H^{target} 和 H^{opinion} 后，使用门控机制进行信息的选择，生成句子针对某个词语的表示 $H^{\text{opinion}(w_i)}$。

$$H^{\text{target}_1} = \text{conv}_1(H^{\text{target}}), \tag{4-11}$$

$$H^{\text{target}_2} = \text{conv}_2(H^{\text{target}_1}), \tag{4-12}$$

$$\text{gate}_i = \text{ReLU}(\text{Linear}_{\text{gate}}(H_i^{\text{opinion}}; H_i^{\text{target}_2})), \tag{4-13}$$

$$H^{\text{opinion}(w_i)} = H^{\text{opinion}} \times \text{gate}_i, \tag{4-14}$$

其中，$\text{conv}_1, \text{conv}_2$ 为卷积层，卷积核的大小均为 3；$\text{gate}_i \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$，$\times$ 为按位乘。

span 抽取

评价对象的抽取和观点评价词的抽取都是通过 span 抽取模块完成，这里统一介绍。对于句子的表示 $H^{\text{task}} \in \mathbb{R}^{d_{\text{model}} \times |\text{seq}|}$，通过两个全连接层预测 span 开始的位置和结束的位置，然后通过 softmax 层得到相应的概率值。

$$\mathbf{g}^{\text{start}} = \text{Linear}_{\text{start}}(H^{\text{task}}), \tag{4-15}$$

$$\mathbf{g}^{\text{end}} = \text{Linear}_{\text{end}}(H^{\text{task}}), \tag{4-16}$$

$$\mathbf{p}^{\text{start}} = \text{softmax}(\mathbf{g}^{\text{start}}), \tag{4-17}$$

$$\mathbf{p}^{\text{end}} = \text{softmax}(\mathbf{g}^{\text{end}}). \tag{4-18}$$

其中，$\mathbf{p}^{\text{start}}, \mathbf{p}^{\text{end}} \in \mathbb{R}^{|\text{seq}|}$。

此外，为了 span 的解码，还预测了句子中包含的 span 数量。span 数量在评价对象的抽取中即为句子中的评价对象数量；在观点评价词的抽取中则为某个词作为评价对象的一部分所对应的观点评价词的数量。

$$h^{\text{num}} = \text{ReLU}(\text{Linear}_{\text{num}_1}(h)), \tag{4-19}$$

$$g_{\text{num}} = \text{Linear}_{\text{num}_2}(h^{\text{span}}), \tag{4-20}$$

$$p^{\text{num}} = \text{softmax}(g^{\text{num}}), \tag{4-21}$$

其中，$p^{\text{num}} \in \mathbb{R}^{n_{\text{span}}}$，$n_{\text{span}}$ 为 span 数目可能的取值；$h, h^{\text{num}} \in \mathbb{R}^{d_{\text{model}} \times 1}$。在评价对象抽取中，$h = H^{\text{target}}_{[\text{CLS}]}$；在观点评价词抽取中，$h = H_i^{\text{target}_2}$。

得到 $\mathbf{p}^{\text{start}}, \mathbf{p}^{\text{end}}$ 以及 p^{num} 之后，通过交叉熵损失函数分别计算损失：

$$\mathcal{L}^{\text{start}} = -\sum_i \mathbf{y}_i^{\text{start}} \log(\mathbf{p}_i^{\text{start}}), \tag{4-22}$$

$$\mathcal{L}^{\text{end}} = -\sum_i \mathbf{y}_i^{\text{end}} \log(\mathbf{p}_i^{\text{end}}), \tag{4-23}$$

$$\mathcal{L}^{\text{num}} = -\sum_j y_j^{\text{num}} \log(p_j^{\text{num}}), \tag{4-24}$$

其中，$\mathbf{y}^{\text{start}}, \mathbf{y}^{\text{end}} \in \mathbb{R}^{|\text{seq}|}, y^{\text{num}} \in \mathbb{R}^{n_{\text{span}}}$ 为 span 的开始位置、结束位置以及数目的

真实值。

情感预测

这里的情感预测是指情感倾向序列的预测。给定在句子表示 H^{polarity}，通过一个全连接层和 softmax 层即可获得情感倾向序列的预测值。

$$\mathbf{g}^{\text{polarity}} = \text{Linear}_{\text{polarity}}(H^{\text{polarity}}), \tag{4-25}$$

$$\mathbf{p}^{\text{polarity}} = \text{softmax}(\mathbf{g}^{\text{polarity}}). \tag{4-26}$$

得到情感倾向序列的预测值之后，即可通过交叉熵损失函数计算损失，记作 $\mathcal{L}^{\text{polarity}}$。如图4-2所示，计算损失时只考虑评价对象和观点评价词对应位置的情感倾向。

模型的训练

除了进行评价对象抽取、词对应的评价词抽取以及情感预测之外，本文还预测句子上的所有观点评价词，作为一个辅助任务。因此，模型整体的损失如下：

$$\mathcal{L} = \mathcal{L}^{(\text{start,target})} + \mathcal{L}^{(\text{end,target})} + \mathcal{L}^{(\text{num,target})} + \mathcal{L}^{(\text{polarity})} + \tag{4-27}$$

$$\mathcal{L}^{(\text{start,opinion})} + \mathcal{L}^{(\text{end,opinion})} + \mathcal{L}^{(\text{num,opinion})} + \tag{4-28}$$

$$\sum_i \frac{\mathcal{L}^{(\text{start,opinion}(w_i))} + \mathcal{L}^{(\text{end,opinion}(w_i))} + \mathcal{L}^{(\text{num,opinion}(w_i))}}{|\text{seq}|}. \tag{4-29}$$

解码过程

总体的解码过程如下：

1. 首先根据评价对象上的 span 预测结果解码出句子中的评价对象。

2. 对评价对象中的每个词，模型都会输一个该词对应的观点评价词上的 span 预测结果。综合一个评价对象所有词上的预测结果，解码出该评价对象对应的观点评价词。

3. 对于根据上面解码出的评价对象和对应的观点评价词，综合考虑情感极性序列上这些对应位置的情感极性预测值，解码出对应的情感极性。

4. 将所得的评价对象、观点评价词、情感倾向组合为一个观点三元组。

其中，span 的解码采用的是在第 3 章描述的启发式的 span 解码算法。而综合一个评价对象上所有词的预测结果，是指每个位置上的得分取所有词上最大的一个。例如，对于评价对象 target，位置 j 为开始位置的得分为：

$$\mathbf{g}_j^{\text{start(target)}} = \max_{w_i \in \text{target}} \mathbf{g}_j^{\text{start}(w_i)}. \tag{4-30}$$

而情感倾向的解码同理。对于评价对象 target 和相应的观点评价词 opinion

(这里不妨假设评价对象对应的观点评价词只有一个)，其情感倾向为 p 的得分为:

$$\mathbf{g}_p^{\text{polarity(target,opinion)}} = \max_{w_i \in \text{target} \cup \text{opinion}} \mathbf{g}_p^{\text{polarity}(w_i)}. \tag{4-31}$$

4.4 数据及实验设置

在 ASTE 任务上所使用的属性级情感分析数据集来自 Peng Haiyun 等人 (2020) 的工作[7], 它的统计信息如表4-1所示。这些数据集来自 SemEval2014、SemEval2015 和 SemEval2016。原始的数据只标注了评价对象和相应的情感倾向，没有标注评价对象对应的观点评价词。其中的观点评价词是 Fan Zhifang 等人[4](2019) 标注的。这些数据集存在的一个问题是，没有考虑一个观点评价词对应多个评价对象的情况，这应是未来的工作需要修复的。

表 4-1 用于 ASTE 任务的数据集的统计信息。

	Datasets	Train			Test		
		sentence	target	opinion	sentence	target	opinion
D1	Restaurant14	1300	2079	2145	496	849	862
D2	Laptop14	920	1283	1265	339	475	490
D3	Restaurant15	593	834	923	318	426	455
D4	Restaurant16	842	1183	1289	320	444	465

本节所使用的 BERT 模型为 `bert-base-uncased`。为了减少训练时间，使用了混合精度的训练方式①，并且设置 opt_level 为 01。优化器则使用了学习率为 5e-5 ② 的 Adam 优化器，并且以批大小为 16 在训练数据上运行了 10 个迭代。

评价指标

为了与前人的工作进行比较，本章使用了三个评价指标对所抽取的观点三元组进行评价：(1) $f1^{\mathcal{TS}}$，即在评价对象及情感倾向上的 f1 值，在此指标中只有评价对象的边界和情感倾向都正确，该评价对象的抽取才算正确；(2) $f1^{\mathcal{O}}$，即在评价词抽取上的 f1 值，该指标是在句子的所有观点评价词上计算；(3) $f1^{\text{pair}}$，所抽取的评价对象和观点评价词对上的 f1 值，注意该评价指标考虑的是评价对象和观点评价词的对应关系，不要求评价对象上的所有观点评价词都正确；(4) $f1^{\text{triple}}$，所抽取的评价对象、观点评价词及其情感倾向上的 f1 值，同样地，不要求抽取评价对象上的所有观点评价词，因此该评价指标不等于观点三元组抽取的 f1 值。此外，

① https://github.com/NVIDIA/apex

② 使用了 warmup 比率为 0.1 的 `cosine_with_hard_restarts_schedule_with_warmup`，随着训练的进行自动地调整学习率。

对比的模型

CMLA[32] 使用注意力机制进行评价对象抽取和评价词抽取之间的交互。为了引入情感倾向，这这里将情感倾向也加入到字标签中。**Li-unified-R**[36] 针对同一个评价对象上情感倾向标签的一致性问题，提出了 BG 和 SC(sentiment consistency) 模块。**Peng-pipeline** 是指 Peng 等人 (2020) 的工作[7]，其在 Li-unified-R 的基础上使用了一个图神经网络来进行评价对象的抽取，并设计了一个 TG(target guidance) 模块来完成评价对象和观点评价词的交互。上述的三个模型在得到评价对象及其情感倾向以及观点评价词之后，将它们两两组对送入分类器中，以确定其对应关系。因此，上述三个模型皆为端到端的模型。

4.5 实验结果及分析

表4-2展示了在 $f1^{TS}$ 上不同模型的比对结果，表中 part-polarity 是指，输出情感倾向时，仅考虑情感倾向序列上评价对象所对应位置的值。从表中，首先可以看出本文所提出的模型在 $f1^{TS}$ 上远高于已有的模型，特别是在数据集 D2 的提高接近 6%。其次可以看出预测情感倾向时，综合考虑评价对象和观点评价词上的倾向预测值会带来更好的 $f1^{TS}$。但也有例外，在数据集 D3 上，part-polarity 上的 $f1^{TS}$ 更高，这或许是因为模型在观点评价词的抽取上表现不佳，带来了额外的噪声。

表 4-2 在 $f1^{TS}$ 指标上不同模型性能的比较。

	指标	CMLA	Li-unified-R	Peng-Pineline	Our Model	
					part-polarity	full model
D1	P	0.6780	0.7315	0.7660	0.7669	0.7755
	R	0.7369	0.7444	0.6784	0.7362	0.7444
	F1	0.7062	0.7379	0.7195	0.7512	**0.7596**
D2	P	0.5470	0.6628	0.6315	0.6906	0.7040
	R	0.5920	0.6071	0.6155	0.6484	0.6611
	F1	0.5690	0.6338	0.6234	0.6688	**0.6819**
D3	P	0.4990	0.6495	0.6765	0.7275	0.7175
	R	0.5800	0.6495	0.6402	0.6831	0.6737
	F1	0.5360	0.6495	0.6579	**0.7046**	0.6949
D4	P	0.5890	0.6633	0.7118	0.7394	0.7416
	R	0.6360	0.7455	0.7230	0.7477	0.7500
	F1	0.6120	0.7020	0.7173	0.7436	**0.7458**

表4-3为在评价词抽取上的性能比较。表中，full-opinion 是指评价词抽取模块的输出结果上计算 $f1^{\mathcal{O}}$ 值，而 target-opinon 是指合并特定词的评价词抽取模块的输出结果后，计算 $f1^{\mathcal{O}}$ 值。从表可以看出，在四个数据集上 full-opinion 上的 $f1^{\mathrm{O}}$ 值均高于基线模型，而在 target-opinion 上却不尽然。这是因为在解码过程中，是先进行评价对象的解码，然后再根据评价对象的解码结果进行观点评价词的解码，这会导致评价对象解码过程中的错误累计到观点评价词的解码上，即一定数量的观点评价词未被解码出来。这导致 target-opinion 上的召回率偏低。但这并不意味着本文所提出的特定词上的评价词抽取模型性能更差，因为在抽取词对应的观点评价词时，同时也获得了评价对象和观点评价词之间的对应关系。此外，可以看出在 D3 数据集 target-opinion 方法上的 $f1^{\mathcal{O}}$ 不仅在四个数据集上最低，而且低于基线方法。这也印证了上一段中对 part-polarity 上 $f1^{\mathcal{TS}}$ 结果的分析。

表 4-3 在 $f1^{\mathcal{O}}$ 指标上不同模型性能的比较。

	指标	CMLA	Li-unified-R	Peng-Pineline	Our Model	
					full-opinion	target-opinion
D1	P	0.6947	0.8120	0.8472	0.8676	0.8761
	R	0.7453	0.8318	0.8039	0.8364	0.8121
	F1	0.7191	0.8213	0.8245	**0.8517**	0.8429
D2	P	0.5180	0.7662	0.7822	0.8035	0.8936
	R	0.6530	0.7490	0.7184	0.7592	0.6510
	F1	0.5770	0.7570	0.7484	**0.7807**	0.7532
D3	P	0.6080	0.7918	0.7807	0.8100	0.8315
	R	0.6530	0.7588	0.7807	0.7868	0.6615
	F1	0.6290	0.7744	0.7802	**0.7982**	0.7368
D4	P	0.7450	0.7984	0.8109	0.8496	0.9066
	R	0.6900	0.8688	0.8667	0.8624	0.7720
	F1	0.7170	0.8316	0.8373	**0.8559**	0.8339

表4-4展示了在 $f1^{\mathrm{pair}}$ 上不同模型的比对结果。可以看到，本文所提出的模型大幅超过基线模型，特别在数据集 D1 上超过基线模型 14 个百分点。这或许是因为数据集 D1 中句子数目较多，特别是一个句子中包含的评价对象和观点评价词较多，而本文所提出的模型可以更好地建模句子中评价对象和观点评价词之间的对应关系。此外，值得注意的是，本文的模型在召回率上均偏低，在精确率上偏高。这仍是因为本文所设计的特定词上的评价词抽取模块，在观点评价词的抽取上更为保守。这也是在未来的研究工作中需要改进的。

表 4-4 在 $f1^{\text{pair}}$ 指标上不同模型的比较。

	指标	CMLA	Li-unified-R	Peng-Pineline	Our Model
					full model
D1	P	0.4517	0.4437	0.4776	0.6723
	R	0.5342	0.7367	0.6810	0.7332
	F1	0.4895	0.5534	0.5610	**0.7014**
D2	P	0.4210	0.5229	0.5000	0.5854
	R	0.4630	0.5294	0.5847	0.5388
	F1	0.4410	0.5256	0.5385	**0.5611**
D3	P	0.4270	0.5275	0.4922	0.5818
	R	0.4670	0.6175	0.6570	0.5626
	F1	0.4460	0.5685	0.5623	**0.5721**
D4	P	0.5250	0.4611	0.5235	0.6358
	R	0.4790	0.6455	0.7050	0.6796
	F1	0.5000	0.5375	0.6004	**0.6570**

表 4-5 在 $f1^{\text{triple}}$ 指标上不同模型的比较。

	指标	CMLA	Li-unified-R	Peng-Pineline	Our Model	
					part-polarity	full model
D1	P	0.4011	0.4144	0.4418	0.6213	0.6266
	R	0.4663	0.6879	0.6299	0.6775	0.6833
	F1	0.4312	0.5168	0.5189	0.6482	**0.6537**
D2	P	0.3140	0.4225	0.4040	0.4922	0.5078
	R	0.3460	0.4278	0.4724	0.4531	0.4673
	F1	0.3290	0.4247	0.4350	0.4718	**0.4867**
D3	P	0.3440	0.4334	0.4097	0.5159	0.5159
	R	0.3760	0.5073	0.5468	0.4989	0.4989
	F1	0.3590	0.4669	0.4679	**0.5073**	**0.5073**
D4	P	0.4360	0.3819	0.4676	0.5553	0.5553
	R	0.3980	0.5347	0.6297	0.5935	0.5914
	F1	0.4160	0.4451	0.5362	**0.5738**	0.5717

表4-5展示了 $f1^{\text{triple}}$ 上不同模型的比对结果。可以看到在 $f1^{\text{triple}}$ 上，本文提出的模型优于现有的模型。同时，我们可以看出，在四个数据集上的 $f1^{\text{triple}}$ 值最高只有 0.6537，这表明在观点三元组抽取任务上仍然存在比较大的提升空间。

4.6 本章小结

目前大部分的端到端的属性级情感分析工作，没有挖掘评价对象和观点评价

词之间的关系，而且没有将观点评价词作为端到端模型输出结果的一部分，只是将评价词抽取视作一个辅助任务。因此，本文关注了观点三元组抽取 (ASTE) 任务。ASTE 任务旨在抽取文本中的所有观点三元组，包含评价对象及其对应的观点评价词和情感倾向。为了使用一个端到端的模型去解决 ASTE 任务，本章使用门控机制来获得针对特定词语的句子表示，进而来抽取该词语对应的观点评价词，在解码阶段则综合评价对象中所有词上的抽取结果。实验表明，本章所提出的模型优于现有的模型。

结　论

本文着眼于属性级情感分析中的观点评价词，并重点关注了观点评价词和评价对象之间的对应关系。本文首先对评价词抽取进行研究，研究了如何对给定的评价对象，抽取相应的观点评价词。随后，对如何将观点评价词应用于属性级情感分类模型中进行了研究。并在最后实现了一个端到端的模型，以抽取文本的观点三元组。

本文的研究贡献主要体现在以下几个方面：

1. 本文提出了一个基于 BERT 的给定评价对象的评价词抽取模型，使用了抽取式问答中的边界预测方法（以下称为 span 抽取）进行评价词的抽取。本文改进了 span 抽取中的解码算法，减少了解码算法中所需设置的超参数。所提出的模型在四个公开数据集上性能超过基线系统达到最优。使用所改进的解码算法，模型的性能得到提高。

2. 本文提出了一种评估某类单词在模型中重要程度的方法，并提出了两个评估某类单词贡献度的指标。该指标在 BERT 的自注意力层上计算，可以定量地反应某类单词对于某个任务的重要程度。

3. 本文提出了一个结合观点评价词的属性级情感分类模型。该模型通过修改了 BERT 模型上的自注意力层，实现了计算注意力分数时考虑单词的类型信息。进而该模型可以通过指示出文本中的观点评价词来提高属性级情感分类的性能。所提出的模型超过基线模型达到最优。

4. 本文提出了一个端到端的观点三元组抽取模型。该模型首先抽取句子中的评价对象，然后沿着 TOWE 方法的思路，对评价对象中的每个词都抽取对应的观点评价词，最后采用一个解码算法解码出评价对象和观点评价词之间的对应关系。所提出的模型在公开数据集上超过基线模型达到最优。

本文未来的工作可以从以下几个方面展开：

1. span 抽取上的改进。span 抽取的性能主要是受限于其解码算法。在解码算法中，使用了一个启发式的评分函数来对候选的 span 进行打分，然后使用一个贪心算法进行解码。本文认为这个评分函数的形式尚有待商榷，贪心算法也可改进。此外，优于解码算法中 span 数目对模型的整体性能影响很大，因此在未来的工作应重点提升 span 数目预测的准确率，并更为细致地研究 span 数目对模型性能的影响。

2. boat 层的其他应用。本文所提出的 boat 层通过修改自注意力层来考虑单词的类型信息。这一想法可以扩展到属性级情感分析之外，作为一种预训练语言模型结合外部知识的方法。如结合词性信息、更高级的实体信息。

3. 观点三元组模型的改进。本文所设计的端到端的观点三元组抽取模型虽然实现了在抽取实体的时候同时抽取关系，但各个任务之间的交互似乎仍不够。如可在进行评价对象对应的评价词抽取的同时，进行观点评价词对应的评价对象抽取。此外，现存数据集中存在的问题也应修复。

参考文献

[1] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval-2014 Task 4: Aspect Based Sentiment Analysis[C/OL] // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland : Association for Computational Linguistics, 2014 : 27-35. https://www.aclweb.org/anthology/S14-2004.

[2] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2015 Task 12: Aspect Based Sentiment Analysis[C/OL] // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado : Association for Computational Linguistics, 2015 : 486-495. https://www.aclweb.org/anthology/S15-2082.

[3] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis[C/OL] // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California : Association for Computational Linguistics, 2016 : 19-30. https://www.aclweb.org/anthology/S16-1002.

[4] Fan Z, Wu Z, Dai X, et al. Target-oriented opinion words extraction with target-fused neural sequence labeling[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 : 2509-2518.

[5] Wang W, Pan S J, Dahlmeier 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. 2016 : 616-626.

[6] Li X, 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. 2017 : 2886-2892.

[7] Peng H, Xu L, Bing L, et al. Knowing What, How and Why: A Near Complete Solution for Aspect-based Sentiment Analysis[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2020.

[8] Tang D, Qin B, Feng X, et al. Effective LSTMs for Target-Dependent Sentiment Classification[C] // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 : 3298-3307.

[9] Vo D-T, Zhang Y. Target-dependent twitter sentiment classification with rich automatic features[C] // Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[10] Zhang M, Zhang Y, Vo D-T. Gated neural networks for targeted sentiment analysis[C] // Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[11] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C] // Proceedings of the 2016 conference on empirical methods in natural language processing. 2016 : 606-615.

[12] Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM[C] // Thirty-second AAAI conference on artificial intelligence. 2018.

[13] Ma D, Li S, Zhang X, et al. Interactive attention networks for aspect-level sentiment classification[C] // Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017 : 4068-4074.

[14] Liu J, Zhang Y. Attention modeling for targeted sentiment[C] // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017 : 572-577.

[15] Cheng J, Zhao S, Zhang J, et al. Aspect-level sentiment classification with heat (hierarchical attention) network[C] // Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017 : 97-106.

[16] He R, Lee W S, Ng H T, et al. Effective attention modeling for aspect-level sentiment classification[C] // Proceedings of the 27th International Conference on Computational Linguistics. 2018 : 1121-1131.

[17] He R, Lee W S, Ng H T, et al. Exploiting Document Knowledge for Aspect-level Sentiment Classification[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018 : 579-585.

[18] Li X, Bing L, Lam W, et al. Transformation Networks for Target-Oriented Sentiment Classification[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 : 946-956.

[19] Tang D, Qin B, Liu T. Aspect Level Sentiment Classification with Deep Memory Network[C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 : 214-224.

[20] Chen P, Sun Z, Bing L, et al. Recurrent attention network on memory for aspect sentiment analysis[C] // Proceedings of the 2017 conference on empirical methods in natural language processing. 2017 : 452-461.

[21] Fan C, Gao Q, Du J, et al. Convolution-based memory network for aspect-based sentiment analysis[C] // The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 : 1161-1164.

[22] Xue W, Li T. Aspect Based Sentiment Analysis with Gated Convolutional Networks[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 : 2514-2523.

[23] Tay Y, Tuan L A, Hui S C. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[24] Nguyen T H, Shirai K. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis[C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 : 2509-2514.

[25] Sun C, Huang L, Qiu X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 : 380-385.

[26] Xu H, Liu B, Shu L, et al. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 : 2324-2335.

[27] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 : 1064-1074.

[28] Xu H, Liu B, Shu L, et al. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018 : 592-598.

[29] Ye H, Yan Z, Luo Z, et al. Dependency-tree based convolutional neural networks for aspect term extraction[C] // Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2017 : 350-362.

[30] Luo H, Li T, Liu B, et al. Improving aspect term extraction with bidirectional dependency tree representation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7) : 1201-1212.

[31] Ma D, Li S, Wu F, et al. Exploring Sequence-to-Sequence Learning in Aspect Term Extraction[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 : 3538-3547.

[32] Wang W, Pan S J, Dahlmeier D, et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms[C] // Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[33] Li X, Bing L, Li P, et al. Aspect term extraction with history attention and selective transformation[C] // Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018 : 4194-4200.

[34] Luo H, Li T, Liu B, et al. DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction[C/OL] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy : Association for Computational Linguistics, 2019 : 591-601. https://www.aclweb.org/anthology/P19-1056.

[35] Ma D, Li S, Wang H. Joint learning for targeted sentiment analysis[C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 : 4737-4742.

[36] Li X, Bing L, Li P, et al. A unified model for opinion target extraction and target sentiment prediction[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 33. 2019 : 6714-6721.

[37] Hu M, Peng Y, Huang Z, et al. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification[C/OL] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy : Association for Computational Linguistics, 2019 : 537-546. https://www.aclweb.org/anthology/P19-1051.

[38] Zhang X, Goldwasser D. Sentiment Tagging with Partial Labels using Modular Architectures[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 : 579-590.

[39] He R, Lee W S, Ng H T, et al. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[40] Wan H, Yang Y, Du J, et al. Target-Aspect-Sentiment Joint Detection for Aspect-Based Sentiment Analysis[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2020.

[41] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 : 4171-4186.

[42] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 : 5998-6008.

攻读硕士学位期间发表的论文及其他成果

（一）发表的学术论文

[1] Zhang Y, Liu J, Fan Y, et al. CN-HIT-IT.NLP at SemEval-2020 Task 4: Enhanced Language Representation with Multiple Knowledge Triples[C]//Proceedings of The 14th International Workshop on Semantic Evaluation. Association for Computational Linguistics.(已投稿)

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《结合观点评价词的端到端的属性级情感分析研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：张文策　　日期：2020年 6 月 26 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：张文策　　日期：2020年 6 月 26 日

导师签名：刘远超　　日期：2020年 6 月 26 日

致　谢

两年时光转瞬即逝。两年前，我仿如步行在迷雾中的大地，只能远远地看见一座又一座山尖；两年后，迷雾些许散开了，视野开始宽阔起来。这两年来，我收获了知识和成长，但在很多地方仍稚嫩浅陋，需要进步。在这两年的学习生活中，许多人为我提供了切实的帮助，在此请我让对他们表示感谢。

首先要感谢我的硕士指导老师刘远超老师对本人的精心指导。在硕士的起步阶段，刘老师为我指明了研究方向，并耐心地为我介绍前人的工作，让我在学术研究的道路上有了一个相对轻松的起步。在整个硕士阶段，刘老师始终给予我非常大的鼓励和期望，让我时刻充满信心和动力。刘老师那认真严谨的治学态度也深深感染着我。此外，感谢实验室的孙承杰老师、单丽莉老师，他们在毕业论文上给了我许多宝贵的建议和帮助。

然后要感谢实验室的同学和室友。感谢崔鹏、王纯宇、许博师兄和胡乐师姐，感谢他们在学术和生活上予以的指导与关心。感谢同一级的张思奇、凌雪、高畅、范炀、戴尚峰和贾荫鹏同学。还要感谢的是我的室友张耀杰和宋治勋，与他们的讨论，常常让我受益匪浅。