**目 录**

[主题词 1](#_Toc33443903)

[第一章 研究背景和意义 2](#_Toc33443904)

[第二章 问题的已有解决方法及其优缺点 3](#_Toc33443905)

[第三章 本文的解决方案 4](#_Toc33443906)

[3.1 整体方案 4](#_Toc33443907)

[3.2 具体技术方案 4](#_Toc33443908)

[3.2.1 语料标注及数据增强 4](#_Toc33443909)

[3.2.2 子句分割算法 5](#_Toc33443910)

[3.2.3 多注意力表示层 5](#_Toc33443911)

[3.2.4 模型训练 6](#_Toc33443912)

[3.2.5 数据分布 7](#_Toc33443913)

[3.2.6 实验设定 7](#_Toc33443914)

[3.2.7 Baseline 7](#_Toc33443915)

[3.3 结论 8](#_Toc33443916)

[3.3.1 实验结果 8](#_Toc33443917)

[3.3.2 分析 8](#_Toc33443918)

[3.3.3 不足之处 8](#_Toc33443919)

[3.4 后续工作 9](#_Toc33443920)

[参考文献 10](#_Toc33443921)

**摘 要**

在电商平台中，产品的评论很多都是以问答对的形式出现的。相对于非交互式的单条评论，这些问答对蕴含着非常丰富的属性方面的信息。

本篇设计报告是对问答对进行属性的单标签分类。

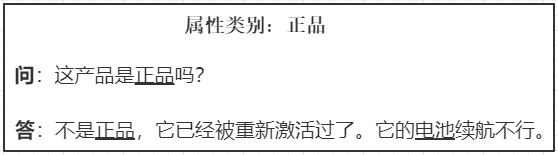
首先，本人爬取了淘宝网的1000多条产品评论，对这些语料进行标注以及数据增强。然后，使用子句分割算法将答句分割成多个子句，对单个问句和多个子句进行编码。接着，通过多注意力表征计算问句和多个子句的匹配程度。最后，解码端通过softmax输出预测概率最高的标签。

实验结果证明了该模型能充分利用答句的信息，并超过了一些baseline。

**关键词**：问答对；属性分类；多注意力

# 主题词

近年来，随着电子商务的快速发展，越来越多的人选择网上购物。不管是售前咨询还是售后评价，用户都会在评论区留言。在淘宝、京东等网站上，都有大量的问答式评价。如下图所示：



对于关心这种产品是否正品的用户而言，他们会到评论区提问。而购买过这种产品的顾客会回答这些问题。因此，交互式的问答便产生了。问答对类型的评论更具有可信度和说服力。

遗憾的是，很少有工作来研究问答对式的属性分类。就来上图这个例子来说，在问句中只有一个属性类别“正品”，但在答句中有两个属性类别“正品”和“电池”。从提问者的角度来看，他只想知道这个产品是否“正品”，而关于“电池”的信息属于噪音。因此，这项研究难点在于如何鉴别出答句中蕴含的与问句中相关的属性信息。

受到该问题的启发，我们可以将答句分割成多个子句，然后计算问句与这些子句的匹配程度。匹配程度最高的即是该问答对的标签信息。

# 研究背景和意义

中国的网民数量已经达到了九亿多，加上近十年来淘宝、京东等电商企业的飞速发展，网购以其独特的优势逐渐成为了老百姓生活的一部分。但由于网上产品质量参差不齐，假货新闻屡见不鲜，消费者的利益受到了极大的损害。

对于普通用户来说，为了了解某个产品的具体信息，除了查看官方给出的参数以及咨询客服外，最有效的办法是到评论区提问。例如“华为P30手机电池续航怎么样啊？”等。出于官方不会供出自己产品的负面信息，而只有那些购买并使用过该产品的消费者的评论才更有可信度和说服力。

对于商家而言，自家产品的优缺点都需要了解。一款产品在某些属性上有优点，在某些属性上有缺点。举手机的例子来说，可能这款手机外观好看，但在续航时间很短。因此，商家以及生产厂家需要对电池进行改进。

综上所述，消费者和商家都需要对产品的属性有详细的了解。更进一步地讲，属性分类任务是属性级情感分析的基础。而问答对式的评论蕴含的属性信息是最丰富的。因此，对问答对进行属性分类极为重要。

# 问题的已有解决方法及其优缺点

属性分类可以用文本分类的方法来解决，例如CNN [2]，LSTM [3]。Toh等人[8]利用sigmoidal前馈网络来训练属性类别分类的二分器。Xue等人[4]基于神经网络来联合学习属性类别分类和属性术语抽取这两个任务。

另外，还有很多学者聚焦于细粒度的观点对象抽取。这种任务和属性分类任务相似，都是给出一个问答对，识别出属性类别。区别在于，属性抽取任务包含两个子任务，分别是从语料中抽取出所有的属性描述语和将近义的属性描述语聚类在一起。

CRF[6]方法需要大量的人工标注。Shu[1]等人提出了一种lifelong CRF模型来抽取属性描述语，该模型利用先验知识来辅助抽取任务。

近年来，LDA[7]模型及其变体使用无监督学习来做属性抽取。但是，LDA系列的模型需要对文档做主题分布计算。He等人提出了基于注意力机制的神经网络模型来突出属性相关的词语。

目前，问答对的属性分类很少有人涉及。但本篇的设计报告基于上述的LSTM、注意力等方法。实验结果也超过了一些baseline。

# 本文的解决方案

## 整体方案

首先我们使用子句分割算法将答句分割成多个子句，对单个问句和多个子句进行编码。接着，将编码后的问句和子句输入LSTM获得隐藏状态。通过多注意力表征计算问句隐状态和多个子句隐状态的匹配程度，即注意力分配。然后将注意力与问句隐状态输入到全连接层。最后，解码端通过softmax输出预测概率最高的标签。

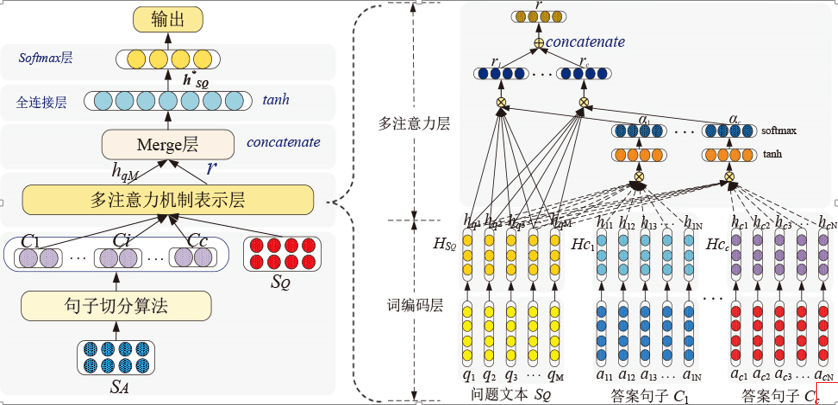
模型结构图如下所示：

图3.1 模型的整体结构

## 具体技术方案

### 语料标注及数据增强

语料的标注格式为：5s运行iOS10真的不卡吗？实际体验的朋友帮忙指导下。 升级系统，真的不卡。我老公在用 系统性能

黑色字体部分为问句，蓝色字体部分为答句，红色字体部分为标注的标签类别。

由于只爬了1000多条评论，数据量还不够。这里的增强方法是基于哈工大同义词词林替换，例如将“老公”替换为“丈夫”等。增强后的数据集达到了3000多条。

### 子句分割算法

正如主题词所述，答句中可能包含多个属性类别，但是只有一条子句是与标注的标签相关的。因此，我们需要将答句进行分割来捕捉蕴含在答句中的有效信息。

伪代码如下：

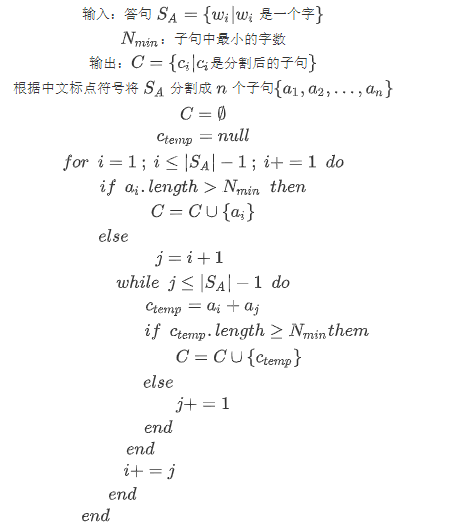


图3.2 子句切分算法

中文标点符号作为分隔符，分割后的子句可能只有一个字，比如“好”。这显然是不合理的，因此这些短子句需要合并。

### 多注意力表示层

本文解决方案的核心就是利用多注意力表征层来捕捉问答对中最有用的信息。

给定一个问答对，假设答句已经被分割成个子句，每个子句包含个字。向量表示第个子句的第个字。问句包含个字，向量表示问句的第个字。表示词向量维度。词向量可以有多种编码方式，比如最近大火的Bert，但本文仍然采用的是传统的Word2Vec编码方式。

首先，我们将编码后的和输入到LSTM中。这样每个字都能学习到与前文的依赖关系，并获得每个时间步的隐状态和。即：

表示问答句中字的个数，表示LSTM隐状态的维度。

接着，我们通过计算和得出注意力分配。分配的注意力越大表示该子句与问句相关度越高，蕴含的标签信息也越多。计算方法如下：

都是可训练的参数。

根据注意力分配我们就可以得到注意力表征：

由于答句被分割成多个子句，我们就可以获得问句的多个注意力表征，将这些表征拼接起来就能得到新的向量：

单标签的问答对主要取决于问句，因此最终的特征表示计算如下：

都是可训练的参数。

最后，解码端采用softmax来获得条件概率分布：

都是可训练的参数。

概率最高的那个标签即是预测的该问答对的属性类别。

### 模型训练

我们的模型使用交叉熵损失函数来训练：

是训练样本的数量，是属性个数，是标注的属性类别，是模型输出，即预测函数，是正则化项来防止模型过拟合。

训练过程中采用Adgrad优化器[2]来优化参数更新。

### 数据分布

|  |  |
| --- | --- |
| **属性类别** | **数量** |
| 性能 | 548 |
| 电池 | 230 |
| IO | 908 |
| 功能 | 346 |
| 质量 | 403 |
| 正品 | 370 |
| 计算 | 269 |

### 实验设定

* 训练集、验证集和测试集的比例为6:2:2
* 词向量使用Skip-Gram生成，维度为768
* 学习率为0.01
* Dropout率为0.5
* ，即子句最少字数，设置为5
* 评价指标采用准确率和F1值

### Baseline

为了验证模型的有效性，我们设计了一些baseline实验来进行对比：

* CNN(Q)：将单条问句作为CNN的输入。
* CNN(Q+A)：将问句和答句拼接在一起作为CNN的输入。
* LSTM(Q)：将单条问句作为LSTM的输入。
* LSTM(Q+A)：将问句和答句拼接在一起作为LSTM的输入。

## 结论

### 实验结果

|  |  |  |
| --- | --- | --- |
| **模型** | **准确率** | **F1值** |
| CNN(Q) | 0.731 | 0.584 |
| CNN(Q+A) | 0.773 | 0.596 |
| LSTM(Q) | 0.806 | 0.661 |
| LSTM(Q+A) | 0.850 | 0.704 |
| 多注意力(本文模型) | 0.865 | 0.818 |

### 分析

模型采用问答对作为输入的效果要比单独采用问句作为输入的效果好很多，这说明了答句可以辅助问句来识别标签。

LSTM模型要比CNN模型效果要好，这是因为CNN文本中的字各自独立，而LSTM文本中的字都能学习到与前文的时序依赖关系。

本文的模型无论是准确率还是F1值都超过了这些baseline，这说明我们的方法能充分捕捉到问答对中与属性最相关的信息。相对于LSTM而言，本文模型首先采用子句分割算法来分割答句，然后使用多注意力机制来去除噪音，并得到与问句属性一致的子句。LSTM虽然能学习到时序关系，但对于每个字分配的权重都是一样的，这跟日常生活中人的直觉也不相符。

### 不足之处

在数据集上，尽管进行了数据增强，但数据量还是太少，只有3000多条。有两种解决办法：第一，从电商网站上爬取更多的语料。要注意的是，有些网站有反爬措施，比如京东网只能爬取100页，而且很多评论是重复的，可能是商家雇佣水军刷的。爬取下来的评论要进行清洗；第二，再次进行数据增强，基于哈工大同义词林的增强策略是：对文本进行分词，然后从这些词中随机挑选若个词语，每个词语在词典中都有若干个同义词。每替换一个同义词就能生成一条文本。循环多次，语料就能得到极大扩充。

实验中设定的超参数没有得到充分验证。比如设置为3，对于其它值没有进行验证。可能在时，实验结果能更好。词向量的维度也是个值得商榷的问题，本文设定的是768，对于小规模的语料，维度过高可能学习不到更多有用的信息。同时，与传统的Skip-Gram相比，最近大火的Bert用了12层transformer，可以充分学习到字与字的依赖关系。

本文设计的Baseline对比性不强。因为只有本文模型使用了Attention机制，还需要设计其它的Baseline，比如Single Attention，Hierarchical Attention等。

## 后续工作

在实际的问答对评论中，多标签频率更高。因为用户想要了解产品的多个方面。例如“亲们这款手机质量怎么样啊！音质好不好电池能用几天啊 老人用着还可以，声音蛮好的。 质量,IO“。

下一步的工作要围绕多标签来展开。如果是基于本文单标签模型设计的话，扩展到多标签也很容易。问句中涉及多个属性，那么我们就将问句进行拆分。拆分后的问句和拆分后的答句做笛卡尔积，后续的操作与单标签模型一致。另外，文本的编码可以采用预训练模型Bert。如果想要编码效果更好，可以使用Bert微调，这样在编码层方面就提升了多标签模型的性能。

上述的多标签模型设想还没得到验证，后续会通过实验来证明多标签模型的有效性。

# 参考文献

1. Chen, Z., Liu: Topic modeling using topics from many domains, lifelong learning and big data. In: International Conference on Machine Learning, pp. 703-711(2014)
2. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP, pp. 1746-1751(2014)
3. Tang, D., Liu, T.: Effective LSTMS for target-dependent sentiment classification. In: COLING, pp. 3298-3307(2016)
4. Xue, Wang, Q.: MTNA: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In: International Joint Conference on Natural Language Processing, pp. 151-156(2017)
5. Mitchell, M., Van Durme, B.: Open domain targeted sentiment. In: EMNLP, pp. 1643-1654(2013)
6. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: ACL, pp. 339-348(2012)
7. Toh, Z., Su, J.: NLANGP: supervised machine learning system for aspect category classification and opinion target extraction. In: International Workshop on Semantic Evaluation, pp. 496-501(2015)
8. J. Dean, S. Ghemawat.MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2018, 51(1): 107-113.