中文新闻文本分类研究

操海洲
haggi.cho@outlook.com
202028016029011

杨旭锐
yangxurui20g@ict.ac.cn
2020E8013282023

王怡宁
wyn_cnic@163.com
2020E8016082028

祝菲
zhufei@iie.ac.cn
202028018629033

陆峰
lukunjie@live.com
2020E8013282001

1 问题描述

文本分类是自然语言处理领域（NLP）中一个经典问题，旨在对一些如句子、段落和文档等文本单元打上标签或分类。其方法目前广泛应用于情感分析、垃圾邮件检测、新闻分类、内容审核等场景中。

文本分类方法在新闻领域有着重要的应用场景，它能够实现对新闻信息的自动归类。近年来，随着互联网的广泛普及，新式词汇和文体的出现速度明显加快，给文本分类带来了新的巨大挑战。同时，网络使得信息的传播渠道得到了极大的扩宽，传播速度迅猛提高，促进了新闻的生产，新闻的数量呈现出爆发式增长的态势。这些都表明智能分类、标引方法在新闻领域有着极大的应用价值。此外，中文文本不同于英文文本，词与词之间没有明显的区分，增加了分词的难度，不同的分词结果甚至影响句子的含义。因此，中文新闻文本分类方法的研究有着重要的意义。

2 相关工作

20 世纪 90 年代之后，主要用机器学习方法解决文本分类问题。而传统的机器学习方法在学习文本的语义特征上表现不佳，深度学习能更好地学习与表达特征，且具有良好的建模能力。2014 年，注意力机制模型首次在机器翻译领域被提出 [?]，其优化了之前的 Encoder-Decoder 模型，在解码时选择性地从输入向量序列中挑选一个子集进行进一步处理。与此同时 TextCNN 模型被提出 [?]，它将卷积神经网络（CNN）应用到文本分类任务中，利用多个不同尺寸的卷积核来提取句子中的关键信息，从而能够更好地捕捉局部相关性。Vaswani 等人于 2017 年提出 Transformer 模型 [?]，它用注意力机制代替了循环神经网络（RNN）搭建了整个模型框架，且提出了多头注意力机制方法，并在之后的编码器和解码器大量使用多头注意力机制。进一步地，2018 年 Google 提出 BERT 模型 [?]，使用 Transformer 模型作为算法主要框架，引入掩码语言模型（MLM）与连贯性判定（NSP）方法预处理目标文本，训练更大规模的数据，使 NLP 达到了一个全新的高度。

3 备选实验方案

总体的实验流程：(1) 读取实验数据；(2) 对文本字符长度以及标签类别进行分析，完成数据预处理工作；(3) 建立中文新闻文本分类模型（包括对中文的分词、Word Embedding 等工作）；(4) 根据测试集数据的评估情况进一步调整及优化模型。

其中，在 (3) 文本分类模型建立部分，我们计划使用三种在文本分类领域中应用较为广泛的深度学习模型，分别为 LSTM、LSTM+Attention 和 BERT，并计划在实验过程中尝试在这些模型的基础上寻找可改进点。

3.1 长短期记忆网络（LSTM）

LSTM[?] 是一种改进的循环神经网络，它极大地改善了原始循环神经网络易于梯度消失或梯度爆炸的问题，常用于处理和预测具有时序关联的数据。在将中文文本转化为词向量，即可使用 LSTM 进行文本分类。

3.2 LSTM+Attention

LSTM 在一些任务中仍有其一定的局限性，比如其性能受限于固定长度的向量表示。而 Attention 打破了这种固定长度向量的限制，并能更好地区分不同信息的重要程度。将 LSTM 与注意力机制结合使用，可能会提高本任务的分类性能。

3.3 BERT

BERT 是一个基于双向 Transformer 的多层 Encoder-Decoder 结构的模型，具有较高的分类性能和较强的泛化性能。在实验时可以使用 Google 开源的 Bert 中文预训练模型对新闻文本数据进行训练和对模型参数的调整，以完成中文新闻的分类任务。