

融合提示学习和迁移学习的舆情分析应用研究

李洪霞

青岛农业大学 理学与信息科学学院

摘要: 随着社交媒体的快速发展,网络舆情随时可能出现,良好的舆情监测机制对维护社会稳定尤为重要。本文从人工智能在网络舆论中的情感分析出发,首先介绍了提示学习、迁移学习和小样本学习;然后在此基础上提出一种融合提示学习和迁移学习的舆情监测方法;最后对舆情分析中需要进一步研究的问题进行了总结和展望。

关键词: 提示学习;预训练;微调;迁移学习;舆情监测

【DOI】10.12252/j.issn.2096-627X.2023.09.132

一、引言

随着微博、微信等社交媒体的迅猛发展,人们可以对社会事件、热点问题等自由地发表自己的言论和看法,倾向性强的言论和观点进而形成网络舆情。而自媒体的出现和迅速发展,使网络舆情传播呈现出舆情主体多元化,参与人员范围广,舆情参与过程全程化的特征^[1]。如何对这些海量数据中分析,对负面舆情进行监测,从而达到社会稳定,是当前面临的重要问题。

近年来,随着人工智能技术的飞速发展,特别是机器学习和深度学习技术等自然语言处理等方面的广泛应用,使得将人工智能应用到舆情监测成为可能;但是,虽然对舆情监测的研究有很多,也有足够大的舆情数据样本空间,而真正进行正确标注的数据则比较少,对这种问题又应该如何应对。

本文从网络舆论中的情感分析入手,回顾了近年来情感分析的研究成果,介绍了传统预训练语言模型和提示学习各自的优势,以及迁移学习和小样本学习的具体应用。在此基础上,提出了一种将迁移学习嵌入到提示学习的舆情监测方法。

本文第2节是相关工作介绍,介绍了情感分析的发展历史及现状,特别介绍了提示学习相对于预训练-微调模型的优势,然后针对标注数据少的情况,介绍了迁移学习和小样本学习;第三节在前面介绍的基础上,给出将迁移和提示学习结合起来的可能性,并给出具体的结合方法来解决舆情监测问题;最后对舆情监测下一步研究面临的问题进行了总结和展望。

二、相关工作

1. 情感分析发展

对网络舆论中的情感分析是人工智能在自然语言处理的(Natural Language Process, NLP)重要应用领域。随着自然语言处理技术的发展,对舆情中情感分析的成果主要分为3类:1)基于规则和词典的方法,这

种方法较少考虑上下文信息,因此有比较大的局限性。

2)应用机器学习,特别是深度学习的方法进行情感分析,主要是应用了循环神经网络(RNN)及其变体(如LSTM)和注意力机制。这类方法能够通过捕获句子的序列信息和上下文信息,进而获得单词之间的语义和语境环境,因此分析相对准确。但这种方法对每一个新出现的数据集都需要重新训练网络权重。3)基于预训练语言模型(Pretrained Language Model, PLM)的情感分析方法,如BERT, RoBERTA等。PLM一般采用预训练-微调模式,即先在大型语料库上进行无监督训练得到高容量的PLM模型,然后根据具体的下游任务对模型中的参数进行微调以解决具体的问题。由于PLM一般采用了注意力机制(Attention)等解决了一词多义、词汇之间、句子之间以及上下文之间的情感特征,又有巨大的语料库、模型规模和参数做助力,因此PLM在情感分析方面取得了较好的效果。

2. 提示学习

传统的PLM模型以学习语言的通用特征为目的,其学习过程和模型参数等均与具体的下游任务无关^[2]。为了能够通过微调来使模型适应各种不同领域(如金融、医学等)的任务,就需要有足够大的模型规模,对语料库中数据的数量、质量和广度以及硬件的要求都会提高。

随着GPT系列产品的异军突起,提示学习(Prompt Learning)的作用也受到了重视。提示学习也是PLM模型的一种,不同于原来的预训练-微调模型,它是在大型语料库先预训练模型,然后通过文本提示的方式,对下游任务进行重构,使下游任务转化成模型可以理解的形式,然后将模型应用到下游任务中。相对于预训练-微调模型,提示学习有其优势:不再使PLM适应下游任务,而是使目标任务接近于原PLM训练所要解决的任务,从而减少模型微调的难度。

3. 小样本学习

在传统的监督学习中，需要丰富的、已加标记的数据来训练大规模语言模型，但无论对大量数据加标记，还是对加了标记的数据进行模型训练，都需要耗费相当多的资源。如何能对少量样本进行学习就可以得到解决问题的模型，这就是小样本学习^[3]。

对小样本学习问题，常采用模型微调，基于数据增强和基于迁移学习等进行解决。模型微调即先在大规模数据上训练模型，然后在目标数据集对模型的某一层或某几层进行参数微调，以适应小样本问题。基于数据增强的方式指充分利用辅助数据或辅助信息，对小样本数据进行扩充或特征增强，以提高样本的多样性，从而达到学习的目的。迁移学习则在借用其他领域的先验知识，来解决目标仅有少量有标签样本数据的情况。

4. 迁移学习

在传统学习中，进行模型训练时的源领域和应用模型的目标领域是相同的，即利用领域A中的部分数据进行模型训练，然后将训练后的模型应用到领域A中的其他数据^[4]。迁移学习与之相对应：利用知识可迁移的特点，将源领域A中学到的知识应用到目标领域B中。如学习了C语言后，再学习其他编程语言，如Python，比从零基础学习Python语言要容易的多。迁移学习非常适合目标领域是零样本或小样本数据的情况。同时，源领域与目标领域之间的关联性越强，迁移学习的效果也越好。

迁移学习当前有多种分类方法，其中根据源领域和目标领域中的数据是否已做标记，可分为归纳迁移学习、直推式迁移学习和无监督迁移学习。其中，归纳迁移学习应用到目标领域中仅有少量标注的数据，源领域和目标领域任务相关但不相同的情况；在直推式迁移学习中，只有源领域数据有标注而目标领域数据无标注，应用到源领域和目标领域之间相关的、但其任务相同的情况；无监督迁移学习应用到源领域和目标领域中的数据均无标注，源领域和目标领域任务相关但不相同的情况。

三、融合提示学习和迁移学习的小样本舆情分析

当前的PLM语料以英文形式的居多，中文形式的语料库相对较小，具体到国内舆情监测，其正确标注的数据更少。若采用预训练-微调模型对舆情进行分析，则在微调时要适应舆情分析要求，即用大量数据适应较少的样本，从而容易形成过拟合。而且因为某种具体任务

重新调整模型，也会造成资源的浪费。提示学习可将舆情分析这一任务，通过设计或选择与预训练任务相契合的模板的方式，将舆情监测的相关数据转成自然语言形式，输入到PLM模型中，从而充分挖掘PLM模型自身对语言的理解能力。

同时，舆情分析作为情感分析的具体应用，与一般的情感分析有相似性，也有其特殊性：一是已标注的数据样本量较少，二是应用场景有其局限性。将情感分析的通用方法、模型和成果迁移到舆情分析中，既可以充分利用情感分析的已有知识，又降低了模型训练的难度，提高了模型训练的效率。由此，得到一种融合提示学习和迁移学习的舆情监测方法。

1. 提示学习的应用

首先，在大规模数据上预训练得到PLM，然后将提示学习应用到模型中。提示学习包括以下三个步骤：

1) Prompt添加；2) Answer搜索；3) Answer 影射。其中：

Prompt添加：指通过合适的提示，将舆情分析转化成PLM可以理解和分析的形式，并在合适的位置上Mask，以充分激发原PLM的能力；

Answer搜索：将Prompt添加后的内容输入到PLM中，并在文本空间中搜索最优的预测词，作为预测语言模型对舆情分析的预测输出；

Answer影射：通过简单或复杂的映射函数，将预测输出映射到舆情分析所需要的类别或标签中，使模型输出符合舆情分析任务的要求。其过程可描述为图1的形式。

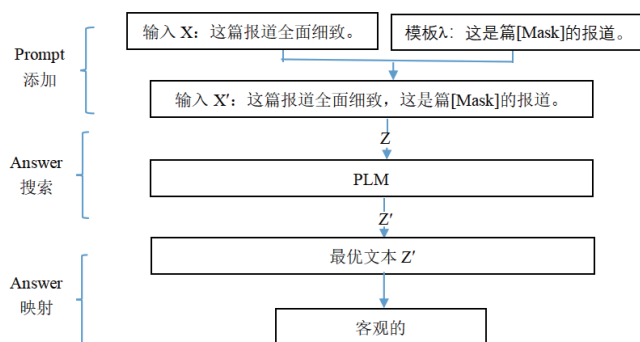


图1 提示学习应用

在上述提示学习的步骤中，Prompt添加中的提示模板是提示学习的根本，它决定着学习的质量。当前Prompt模板主要有硬提示和软提示两大类模板。其中，硬提示模板一般将事实表述成“完型填空”的形式，以助于模型理解下游任务。“完型填空”可以是人工设计，也可以是从离散空间中搜索得到。其中，人工设计

的模板灵活性高，可针对具体的任务进行设计，但需要有针对下游任务的专业背景和经验，而且耗费人力和资源；从离散空间中搜索指在从现有文本语料库中利用同义语替换、寻找输入输出之间的路径等方式来构造模板。无论哪种硬提示模板方式都考虑到了模板的可读性和可解释性。软提示模板则不追求模板的可解释性，而是将离散标记优化为一系列可学习的参数。软提示模板可以在耗费计算资源较少的情况下，达到与预训练-微调方式类似的效果。

在舆情分析时，采用软-硬提示模板相结合的方式。先用硬提示模板（人工或离散空间搜索）给出粗糙的原始模板，然后利用软模板对各离散标记进行细化；最后再由专家对模板进行打磨，确定最终的模板。这种软-硬结合的提示模板方式，一是可以综合各种模板的优势，二是考虑国情和民意的特殊性，最终提示模板由行业专家来确定。

将提示学习应用于舆情分析后，模型结构包括PLM和提示模板两种参数。根据这两类参数是否都需更新，可分成四种训练策略，一是PLM参数和提示模板参数均不调整，直接应用PLM预测结果，适用于零样本学习问题；二是冻结PLM参数、更新提示模板参数，这种策略可使模板对下游任务更有针对性；三是更新PLM参数、冻结提示模板参数，这种策略确保模型对提示的敏感性；四是PLM参数和提示模板参数均加以更新，这种更新对样本数据和计算资源的要求更高。根据舆情分析数据的小样本特点，同时考虑模板的敏感性，此处采用只更新PLM参数而冻结提示模板参数的形式。

2. 提示学习与迁移学习的融合

将提示学习应用到从大规模自然语言中训练得到的PLM有助于舆情的分析。由于当前对情感分析的成果较多，将情感分析中其他细分领域的模型迁移到舆情分析中，将比直接在大规模自然语言中训练得到的模型更有针对性。

将当前情感分析的三个子应用：产品评价、情绪分析和文本识别中的模型迁移到舆情分析中进行小样本学习。因这三种应用均属于情感分析大类，因此源域和舆情分析的目标域距离较小，可以进行迁移学习。另外，根据舆情分析只有少量标注数据的情况，采用归纳式迁移学习。做法如图2所示，具体可描述如下：

对情感文本的三个子应用利用已有的软或硬提示模板P进行提示学习得到预训练模型M；

将提示模板P采用归纳式方式迁移到舆情分析任务，得到目标提示模板P'；

将现有知识经提示模板P' 重构后，输入到预训练模型M中，得到舆情分析要得到的结果。

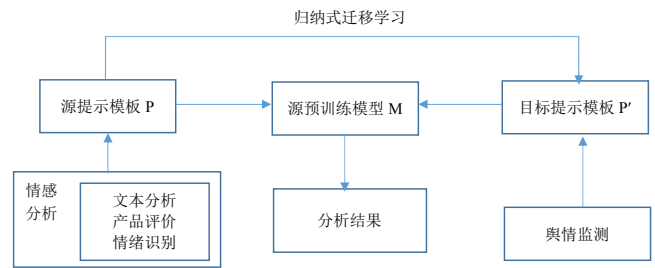


图2 提示学习与迁移学习的融合

四、总结与展望

本文针对舆情分析这种小样本问题，给出了将迁移学习嵌入到提示学习中的解决方法。提示学习通过引入任务描述或问题来指导模型，可以弥补传统模型针对性不足的缺点。而迁移模型是专门针对小样本问题设计，将已知领域知识应用到未知领域中。基于现有情感分析类应用的成果，以及提示学习和迁移学习的特点，本文将情感分析的成果应用到舆情监测这一具体领域中，提出一种融合提示学习和迁移学习的舆情分析方法。

当然，要实验舆情精准监测仍然需要进一步的研究，如情感分析类成果仍有完善空间，生成式人工智能数据的出现为训练模型带来了更大的难度。

参考文献

[1] 付可, 孙源. 社交媒体——舆情监测的“前哨站”——以抖音平台为例[J]. 西部广播电视, 2022, 43 (S01): 57-65.

[2] 顾勋勋, 刘建平, 邢嘉璐, 任海玉. 文本分类中Prompt Learning方法研究综述[J/OL]. 计算机工程与应用. <https://link.cnki.net/urlid/11.2127.TP.20231211.1556.004>

[3] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2021.

[4] 庄福振, 中国科学院智能信息处理重点实验室, 庄福振, 等. 迁移学习研究进展[J]. 软件学报, 2015, 26 (1): 14.

项目：2023年度青岛市社会科学规划研究项目（QDSKL2301305），青岛统一战线智库2023年度立项课题（QDTZZK2023032）

作者简介：李洪霞，青岛农业大学理学与信息科学学院，博士，讲师。