大模型背景下的“自然语言处理”课程与智能问答辅助教学

訾红英 刘 涛 张坤丽[通讯作者] 王新荣
（郑州大学，河南 郑州 450001）

摘要：对于社会对智能信息处理人才的迫切需求和目前自然语言处理快速的更迭中教育教学所面临的困难，基于大规模语言模型和自然语言处理的知识，提出转变教学、更新教学内容、丰富教学方法，设计智能问答来搭建智能答疑平台实现课堂的翻转教学，最后应用于自然语言处理课程的教学。

关键词：自然语言处理；大规模语言模型；智能问答

2022 年，科技部、工业和信息化部等六部门印发了《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，以落实《新一代人工智能发展规划》，系统指导各地方和各主体加快人工智能场景应用。随后，2023 年发布了《北京市人工智能行业大模型创新应用白皮书（2023 年）》，强调将发展人工智能置于全局工作的统筹谋划中，加速布局和发展智能产业。

自然语言处理一直是计算机科学领域与人工智能领域中的一个重要研究方向之一，旨在探索实现人与计算机之间用自然语言进行有效交流的理论与方法。近年来，随着大规模模型成为人工智能领域的热点话题，自然语言处理任务得到了更加准确和深入的处理。例如，OpenAI 的 GPT 系列和 Google 的 Bert 系列等模型在 NLP 领域应用非常广泛，显著地提升了信息抽取、情感分析、文本摘要、机器翻译、知识图谱和智能问答等任务的效果。大模型应用于自然语言处理逐渐成为人工智能最核心的地位，其进步必将推动人工智能整体的发展。

因此，在大模型的背景下，“自然语言处理”课程正在被高校引入，作为计算机、人工智能相关专业本科生或研究生的必修和选修课程。然而，在学习过程中，学生由于缺乏兴趣，难以有效领会大模型和自然语言处理的相关专业知识。为了解决这一问题，本文将利用计算机与人工智能的便捷性，采用自然语言处理中的智能问答系统，致力于线上教学，为学生提供个性化的学习辅助。系统可以根据学生的学习进度和理解程度，提供定制化的问题和解答，帮助他们更好地理解课程内容。同时在课堂上使用问答系统进行实时互动，发挥专业课程的集群效应和线上线下的互补优势。

一、学科背景

（一）自然语言处理的发展

自然语言处理的研究内容十分庞杂，整体上可以分为基础算法研究和应用技术研究。从语言单位的角度来看，涵盖了字、词、短语、句子、段落、篇章等不同粒度。从语言学角度看，涉及形态学、语法学、语义学等不同层面。针对特定的自然语言处理任务，以有监督学习（Supervised Learning）、无监督学习（Unsupervised Learning）、半监督学习（Semi-supervised Learning）、强化学习（Reinforcement Learning）等不同的机器学习算法为基础进行构建。目前基于机器学习和深度学习的自然语言处理算法主要集中在形态、语法、语义三个层面。我们从语言单元粒度和语言学研究层次两个维度，对自然语言处理的主要研究内容进行了归类。如图 1 所示：

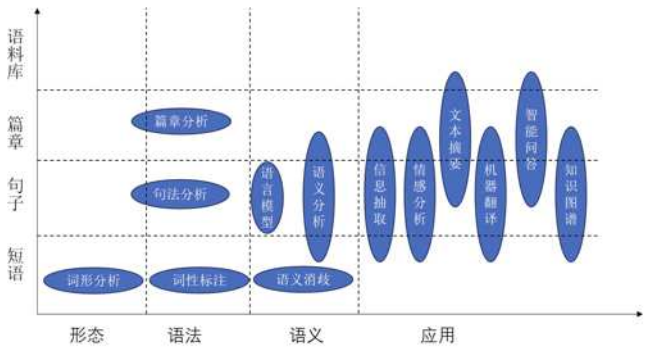


图 1 自然语言处理的主要研究内容

“自然语言处理”的发展经历了基于规则、基于统计学习的方法，到基于深度学习的方法，再到目前大模型的方法，经历了四个阶段的发展。

基于规则的自然语言处理方法的主要思想是通过词汇、形式文法等制定的规则引入语言学知识，从而完成相应的自然语言处理任务，其中主要包括数据构建、规则构建、规则使用、效果评价四个部分。

基于机器学习的自然语言处理方法绝大部分采用有监督分类算法，将自然语言处理任务转化为某种分类任务，在此基础上根据任务特性构建特征表示，并构建大模型的有标注语料，完成模型训练。通常分为四个步骤：数据构建、数据预处理、特征构建、模型学习。对于复杂的自然语言处理任务而言，需要参与的人工参与和选择的环节非常多，从特征设计到模型，再到优化方法和超参数，这些选择非以往经验，缺乏有效的理论支持。其次，多模块的级联会造成错误传播，前一步的错误会影响后续的模型，这些问题都提高了基于机器学习的方法实际应用的难度。

基于深度学习的方法通过构建具有一定“深度”的模型，将特征学习和预测模型融合，通过优化算法使模型自动地学习到好的特征表示，并基于此进行结果预测。随着深度学习研究的不断深入和计算能力的快速发展，模型深度从早期的 5–10 层增加到现在的数百层，从而使得预测部分更加简单，预测也更加容易。

自 2018 年 ELMo 模型被提出后，基于深度学习的自然语言处理又进一步演进为预训练微调范式。预训练模型在模型网络结构上可以采用 LSTM、Transformer 等具有较好序列建模能力的模型，预训练任务可以采用语言模型、掩码语言模型（Masked Language Model）、机器翻译等自监督或有监督的方式，还可以引入知识图谱、多语言、多模态等扩展任务，取得了非常好的效果。但仍然面临模型稳健性提升、模型可解释性等诸多问题亟待解决。

2020 年 OpenAI 发布的 GPT-3 模型的规模达到了 1750 亿个参数，这种参数级的语言模型很难再延续此前针对不同的任务而使用的与训练微调范式。2022 年 ChatGPT 所展现出的通用任务理解能力和未知任务泛化能力，使得未来自然语言处理的研究范式可能进一步发生变化。大规模语言模型的自然语言处理方法的基本流程为大模型语言模型构建、通用能力注入和特定任务使用。如图 2 所示：

图 2 大模型构建的基本流程

大规模语言模型构建阶段，大量文本内容训练对长文本的建模能力，使得模型具有语言生成能力。在通用能力注入阶段，利用包括阅读理解、情感分析、信息抽取等现有任务的标注数据，使得模型具有很好的泛化能力。特定任务使用阶段变得非常简单，模型具备了通用任务能力，只需要根据任务需求设计任务指令。如果该范式在非常多的任务上达到了基于与训练微调范式的结果，会使得自然语言处理产生质的飞跃。

（二）课程教学现状分析

自然语言处理属于典型的交叉学科，涵盖语言学、计算机科学、数学等学科知识。以前的自然语言处理课程覆盖广泛，偏重于统计的处理方法，滞后于目前人工智能技术的发展水平与实际应用需求。深度学习在该领域取得了显著的成果，大语言模型也横空出世。各种居于前列的智能信息处理大部分是大模型和深度学习模型，教学内容的更新和优化十分必要。

以往的课程教学评价存在“重理论、轻实践”以及“重考试结果、轻动手过程”的弊端，往往通过期末考试 + 大作业考核的方式评定最终的学习效果，评价结果并不能反映学生将理论和实践相结合后的学习成果。学生的学习也比较盲目，缺乏持久的兴趣和动力。

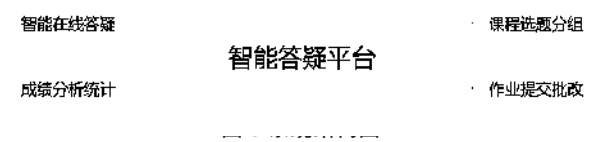
基于对现状的分析，结合自然语言处理发展的现状与社会对智能信息处理人才应具备能力的硬性需求，转变传统教学理念。将理论课程、比赛实践、辅导系统贯穿于整个教学系统。核心围绕于学生获得前沿知识，参与公开算法比赛来提升自我解决能力和综合知识运用能力，并利用智能问答系统全程辅导学生的学习。与传统的教学相比，教学内容突出前言知识的理论和应用，侧重于深度学习算法的理解与经典大模型的运用，实践项目选取来自阿里云天池平台，并引导学生参与比赛并打榜，鼓励学生探究和创新自然语言技术应用。

目前该领域的研究内容分为基础性技术研究、智能应用技术研究、大模型技术研究。基础技术涉及词汇分析、句法分析、语法分析、篇章分析、语言模型。智能应用技术包括信息抽取、机器翻译、情感分析、智能问答、文本摘要、知识图谱等，内容涉及相当广泛。大模型技术包括模型的稳健性、大规模语言模型、大规模预训练模型等。

为了贴合行业对人才与时俱进的要求，适应人工智能发展的迅猛劲头。在自然语言处理课程教学内容的设计和选取上，我们结合国内 2024 年 1 月出版的最新两本教材：自然语言处理导论和大规模语言模型。在课程内容上做到取舍、有重点、有层次地安排教学课程内容，兼具广度和深度。我们首先弱化部分理论性、公式繁多且抽象的内容，如决策树，K 邻近算法等传统机器学习。强化神经网络与深度学习知识与大语言模型知识，以及大模型微调技术。做到人工智能的发展的历史和前沿，培养宏观与微观相结合的科学思维能力与创新思维能力。

二、智能答疑平台建设

传统教学模式条件苛刻，要求教师和学生必须在学校课堂上完成教学任务和学校任务，导致师生之间的交流较少。为解决这一问题，我们引进网络信息技术，搭建智能答疑平台。帮助教师和学生开展学习交流活动。针对于课堂中不解以及疑惑的知识点，学生可通过该平台向老师和智能辅助系统求助。智能答疑平台主要应用于“自然语言处理”课程。针对自然语言处理专业课，通过课程选题分组、作业提交批改、智能在线答疑、成绩分析统计四个环节，建立该专业课的智能答疑平台。如图 3 所示：



（一）平台主体

从系统设计的角度，该答疑平台分为教师子系统和学生子系统。教师管理子系统包括学生信息管理、人工答疑管理、作业批改管理、智能答疑数据管理、学生成绩分析。学生子系统包括小组分组管理、作业提交下载、智能答疑管理、课后提问管理等模块。

2、智能答疑

智能答疑子系统。该子系统包括模式匹配、信息抽取、问答推理三大模块。从文句中提取关键的词语，用信息检索的方法找出包含候选答案的段落或句子，然后基于问答类型用信息抽取的方法从备选答案中提取出最佳的答案。检索过程：段落或者句子级排序，利用不同类型关键词的加权组合。答案抽取过程：根据问答类型从排序后的段落或句子中抽取答案对于某些提问类型。模式匹配中：对于某些提问类型（某人的出生日期、原名、别称等），问句和包含答案的句子之间存在一定的答案模式，该方法在信息检索的基础上根据这种模式找出答案。如图 4 所示：

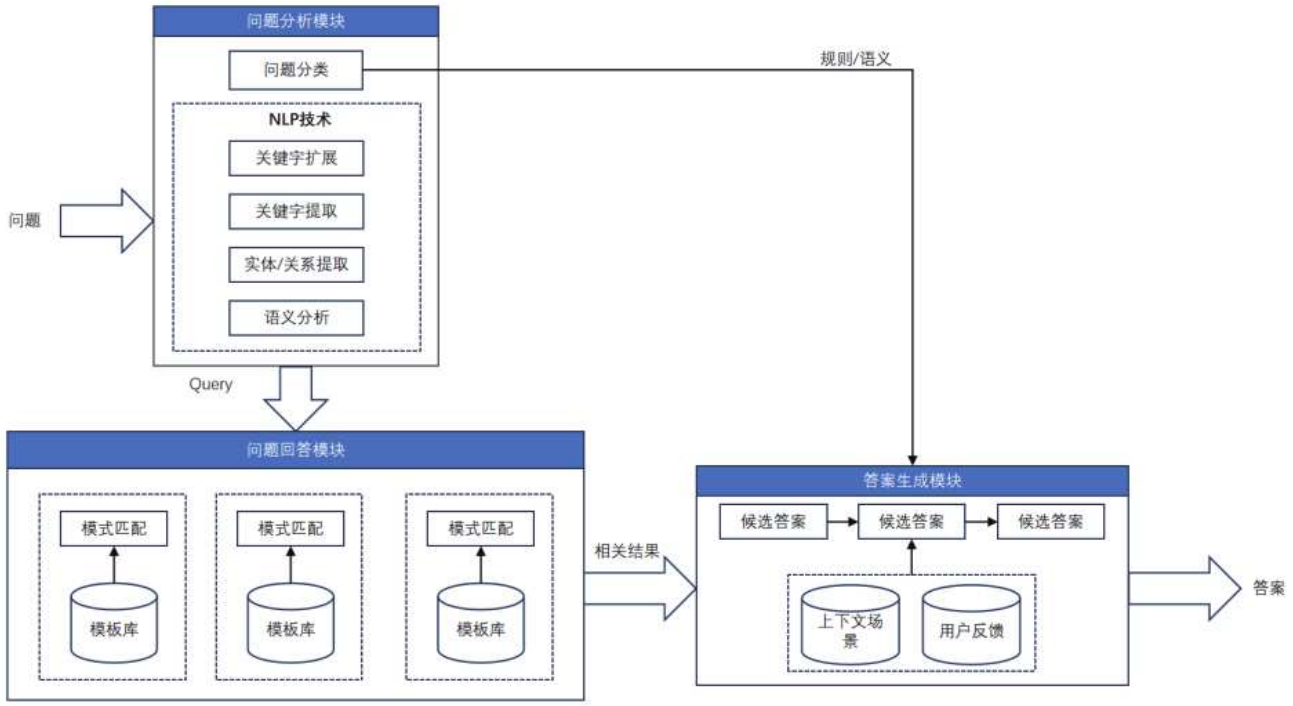


图 4 答案生成过程图

给定自然语言处理问题，通过对问题进行语义理解和解析，利用知识库进行查询、推理得出答案。其特点是，回答的答案是知识库中的实体。答案的生成是问答模型返回的简短的答案。在训练过程中我们还对用户的行为进行了反馈，根据用户的结果去指导我们去做更语言处理为众多的智能用户与智能设备提供了关键的技术，大模型也受到了越来越广泛地关注和重视。各大高校也在高年级本科生与研究生阶段开设自然语言处理课程，并作为一门重要好对话模型的理解。

三、结语

笔者结合自然语言处理的发展，梳理教学知识，以及调研行业前沿智慧应用技术来满足学生实际的学习需求，建立一个可行的、合理的、以具体成果为导向的教学课程。在教学内容的设计和选取上兼具广度和深度，以及侧重基于深度学习的自然语言处理技术内容，加快课程内容知识的更新；在不同教学章节中运用问题引导式、讨论式、项目驱动式等教学方法；并利用人工智能的技术，设计智能答疑平台为学生的学习扫清障碍，翻转课堂。体现了人工智能类课重实践、重过程、重自学能力的特点。

参考文献：

[1] 张博，董瑞海 . 自然语言处理技术赋能教育智能发展——人工智能科学家的视角 [J]. 华东师范大学学报（教育科学版），2022，40（9）：19.

[2] 赵朝阳，朱贵波，王金桥 .ChatGPT 给语言大模型带来的启示和多模态大模型新的发展思路 [J]. 数据分析与知识发现，2023，7（3）：26-35.

[3] 李红，林珊，欧阳勇 . 基于深度学习的自然语言处理课程教学探索与实践 [J]. 计算机教育，2021.

[4] 张宜浩，刘小洋 . 新工科背景下自然语言处理课程教学改革 [J]. 计算机教育，2023.

[5] 王志敏 . 计算机网络技术中人工智能的有效应用分析 [J]. 数字通信世界，2020（4）：89-89.

[6] 李葆嘉 . 当代语言学理论的追溯 [J]. 华东师范大学学报（哲学社会科学版），2021，53（6）：77.

本研究得到了河南省研究生精品在线课程项目的支持，项目编号 YJS2022ZX05，自然语言处理。感谢自然语言处理课程提供的宝贵数据和匿名审稿人对本文提出的宝贵意见和建议。