



《自然语言处理专业概述》

|  |  |
| --- | --- |
| **专 业** | **人工智能（未来技术）** |
| **姓 名** | **刘栩祺** |
| **日 期** | **2022/9/6** |

**摘 要**

自然语言处理专业方向属于人工智能专业（专业代码080717T），是一门集语言学、数学、计算机科学与认知科学等于一体的综合性交叉学科。近年来，随着计算机网络技术和通信技术的迅速发展和普及，自然语言处理的需求急剧增加，人们迫切需要使用的自然语言处理技术来帮忙打破语言屏障，为人际之间、人机之间的信息交流提供便捷、自然、有效的人性化服务。本文通过对自然语言处理相关内容进行概述，使读者可以初探“人工智能皇冠上的明珠”——自然语言处理。

**关键词：**自然语言处理 人工智能 计算机科学与技术

**目 录**

[第1章 专业研究内容介绍 - 1 -](#_Toc14478)

[1.1 定义 - 1 -](#_Toc8865)

[1.2 专业研究内容分类 - 1 -](#_Toc8865)

[第2章 专业主要课程内容 - 3 -](#_Toc14478)

[2.1 专业基础课 - 3 -](#_Toc8865)

[2.2 专业核心课...............................................................................................................- 3 -](#_Toc8865)

[第3章 专业学习建议 - 5 -](#_Toc14478)

[第4章 专业前沿研究方向....................................................................................................- 6 -](#_Toc14478)

[第5章 专业学生毕业去向....................................................................................................- 7 -](#_Toc14478)

[结 论 - 7 -](#_Toc5560)

[参考文献 - 8 -](#_Toc1841)

**第1章 专业研究内容介绍**

## 1.1 定义

自然语言处理(Natural Language Processing，NLP)是人工智能的一个子领域，其产生可追溯到20世纪50年代，它是一门集语言学、数学、计算机科学和认知科学等于一体的综合性交叉学科，被誉为“人工智能皇冠上的明珠”。自然语言处理利用计算机等工具对人类所特有的语言信息（包括口语信息和文字信息）进行各种加工，并建立各种类型的人-机-人系统。自然语言的核心是自然语言理解，其中包括语音和语符的自动识别以及语音的自动合成。

### 1.2 专业研究内容分类

自然语言处理的研究内容是以任务为导向的，从宏观上说有四大基本任务，分别是序列标注、分类任务、句子关系判断和生成式任务。具体到下游应用时又细化为：

信息抽取：从给定文本中抽取重要的信息，比如时间、地点、人物等。即了解谁在什么时候、什么原因、对谁、做了什么事、有什么结果。

文本生成：让计算机像人一样使用自然语言进行表达和写作，主要包括数据到文本生成和文本到文本生成。数据到文本生成是指将包含键值对的数据转化为自然语言文本；文本到文本生成是对输入的文本进行转化和处理从而产生新的文本。

文本分类：文本分类是利用计算机系统对大量的文档按照一定的分类标准（比如主题、内容等）实现自动归类。近年来，情感分类成为研究的热点。

问答系统：人类用自然语言提出问题，问答系统给出一个精准的答案。这项任务需要计算机能对自然语言描述的问题进行语义分析，然后到知识库中查找可能的候选答案并选出最佳答案。

对话系统：人类用自然语言与计算机进行一系列的对话，包括聊天、回答、完成人类指定的某一项任务。

语音识别和生成：语音识别既包含对输入计算机的语音信号进行识别转换成书面语表示，又包含将书面符号转化为语音信号。语音生成又称文语转换、语音合成。

机器翻译：机器翻译是自然语言处理最早的应用，它是把输入的源语言文本通过自动翻译获得另外一种语言的文本。目前机器翻译方法主要基于神经网络的编码-解码模型，广为人知的应用有百度翻译、谷歌翻译等。

信息检索：对于用户的查询，索引大量的文档。在查询的时候，对输入的查询表达式（比如一个检索词或者一个句子）进行分析，然后在索引里面查找匹配的候选文档，再根据一个排序机制把候选文档排序，最后输出排序得分最高的文档。

隐喻计算：“隐喻”是指用乙事物或其某些特征来描述甲事物的语言现象。隐喻计算就是研究自然语言语句或篇章中隐喻修辞的理解方法。

**第2章 专业主要课程内容**

## 2.1 专业基础课

与其他计算机专业一样，人工智能专业基础课有《高级语言程序设计》、《数据结构与算法》、《算法设计与分析》、《计算机系统》、《计算机组成原理》、《计算机网络》等，除此之外还有：

《自动控制原理》：主要教授自动控制系统的构成（基本环节、传递函数、系统框图）、工作原理和控制系统数学模型的方法（时域法、根轨迹法、频率法），掌握计算机控制系统的工作原理以及分析和综合的方法。

《信号与系统》：主要教授信号的时频相频特性、线性时不变系统的特性、周期及连续信号的傅里叶变换等，掌握用系统的观点和方法分析求解电子系统的特性，从而为后续《视听觉信号处理》、《数字图像处理》的课程奠定基础。

### 2.2 专业核心课

本专业的专业核心课共有六门，分别是《机器学习》、《人工智能》、《自然语言处理》、《信息检索》、《语言与认知》和《中文信息处理》。其中：

《机器学习》：主要教授机器学习的基本问题和基本算法（比如聚类、回归、分类、标注相关算法）；理解神经网络模型、深度学习模型及其应用方法；掌握距离度量、模型评价、过拟合、最优化、特征工程、降维与超参数调优等机器学习工程应用方法。

《人工智能》：主要教授人工智能的基本原理和方法（知识表示技术、启发式搜索策略、确定及非确定性推理技术）、人工智能的三个重要研究领域（机器学习、神经网络学习和自然语言理解）、人工智能的两个重要应用领域（专家系统和智能决策支持系统）。

《自然语言处理》：主要教授计算机处理人类语言的基本原理、利用统计和其他机器学习方法对自然语言建模和计算（如统计语言模型、序列标注方法）以及设计、开发面向自然语言的人工智能应用系统（如词法分析、词义分析等）。

《信息检索》：作为承上启下的课程，本课程聚焦传统文本检索技术的基本知识与互联网时代信息检索技术的发展状况、信息检索技术的基本概念、信息检索系统的基本原理和常用方法以及利用人工智能、机器学习、语言处理技术对信息进行加工处理的实践。

《语言与认知》：主要讲授近年来认知科学中与语言处理相关的新理论新进展，鼓励学生在实践中运用认知科学理论与方法解决语言理解相关问题以及设计常见的语言认知系统。

《中文信息处理》：主要教授计算机处理中文语言文字的方法以及汉字处理中的智能识别技术（如模式识别），比如中文信息系统的构成、中文字符的编码方法、主流的输入法的原理及设计、汉字的字形技术及压缩方法、中文操作系统的原理和汉字识别技术的基本原理和方法等。

**第3章 专业学习建议**

自然语言处理是最近几年出现并流行的研究领域，其入门大致分为四步。第一步应该打牢自然语言处理基础。在数学预备知识方面，需要掌握微积分、线性代数、概率论等相关知识。之后需要掌握NLP的基本模型和通用方法，比如学习词嵌入、神经网络、预训练模型等，从而为后续阅读和复现相关论文提供坚实的基础。第二步是阅读NLP经典论文，通过阅读原论文，会对经典的模型和方法有更深入细致的理解，同时对当前的研究方向有一个大致的了解。第三步是构建模型，根据一个带注释的代码框架，学习基于深度学习的NLP模型，以及如何进行数据处理、训练和预测。在这一部分后，应当对如何实现模型有明确的思路。最后一步是项目实践，通过对特定的自然语言处理任务进行实操，从中锻炼设计和实现模型的能力。

**第4章 专业前沿研究方向**

Transformer扩展：Transformer是一种最先进的 NLP 模型，并已被成功地应用于许多其它的领域。对Transformer 进行扩展（比如让transformer能够兼容多个输入，增加其支持的最大输入长度），会出现很多更高效的变体，比如使用了参数共享的 ALBERT。

通用无监督预训练：无监督预训练是指用来训练的数据不包含输出目标，需要学习算法自动学习到一些有价值的信息。无监督预训练使得在训练模型时对已标注数据的需求量大大减少。通过这种技术，可以对那些无法收集大量数据的领域进行研究，并朝着多模态紧密融合的方向发展。

解决自然语言生成（NLG）中的解码误差：尽管自然语言生成（natural language generation, NLG）领域的模型越来越强大，但是它们仍然经常生成重复或者毫无意义的句子。最新的研究表明要研究新的采样方式（如原子核采样）或新的损失函数来解决这个问题。

增强预训练模型：目前的NLP模型参数越来越多，功能也越来越强大，但仍有许多知识是模型无法仅从文本中学习的。特别是在现实生活中，当处理更复杂的任务时，如果可用的数据有限且无法使用事实或常识进行显式的推理，那么就需要更强的归纳偏置。

FinQA：当前，大量的财务报表使得人类很难访问和分析企业的财务状况。该技术旨在实现大型财务报表的自动化分析。比如，向模型输入一个企业全年的财务报表，并查询该企业的建议派发股息。模型需要用数学推理、语义理解等技术对财务报表及用户查询进行综合分析，最终给出答案。

HybridQA：现有的问答系统侧重于处理同类信息，要么都是文本，要么都是知识库/表格信息，而人类知识分布在不同的形式上。HybridQA能够对异构信息进行推理，即对于用户提出的每个问题，HybridQA都可以在文本、数据库、图片和表格中进行查找，并汇总结果返回给用户。

Text-to-SQL：当今世界，许多信息存储在结构化和半结构化的数据库中，访问这些数据库需要熟练地使用编程语言（如SQL）。Text-to-SQL技术能够将自然语言问题转化为可在数据库上直接执行的编程语言（如SQL），它不仅降低了用户使用结构化数据库的门槛和成本，而且提高了结构化数据的价值和效率。

**第5章 专业学生毕业去向**

南京大学人工智能学院于2018年建立，是我国高校中最早建立的人工智能学院，也是目前人工智能专业的首批毕业生。以南京大学“AI一期”为例：南京大学人工智能学院2022届毕业生共有105名，其中本科生72名，硕士生33名。总体就业率 96.19%，本科95.83%，硕士生96.97%。其中，本科毕业生68% 升学，其中，国内升学占比90%，均为双一流高校；海外升学占比10%，其中QS前100高校占比 80%。企业就业本科生中，40%签约字节跳动等500强或荣耀等独角兽企业，以及TP-link、航天系企业等，平均年薪30.38万元，中位数30+万元。硕士毕业生 12% 升学，其中75%选择在国内双一流高校继续深造；25%选择海外深造，海外深造的高校均为QS前100。企业就业的硕士生93%签约500强或独角兽企业，平均年薪48.21万元，中位数45+万元。随着更多高校开设人工智能专业，本专业毕业生去向将更加多样化。

**结 论**

自然语言处理专业正随着互联网、移动通信和计算技术的迅猛发展而日新月异。在不久的将来，必将有更多新的自然语言处理理论方法和应用系统成为热点，自然语言处理技术也必将在社会生活的各个领域得到越来越广泛的应用。

**参考文献**

[1] 宗成庆. 统计自然语言处理. 北京：清华大学出版社，2013.

[2] 苑春法. 统计自然语言处理基础. 北京：电子工业出版社，2005.

[3] 冯志伟. 自然语言处理综论. 北京：电子工业出版社，2005.