Natural Language Processing
自然语言处理基础知识
NLP 基础

2021 年 11 月 13 日

Date Performed: November 4, 2021
Reference I: 《Python 自然语言处理实战：核心技术与算法》
《机器阅读理解：算法与实践》
《自然语言处理：从入门到实战》
Reference II: 维基百科 Wikipedia
百度百科 Baidu Encyclopedia

1 什么是 NLP

1.1 NLP 的概念

NLP(Natrual Language Processing，自然语言处理) 是计算机科学领域以及人工智能领域中的一个重要研究方向，它研究用计算机来处理、理解以及运用人类语言（如中文、英文等），达到人与计算机之间进行有效通讯。NLP 基本上可以分成两个部分：**自然语言理解**和**自然语言生成**。

a. **自然语言理解**：自然语言的理解是一个综合的系统工程，它包含着很多细分的学科：音系学（指代语言中发音的系统化组织）、词态学（研究单词构成以及相互之间的关系）、句法学（研究给定文本的哪部分是语法正确的）、语义学（研究给定文本的含义是什么）和语用学（研究文本的目的是什么）。

b. **自然语言生成**：从结构化数据中以读取的方式自动生成文本，过程包括：文本规划（完成结构化数据中的基础内容规划）、语句规划（从结构化数据中组合语句来表达信息流）、实现（产生语法通顺的语句来表达文本）。

1.2 NLP 的研究任务

机器翻译 计算机具备将一种语言翻译成另一种语言的能力。

情感分析 计算机能够判断用户评论是否积极。

智能问答 计算机能够正确回答输入的问题。

文摘生成 计算机能够准确归纳、总结并产生文本摘要。

文本分类 计算机能够采集各种文章，进行主题分析，从而进行自动分类

舆论分析 计算机能够判断目前的舆论导向

知识图谱 知识点相互连接而成的语义网络

2 NLP 的发展历程

具体发展历程可以参考维基百科：NLP 发展历程

3 NLP 的一些基本术语

为了更好地理解和学习 NLP，这里会一一介绍一些基本的 NLP 领域中的专业术语（包括但不仅限于这些），以下仅仅只是部分常用的入门术语，深入到更加具体的不同领域时，会有更多其他这里未提及的术语。

a. **分词（Segment）**：单词是语言中最重要的元素。一个单词可以代表一个信息单元，有指代名称、功能、动作、性质等作用。在 NLP 中，理解单词对于分析语言结构和语义有着重要作用，尤其是在**机器阅读理解（MRC）**领域。解决歧义性是分词任务中的一个大问题，如：“美国会通过对台售武法案”，这句话可以理解为：“美”“国会”“通过对台售武法案”或者“美国”“会”“通过对台售武法案”。为了解决这一问题，许多有

效算法都被提出，并在实践中取得了很好的效果，具体实现将在未来学习中展现。

b. **词性标注（part-of-speech tagging）**：词性一般是指动词、名词、形容词等。标注的目的是表征词的一种隐藏状态，隐藏状态构成的转移就构成了状态转移序列（是不是看不懂？没事，跳过，之后会详细讲解）。如：我/r 爱/v 南京/ns 河海大学/ns。其中，/r 是代词，/v 是动词，/ns 是名词，s 代表地名。中文标注可以采用 **Python** 中的 **jieba、NLTK.....** 包，英文则可以使用 **spaCy、NLTK、jieba.....** 包。

c. **命名实体识别（NER，Named Entity Recognition）**：NER 是指从文本中识别具有特定类别的实体（通常是名词），例如：人名、地名、机构名、专有名词等。其过程是从是非结构化文本表达式中产生专有名词标注信息的命名实体表达式，目前 NER 有两个显著的问题，即识别和分类。例如，“奥巴马是美国总统”的“奥巴马”和“美国”都代表一个具体事物，因此都是命名实体。而“总统”不代表一个具体事物，因此不是命名实体。

d. **句法分析（syntax parsing）**：句法分析往往是一种基于规则的专家系统，最初的时候是利用语言学专家的知识来构建的。句法分析的目的主要是分析各个成分的依赖关系，算法的结果一般是**树**结构。句法分析可以解决传统词袋模型不考虑上下文的问题。比如，“小李是小杨的班长”和‘小杨是小李的班长”，这两句话，用词袋模型是完全相同的，但是句法分析可以分析出其中的主从关系，真正理清句子的关系。

e. **纠错（Correction）**：自动纠错在搜索技术中利用得很多。由于用户的输入出错的可能性比较大，出错的场景也比较多。所以，我们需要一个纠错系统。具体做法有很多，可以基于 N-Gram 进行纠错，数据结构上，字典树、有限状态机可以考虑。

f. **指代消解（anaphora resolution）**：它的作用简单来说就是用来表征前文出现过的人名、地名等等。如：我在南京学习，这座城市很漂亮。其中，“这座城市”指代的就是前面的南京，出于中文的习惯，而不会再次以“南京”出现。

g. **实体关系抽取（Entity relation extraction）**：实体关系抽取是自动识别非结构化文档中两个实体之间的关联关系，属于信息抽取领域的

基础知识之一。一般情况下，我们将关系定义为两个或多个实体间的某种联系，而实体关系抽取旨在自动发现实体间存在的某种语义关系。如：在这段话“乔布斯和沃兹尼亚克在 1976 年共同创立了苹果公司”中，我们可以判别出“乔布斯（人）”和“苹果（公司）”之间有一种创始人关系。

4 NLP 语料库

语料库一词在语言学上意指大量的文本，通常经过整理，具有既定格式与标记；或是说经科学取样和加工的大规模电子文本库，其中存放的是在语言的实际使用中真实出现过的语言材料。简单来说，语料库就是 NLP 的模型训练的数据集。

常用的语料库：中文维基百科、搜狗新闻语料库、IMDB 情感分析语料库、中科院自动化所的中英文新闻语料库等等.....

5 NLP 知识结构体系

作为一门综合学科，NLP 是研究人与机器之间用自然语言进行有效通信的理论和方法。这需要很多跨学科的知识，需要语言学、统计学、最优化理论、机器学习、深度学习以及自然语言处理相关理论模型知识做基础。作为一门杂学，NLP 可谓是包罗万象，体系化与特殊化并存，这里简单罗列其知识体系:

a. **句法语义分析**：针对目标句子，进行各种句法分析，如分词、词性标记、命名实体识别及链接、句法分析、语义角色识别和多义词消歧等。

b. **关键词抽取**：抽取目标文本中的主要信息，比如从一条新闻中抽取关键信息。主要是了解是谁、于何时、为何、对谁、做了何事、产生了有什么结果。涉及实体识别、时间抽取、因果关系抽取等多项关键技术。

c. **文本挖掘**：主要包含了对文本的聚类、分类、信息抽取、摘要、情感分析以及对挖掘的信息和知识的可视化、交互式的呈现界面。

d. **机器翻译**：将输入的源语言文本通过自动翻译转化为另一种语言的文本。根据输入数据类型的不同，可细分为文本翻译、语音翻译、手语翻译、图形翻译等。机器翻译从最早的基于规则到二十年前的基于统

计的方法，再到今天的基于深度学习 (编解码) 的方法，逐渐形成了一套比较严谨的方法体系。

e. **信息检索**：对大规模的文档进行索引。可简单对文档中的词汇，赋以不同的权重来建立索引，也可使用算法模型来建立更加深层的索引。查询时，首先对输入比进行分析，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。

f. **问答系统**：针对某个自然语言表达的问题，由问答系统给出一个精准的答案。需要对自然语言查询语句进行语义分析，包括实体链接、关系识别，形成逻辑表达式，然后到知识库中查找可能的候选答案并通过一个排序机制找出最佳的答案。

g. **对话系统**：系统通过多回合对话，跟用户进行聊天、回答、完成某项任务。主要涉及用户意图理解、通用聊天引擎、问答引擎、对话管理等技术。此外，为了体现上下文相关，要具备多轮对话能力。同时，为了体现个性化，对话系统还需安基于用户画像做个性化回复。

6 本章小节

本章介绍了 NLP 相关的些基础知识，主要面向 NLP 刚刚入门的读者（比如：我这样的菜鸟）。首先介绍了 NLP 的概念、应用场景和发展历程，在学习 NLP 技术之前，有必要了解这些宏观的内容；接着讲解了 NLP 的基本术语、知识结构，以及之后学习将会用到的语料库，告诉读者在学习 NLP 的最初，应该做好哪些技术储备；最后希望大家一起努力，打好 NLP 的基础。后续将介绍通过 Python 处理 NLP 中的一些关键库以及 NLP 日常处理中需要掌握的技术。冲!!!

本章节内容主要以《Python 自然语言处理实战：核心技术与算法》为主，其他参考材料见开头的 Reference I 与 II