# NLP简介

NLP（Natural Language Processing），即自然语言处理。是一门用机器去理解自然语言、处理与自然语言和知识相关的任务的技术。其方法经历了规则、统计和机器学习这几个阶段，目前在深度学习加持之下如火如荼地发展。

NLP的基本应用领域包括文本分类、机器翻译、摘要生成、对话系统和知识图谱等；其基本任务包括分词、句法分析、词性标注、语义角色标注、命名实体识别、关系抽取、事件抽取、指代消歧等。

NLP目前的主要研究范式是神经网络和预训练语言模型。

# NLP基础任务

语言是一种复杂的、冗余的，但鲁棒的、具有强大表征能力的符号系统，它利用词的排列组合来表达人的知识、情感、意图等等。语言的复杂和语料的稀疏导致NLP任务通常难以通过简单的端对端建模取得完美的结果。因此实际中要么通过设计特殊结构的模型，要么把任务拆解成pipline，来解决特定的NLP任务。对于pipline建模，我们会先针对语料做一些语言学相关的基础任务（包括语法方面和语义方面），并把结果作为预处理的结果或者特征工程的结果。基于这些结果，再对下游任务进行建模。

第一节 分词

中文NLP需要先分词，再做其他；英文NLP则不用，因为英文是天然分词的。中文分词的方法有很多，我们将在专门的笔记中介绍它们。

第二节 句法分析

句法分析包括句法结构分析和依存句法分析。前者是对句子按照句法构建一棵语法树，用来指明句子的成分和成分之间的关系；后者是解析出句子中各成分之间的依存关系。

第三节 词性标注（POS）

就是标注句子中每个词的词性。可以用HMM实现。

第四节 语义角色标注

语义角色标注是实现浅层语义分析的一种方式。在一个句子中，谓词是对主语的陈述或说明，指出“做什么”、“是什么”或“怎么样，代表了一个事件的核心，跟谓词搭配的名词称为论元。语义角色是指论元在动词所指事件中担任的角色。主要有：施事者（Agent）、受事者（Patient）、客体（Theme）、经验者（Experiencer）、受益者（Beneficiary）、工具（Instrument）、处所（Location）、目标（Goal）和来源（Source）等。

语义角色标注（Semantic Role Labeling，SRL）以句子的谓词为中心，不对句子所包含的语义信息进行深入分析，只分析句子中各成分与谓词之间的关系，即句子的谓词（Predicate）- 论元（Argument）结构，并用语义角色来描述这些结构关系，是许多自然语言理解任务（如信息抽取，篇章分析，深度问答等）的一个重要中间步骤。在研究中一般都假定谓词是给定的，所要做的就是找出给定谓词的各个论元和它们的语义角色。

<https://www.cnblogs.com/cmybky/p/11767723.html>

第五节 命名实体识别（NER）

识别句子中出现的所有命名实体。命名实体是指人名、地名、组织机构等实体。

第六节 关系抽取

关系可以用SPO结构的三元组来表示，即 (Subject, Predication, Object)。

<https://www.cnblogs.com/sandwichnlp/p/12020066.html>

<https://www.cnblogs.com/vpegasus/p/re.html>

<https://baijiahao.baidu.com/s?id=1650526419042842719&wfr=spider&for=pc>

第七节 事件抽取

事件抽取技术是从非结构化信息中抽取出用户感兴趣的事件，并以结构化呈现给用户。事件抽取任务可分解为4个子任务: 触发词识别、事件类型分类、论元识别和角色分类任务。其中，触发词识别和事件类型分类可合并成事件识别任务。事件识别判断句子中的每个单词归属的事件类型，是一个基于单词的多分类任务。论元识别和角色分类可合并成论元角色分类任务。角色分类任务则是一个基于词对的多分类任务，判断句子中任意一对触发词和实体之间的角色关系。

<https://blog.csdn.net/muumian123/article/details/81746583>

第八节 指代消歧

# 语义分析

1、语义表示：

（1）词表征和句表征：数据建模需要把数据进行数值化，然后才能输入到模型。NLP也是如此，它需要将语料进行数值化。语料的数值化分为词粒度的和句子粒度的，分别是词表征和句子表征。我们将在其他笔记中专门介绍它们。

# 第四章 应用领域

NLP的主要应用领域为：文本分类、机器翻译、摘要生成、对话系统、知识图谱等。

第一节 文本分类

文本分类主要有三个场景：情感分类、情绪分类、主题分类。

第二节 机器翻译

机器翻译是非常古老的一个NLP任务。它是典型的seq2seq任务。

第三节 摘要生成

摘要生成也属于seq2seq任务，它主要用于文章摘要生成、商品信息摘要生成等。

第四节 对话系统

从任务的特点来划分，对话系统可以分为聊天系统、任务型系统和问答型系统。从对话论数来分，可以分为单轮对话和多轮对话。

第五节 知识图谱

知识图谱是大量的实体之间关系的表示，实体的关系可以用SPO三元组表示。若要从大量无结构的数据中提取SPO信息，需要信息提取的技术，包括NER和关系抽取。

知识图谱的应用需要解决的问题有：知识抽取、知识表示、知识融合、知识推理。

# 模型架构

针对NLP的不同任务的特点，我们会使用不同的架构。

NLP任务中最常见的模块就是RNN/LSTM，因为它们是专门用于序列建模的。对于输入序列输出序列的问题，即seq2seq任务，常常用encoder-decoder结构建模，encoder和decoder的结构可以是RNN、CNN，也可以用self-Attention。seq2seq的建模还常常使用Attention机制。对于文本生成任务，在encoder-decoder模型的基础上，可以加入copy机制、coverage机制等，以改善模型的效果。

# 预处理、特征工程及文本数据增强

第一节 预处理

包括语料的清洗、分词、去停用词、整理成任务要求的格式等。

第二节 特征工程

NLP的特征工程有非常多的花样。其样本是一段文本，因此NLP的特征工程就是对一段文本提取其特征。特征工程的结果就是文本的特征，它会直接被输入到模型中。

首先，句子表征就是最重要的特征。比如词袋、tf-idf、主题模型、词向量的平均、或者用神经网络模型学出来的基于上下句关系的句向量。

其次，基于NLP基础任务得到的结果，如POS、依存句法分析等的结果。

然后，统计句子的一些简单的语言特征，比如句子长度、标点符号的个数、句子中首字母大写的词的个数等。

这些特征可以拼接起来，作为最终的特征向量，输入到模型中去。

第三节 文本数据增强

文本数据增强的主要方法有：近义词替换、回译等。

<https://github.com/425776024/nlpcda>