# 自然语言处理课程大纲

目录

[说明](#_Toc30461_WPSOffice_Level1) [3](#_Toc30461_WPSOffice_Level1)

[一、 自然语言处理相关概述](#_Toc2154_WPSOffice_Level1) [3](#_Toc2154_WPSOffice_Level1)

[1.1 自然语言处理概念](#_Toc2154_WPSOffice_Level2) [4](#_Toc2154_WPSOffice_Level2)

[1.2 自然语言处理研究内容与现状](#_Toc27560_WPSOffice_Level2) [4](#_Toc27560_WPSOffice_Level2)

[1.3 自然语言处理目前业界发展状况](#_Toc31184_WPSOffice_Level2) [4](#_Toc31184_WPSOffice_Level2)

[二、 自然语言处理与数学](#_Toc27560_WPSOffice_Level1) [4](#_Toc27560_WPSOffice_Level1)

[2.1 隐马尔科夫模型 (HMM)](#_Toc8223_WPSOffice_Level2) [4](#_Toc8223_WPSOffice_Level2)

[2.2 最大熵马尔科夫模型（MEMM）](#_Toc4497_WPSOffice_Level2) [4](#_Toc4497_WPSOffice_Level2)

[2.3 条件随机场 (CRF)](#_Toc19471_WPSOffice_Level2) [4](#_Toc19471_WPSOffice_Level2)

[三、自然语言处理与深度学习](#_Toc31184_WPSOffice_Level1) [4](#_Toc31184_WPSOffice_Level1)

[3.1 CNN](#_Toc31812_WPSOffice_Level2) [5](#_Toc31812_WPSOffice_Level2)

[3.2 RNN及其变种网络](#_Toc26630_WPSOffice_Level2) [5](#_Toc26630_WPSOffice_Level2)

[3.3 lstm + crf模型架构介绍](#_Toc7504_WPSOffice_Level2) [5](#_Toc7504_WPSOffice_Level2)

[四、自然语言处理基础技术说明](#_Toc8223_WPSOffice_Level1) [5](#_Toc8223_WPSOffice_Level1)

[4.1 语料库与语言知识库](#_Toc2737_WPSOffice_Level2) [6](#_Toc2737_WPSOffice_Level2)

[4.2 词法分析](#_Toc30214_WPSOffice_Level2) [6](#_Toc30214_WPSOffice_Level2)

[4.2.1 中文分词](#_Toc2154_WPSOffice_Level3) [6](#_Toc2154_WPSOffice_Level3)

[4.2.2 词性标注](#_Toc27560_WPSOffice_Level3) [6](#_Toc27560_WPSOffice_Level3)

[4.2.3 命名体识别](#_Toc31184_WPSOffice_Level3) [6](#_Toc31184_WPSOffice_Level3)

[4.2.4 同义词识别](#_Toc8223_WPSOffice_Level3) [6](#_Toc8223_WPSOffice_Level3)

[4.2.5 汉字拼音转换](#_Toc4497_WPSOffice_Level3) [6](#_Toc4497_WPSOffice_Level3)

[4.3 句法分析](#_Toc18445_WPSOffice_Level2) [6](#_Toc18445_WPSOffice_Level2)

[4.3.1 完全句法分析](#_Toc19471_WPSOffice_Level3) [6](#_Toc19471_WPSOffice_Level3)

[4.3.2 浅层句法分析](#_Toc31812_WPSOffice_Level3) [6](#_Toc31812_WPSOffice_Level3)

[4.3.3 依存句法分析](#_Toc26630_WPSOffice_Level3) [6](#_Toc26630_WPSOffice_Level3)

[4.3.4 歧义问题](#_Toc7504_WPSOffice_Level3) [6](#_Toc7504_WPSOffice_Level3)

[4.3.5 其他句法概念详解](#_Toc2737_WPSOffice_Level3) [6](#_Toc2737_WPSOffice_Level3)

[4.4 语义分析](#_Toc19094_WPSOffice_Level2) [6](#_Toc19094_WPSOffice_Level2)

[4.4.1 词义消歧](#_Toc30214_WPSOffice_Level3) [6](#_Toc30214_WPSOffice_Level3)

[4.4.2 共指消解](#_Toc18445_WPSOffice_Level3) [7](#_Toc18445_WPSOffice_Level3)

[4.4.3 语义角色标注](#_Toc19094_WPSOffice_Level3) [7](#_Toc19094_WPSOffice_Level3)

[4.4.4 语句边界消歧](#_Toc31501_WPSOffice_Level3) [7](#_Toc31501_WPSOffice_Level3)

[4.4.5 深层语义分析](#_Toc7076_WPSOffice_Level3) [7](#_Toc7076_WPSOffice_Level3)

[4.4.6 语义相似度计算](#_Toc21848_WPSOffice_Level3) [7](#_Toc21848_WPSOffice_Level3)

[4.4.7 语义依存分析](#_Toc389_WPSOffice_Level3) [7](#_Toc389_WPSOffice_Level3)

[4.4.8 依存句法分析和语义依存分析比较](#_Toc11791_WPSOffice_Level3) [7](#_Toc11791_WPSOffice_Level3)

[五、语言模型](#_Toc4497_WPSOffice_Level1) [7](#_Toc4497_WPSOffice_Level1)

[5.1 传统语言模型](#_Toc31501_WPSOffice_Level2) [7](#_Toc31501_WPSOffice_Level2)

[5.2 神经序列模型](#_Toc7076_WPSOffice_Level2) [7](#_Toc7076_WPSOffice_Level2)

[5.3 语言模型评估](#_Toc21848_WPSOffice_Level2) [7](#_Toc21848_WPSOffice_Level2)

[六、知识图谱](#_Toc19471_WPSOffice_Level1) [7](#_Toc19471_WPSOffice_Level1)

[6.1 知识图谱概述](#_Toc389_WPSOffice_Level2) [8](#_Toc389_WPSOffice_Level2)

[6.2 知识建模](#_Toc11791_WPSOffice_Level2) [8](#_Toc11791_WPSOffice_Level2)

[6.3 知识抽取](#_Toc7793_WPSOffice_Level2) [8](#_Toc7793_WPSOffice_Level2)

[6.4 知识挖掘](#_Toc4031_WPSOffice_Level2) [8](#_Toc4031_WPSOffice_Level2)

[6.5 知识融合](#_Toc18215_WPSOffice_Level2) [8](#_Toc18215_WPSOffice_Level2)

[6.6 知识推理](#_Toc18155_WPSOffice_Level2) [8](#_Toc18155_WPSOffice_Level2)

[6.7 知识存储](#_Toc15832_WPSOffice_Level2) [8](#_Toc15832_WPSOffice_Level2)

[6.7.1 图数据库](#_Toc7793_WPSOffice_Level3) [8](#_Toc7793_WPSOffice_Level3)

[6.8 语义搜索](#_Toc32208_WPSOffice_Level2) [8](#_Toc32208_WPSOffice_Level2)

[6.9 知识图谱应用](#_Toc31068_WPSOffice_Level2) [8](#_Toc31068_WPSOffice_Level2)

[七、项目](#_Toc31812_WPSOffice_Level1) [8](#_Toc31812_WPSOffice_Level1)

[7.1 多标签文本分类](#_Toc5964_WPSOffice_Level2) [8](#_Toc5964_WPSOffice_Level2)

[7.1.1 项目说明](#_Toc4031_WPSOffice_Level3) [8](#_Toc4031_WPSOffice_Level3)

[7.1.2 实现思路](#_Toc18215_WPSOffice_Level3) [9](#_Toc18215_WPSOffice_Level3)

[7.2 写诗机器人](#_Toc753_WPSOffice_Level2) [10](#_Toc753_WPSOffice_Level2)

[7.2.1 项目说明](#_Toc18155_WPSOffice_Level3) [10](#_Toc18155_WPSOffice_Level3)

[7.2.2 实现思路](#_Toc15832_WPSOffice_Level3) [10](#_Toc15832_WPSOffice_Level3)

[7.2.4 总结](#_Toc32208_WPSOffice_Level3) [10](#_Toc32208_WPSOffice_Level3)

[7.3 信息抽取](#_Toc18967_WPSOffice_Level2) [11](#_Toc18967_WPSOffice_Level2)

[7.3.1 项目描述](#_Toc31068_WPSOffice_Level3) [11](#_Toc31068_WPSOffice_Level3)

[7.3.2 子任务](#_Toc5964_WPSOffice_Level3) [11](#_Toc5964_WPSOffice_Level3)

[7.3.3 技术架构](#_Toc753_WPSOffice_Level3) [11](#_Toc753_WPSOffice_Level3)

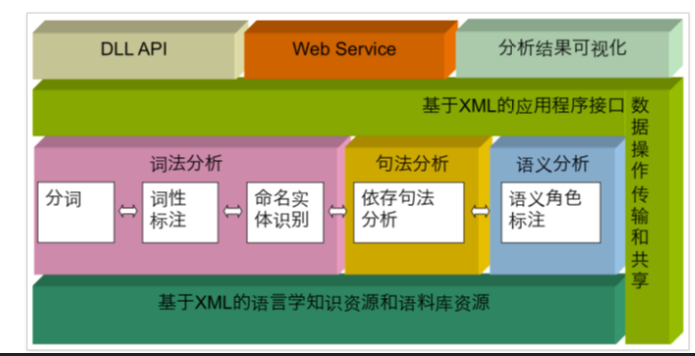
# 说明

自然语言处理（Nature Language Process, NLP）是大数据(数据挖掘)下的细分领域，目前随着企业数据量的增大、数据维度的增加以及企业业务的增多，人们越来越重视非结构化数据的处理，（此处我们主要针对文本数据）。由于汉语言属于意形语言，博大精深，目前自然语言处理中还有很多任务的准确性还有待提高，比如语义分析等，很多学者还在不断钻研，每年新发的论文也很多。这也是目前很多高校实验室的重点研究方向。

本课程讲从自然语言处理的起源、相关概念以及针对汉语的特殊情况(如英语不需要分词，中文在某些情景中有分词需要)和目前自然语言技术在工业界的应用展开说明。然后再结合这些应用进行技术的细分讲解和工程的具体实现。

以下章节目录仅仅为本课程的知识点罗列，并且只做了简要说明，在此文档中没有做出详细说明。

课程整体结构如下图：



1. 最底层介绍语料库和语言知识库。
2. NLP中的词法分析、句法分析、语义分析。
3. 将上述的模型输出结果以Restful风格以接口化形式封装。
4. 本课程主要编程语言采用python，深度学习框架采用tensorflow。

### 自然语言处理相关概述

本章主要介绍自然语言处理的发展过程以及相关概念以及目前NLP技术在中文领域的发展状况、目前所在的局限性进行说明。让同学们对NLP有一个比较宏观且较为深入的认识。

### 1.1 自然语言处理概念

### 1.2 自然语言处理发展历程与现状

### 1.3 自然语言处理目前业界发展状况

## 自然语言处理与数学

本章介绍自然语言处理技术中运用到的数学知识，由于NLP技术对数学基本功要求较高，且数学是一门及其繁杂且深奥的学科。所以在这里并不是讲解所有运用到的数学知识，而是着重讲解NLP中运用比较多且比较难理解的概率图模型([Probabilistic Graphical Models](http://www.baidu.com/link?url=y26LTgmUjfPxUSDt8USg40CqI9EaCrjl00fPEN1e0x54_c5XB5j-ctAYCy-oi1idqlgPkMS43Ae4qJPQ3mU6zh7bQY0Fwxwi4Js3OgiWjRA4K-6jWr_A60yr0lhyKPDZdqyaT64_ZdQ5xvV8SA5YM_" \t "https://www.baidu.com/_blank))。

主要包括隐马尔科夫模型、高斯隐马尔科夫模型、条件随机场这三种模型的原理、公式推导及性能的优劣势比较。

### 2.1 隐马尔科夫模型 (HMM)

### 2.2 最大熵马尔科夫模型（MEMM）

### 2.3 条件随机场 (CRF)

## 三、自然语言处理与深度学习

本章说明目前最火热的深度学习在NLP领域的应用。包括卷积神经网络（CNN）、循环神经网络（RNN）及其变种网络（LSTM、bi-LSTM）等。主要讲解神经网络的结构、公式推导、各个模型层的原理。同时，会在后续的内容中穿插说明各种模型细节（比如Attention机制）以及参数优化方法。

### 3.1 CNN

### 3.2 RNN及其变种网络

### 3.3 lstm + crf模型架构介绍

## 四、自然语言处理基础技术说明

本章主要自然语言处理的技术细节。包括语言知识库、语料库的制作、开源语料库的获取。然后从词法分析、句法分析、语义分析三个大的方面详细介绍NLP的各个知识点。这一章是本课程最核心的部分。

### 4.1 语料库与语言知识库

### 4.2 词法分析

#### 4.2.1 中文分词

#### 4.2.2 词性标注

#### 4.2.3 命名体识别

#### 4.2.4 同义词识别

#### 4.2.5 汉字拼音转换

### 4.3 句法分析

#### 4.3.1 完全句法分析

#### 4.3.2 浅层句法分析

#### 4.3.3 依存句法分析

#### 4.3.4 歧义问题

#### 4.3.5 其他句法概念详解

### 4.4 语义分析

#### 4.4.1 词义消歧

#### 4.4.2 共指消解

#### 4.4.3 语义角色标注

#### 4.4.4 语句边界消歧

#### 4.4.5 深层语义分析

#### 4.4.6 语义相似度计算

#### 4.4.7 语义依存分析

#### 4.4.8 依存句法分析和语义依存分析比较

## 五、语言模型

本章主要介绍语言模型。包括传统的n-gram语言模型、深度学习语言模型二者的对比关系包括word2vec、doc2vec、cw2vec、glave等等重要知识点。

## 5.1 传统语言模型

## 5.2 神经序列模型

## 5.3 语言模型评估

## 六、知识图谱

计算机的诞生给我们的生活带来了很多方便，但是我们都知道，计算机的计算原理是二进制的，只认识0和1，所以计算机的发展一直面临这样的困境--无法获取文本的语义信息。尽管近年来AI发展迅速，但是计算机的“智能化”离我们的要求还差距甚远。如我们输入“罗纳尔多”(在没有互联网的情况下)，计算机并不能理解我输入的内容是什么，这时想要使机器理解我们输入的真实意义，就需要我们对可描述的实体(Entity)进行建模，填充一系列属性，拓展它和其他实体的联系，即构建机器的先验知识。此时，知识图谱(Knowledge Graph)应运而生。

本章将从数据获取（爬虫）、知识建模、知识抽取、知识融合、知识推理知识存储（主要介绍图数据库Neo4J）等等方面详细介绍知识图谱的构建和应用。

### 6.1 知识图谱概述

### 6.2 知识建模

### 6.3 知识抽取

### 6.4 知识挖掘

### 6.5 知识融合

### 6.6 知识推理

### 6.7 知识存储

#### 6.7.1 图数据库

### 6.8 语义搜索

### 6.9 知识图谱应用

## 七、项目

### 7.1 多标签文本分类

#### 7.1.1 项目说明

分类问题是机器学习中所面临的比较普遍的问题，在我们的生活中很多问题也可以归结为分类问题。文本分类也就是根据文本的内容将其分为合适的类别，比如读新闻的时候可以看到其明显的分类为社会新闻、财经新闻等等。他是NLP领域一个很重要的问题。文本分类主要应用于信息检索、机器翻译、自动文摘、信息过滤、邮件分类等下游任务中。所以说文本分类的准确高低也直接影响到后续工作的准确性。

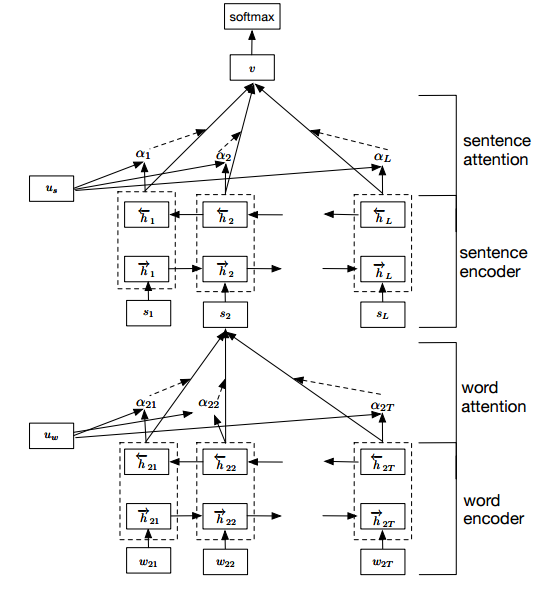
#### 7.1.2 实现思路

分类器基本上都是统计分类方法，基本上大部分机器学习方法都在文本分类中有所应用，比如朴素贝叶斯算法、KNN、SVM、最大熵模型等等，而且SVM的分类准确率也已经达到了95%左右。但是传统机器学习算法不是我们要说明的重点，所以在此忽略。我们这个项目采用深度学习的方法对文本进行分类。

应用深度学习解决大规模文本分类的问题主要是解决文本表示，然后再利用CNN/RNN等网络结构自动获取特征表达能力，去掉繁杂的人工特征工程，采用End-to-End的思想解决问题。

主要实现思路：

1. 数据预处理。本项目采用的是知乎公开的300W条带有标题、问题的描述和问题的话题的数据（经过了脱敏处理）。首先我们要观察数据的特征，将其转化为numpy格式，并且对于不同长度的问题文本，太长的可以适当进行分割截取，太短的进行补空格等。其次对数据进行数据增强，这里采用shuffle和drop两种方法。（重点会讲解一下数据增强的原理和操作方法）。
2. 文本的分布式表示。其基本思想是将每个词表达成n维稠密、连续的实数向量。与之相对的是one-hot表示。
3. 构建分类模型。这里我们主要采用 TextRNN + Attention结构。RNN在序列模型中有着很好的表现，它能够更好的表达上下文信息和记忆长期信息（这里会重点讲解一下Memory Network）。针对在文本分类中，双向RNN从某种意义上可以理解为捕获变长的双向n-gram信息。而attention机制（重点讲Attention机制在NLP中的应用）。在我们的文本数据中，下文中某个词的意义是重点受上文中某个词的意义的影响。换句话说也就是 不同的词所占的权重不同。此时加入Attention机制最大的好处就是能够直观的体现出各个句子和词对分类类别的重要性。基本模型图如下所示：



1. 代码实现，参数调整。

### 7.2 写诗机器人

#### 7.2.1 项目说明

自然语言处理的有意重大功能是文本生成，目前工业界应用最多的就是应用于对话系统。如天猫精灵、siri和笔者目前正在开发的Chatbot\_CN客服系统。本项目讲解基于RNN 和 seq2seq模型的文本生成模型--生成唐诗。

#### 7.2.2 实现思路

1、数据处理。本项目采用4万首唐诗作为训练数据集。首先我们需要对数据进行处理，进行字符-数字映射和数字-字符映射，然后对原文进行encoder。

2、构建模型。模型主要分为输入层、LSTM层、输出层、loss（损失函数）、optimizer（优化器）等部分组成。

3、模型训练。模型构建好以后就可以设置参数，开始进行训练，直到loss收敛。

4、文本生成。模型训练好以后，我们就可以利用保存下来的参数进行文本生成，当我们输入一首诗的第一个字，他便会自动生成一首诗。输出如下：



在这里同时要讲一个而非常重要的序列模型：seq2seq。他是对话系统中应用比较广泛。

#### 7.2.4 总结

这个项目为一个比较简单的文本生成项目，主要目的为熟悉模型的结构和nlp项目的开发流程。后续可以讲解比较复杂的对话系统。

### 7.3 信息抽取

#### 7.3.1 项目描述

信息抽取是自然语言处理中一个非常重要模块，目前在工业界已经有很广泛的应用。

信息抽取是指从文本数据（非结构化）中抽取出特定的事实信息，比如从新闻中抽取出机构名称（ORG）、人名（PER）、地址（LOC）、时间（TIM）、事件（Event）、人物关系（Relation）等等。这些被抽取出来的信息通常以结构化的形式直接存入数据库（或者es等），可以提供用户查询及进一步分析使用，为之后构建知识库、只能问答等提供数据支撑。[这是一个将非结构化数据转化为结构化数据的过程。]

信息抽取对于构建大规模的知识库有着重要的意义，但是目前由于自然语言本身的复杂性、歧义性等特征，而且信息抽取目标只是规模巨大、复杂多样等问题使得信息抽取技术还不是很完善，但是在目前深度学习和只是图谱及语言学的深入研究下，信息抽取技术将会进一步完善和更广阔的前景。

#### 7.3.2 子任务

信息抽取项目可以细分为以下三个子任务：

1、实体抽取：包括时间、地点、人名、机构名等。

2、关系抽取：发现文本中实体的关系，并用三元组表示出来。

3、事件抽取：抽取出文本中主要发生的事件信息，包括事件主体和事件触发词抽取，并且与上面的实体抽取和关系抽取相结合。

4、实体消歧：此任务主要消除抽取出的实体之间的歧义问题。如在真实场景中出现“回民街56号”，第一眼觉得这是一个地址，但是在曾经遇到的场景中，这是一家饭店的名字。

#### 7.3.3 技术架构

信息抽取涉及的主要技术包括实体识别、句法分析、篇章分析及知识库等。目前业内使用的主流技术是深度学习 + 统计规则模式。对于部分规律性比较强的数据则可以利用规则的方法来提取，无规律性的数据则用深度学习模型的方法来提取。

主要实现思路：

1. 训练语料的准备，由于针对不同行业需要准备不同的训练语料。所以目前企业中都会准备自己特定的语料库。我们首先需要对语料进行标注，利用IOB标注策略。（标签矩阵为B-PER，I-PER等等）。语料的好坏直接影响到模型输出的结果，所以准备一份优质的训练集是非常有必要的。数据标注是一个重复枯燥的工作。
2. 构建模型。此处的主要模型结构如下图，比模型结构为bi-LSTM + CRF。首先对输入文本进行分布式向量表表示，然后由双向长短期神经网络进行编码和解码，最后由条件随机场的线性链条件随机场解码输出。在事件抽取中，模型会更加复杂，加入了词性标注、句法分析、语义分析、槽填充等相关内容，并且利用到了Bi-TreeLSTM等树形结构。
3. 模型测评。改效果的测评还是依据准确率、召回率和F1-score来进行评测。
4. 模型优化。针对模型输出结果的不准确的数据进行分析，适当进行调参及语料库的扩充等等。

