



# Natural Language Processing

## 第四周 自然语言处理介绍

庞彦

yanpang@gzhu.edu.cn

# Overview



## CONTENTS

01

自然语言处理介绍

02

系统搭建



01

# Natural Language Processing

## 自然语言处理介绍

Spring 2023

# Why is NLP important?



Why is NLP important? 为什么自然语言处理这么重要?

66

"Language understanding is the crown jewel in the field of artificial intelligence"  
语言理解是人工智能领域皇冠上的明珠

Bill Gates

# Why is NLP important?



结构化数据

Excel 数据库



非结构化数据

文本 图片 视频

# Language



不同物种有自己的沟通方式



0100011101010



你好  
hello



汪汪汪

NLP就是人类和机器之间沟通的桥梁



# Why is "natural language" processing?



66

Natural language is the expression commonly used in daily life, which is what people usually mean by “speaking.” 讲人话

Natural language: I have a little camel on my back. 我驼背。  
(unnatural language: my back is curved) 背部呈弯曲状

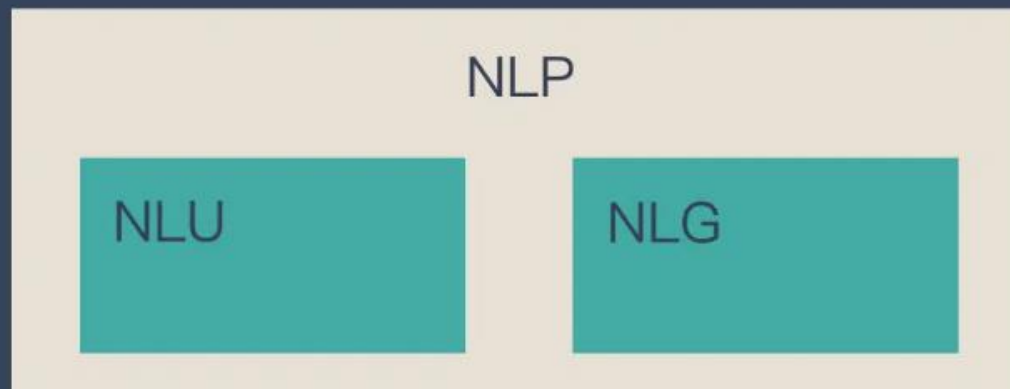
Natural language: The baby's agent talk to baby's baby.  
(a lot of this kind on Weibo) 微博上这种段子一大把



# Core Task



NLP 的2个核心任务



Natural Language Understanding – NLU 自然语言理解  
Natural language generation – NLG 自然语言生成

# Natural Language Understanding



自然语言理解就是希望机器像人一样，具备正常人的语言理解能力

# Challenge of NLU



## 自然语言理解的5个难点：

- |                                  |         |
|----------------------------------|---------|
| 1. Language diversity            | 语言的多样性  |
| 2. Language ambiguity            | 语言的歧义性  |
| 3. Language robustness           | 语言的鲁棒性  |
| 4. Language knowledge dependence | 语言的知识依赖 |
| 5. Language context              | 语言的上下文  |

# Challenge of NLU



## 基于规则的意图判断

机器通过「订机票」这个关键词来识别意图

我想订机票



我要出差



帮我看看去北京的航班



# Application of NLU



## 基于 NLU 的应用



机器翻译



机器客服



智能音箱

# Natural Language Generation



NLG – 将非语言格式的数据转换成人类可



Language

文字语言

# Natural Language Generation



自然语言生成 – NLG 有2种方式:

1.text – to – text: 文本到语言的生成

2.data – to – text : 数据到语言的生成

The diagram illustrates the process of data-to-text generation. On the left, a table lists financial data for various companies. The row for Apple, Inc. is highlighted in blue, and a blue arrow points from this row to a news article on the right. The news article, titled "Apple tops Wall Street 1Q forecasts", contains text generated from the data in the highlighted row of the table.

1	Company	Q1 Net Income	Earnings Per Share	Total Revenue
2	Nike Inc.	1,200,000,000	.013424	8,400,000,000
3	Apple, Inc.	18,020,000,000	.030643	74654284021
4	Amazon.com	513,000,000	.010723	29,130,000,000
5	AT&T	3,800,000,000	.006134	40,530,000,000
6	PepsiCo Inc.	2,010,000,000	.013825	15,400,000,000
7	Exxon Mobil	1,810,000,000	.004345	48,710,000,000
8	Microsoft Co	4,600,000,000	.005724	20,400,000,000
9	Facebook Inc	2,229,000,000	.007732	5,380,000,000

Apple tops Wall Street 1Q forecasts

CUPERTINO, Calif. (AP) - Apple, Inc. (AAPL) on Tuesday reported fiscal first-quarter net income of \$18.02 billion. The Cupertino, California-based company said it had profits of \$3.06 per share. The results surpassed Wall Street expectations.

The average estimate of analysts surveyed by Zacks Investment Research was for earnings of \$2.60 per share. The maker of iPhones, iPads and other products posted revenue of \$74.6 billion in the period, also exceeding Street forecasts. Analysts expected \$67.38 billion, according to Zacks. For the current quarter ending in March, Apple said it expects revenue in the range of \$52 billion to \$55 billion. Analysts surveyed by

# Step of NLG



## NLG 的6个步骤





# Application of NLG



自动写新闻



聊天机器人



BI报告生成

# Summary



## NLP 的5个难点

1

没有规律

2

自由组合

3

开放集合

4

知识依赖

5

上下文

# Traditional NLP



## 传统机器学习的 NLP 流程



# Deep Learning NLP



## 深度学习的 NLP 流程



# English NLP corpus preprocessing



## 英文语料预处理的6个核心步骤

- ① 分词
- ② 词干提取
- ③ 词形还原
- ④ 词性标注
- ⑤ 命名实体识别
- ⑥ 分块

# Chinese NLP corpus preprocessing



## 中文语料预处理的 4 个核心步骤

① 分词

② 词性标注

③ 命名实体识别

④ 去除停用词

# Chinese vs English



## 中英文分词的 3 大区别



分词方式不同, 「中文」更难



「英文」单词有多重形态



「中文」需要考虑分词粒度

# Chinese vs English



## 中文分词的 3 大难点



没有统一的标准



歧义如何切分?



新词如何识别?





# 02 | Install packages 系统搭建

Spring 2023

- Install Ubuntu 20.04 **alongside** Windows <https://ubuntu.com/>;  
安装双系统

Or

- Install Ubuntu 20.04 inside Windows (started from the APP store)  
Windows系统里面直接安装Ubuntu

# Terminal



- Type Ctrl-Alt-T to open your terminal. 打开终端

- Ref. <https://www.linkedin.com/pulse/install-keras-based-tensorflow-gpu-caffe-gpu-ubuntu-1804-yan-pang/>

Select Target Platform

Click on the green buttons that describe your target platform. Only supported platforms will be shown. By downloading and using the software, you agree to fully comply with the terms and conditions of the [CUDA EULA](#).

Operating System

LinuxWindows

Architecture

x86\_64ppc64learm64-sbsa

Distribution

CentOSDebianFedoraOpenSUSERHELSLESUbuntuWSL-Ubuntu

Version

16.0418.0420.04

Installer Type

deb (local)deb (network)runfile (local)

# Conda



- Install Anaconda or Mini Conda under your Ubuntu System  
安装Anaconda或者Mini Conda
- Ref. <https://www.anaconda.com/>



# Create a new environment



- Create a new environment under the installed Anaconda  
在conda下面创建环境
- Ref. <https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html>



# Python



- Install Python 3.9 or 3.10 under created environment.



# PyTorch



- Use commands, pip and conda, to install Pytorch (version depends on your Graphic Card), and other tools under the same environment.  
利用pip或conda命令安装PyTorch（注意选择好版本）
- Ref. <https://pytorch.org/get-started/locally/>

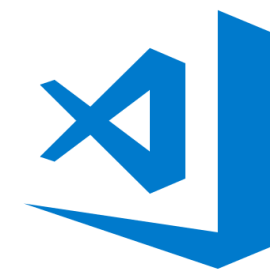




# VS Code



- Install VS Code under the Ubuntu System. You will use this IDE to build all projects in the future. 安装VS Code编译器（建议）
- Ref. <https://code.visualstudio.com/download>



Visual Studio Code

# Q&A



Spring 2023