《自然语言处理实战》

——段方

大数据（分析） 总设计师

教授　北京大学博士后

1 自然语言处理概述

1.1 什么是自然语言处理（NLP）

1.2 NLP 的发展历史

1.2.1 1956 年以前的萌芽期

1.2.2 1957-1970 年的快速发展期

1.2.3 1971-1993 年的低谷期

1.2.4 1994 年至今的复苏其

1.3 NLP 发展的原因

1.3.1 互联网提供了大量的语料库

1.3.2 深度学习算法提升了精度

1.3.3 场景更加丰富

1.4 NLP 的价值

1.4.1 语言是信息的载体

1.4.2 构建人机自然交互接口

1.4.3 语言翻译、信息检索等价值凸显

1.8.1 基本概念

1.8.2 搜索引擎

1.8.3 【例】谷歌搜索引擎算法

1.9 自动问答

1.9.1 基于知识图谱的问答系统

1.9.2 BERT 算法

1.10 NLP 产品举例

1.10.1 Google translate

1.10.2 Bing translate

1.10.3 语音输入法

1.10.4 Siri 问答

1.11 【例】 **NLP** 在（电信） 客服中的应用

1.11.1 部分替代人工

1.11.2 解决部分业务问题

1.11.3 中国移动的例子

1.12 【例】微软小冰的问答原理

2 自然语言处理的基本任务

2.1 语言分析

2.2 词法分析

2.3 句法分析

2.4 语义分析

2.5 语言生成

2.6 多语言处理（机器翻译、跨语言检索）

2.7 NLP 在各个行业的应用

2.7.1 教育

2.7.2 医疗

2.7.3 金融

2.7.4 法律等

2.8 【例】分词工具——结巴、 NLPIR 、IKAnalyzer（开源）

2.9 【例】附件-IBM 沃森介绍

3 自然语言处理的基本策略和实现方法

4.4.1 词性标注方法

4.5 汉语分词

4.5.1 切分歧义及歧义字段的种类

4.6 分词方法

4.6.1 基于词库

4.7 句法分析

4.7.1 基于上下文无关语法（**CFG**） 的表示

4.7.2 基于 **CFG** 算法及过程

4.7.3 搜索策略

4.7.4 自顶向下句法分析

4.7.5 自底向上句法分析

chart parsing

4.7.6 句法分析与逻辑程序设计

4.7.7 传统 **CFG** 在描述自然语言的问题

4.7.8 基于特征的扩展 **CFG**

合一文法

合一运算

chart parsing 举例

4.8 语义分析

4.8.1 词汇语义

4.8.2 语义类

4.8.3 词义间关系

4.9 句义分析

4.9.1 句义表示与语义组合

4.9.2 论旨角色与格语法

4.9.3 格语法

4.9.4 基于格语法的语义分析

4.10 【例】附件**-Siri** 的介绍

5 机器翻译

5.1 机器翻译的历史

5.2 机器翻译的基本策略

6.1 深度学习算法简介

6.2 词向量模型

6.2.1 原理

6.2.2 Word2vec

6.2.3 ELMo

6.2.4 OpenAI GPT

6.3 BERT 词向量模型

6.4 信息抽取

6.4.1 实体识别与抽取

隐马尔可夫模型 **HMM**

最大熵马尔可夫模型

MEMM 条件随机场算法

CRF

6.4.2 开放式实体抽取

6.4.3 命名实体消歧

6.4.4 关系抽取

传统方法

基于特征向量

基于核函数

基于神经网络

6.5 情感分析

6.5.1 情感分析的层次

6.5.2 句子级

6.5.3 词语级

6.5.4 情感信息抽取

6.5.5 情感分析的方法

6.6 语义分析

6.6.1 词汇级语义分析

6.6.2 词义消歧

基于规则

基于词典

6.6.3 词汇级语义分析

有监督

无监督

6.6.4 句子级语义分析

句义分析

句子语义相似度分析

6.7 【例】附件-NLP 的调用工具（百度等）

7 自然语言处理的未来发展

7.1 下一代信息检索

7.1.1 当前搜索引擎的问题

7.1.2 垂直搜索

7.1.3 智能搜索

7.1.4 个性化搜索

7.1.5 跨语言信息检索

7.1.6 多媒体信息检索

7.2 物联网与 NLP

7.2.1 5G 开启物联网