第四章 自然语言处理 复习

目录

汉语自然语言处理

- 核心技术：

  中文词汇自动切分（中文分词技术）

◆ 中文分词技术

两大难题：切分歧义，新词(未登陆词)识别

切分歧义

1. 交集型歧义
2. 组合型歧义
3. 混合型歧义

新词识别包括：

1. 数字识别
2. 命名实体识别
3. 形式词、离合词识别

汉语自然语言处理

- 分词流派

◆ 1）机械分词法（基于词典）：简单实用，但严重依赖于词典，分词效果得不到保障；

匹配法

◆ 2）基于语法和规则分词法：尚无明确标准能很好分词，还处在试验阶段；

◆ 3）基于统计的分词法：

最大概率法

标注法

机器学习算法

支持向量机（SVM）
最大熵（Maximum Entropy）
隐马模型（HMM）
最大熵隐马模型（MEMM）
条件随机场（CRFs）

词表征与词嵌入

- TF-IDF方法

  字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

- TF（Term Frequency，词频），指一篇文章中关键词出现在所有文档的频率

$$TF = N / M$$

N： 在文档中出现的频率
M： 文档中所有词汇出现的频率总和

- IDF（inverse document frequency，逆文本频率），用于衡量关键词权重

$$IDF = \log(\frac{D}{D_w})$$

D： 语库中的文档总数
Dw： 包含词w的文档总数

- TF-IDF的公式
- TF-IDF值越大说明这个词越重要，也可以说这个词是关键词。

$$TF - IDF(w) = TF(w) * IDF(w)$$

缺点： 依旧损失了词语之间的共现关系！认为一个词出现的可能性与其他词出现的可能性无关，词语的出现是相互独立的。

词表征与词嵌入

- 词嵌入（word embeddings）
- 自然语言处理（NLP）中语言模型与表征学习技术的统称
- 把一个维数为所有词的数量的高维空间嵌入到一个低的连续向量空间中，每个单词或词组被映射为实数域上的**向量**

- 不考虑语序：

  词袋模型（bag of words，BOW）

  TF-IDF方法
- 考虑语序

  词嵌入模型：基于CBOW和Skip-Gram算法的神经网络模型。