bobby · 更新于 2022-11-16

上一节 10. Elasticsearc... 12. ik分词器安... 下一节

文本分词

单词是语言中重要的基本元素。一个单词可以代表一个信息单元，有着指代名称、功能、动作、性质等作用。在语言的进化史中，不断有新的单词涌现，也有许多单词随着时代的变迁而边缘化直至消失。根据统计，《汉语词典》中包含的汉语单词数目在37万左右，《牛津英语词典》中的词汇约有17万。

理解单词对于分析语言结构和语义具有重要的作用。因此，在机器阅读理解算法中，模型通常需要首先对语句和文本进行单词分拆和解析。

分词（tokenization）的任务是将文本以单词为基本单元进行划分。由于许多词语存在词型的重叠，以及组合词的运用，解决歧义性是分词任务中的一个挑战。不同的分拆方式可能表示完全不同的语义。如在以下例子中，两种分拆方式代表的语义都有可能：

代码块

```
1  南京市|长江|大桥
2  南京|市长|江大桥
```

分词的意义 - nlp

1.将复杂问题转化为数学问题

在 机器学习的文章 中讲过，机器学习之所以看上去可以解决很多复杂的问题，是因为它把这些问题都转化为了数学问题。

而 NLP 也是相同的思路，文本都是一些「非结构化数据」，我们需要先将这些数据转化为「结构化数据」，结构化数据就可以转化为数学问题了，而分词就是转化的第一步。

2.词是一个比较合适的粒度

词是表达完整含义的最小单位。

字的粒度太小，无法表达完整含义，比如”鼠“可以是”老鼠“，也可以是”鼠标“。

而句子的粒度太大，承载的信息量多，很难复用。比如”传统方法要分词，一个重要原因是传统方法对远距离依赖的建模能力较弱。”

**

中英文分词的3个典型区别

区别1：分词方式不同，中文更难

英文有天然的空格作为分隔符，但是中文没有。所以如何切分是一个难点，再加上中文里一词多意的情况非常多，导致很容易出现歧义。下文中难点部分会详细说明。

区别2：英文单词有多种形态

英文单词存在丰富的变形变换。为了应对这些复杂的变换，英文NLP相比中文存在一些独特的处理步骤，我们称为词形还原（Lemmatization）和词干提取（Stemming）。中文则不需要
词性还原：does，done，doing，did 需要通过词性还原恢复成 do。
词干提取：cities，children，teeth 这些词，需要转换为 city，child，tooth”这些基本形态

区别3：中文分词需要考虑粒度问题

例如「中国科学技术大学」就有很多种分法：

- 中国科学技术大学
- 中国 \ 科学技术 \ 大学
- 中国 \ 科学 \ 技术 \ 大学

粒度越大，表达的意思就越准确，但是也会导致召回比较少。所以中文需要不同的场景和要求选择不同的粒度。这个在英文中是没有的。

中文分词的3大难点

难点 1：没有统一的标准

目前中文分词没有统一的标准，也没有公认的规范。不同的公司和组织各有各的方法和规则。

难点 2：歧义词如何切分

例如「兵乓球拍卖完了」就有2种分词方式表达了2种不同的含义：

- 乒乓球 \ 拍卖 \ 完了
- 乒乓 \ 球拍 \ 卖 \ 完了

难点 3：新词的识别

信息爆炸的时代，三天两头就会冒出来一堆新词，如何快速的识别出这些新词是一大难点。比如当年「蓝瘦香菇」大火，就需要快速识别。

3种典型的分词方法

分词的方法大致分为 3 类：

1. 基于词典匹配
2. 基于统计
3. 基于深度学习

给予词典匹配的分词方式

优点：速度快、成本低

缺点：适应性不强，不同领域效果差异大

基本思想是基于词典匹配，将待分词的中文文本根据一定规则切分和调整，然后跟词典中的词语进行匹配，匹配成功则按照词典的词分词，匹配失败通过调整或者重新选择，如此反复循环即可。代表方法有基于正向最大匹配和基于逆向最大匹配及双向匹配法。

基于统计的分词方法

优点：适应性较强

缺点：成本较高，速度较慢

这类目前常用的是算法是 **HMM、CRF、SVM、深度学习 **等算法，比如stanford、Hanlp分词工具是基于CRF算法。以CRF为例，基本思路是对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有良好的效果。

基于深度学习

优点：准确率高、适应性强

缺点：成本高，速度慢

例如有人员尝试使用双向LSTM+CRF实现分词器，其本质上是序列标注，所以有通用性，命名实体识别等都可以使用该模型，据报道其分词器字符准确率可高达97.5%。

常见的分词器都是使用机器学习算法和词典相结合，一方面能够提高分词准确率，另一方面能够改善领域适应性。

中文分词工具

下面排名根据 GitHub 上的 star 数排名：

1. jieba
2. Hanlp
3. IK
4. Stanford 分词
5. ansj 分词器
6. 哈工大 LTP
7. KCWS分词器
8. 清华大学THULAC
9. ICTCLAS

英文分词工具

1. Keras
2. Spacy
3. Gensim
4. NLTK

总结

分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。

分词的原因：

1. 将复杂问题转化为数学问题
2. 词是一个比较合适的粒度
3. 深度学习时代，部分任务中也可以「分字」

中英文分词的3个典型区别：

2. 英文单词有多种形态，需要词性还原和词干提取
3. 中文分词需要考虑粒度问题

中文分词的3大难点

1. 没有统一的标准
2. 歧义词如何切分
3. 新词的识别

3个典型的分词方式：

1. 基于词典匹配
2. 基于统计
3. 基于深度学习

10. Elasticsearch Analyze 上一节 下一节 12. ik分词器安装和配置

我要提出意见反馈