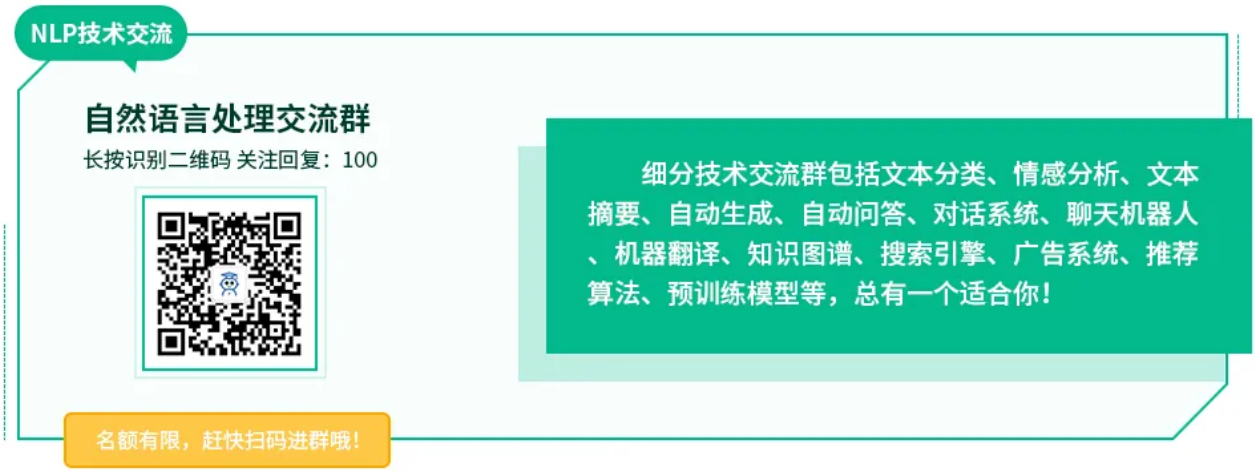
NLPer 如何做关键词抽取

AINLP 1周前

以下文章来源于NLP从入门到放弃 ， 作者DASOU

NLP从入门到放弃

积累一些平时的工作经验和思考，主要是关于NLP，搜索和推荐，只写干货！



关键词的提取，也可以称之为文本标签提取。

比如说，”今天这顿烧烤是真不错啊“，在这句话中，”烧烤“这个词就可以被认为是一个关键词，或者说这个句子的一个标签。

这个标签在一定程度上能够表现出这个句子的含义，比如这个”烧烤“，如果用在文本分类任务中，可以隐含带有”美食“这个类别的信息。

这些标签有些时候也可以用在推荐系统的召回，比如直接按照”烧烤“这个标签做一路召回。

对于关键词的提取一般来说分为抽取式和生成式。其实类比到摘要，其实也是分为抽取式和生成式。

生成式有一个缺点就是有些结果不可控，这其实还挺要命的。

对于抽取式，就是从现有的数据中拿出来词组。最差的结果也就是拿出的单词并不重要，不是我们想要的。

我们的重点是在抽取式提取关键词。

关键词的提取可以分为两个步骤：召回+排序

1.召回

召回就是得到文本中的候选关键词，也就是得到这个句子中有可能是关键词的词汇。

这一步，可以做的方法有很多，比如

1. 我们有积累的关键词词库，在这里直接匹配出来。
2. 一些符合的词性的候选词，比如我挑选出名词作为候选词
3. 还可以基于一些统计特征提出候选词，比如TF-IDF（有些时候统计特征也会用在排序中作为特征）
4. 基于一些规则，比如一个句子出现了人名地名，书名号中词，这些很有可能就是关键词

召回其实是一个很重要的部分，在这一步骤，尽可能的召回有用的词汇。我自己的标准是宁可多不能少。如果多了，无非就是增加了资源消耗，但是少了，可能在排序阶段就是无米之炊了。

2.排序

排序阶段，我们可以将方法大致的分为有监督和无监督的方法

2.1无监督抽取关键词

对于无监督，我们分为基于统计和基于图。基于统计就是TF-IDF和各种变种。基于图最常见的就是TextRank。

关键词提取的一个baseline就是 TF-IDF 提取，这种方法效果已经很好。投入产出比很高，我们一般需要去掉常用的停用词，保留重要的词语。

TF-IDF基于统计，易于实现，但是缺点就是没有考虑词与词，词与文档之间的关系。是割裂的。

另一个baseline就是基于图的TextRank, TextRank 由 PageRank 演变而来。

相比于TF-IDF，TextRank考虑了词与词之间的关系（提取思想就是从窗口之间的词汇关系而来），但是缺点是它针对的是单个文本，而不是整个语料，在词汇量比较少的文本中，也就是短文中，效果会比较差。

随着数据量的积累，我们需要把模型更换到有监督模型加上。一般来说，有监督分为两种，一种是看做序列标注，一种是看做二分类的问题。

2.2有监督之二分类

先说二分类问题，比较简单，就是找到词汇的各种特征，去判断这个词汇是不是这个文本的关键词。

我大概罗列一些可能会用到的特征。

1. 位置特征：

使用位置特征是我们基于文本关键词出现的位置是在大量数据的情况下是有规律可言的，比如微博文本中出现在##符号中部分词汇有很大概率就是文本的一个关键词。

是否出现在开头，是否出现在中间部分，是否出现在末尾，出现的位置（具体是第几个单词）；相对于整个文本的位置；是否出现在##符号中…

2. 统计特征：

共现矩阵信息；词频；逆词频；词性；词跨度；关键词所在句子的最大长度/最小长度/平均长度;

3. 向量特征：

关键词词向量和文档向量的相似性

2.3有监督之序列标注

关键词的提取，就是一个典型的序列标注的问题。判断句子中关键词的开头中间结尾的位置。

序列标注最基础的就是HMM和CRF方法，但是特征工程比较复杂。

为了解决特征工程复杂的问题，我们使用深度学习模型序列标注。

关于序列标注，大家可以参考我这个文章内容：

工业级命名体识别经验+代码总结

3.新词发现

还会出现一个问题，如果我们使用二分类判定关键词，上述的过程我们都是基于我们的分词器来做的。有可能会出现一些新词，由于分词错误，不能及时的出现在你的候选词库中，比如”爷青结“。

这个时候，我们需要一个新词发现系统，持续不断的补充到词库中，在召回阶段可以提升召回率。

对于新词发现来说，基操就是从文本的自由程度和凝固程度来判断是否是新词，这样的问题就是阈值不好调整从而导致召回和精准不好平衡。

我们还可以通过别的方法离线挖掘实体词补充道词库中，之前有借鉴美团ner的文章实现了一下，效果还不错，在这里，大家可以参考我这个文章：实体库构建：离线大规模新词实体挖掘

有兴趣的去github看更多相关文文章：

https://github.com/DA-southampton/NLP_ability

推荐阅读

这个NLP工具，玩得根本停不下来

征稿启示| 200元稿费+5000DBC（价值20个小时GPU算力）

完结撒花！李宏毅老师深度学习与人类语言处理课程视频及课件（附下载）

从数据到模型，你可能需要1篇详实的pytorch踩坑指南

如何让Bert在finetune小数据集时更“稳”一点