关键词提取方案

2018 年 10 月 11 日

陈垚鑫

作者介绍

陈垚鑫，就读于东南大学电子科学与工程学院，研究生二年级。

背景

2018 年神策杯高校算法大师赛， 作者于该比赛中 A 榜排名 3/591，B 榜排名 3/591。

分词

分词对本次比赛至关重要，对于中文分词而言，存在着大量未登录词，对于时事新闻而言更是如此。Jieba 分词工具没法将“霍建华”区分开来。因此，需要添加用户词典，帮助 jieba 工具正确分开这些词语。

我在本次比赛中没有使用外部数据，而是从给定 10 万条文档中自动化发现新词。从内部凝聚程度和自由运行程度来考察一个词语是否是个新词。例如词语“杨幂”,内部凝聚程度指的是 p(“杨幂”)/p(“杨”)p(“幂”)，这个值较大时，说明这两个词组合在一起不是随机事件，他们有可能组成特定短语，是一个新词。然而只看凝聚度是不够的，很容易只找出一半的词，例如“关键词”中的“键词”也是经常共同出现的，但这肯定不是一个新词。所谓自由运行程度，是指这个词能否自由地运用在句子的各个部分，而不是必须要跟别的词组成特定搭配。在这里我看的是该短语左边出现词集个数和右边出现词集个数的最小值。

我首先将每个文本里出现次数超过 5 次，并且自由度大于 3 的词语提取出来，然后全部提取完之后，计算凝聚度，最终得到用户词典。另外，我还将书名号内的词语都当成了特有词语，还有满足英文人名拼写的也当成特定词集。

结果：在添加用户词典之前，给定的关键词中只能切分准确切分出 72.82%的词语。而添加了用户词典之后，切分准确率达到了 85.62%，效果拔群。

算法流程

流程：

1、 首先用 tfidf 和 doc2vec、textrank 的得分进行加权平均，得到排名前 30 的候选关键词，然后把这当成二分类去做，将关键词标记为 1，非关键词标记为 0，同时提取 tfidf、textrank、词性、位置、tf、lda、word2vec、doc2vec、idf 等作为该词语的特征。用 2 层 MLP 在训练集上训练，最后在测试集预测，取每篇文档的概率前 2 名作为关键词。最后 2 层 MLP 可能融合了多个随机种子及参数的结果，避免由于数据太少导致的网络过拟合。

2、 通过这 10 万条文本，我们得到了接近 20 万个关键词，利用搜索指数的思想，某些词语如果经常成为别的文档的关键词，那么它是该篇关键词的概率也就越大。tfidf、textrank 倾向于输出词频大的词作为关键词，而有一些关键词在该文档中出现次数不多。如果它在别的文档中也出现比较频繁，那么通过这种方法我们也能判断出来。因此，我做了第

二遍选取候选关键词，通过 tfidf、doc2vec、textrank 和关键词频率加权平均，共同选择候选关键词。在训练集中，只用 tfidf 作为筛选标准，只有 1800 多条关键词能被成功筛选出来，用 tfidf+textrank+doc2vec，只有 2520 多条关键词被成功筛选出来。而使用 tfidf+textrank+doc2vec+关键词频率，就能筛选出 2670 多条关键词，可见其是有效果的。

3、 筛选出关键词后，又重复 1 的步骤，再次做 2 分类。

4、 由于 1 中提取出的关键词是比较不准确的，而经过 2 和 3 之后，提取的关键词相对比较准确，因此又重复了一遍 2 和 3，这里使用的关键词频率为经过 3 之后计算出的关键词频率。最后得到输出结果。

特征介绍

Tfidf：idf 是从 jieba 自带的词库里面提取的。Tfidf 将标题和 doc 分开提取。Tf 特征同理

词性：使用 jieba 自带的分词词性和 hanlp 的词性，jieba 中切出来很多词性都是无法识别的”x”，而 hanlp 切分出来的词性较好，区分度较明显，因此引入 hanlp。

Textrank：使用 jieba 自带的 textrank 提取，同样带标题和不带标题。

Idf：对所有文档的候选关键词提取 idf。对 idf 取 log 处理

关键词频率：对所有文档预测的关键词进行数目统计，同样取 log 处理。

Lda：使用主题模型，计算该文档主题分布和该候选关键词主题分布的相似性

Word2vec：使用 skip-gram 模型训练词向量，计算每个候选关键词对于其他候选关键词的概率 p(wo/wi) ，进行累加得到该候选关键词的得分。

Doc2vec：有两种训练方式，分别是 dbow 和 pv-dw 的方式。计算该词语和文档向量的相似度。

位置特征:该词语第一次出现的位置和最后一次出现的次数。

其他特征：书名号、双引号里的内容、开头是否为中国的姓氏等

其他未尝试的想法

1、 基于 seq2seq 的想法：使用 seq2seq 自动生成标题，使用 attention 机制得到每个词语的权重。

2、 使用 lda，对相似主题文本进行聚类，将这些文档一起使用 textrank。这个方法和刘知远博士论文中的 expandrank 相似。

3、 提取候选关键词附近词语的信息量作为该候选关键词的特征。因为关键词的附近很有可

能还有关键词。

4、 依存句法分析。通过依存句法分析，主谓关系和动宾关系、核心关系词通常为关键词，其他的状语定语都不是关键词。但是句法分析太费时，因此我没有尝试。

5、 聚类方法，词向量聚类，聚类中心作为关键词