NLP.TM[20] | 词权重问题

原创 机智的叉烧 CS的陋室 2019-11-09

Shotgun Senorita

Blue Stahli - Antisleep Vol. 01

【NLP.TM
】

本人有关自然语言处理和文本挖掘方面的学习和笔记，欢迎大家关注。

往期回顾：

- NLP.TM[13] | 命名实体识别基线 BiLSTM+CRF（上）
- NLP.TM[14] | 命名实体识别基线 BiLSTM+CRF（下）
- NLP.TM[16] | SIGIR2019: 深度NLP在搜索系统中的应用
- NLP.TM[18] | 搜索中的命名实体识别
- NLP.TM[19] | 条件随机场知识整理（超长文！）

NLP领域，大家的目标可能都集中在语言模型、文本分类、命名实体识别等热点任务上，且由于NLP的自动特征工程的特性，词权重问题显得就不那么在大家的目光中，但实际上，他却在很多领域里产生了重要作用，个人感觉词权重更像是一个支持性的任务，能给很多具体任务提供参考吧，而且这种支持因为简单快捷，效果显著，甚至能代替大体量模型完成基线任务，所以在工业界其实用处不少，但是又由于其工作简单渺小，且任务在应用场景的特异性强，所以不容易形成成果，导致不为大部分人所知。

今天，我来谈谈我对词权重的理解，并给大家介绍词权重的应用场景和方法。

词权重问题的定义

词权重问题具体是什么，可以参考关键词提取问题，我从关键词提取开始说起。

关键词提取是自动摘要的降级版本，从一段文本中，抽取比较重要的关键词出来这些关键词能一定程度代表文章，这种任务就被称为关键词提取。抽象的，其实可以理解为，我们的目标是得到一串和句子长度相同的01序列，对这个01序列，为1的位置对应句子的位置的词汇就是关键词，为0的则为为关键词，这样就很好理解了。

那对于这个序列而言，我们可以进一步复杂化，现在只有01，我们可以升级为分等程度，例如分成5个级别，01234，4表示最重要，3次之，以此类推，形成一个分等级的词重要性分析。

不够，再来复杂一点，分等级满足不了我胖虎，我要用连续的值来比对句子中的词汇的重要性，这就是更加复杂的问题了。

这么说一轮下来，大家就可以理解词权重了，就是给句子中每个词汇打分，体现他们的重要性，这种问题就被称为词权重问题。

词权重问题的处理

词权重问题的基线，即初版本一般不用有监督学习去做，主要因为标注样本不好构建（这点很考验算法的个人能力，而且我想说，往往是这些细节能力才是算法工程师分高下的关键），所以常用无监督做基线，然后后续迭代的时候加上有监督学习，迭代提升。

无监督方法

首先最为简单的基线方法就是TF-IDF了，这个经典的词袋模型，哪怕是现在预训练模型称霸，仍有一席之地，就在于其简单而且效果还真的不错。引用刘知远教授在知乎中对TF-IDF的评价。

> TFIDF是很强的baseline，具有较强的普适性，如果没有太多经验的话，可以实现该算法基本能应付大部分关键词抽取的场景了

所以不要嫌弃TF-IDF方法，做基线的效果真的挺不错的。至于有关TF-IDF的计算，下面提供两个比较简单的方案：

- jieba中本身就有关键词提取功能，具体可以去看TF-IDF的文档，无论是c++，还是python都有。
- sklearn中有在TfidfTransformer和TfidfVectorizer，可以自己使用语料训练。

当然的，说到TF-IDF，其实就会说到TextRank了，作为从PageRank迁移来的舶来品，且对语料依赖性不强，一直受到不少人的喜爱，jieba、HanNLP中都有集成，但是有一个比较多人诟病的问题在于计算复杂度比较高，结果提升相比tfidf虽有，但是时间代价比较大，这也导致TF-IDF被更多项目选择。

当然了，上述term-weighting的方法其实都是非常常规的，但这也正好是不可干预的，针对不同的问题其实是需要一些特异性干预措施的，下面给大家一些建议，大家可以参考。

首先，先说怎么干预，干预说白了就是去调整原始的term weighting计算，使其达到自己的预期，那么调整的手段，主要有两个：

- 加性调整。即满足某些条件下，给term weighting加减分数，这种调整往往给一些词汇结果带来质变。
- 乘性调整。即满足某些条件下，让term weighting乘一个数值，很容易可以看到乘以大于1的数其实就是在提升重要性，乘以小于1的数其实就是在降低重要性。

那么可以根据什么来做这些干预呢，下面给一些信息：

- 词性，非常重要的，在一般场景下，名词、动词等实词往往是关键词，语气词、疑问词、虚词则基本上不是关键词，因此我们可以给予调整，词性标注可以参考jieba。
- 实体识别。在有实体识别的帮助下，特定领域下上游会有实体识别计算，此时可以根据实体识别结果来给予提降权。这块的效果是最明显的，但也是最可遇不可求的，毕竟你上游不一定有实体识别。
- 位置。根据中文的语言特性，句子中的特殊位置往往有特定的重要性，一般地，句子靠前或者靠后位置往往跟个容易出关键词，可以通过指数衰减的方式进行提降权，在语言模型里面，其实也有position embedding的说法，可见其重要性。
- 左右熵等指标，结合TFIDF会有一些特别的效果。
- 上下文的信息。有时候上下文的词性、实体识别结果其实都会对本位点的预测有好处。

由于是无监督学习，加上调权是人工干预，无法探测大规模的结果，只能通过case分析，所以一般的操作是，针对少量case，例如100个，去分析和调整权重，结果调整好后，再用另一批少量样本，例如100个去做测试，查看召准。

另外补充一份不能错过的材料——刘知远老师的博士论文。他在知乎里面公开了，有兴趣的可以读一读，无论是文献综述，还是他的主体方法，都应该对大家很有启发：

刘知远：基于文档主题结构的关键词抽取方法研究

有监督方法

有监督方法，其实就会比较多样了，小到用基础统计特征做机器学习，序列标注下的HMM、CRF，大到用语义模型做深度学习，其实都有不错的效果。

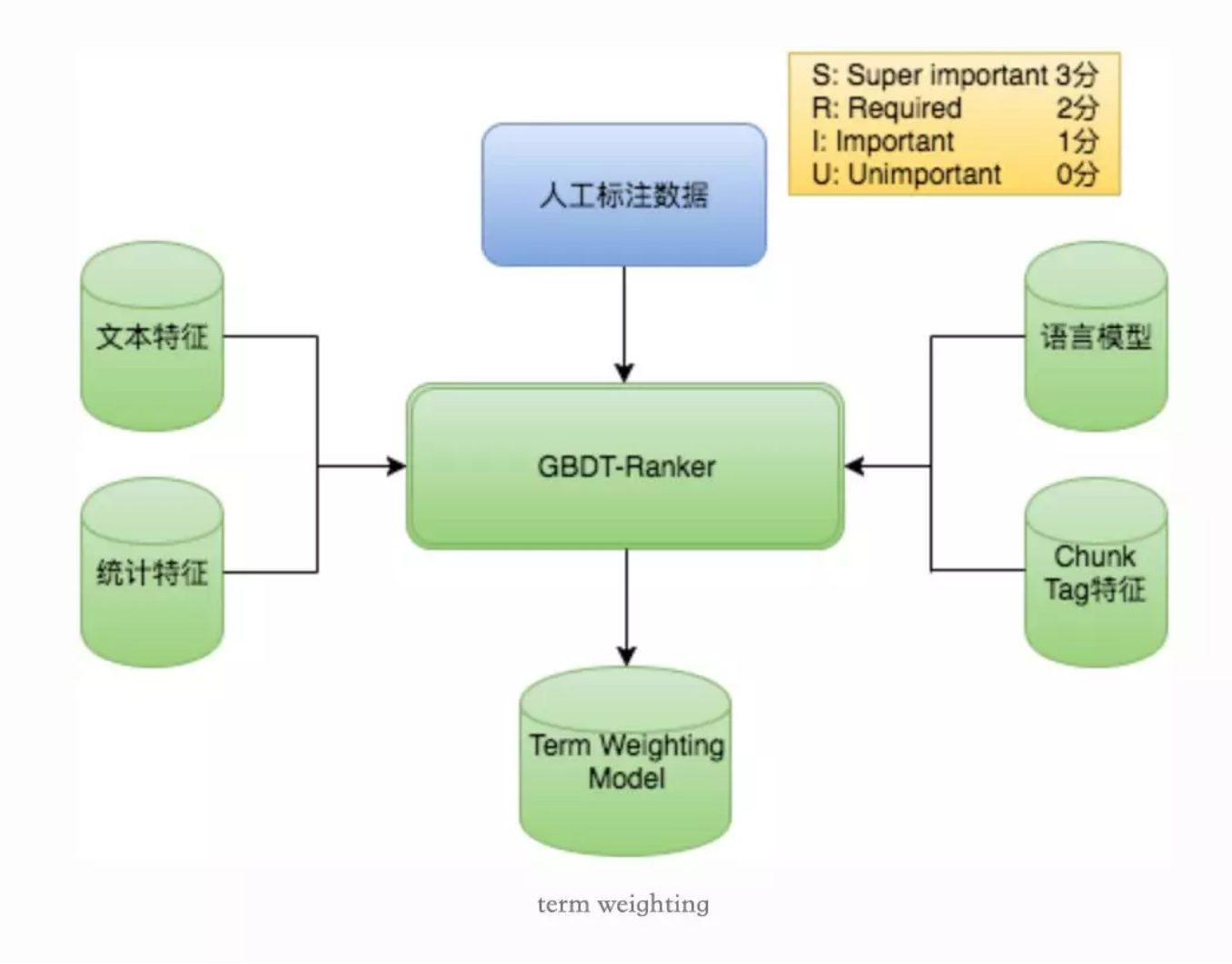
我这里重点谈小型机器学习方法，这似乎也是目前工业界常用的。LR和GBDT体系是目前浅层学习的重要方法，当然序列标注的CRF和HMM也可以参考，因此在模型选型上，主要就是这些，那么，剩下的问题就是特征怎么放了。

常用的特征如下，这个和上面提到的可能会重复。

- 词性。
- 上游实体识别。

- 位置。
- 左右熵的指标。
- TF、IDF、TFIDF
- 上下文的上述特征。
- （如果有）语言模型特征，例如放个word2vector进去。

根据上述特征，批量扔入机器学习模型，其实就已经有比较好的效果。来一份美团的实践经验看看：



具体的链接在这里：https://tech.meituan.com/2017/06/16/travel-search-strategy.html。

另外在《美团机器学习实践》的搜索章节中也有提到，翻翻看。

机器学习能够做到的事情，深度学习当然也要尝试做到，在当前语言模型如此丰富的情况下，这个任务其实并没有什么难度，但是由于现在深度学习模型的体积比较大，计算时间比较长，所以难以有线上模型出现，这种模型一般用在离线，且实际效果并不比上述方法好很多，至少从性价比来说不够高吧，此处不赘述啦。

小结

上述有关term weighting的操作，可以看到一个算法工程师，从工匠角度去打磨和分析一个问题的低姿态，大模型固然能够有好的结果，但是现实中，一方面大模型并没有足够的条件，数据上（没有足够的标注数据）、需求上（响应时长需求）等等，另一方面，构建难易度性价比上，再者，大模型往往伴随的是难以干预的问题，这些问题会导致你其实并不需要用大模型，小而精的方法，甚至是规则、词典，可能就有好的效果，我们更应该花时间在算法流程设计上，而非大模型的开发、模型调参上。

参考文献

- 刘知远：基于文档主题结构的关键词抽取方法研究。另附知乎：https://www.zhihu.com/question/21104071/answer/121576297
- 搜索中词权重计算及实践：http://www.bubuko.com/infodetail-2859295.html
- NLP之关键词提取：https://blog.csdn.net/qq_38923076/article/details/81630442
- 机器之心 | 如何做好文本关键词提取？从三种算法说起：https://www.jiqizhixin.com/articles/2018-11-14-17
- 学习笔记 — 关键词提取：https://www.jianshu.com/p/837539f116d8
- query term weight计算：https://blog.csdn.net/madman188/article/details/51855265
- 美团点评旅游搜索召回策略的演进：https://tech.meituan.com/2017/06/16/travel-search-strategy.html
- 《美团机器学习实践》：P134，8.3.5词权重与相关性计算