
Date：2018.06.22

Author：baby_qian

实验代号：标签词识别实验

标签词识别实验

1、实验内容

现有的常用基础标签词集 A 规模较为局限（数量约 1000+），现今可从微博平台爬取用户的文本评论资料和自贴的标签词，将用户自贴的标签词组合在一起得到一个候选的标签词集 B，从 B 中筛选出满足要求的新的标签词添加到 A 中，从而达到动态更新基础标签词集的目的。

2、新标签词应具备的特点

（1）新发掘的标签词首先应当是有一定的语料支持，即能概括（或高度抽象概括）一定数目的语料的主题等涵义；

（2）新发掘的标签词应当与基础标签词集 A 中的标签词所涵盖的领域差异较大，即最好能代表单独的一个垂直领域，因为基础标签词集 A 涵盖的领域范围相对较为全面，因此期望新的标签词能与 A 中的词不那么相似，即满足“好而不同”的特点；

（3）新发掘的标签词集 B 在实际中应当能被使用上，并且 B 中的词之间相似度具有一定的差异，这样能够达到一个语义去重的目的。

3、评价新添加的标签词的质量

（1）相似度阈值的选取，选出来的 K 个词。

a. 思路：

首先从爬取的微博关键词集中过滤掉词频较低的关键词，得到候选标签词集 B。接着依次计算每个候选标签词语义最相近的 N 个词，然后统计 N 个词中出现基础标签词的数量 M，若 M 大于某一阈值，则表示该候选标签词语义最相近的 N 个词中包含了一定数量的基础标签词，即该候选标签词与基础标签词集 A 中的部分词语义存在重复的现象，则不适合作为新的标签词，否则，该候选标签词适合作为新标签词。

图 1 表示选取相似度阈值，统计候选标签词在全语料中搜索相似词个数和相似度阈值的关系（红线）；以及搜索出的相似词中含基础标签词个数与相似度阈值的关系（蓝线）。

b. 筛选方法：

结合图 1，可选取相似度阈值和需要获取标签词数量进行计算。

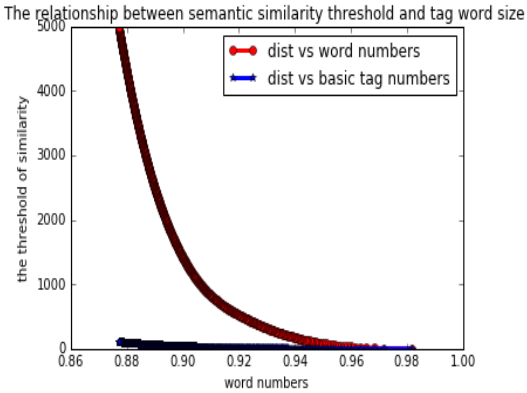
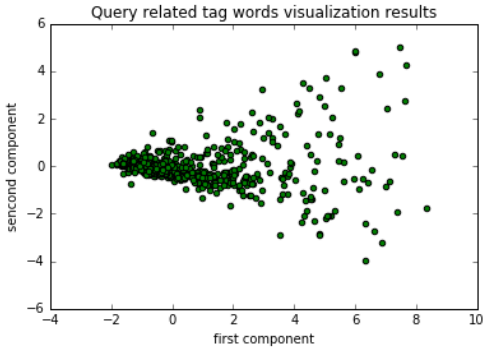


图 1 标签词数—语义相似度关系图

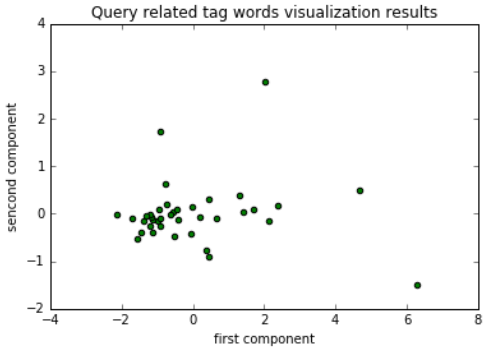
（2）无监督的方法进行评价

假设基础标签词集的中心点为 K1，新标签词集的中心点为 K2；计算 K1 与 K2 两个中心点之间的距离 D1（采用余弦距离），同时计算基础标签词集所有词到 K1 的平均距离 D2，计算新标签词集所有的词到 K2 的平均距离 D3。比较 D1，D2，D3 的大小。

出发点：新添加的标签词与基础标签词所属为不同的垂直领域最优，并且在语料库中有一定量的单词支撑。



基础标签词集二维可视化效果



新标签词集二维可视化效果

实验结果：

（D1，D2，D3）=（0.98，0.66，0.89）

分析：D1>D2，表明新标签词集的中心点距离基础标签词集中心点较远，符合新添加的标签词与基础标签词所属为不同的垂直领域最优的需要。

（3）有监督的方法进行评价（通过具体的任务来体现新标签词的质量）

构建文本—标签词对数据，将基础标签词集与新标签词集合并成一个标签集合 T；设计算法从 T 中为文本推荐标签词，然后计算准确率和召回率。

目标：重点观察召回率的值，新添加的标签词集中的词应当能被选中作为标签词。

算法：尝试计算与某一条文本所有词的平均向量的相似度来为该文本推荐标签词，然后计算准确率和召回率等指标进行评价。

绝对使用率（absolute_usage）= 新标签词中被推荐使用的数量/（基础标签词集数目+新标签词集数目）

相对使用率（relative_usage）= 新标签词中被推荐使用的数量/（新标签词集数目.

图 2(a)表示给指定文本预测标签任务的准确率-召回率曲线，每个点分别表

示推荐不同数目标签词时的评价结果；图 2(b)表示给指定文本推荐不同数目标签词时新标签词被使用的情况，红线表示新标签词集的绝对使用率随着推荐不同数目标签词规模的变化情况，篮线表示新标签词集的相对使用率随着推荐不同数目标签词规模的变化情况。（使用率越高，从一定程度上反映了该新标签词集的质量越好.）

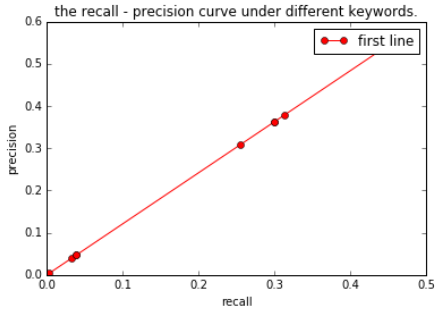


图 2(a) P-R 曲线

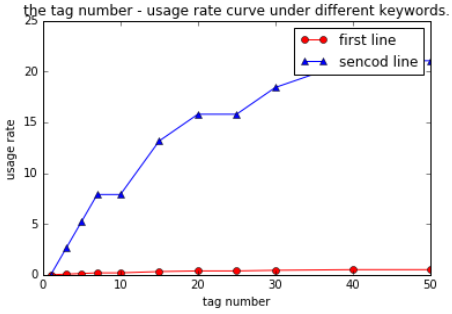


图 2(b)标签词使用率-标签词数曲线

4、后续改进工作

（1）缺陷

本次实验采用的有监督式评价方法存在一定的缺陷，即无法与 baseline（比如人工对照基础标签词集+新标签词集打标的 P-R 曲线）进行比较；另外，由于本实验采用词向量将文本进行表示(用一条文本所有单词的词向量加权求和表示该文本)，并且直接采用余弦相似度来作为文本标签词选取的依据，很大程度上依赖于词向量训练的质量。

（2）改进思路

a. 同样将给文本选定标签词转化为多分类任务，可采用 FastText 实现文本的分类任务，同时能够在该文本语料的条件下获得词向量的表示，并且对于未登陆词也具有一定的词向量表示能力。

b. 尝试采用 TextRank 的方法给文本选定标签词，并计算相应的 P-R 曲线和使用率等；TextRank 能够在一篇文章中提取关键字，从而使用它来作为该文本的标签词。

c. 尝试人工建立一个 baseline，这样能够在实际的任务中对比新标签词的性能或质量。

d. 尝试采用深度学习的方法，为文本选定标签词。