读书报告

09118119 黄一凡

2020 年 8 月 10 日

1 别人提出的问题

1. 为什么 PMI 为负值会造成影响？后面说到，unless our corpora are enormous，语料库大为什么能一定程度上解决这种问题？

这个问题我可能主要因为是负值的话不是很明显，比如书上的例子可以看出判断两个单词是不是真的比预计的出现概率小可能会出现 10e-12 次方的情况，而计算这种概率必须有较大的语料库才能比较准确的判断出来其数值是不是能达到这么小，而小语料库很难分辨其区别，并且对于人来说，不太同时出现的词语实际上也很难判断其是会随机出现还是会出现的情况更小，因为对于人的判断来说这并不明显。

2. gender stereotype 是不是指在学习语义的时候，可能出现性别偏见？a property of human reasoning 是不是说学习的时候可能产生人种的歧视？

gender stereotype 的意思应该是性别偏见，因为学习的材料中存在歧视的现象，比如医生这个职位或者科研工作者一般都联想到男性，女性的话则是护士，或者加上女医生的前缀。a property of human reasoning 可能是说人类的联想，比如说美国人喜欢把负面的词汇和非裔美国人联系起来，这其实属于一种根深蒂固的偏见，而其在测试中反映出来，这种联想不一定就是歧视，比如说它也会把一些好的词语和某些人联系起来。

3. sparse vector 和 dense vector 表示的学习的共通点是什么？

两者的共同点在于都是用了 embedding 的方法，即用向量的形式来表示一个词。不同之处在于效率不同，往往 dense vector 的效果更好。

4. 基于共现矩阵的表示方法的优势是什么？基于 word2vec 的窗口式的学习方法的优势是什么？

共现矩阵表示方法思想较为简单，表示较为直观，可以直接通过内积方式表示词的相似性，但是后面书中也指出这种方式也存在缺陷，因此使用 TF-IDF 的方法。基于 word2vec 的窗口式的学习方法书中在 P18 页说明了其优势，第一，该方法将任务化简为二分类问题，第二，较为简单，仅仅使用了逻辑回归，而不是复杂的神经网络。

5. 负采样 (negative sampling) 为什么会有很好的训练效果?

因为在 SGNS 中，将问题转化为判断一个词有多大概率会出现在这个词附近，因此相当于是一个二分类问题，而从语料库中我们只能找出正例，所以为了训练这个二分类器，我们就只好自己生成反例，所以生成反例不仅是有用的，而且是必须的。

2 读书计划

本周所读： 《Speeach and Language Processing》Chapter 6

下周计划： 《Speeach and Language Processing》Chapter 7