NLP基础|中英文词向量评测理论与实践

原创 bamtercelboo 深度学习自然语言处理 2018-05-13

> 阅读大概需要5分钟
> 跟随小博主，每天进步一丢丢

导读

最近在做词向量相关工作，训练的词向量如何进行评测？本文将从业界使用最广泛的两个评测任务进行阐述，包括相似度任务（word similarity task）和词汇类比任务(word analogy task)，这里已经写好了相关评测脚本

Word_Similarity_and_Word_Analogy

https://github.com/bamtercelboo/Word_Similarity_and_Word_Analogy

包括中文词向量评测脚本和英文V词向量评测脚本，方便大家使用。

相关知识

对于词向量好坏的评测，业界最常用的也是**最快**的评测方式是**计算词之间的相似度任务（word similarity task）**和与之相关的是**词汇类比任务（word analogy task）**，然而，近两年来，词向量仅仅在这两个任务上进行评测已经不再得到公认，要想得到公认，词向量的好坏需要应用到具体任务中进行评测，包括**句子分类，文本分类，词性标注(Part-of-Speech tagging)，命名实体识别（NER）**等，但是这两个任务还是最基本的评测，词向量的相关论文中也会进行这部分的实验。

Word Similarity Task

什么是Word Similarity?

这个任务的目的是评估词向量模型在两个词之间的语义紧密度和相关性的能力，例如男人与女人，男孩与女孩，中国与北京这些词对之间的相似度。

评价指标

在词相似度任务上，一般采用斯皮尔曼等级相关系数（ρ）（Spearman's rank correlation coefficient）作为**评价指标**，简写为 rho，它是衡量两个变量的依赖性的指标，它利用单调方程评价两个统计变量的相关性。 如果数据中没有重复值， 并且当两个变量完全单调相关时，斯皮尔曼相关系数则为 +1 或 −1 。对于样本容量为 n 的样本，相关系数 ρ 的计算如下图：

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

评价指标计算

1、首先，我们有一个金标文件（wordsim-240.txt），这份文件标注了两个词之间的相似度分数，是由人工标注的，类似于下面：

大学生 就业 7.45
图片 照片 7.45
北京 中国 7.4
能源 石油 7.4
电台 音乐 7.4

2、我们根据词向量计算两个词之间的 **余弦值（cos）** 作为词的**相似度分数**，然后计算金标分数与余弦值分数之间的斯皮尔曼相关系数。

3、代码：

```
def Word_Similarity(self, similarity_name, vec):
    pred, label, found, notfound = [], [], 0, 0
    with open(similarity_name, encoding='utf8') as fr:
        for i, line in enumerate(fr):
            w1, w2, score = line.split()
            if w1 in vec and w2 in vec:
                found += 1
                pred.append(self.cos(vec[w1], vec[w2]))
                label.append(float(score))
            else:
                notfound += 1
    file_name = similarity_name[similarity_name.rfind("/") + 1:].replace(".txt", "")
    self.result[file_name] = (found, notfound, self.rho(label, pred))
```

Word Analogy Task

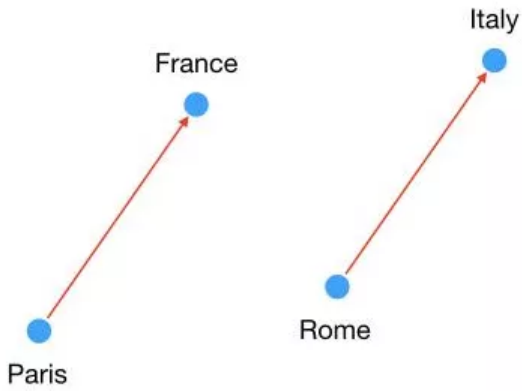
什么是Word Analogy?

这个任务考察了用词向量来推断不同单词之间的语义关系的能力。在这个任务中，三个单词 a ， b 和 s 被给出，目标是推断出第四个单词 t 满足"a 是 b，t 和 s 是相似的"。

如何得到类似的词

例如，我们要完成一句话，巴黎到法国像罗马到（）？巴黎与法国，这两个词之间是有语义关系的（巴黎是法国的首都），那么，根据第三个词（罗马），我们可以推断出是意大利吗?

我们可以根据矢量的加减来做到这一点！这是因为这些单词在空间上具有特定的关系，如下图：



从上图我们可以发现：

$$vec(法国) - vec(巴黎) = answer\_vector - vec(罗马)$$

由此我们可以根据词向量得到类似词汇：

$$answer\_vector = vec(法国) - vec(巴黎) + vec(罗马)$$

Demo

上文已经把简单的理论进行了阐述，相关代码在这里

Word_Similarity_and_Word_Analogy

https://github.com/bamtercelboo/Word_Similarity_and_Word_Analogy

这份代码包括中文词向量评测脚本以及英文词向量评测脚本，更多的细节，可以查看README。

对于英文词向量，Faruqui, Manaal, 和 Chris Dyer 建立了一份词向量评测系统

Word2vec Demo

http://www.wordvectors.org/

可以在这份系统上进行评测，我已经把这个系统的后台代码进行完善封装放在了github上面

en_embedding_similarity

https://github.com/bamtercelboo/Word_Similarity_and_Word_Analogy/tree/master/en_embedding_similarity

可以直接使用这份脚本评测。

References

[1] Word2vec Demo

http://www.wordvectors.org/

[2] Faruqui, Manaal, and Chris Dyer. “Community evaluation and exchange of word vectors at wordvectors. org.” Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.