How to generate a good word embedding

1. 动机

1.如何生成好的词向量？（哪些因素、如何影响词向量？）

2.怎样评价一个词向量好还是不好？（多个角度：词向量的语义特性、作为特征、作为模型的输入）

二、评价方法

1.词向量的语义特性

各种词向量模型均基于分布假说设计而成，即具有相似上下文的词，拥有相似的语义，并且词向量的空间距离更接近。本文通过语义相关性、同义词判别、概念分类和类比等实验论述了词向量具有不同的语义学特性。

1. 词向量用作特征

词向量可以从无标注文本中学习到句法和词法特征，很多现有工作直接使用词向量作为机器学习系统的特征，并以此提高系统的性能。本文选取了两个任务：将词向量作为唯一特征，完成文本分类任务；将词向量作为现有系统的额外特征，完成命名实体识别任务。

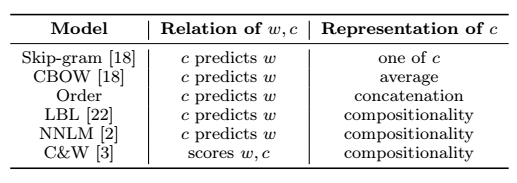
1. 词向量用作神经网络的初始值

恰当的选取神经网络的初始值，可以让神经网络收敛到更好的局部最优解。在自然语言处理任务中，基于神经网络的方法一般会使用词向量作为其输入层的初始值。在将词向量作为特征中，词向量是模型的固定输入值，在模型的训练过程中，输入值不会改变，只有模型中的参数会改变。然而，将神经网络的初始值赋值为词向量之后，神经网络在训练过程中会改变设置的初始值，因而两者是本质上是不同的。本文使用了两个任务：基于卷积神经网络的文本分类和词性标注。

三、本文工作

分析了词向量训练中3个重要的部分。

1. 模型



所有的词向量训练都是基于相同的分布假说的：出现在相似上下文的词有相似的词义。本文从两个角度总结模型:目标词和上下文的关系、上下文的表示。对于目标词和上下文的关系，大多数模型是根据上下文预测目标词，而C&W是将目标词和上下文级联在一起进行打分。实验发现，通过上下文预测目标词更能找到替换关系。

在上下文表示方面，Skip-gram选取了上下文中的一个词，CBOW使用了上下文的词向量的平均，这两者并没有加入词序。Order使用了上下文词的级联，保留了词序。其余则是增加了一个隐藏层。实验结果表明，简单的模型在小语料下表现好，复杂的模型在大语料下略有优势。

1. 语料

本文提出语料不仅是越大越好，也和语料的领域相关，不仅如此，语料的领域往往更重要。有时语料选的不对，越多反而会影响性能，而在确定某一领域后，语料越大越好。

1. 训练参数

（1）迭代次数：在很多机器学习中，迭代停止指标是看验证集的损失是否达到峰值。按照这个方法，我们可以从训练语料中分出一个验证集看损失函数的变化。但是实验中我们发现，这种策略并不好。主要原因就是，训练词向量的目标是，尽可能精确地预测目标词，这一目标和实际任务并不一致。所以更好的方法是，直接拿实际任务的验证集来做终止条件，比如命名实体识别等。

（2）词向量的维度：做词向量语义分析任务的时候，一般维度越大效果越好。对于NLP任务，50维就足够了。

四、总结

尽管本文并没有提出一种新的词向量的训练方法，但是本文系统的从多个方面总结了各种模型，对我们今后的工作有指导性意义。