文章编号:1007-1423(2019)08-0050-03 DOI:10.3969/j.issn.1007-1423.2019.08.010

词向量发展综述

严红

(四川大学计算机学院,成都 610065)

摘要:
随着深度神经网络在自然语言处理领域的应用,仅仅用独热编码等向量空间模型表示单词的方式已经不能满足模型理解文本的需求。自从词向量和深度神经网络结合的模型在自然语言处理领域的应用,提升很多子任务的准确率,从而也使得词向量的研究数剧增。词向量的发展和意义,值得研究和归纳总结。

关键词:
词向量;词表示;独热编码;自然语言处理

0 引言

在自然语言处理领域,文本作为非结构化的字符数据,首先需要转化为可计算的数值数据,所以首先将文本分割为单独的单词,将单词作为文本的原子单位。而每个单词则被表示为词汇表中的一个索引或者只有对应索引位置为 1 其余为 0 的独热编码向量。这样的表示方法具有简单性和健壮性的优点,然而单词表示之间没有相似性,互相没有联系,不包含任何语义语法信息。独热编码是稀疏向量,如果在词汇表特别大的情况,会使模型的计算量剧增造成维数灾难。所以针对这些问题,有人提出了词的分布式表示法——词向量。词向量是一个维度相对来说较低的稠密向量,也就是说它的每个维度都有实数,而非大多数为 0。自从词向量被提出并结合神经网络应用在自然语言处理子任务中,例如命名实体识别、事件抽取、病历去识别化、机器翻译和自动问答等,许多任务的准确率得到很大的提升,可见它对于现有自然处理领域的重要性。

词向量作为词的分布式表示方法自从 1986 年被 Hinton[1]提出后,经过多年的研究,产生了非常多的词向量的生成模型。不同的模型由于其输入输出的不同,使得词向量具有不同含义和影响。例如 Skip-Gram 模型[5]中的词向量,词向量之间可以做简单的算术运算来类比词之间的相似性,例如 vector("King") - vector("Man") + vector("Woman")的结果近似于 Queen 的词向量。从词向量的这种特性推测,它可能在训练的过程中从语料学习到词在上下文中的语义。随着这种特性的发现,也吸引着越来越多研究者投入到词向量的研究中。

1 技术方法

1.1 词向量生成模型

词的表示法最开始一般是潜在语义学 LSA 中所代表的具有统计信息的表示方法,例如独热编码、TF-IDF 向量等,不包含语义信息。词的分布式表示概念首先由 Hinton[1]在 1986 年提出,Bengio[2]接着提出了一种 N-Gram 神经概率语言模型,在训练这个模型的过程中,顺带生成了词向量,词向量的研究就此展开。

Bengio 首先将词表示成单词表中的一个索引,用一个映射矩阵,将其转换为 D 维的向量,也就是词向量,然后将 C 个上文词汇的词向量串联起来,通过多层前馈神经网络进行学习,预测为在该上文的情况下中心词为当前词的条件概率。模型拟合过程中,优化的目标是使得预测概率最大似然化。其中,词向量映射矩阵作为参数存在,在训练这个神经概率语言模型时,词向量也在不断地被训练,使得它们最贴近语料中的

语义。最后得到了语言模型和词向量。然而，Bengio提出的模型中使用Softmax作为模型的输出层，一旦词汇表的数量过大，模型的复杂度难以估量，训练难度倍增。例如，只有10个词汇的词汇表，Softmax层为10维，它和上一层的参数数目为10*W，映射矩阵参数为10*D，假设先忽略隐含层的参数，那么参数数目为10*(W+D)；而实际的词汇表中至少也上万，参数数目至少10000*(W+D)，可见实际模型的训练难度。因此，Morin和Bengio[4]又进行了优化，将Softmax输出层转换为与哈夫曼树结合的分层Softmax，将原有的Softmax维度S降低到了$2*\log S$，大大降低了模型的复杂度，使得我们能以更少的时间和更少的计算资源训练词向量。Mikolov等人[5]接着也提出了两种词向量的生成模型Skip-Gram和CBOW，前者将上下文作为输出，中心词作为输入，后者反之。相比之间的模型[3-4]，Mikolov等人去除了前馈式神经网络的隐含层，使得模型的计算复杂度大大降低，与此同时，并强调了生成的词向量具有语义语法的含义，尤其是词向量之间的加减可以看做一种类比，使得不同的词之间可以通过词向量的运算得到它们的联系。综合来看，CBOW相对来说有着更好的效果。Mikolov等人在之后的研究[6]中，还进一步提出子采样和负采样的方法，他们证明了在训练过程中对频繁单词的子采样可以显著提高速度（约为2x-10x），并提高稀有单词表示的准确性。提出的负采样方法可以更快地训练频繁单词，并得到更好的向量表示。我们可以看出，词向量生成模型在不断地进行优化，一方面是词向量的生成方式使其蕴含更多的含义，更一方面能更快地训练词向量，毕竟越大的词汇表，训练的成本就越高，而实际运用中更注重效率。

除了概率语言模型的方式生成词向量，Collobert和Weston[7]提出了另一种词向量的生成模型，在这个模型当中输入是一个窗口为C的几个词汇，包括一个中心词和同样数量的上文和下文词汇，将其通过映射矩阵映射成词向量，然后也是通过前馈神经网络，但输出层只有一个神经元，用于给中心词和上下文之间的联系进行打分。这种方式中，如果全部输入语料库中的上下文，那么无法给模型打分，所以将语料中的窗口上下文作为正例，将替换掉中心词的上下文作为负例，从而可以训练得到词向量。

从上面提及的方法中可以看出，词向量基本是通过固定窗口的上下文生成的，忽略了每个词汇在全局的统计学意义，所以Pennington等人提出了结合统计和词义的一种词向量生成方法，首先计算词的共现矩阵，然后在局部的上下文窗口使得中心词的预测概率和词共现矩阵中包含上下文几率的误差最小化，从而得到最优的词向量。这种方法，相对来说更强调文本中的统计学，但也同时蕴含了局部的上下文信息，实验证明了它在词相似性、词可类比性上和命名实体识别任务上的优越。

之后，词向量不仅继续应用在NLP领域上，还扩展到了其他领域，将Embedding作为一种思想和表示学习方法，应用到其他数据的表示上，例如Network Embedding[8]、Graph Embedding等。它的研究也越来越多样。

1.2 语料和训练

能影响词向量的不仅有生成方法，还有训练时的变量，包括语料类型、语料规模、词向量维度和迭代轮数。词向量训练从来都是无监督的，不需要标记的数据集，所以为了更好地还原词在上下文的语义，使其更具有普遍性，通常选择规模比较大的语料，比如维基百科、Gigaword或者语言数据联盟（Linguistic Data Consortium，LDC）中的语料作为训练语料，当然也有自采集的网络文本作语料。语料的规模从百万词到百亿词不等，一般情况语料库较大比较好，但并非绝对的。词向量维度在训练时指定，通常为100的倍数或者2的幂，不同的词向量维度会使基于它构建的自然语言处理模型有不同效果，一般而言词向量维度高比低效果要好一些，但是过高的维度会大大提升模型复杂度，所以需要选择适当维度来应用到其他的自然语言处理任务中。虽然词向量是无监督学习方式产生的，但训练的迭代轮数并非越高越好，在文献[3]中就有实验描述了迭代轮数对不同词向量的影响，可能轮数过多会导致过拟合。

2 评价方式

评价词向量的维度主要分为两个：词向量生成模型的复杂度和质量。模型的复杂度包括训练词向量的轮数、词向量模型的参数数目和词汇表大小。词向量质量的评价方式也有两种：一种是主观评价，从其本身的语义和语法方面衡量；一种是客观评价，将不同的词向量运用到自然语言处理子任务中，对比同一任务下词向量造成的影响。

在主观评价中，用到的数据集有SemEval-2012 Task 2的数据集，该数据集包含79个细粒度的单词关

系，每个关系都有 3 到 4 个单词对。给定一组假定具有相同关系的词对，识别关系相似度的连续范围，我们可以看作是类比问题。这种数据集的单词类比关系能够一定程度地衡量词向量包含的语义信息。通过不同词向量在该数据集上的得分，可以评价词向量的质量——是否捕获到单词在上下文中的语义等信息。

在客观评价中，词向量则被运用到命名实体识别、词性标注等任务中，和其他词向量进行对比。评价指标为该任务的 F1 值，比较使用不同词向量的同一任务的准确率、召回率和 F1 值，间接得到词向量的质量评价。如果在同一任务模型上，F1 值越高，说明这种词向量能较好地代表该词的语义信息，从而改善该任务的准确率。

3 结语

词向量和神经网络结合的模型给自然语言处理带来了突破，因为其使得词之间的关系可以通过计算发现关联，大量研究针对它展开，同时也使得“Embedding”思想应用到其他领域，带来新的进展。本文主要按照词向量的发展史，介绍了它相关的研究和评价方法。而不像图像可以可视化，作为高级抽象的文本信息难以被具现化，词向量中到底包含了什么信息，还值得继续探索。

参考文献：

[1]Hinton G E. Learning Distributed Representations of Concepts.[C]. Eighth Conference of the Cognitive Science Society,1989.

[2]Y. Bengio,R. Ducharme,P. Vincent. A Neural Probabilistic Language Model[J] Journal of Machine Learning Research,2003,3:1137-1155.

[3]Pennington J,Socher R,Manning C. Glove: Global Vectors for Word Representation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP),2014:1532-1543.

[4]Morin F,Bengio Y. Hierarchical Probabilistic Neural Network Language Model[C]. Aistats,2005,5:246-252.

[5]Mikolov T,Chen K,Corrado G,et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint:1301.3781,2013.

[6]Mikolov T,Sutskever I,Chen K,et al. Distributed Representations of Words and Phrases and Their Compositionality[C]. Advances in Neural Information Processing Systems,2013:3111-3119.

[7]Collobert R,Weston J,Bottou L,et al. Natural Language Processing(Almost) from Scratch[J]. Journal of Machine Learning Research,2011.

[8]Perozzi B,Al-Rfou R,Skiena S. DeepWalk: Online Learning of Social Representations[J],2014.

作者简介：

严红（1994-），女，四川广元人，硕士，研究方向为命名实体识别、意见挖掘

收稿日期：2018-12-27　　修稿日期：2019-01-11

Overview of Word Embedding Development

YAN Hong

（College of Computer Science, Sichuan University, Chengdu 610065）

Abstract：

With the application of deep neural network in the field of natural language processing, the representation of words by vector space models such as one-hot coding can no longer meet the needs of models to understand text. Since the application of the model combining word embedding and deep neural network in the field of natural language processing, the accuracy of many subtasks has been improved, which also leads to a sharp increase in the research of word embedding. The development and significance of word vectors are worth studying and summarizing.

Keywords：

Word Embedding; Word Representation; One-Hot; Natural Language Processing