基于强负采样的词嵌入优化算法

王雨晨
MobTech / 上海
wangych@yoozoo.com

林淼哲
MobTech / 上海
linmzh@yoozoo.com

詹杰凡
MobTech / 上海
zhanjf@yoozoo.com

摘要

word2vec是自然语言处理领域重要的词嵌入算法之一，为了解决随机负采样作为优化目标可能出现的样本贡献消失问题，提出了可以应用在CBOW和Skip-gram框架上的以余弦距离为度量的强负采样方法：HNS-CBOW和HNS-SG。将原随机负采样过程拆解为两个步骤，首先，计算随机负样本与目标词的余弦距离，然后，再使用距离较近的强负样本更新参数。以英文维基百科数据作为实验语料，在公开的语义-语法数据集上对优化算法的效果进行了定量分析，实验表明，优化后的词嵌入质量显著优于原方法。同时，与GloVe等公开发布的预训练词向量相比，可以在更小的语料库上获得更高的准确性。

关键词： 自然语言处理；词嵌入；强负采样

Word Embedding Optimization Based on Hard Negative Sampling

Wang Yuchen
MobTech / Shanghai
wangych@yoozoo.com

Lin Miaozhe
MobTech / Shanghai
linmzh@yoozoo.com

Zhan Jiefan
MobTech / Shanghai
zhanjf@yoozoo.com

Abstract

Word2vec is the important algorithms of word embedding in natural language processing. Traditional training method regards random negative sampling as the objective of optimization, which may cause some samples to lose their contribution in the process of training. To solve this problem, this paper proposes a hard negative sampling method based on cosine distance. This sampling method can be applied to CBOW and Skip-gram, which we call HNS-CBOW and HNS-SG. In this paper, the original random negative sampling process is divided into two steps. First, we calculate the cosine distances between the negatives and the target, and then update parameters by using the hard negatives which close to the target. We use a Wikipedia dump as corpora and conduct experiments on the Semantic-Syntactic Word Relationship test set. After analysis and experimental verification, the quality of word embedding based on hard negative sampling is significantly better than the original word2vec. We also calculate the accuracy of some pre-trained word vectors that have been published, such as GloVe, on the Semantic-Syntactic Word Relationship test set. The HNS models outperform all other baselines, often with smaller corpora.

Keywords: natural language processing , word embedding , hard negative sampling

第十九届中国计算语言学大会论文集，第208页-第215页，海南，中国，2020年10月30日至11月1日。

1 引言

词嵌入技术是自然语言处理（Natural Language Processing，NLP）领域的一项基础工作，它将一个词表达成了语义空间中的实值向量，解决了传统词袋模型的高维、稀疏等问题。把嵌入后的向量当做特征可以应用于一系列NLP下游问题，比如，信息检索(Ye et al., 2016)、文本分类(Miyato et al., 2017; Lilleberg et al., 2015)、句法分析(Bansal M et al., 2014)、命名实体识别(Habibi M et al., 2017; Katharina, 2015)等。

目前广为流行的词嵌入技术是Mikolov(2013; 2013)提出的word2vec算法。该算法以神经网络语言模型为基础，主要针对负采样（Negative Sampling，NS）进行优化，其思想是对于每一个目标词，都会按照词频概率随机地抽取一部分负样本用于参数更新。但随机负采样的一个缺陷是，有的样本在学习多次后会出现贡献消失的情况，学习这类样本不仅会浪费计算资源，也不利于生成高质量的词嵌入。

因此，本文提出了一种基于强负采样（Hard Negative Sampling，HNS）的优化算法。本文的主要贡献有：1)将强采样的思想引入到词嵌入的训练过程中，并详细解释了具体方法；2)对强负采样的有效性及其参数的影响进行了验证。

本文的结构如下：第一部分介绍了词嵌入和强采样的相关工作；第二部分详细解释了通过强负采样改进词嵌入训练的具体方法；第三部分通过实验证明了强负采样的有效性并对其参数进行了分析；最后一部分总结了本研究的改进方向。

2 相关工作

Mikolov(2013; 2013)提出的word2vec框架，将词的独热编码映射成连续的向量表示，在向量空间中，许多语言规律具有了线性平移的性质，比如，“King” - “Man” + “Woman” = “Queen”。不同于传统的神经网络语言模型，word2vec框架剔除了非线性隐藏层，只保留含有输入层、投影层和输出层的3层结构，由于不涉及密集矩阵乘法，该框架实现了更低的计算成本。论文中也指出，word2vec框架相较于传统的神经网络语言模型可以获得更高质量的词嵌入。为了进一步加快训练速度，word2vec在训练时使用了负采样和高频词亚采样的策略。两种方法的采样概率都是词频的函数，词频越高，被采样的概率越大。

word2vec的方法被提出后，许多研究在其基础上进行了拓展和衍生。斯坦福大学的Pennington(2014)提出了GloVe方法，结合全局矩阵分解和局部上下文窗口方法的优点，最终得到一个全局对数双线性模型，该模型生成的词向量质量优于传统的word2vec方法。GloVe也是目前流行的词嵌入算法之一，算法生成预训练词向量已经发布在斯坦福网站上。Wang(2015)修改了word2vec模型使其更关注上下文的相对位置，改进后的模型在词性标注和语法分析任务上的表现均有提升。Ji(2016)通过使用小批量和负样本共享，提高了word2vec算法中各种数据结构的重用性，但该方法主要关注的是训练效率而非词向量质量。

Schroff(2015)提出的FaceNet框架，提供了人脸图像到欧式空间的映射，在嵌入空间中，L2距离的平方直接对应人脸的相似度。论文提出了一个新的三元损失函数，该损失函数同时考虑了在嵌入空间中某点到相同以及不同个体的距离，以期实现最小化相同个体之间距离的同时最大化不同个体之间的距离。论文认为，随机选择正负样本会导致收敛速度变慢，因此文中提出了强正样本和强负样本的概念，用以避免容易满足约束条件的样本过多地进入到训练过程中。论文同时指出，直接优化与任务相关的损失可以提升模型性能。

本文将结合强采样的思想针对词嵌入的负采样过程进行优化。

3 方法

3.1 word2vec的上下文表示方法

word2vec包含两种不同的上下文表示方法：CBOW和Skip-gram，这两种方法都是为了获得目标词和上下文的关系。不同的是，CBOW是给定上下文$w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$来预测目标词$w_t$出现的概率，而Skip-gram则是给定目标词$w_t$预测上下文$w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$。

在CBOW方法下，对于目标词w和w对应的上下文$Context(w)$，构造条件概率函数$p(w \mid Context(w))$，优化目标是获得参数θ使得公式(1)的概率最大化。

$$\arg\max_{\theta} \prod_{(w,Context(w))\in D} p(w \mid Context(w);\theta) \tag{1}$$

其中，D是语料库中所有目标词w与上下文$Context(w)$组合的集合。
与之相对，Skip-gram的优化目标是公式(2)的概率最大化。

$$\arg\max_{\theta} \prod_{(w,Context(w))\in D} p(Context(w) \mid w;\theta) \tag{2}$$

3.2 负采样的优化目标

负采样是一种能够提高词嵌入训练速度并且有效改善词嵌入质量的方法(Mikolov et al., 2013)。该方法定义$p(D=1 \mid w, Context(w);\theta)$是目标词$w$与上下文$Context(w)$组合出现在语料库中的概率，相对的，$p(D=0 \mid w, Context(w);\theta)$是语料库不包含目标词$w$与上下文$Context(w)$组合的概率(Goldberg and Levy, 2014)。根据定义可知，$p(D=1 \mid w, Context(w);\theta) + p(D=0 \mid w, Context(w);\theta) = 1$。与3.1节相同，此处假设两组概率都是参数$\theta$的函数。

此时，词嵌入的优化目标可以表示为寻找参数θ，使得语料库中存在的目标词w与上下文$Context(w)$组合出现的概率最大化，对其取对数后得到公式(3)。

$$\begin{aligned}
&\arg\max_{\theta} \prod_{(w,Context(w))\in D} p(D=1 \mid w, Context(w);\theta) \\
=&\arg\max_{\theta} \log \prod_{(w,Context(w))\in D} p(D=1 \mid w, Context(w);\theta) \\
=&\arg\max_{\theta} \sum_{(w,Context(w))\in D} \log p(D=1 \mid w, Context(w);\theta)
\end{aligned} \tag{3}$$

同时，若将语料库中不存在的目标词w与上下文$Context(w)$组合定义为负样本集合D'，我们希望最小化这种组合出现的概率，从形式上相当于最大化$p(D=0 \mid w, Context(w);\theta)$，同样取对数后得到公式(4)。

$$\begin{aligned}
&\arg\min_{\theta} \prod_{(w,Context(w))\in D'} p(D=1 \mid w, Context(w);\theta) \\
=&\arg\max_{\theta} \prod_{(w,Context(w))\in D'} p(D=0 \mid w, Context(w);\theta) \\
=&\arg\max_{\theta} \prod_{(w,Context(w))\in D'} (1 - p(D=1 \mid w, Context(w);\theta)) \\
=&\arg\max_{\theta} \sum_{(w,Context(w))\in D'} \log(1 - p(D=1 \mid w, Context(w);\theta))
\end{aligned} \tag{4}$$

在公式(3)和公式(4)中，参数θ相当于目标词w与上下文$Context(w)$的词向量v_w和v_c，$(v_w, v_c) \in \mathbb{R}^d$，$d$是向量长度。使用softmax函数，$p(D=1\Theta w, Context(w);\theta)$可以转化为公式(5)。

$$p(D=1 \mid w, Context(w);\theta) = \frac{1}{1+e^{-v_c \cdot v_w}} \tag{5}$$

因此，负采样的最终优化目标变为公式(6)。

$$\begin{aligned}&\arg\max_{\theta}\sum_{(w,Context(w))\in D}\log\frac{1}{1+e^{-v_c\cdot v_w}}+\sum_{(w,Context(w))\in D'}\log(1-\frac{1}{1+e^{-v_c\cdot v_w}})\\&=\arg\max_{\theta}\sum_{(w,Context(w))\in D}\log\frac{1}{1+e^{-v_c\cdot v_w}}+\sum_{(w,Context(w))\in D'}\log\frac{1}{1+e^{v_c\cdot v_w}}\end{aligned}\tag{6}$$

3.3 强负采样方法

在基于负采样的词嵌入方法中，如何选择负样本是非常重要的环节。不同负样本对于参数更新的贡献是有差别的，在多轮迭代后，某些负样本对于参数更新的贡献可能变得很小，学习这类负样本会减缓收敛的速度，同时也影响了最终的词嵌入质量。

为了保证在快速收敛的同时获得更高质量的词嵌入，在每次负采样时，需要选择能够为参数更新提供更多贡献的词作为负样本。我们提出了一个假设，负样本的贡献大小与其在当前的向量空间中和目标词的距离是相关的，距离目标词越近，负样本能够提供的贡献就越大。

因此，按照与目标词的距离远近进行采样的方法我们称之为强负采样。本文选择余弦距离来衡量向量空间中词与词之间的距离，强负样本的定义见公式(7)。

$$\arg\min_{x_i^n}(1-\frac{f(x_i^t)\cdot f(x_i^n)}{\parallel f(x_i^t)\parallel_2\parallel f(x_i^n)\parallel_2})\tag{7}$$

其中，词嵌入表示为$f(x)$，意思是将词x映射到向量空间中，强负采样时，我们希望选择的负样本x_i^n距离目标词x_i^t足够近。

在进行强负采样时，如果对于每一个目标词，都在整个词典上搜索$\arg\min$，这样做的计算成本是无法负担的，同时，全局搜索强负样本也可能导致模型过早地陷入局部最优。为此，一个显而易见的解决方案是，每次从词典中选择一个小型的批数据（mini-batch）作为候选负样本集合，目标词的强负样本只从这个集合中产生。词被挑选为候选负样本的概率通过公式(8)计算得到。

$$P_n(w)=\frac{U(w)^\alpha}{\sum_{u\in D}U(u)^\alpha}\tag{8}$$

其中，$P_n(w)$是词典D中第n个词被选为候选负样本的概率，$U(w)$是词w在语料库中出现的次数，α是用于平滑的固定值参数，一般取值是0.75，平滑参数的作用是提高词频较少的词的权重。显然，词频越高，词w越有可能被选为候选负样本。

在生成候选负样本集合时，集合的容量也会影响到训练速度和准确率。容量越小，寻找强负样本所需要的计算成本也越低；而容量越大，则越有可能找到符合全局最优的强负样本，使其对参数更新的贡献更多。在实际计算时，需要考虑平衡以上两点。

3.4 强负采样的复杂度

与随机负采样相比，强负采样算法产生的额外开销包括计算余弦距离以及根据距离对负样本排序。在复杂度方面，假设Q是训练每个词的复杂度，N是目标词上下文的数量，C是上下文距离目标词的最大距离，D是词向量长度，k是随机负采样方法中的负样本数量，$neg1$是候选负样本数量，$neg2$是强负样本数量。CBOW在随机负采样下的时间复杂度是公式(9)，强负采样下的时间复杂度是公式(10)。

$$Q=N\times D+D\times k\tag{9}$$

$$Q=N\times D+D\times neg2+D\times neg1+neg1\times\log(neg1)\tag{10}$$

由于在本文算法中，$k=neg2$，因此，在CBOW方法中，强负采样比随机负采样多出的时间复杂度为$D\times neg1+neg1\times log(neg1)$，两项分别表示计算距离的复杂度和排序取前$neg2$个强负样本的复杂度。

而Skip-gram在随机负采样和强负采样下的时间复杂度分别是公式(11)和公式(12)。同理，Skip-gram的强负采样方法距离计算的复杂度是$C \times D \times neg1$，排序的复杂度是$C \times neg1 \times log(neg1)$。

$$Q = C \times (D + D \times k) \tag{11}$$

$$Q = C \times (D + D \times neg2 + D \times neg1 + neg1 \times \log(neg1)) \tag{12}$$

4 实验与结果分析

4.1 语料库

本文使用的语料库来自维基百科2019的转储数据，全部文件共包含26亿个词例。采用gensim库提供的语料库处理工具，可以对维基百科原始文件中的HTML元数据、超链接等冗余标签做预处理，同时，对语料库分词和小写化。因为gensim一次只允许一条记录驻留在内存中，所以理论上gensim可以处理任意大的语料库。特别注意，在处理原始语料时，不需要做词形还原。

4.2 评价方法

在评价词嵌入质量方面，本文采用Mikolov(2013)在论文中整理的语义-语法词相关测试集进行实验，该测试集也是业内通用的词嵌入质量评价数据集。测试集共包含19544组相关词对，涵盖5类语义（semantic）问题和9类语法（syntactic）问题。评价方法是，给出一组相关词对中的前3个词，嵌入后的词向量通过计算如果能够准确回答出第4个词则得分。回答正确的词对越多，则认为词嵌入的质量越高。

这种测试可以描述成“当a对应b时，c对应什么？”。例如，描述实体之间类比关系的语义问题：“当brother对应sister时，grandson对应什么？”，或描述时态、形态变化的语法问题：“当big对应biggest时，small对应什么？”。在一个理想的词嵌入空间中，上述问题通过向量的代数运算就可以回答。首先计算向量$X = vector(w_a) - vector(w_b) + vector(w_c)$，然后在$X$附近寻找余弦距离最近的词作为答案。

词向量计算的结果必须与测试集给出的第4个词完全匹配，同义词在本实验中不得分。另外，如果是自定义的语义-语法测试集，则需要注意测试集中的相关词对必须具有方向性，否则无法进行有效的向量运算。

4.3 实验结果

影响词嵌入质量的因素有很多，为了得到一个比较可信的HNS性能，对于一些通用的词嵌入训练参数，本次实验中将对其进行统一设置：上下文窗口大小为8，初始学习率为0.05，亚采样的概率为1e-4，迭代训练2次。大部分参数是word2vec工具的默认值，我们相信这些默认值可以带来一个相对优异的结果。对于HNS的参数，设置候选负样本个数为100，强负样本个数为15。

表1中我们比较了HNS方法与部分已经发布的预计算词向量在语义-语法测试数据上的准确率。基于HNS方法的CBOW模型称之为HNS-CBOW，基于HNS方法的Skip-gram模型称之为HNS-SG，HNS-CBOW和HNS-SG都训练了完整的维基百科26亿词例的语料库。其他模型结果：CBOW和SG的准确率结果来源于论文[9]，文中使用的测试集和本文相同；GloVe的词向量公开发布在斯坦福网站[0]，根据文档介绍该词向量在60亿词例的语料库上迭代训练了50次，其准确率由本文下载后测算得到。另外，我们控制所有词嵌入的词典大小都是由40万个最常出现的词组成，避免了词典大小对于准确率的影响。

结果表明，HNS模型使用更小的语料库和更少的迭代次数就能够达到比其他模型更高的整体准确率。当向量长度为100维和300维时，HNS-CBOW的语义准确率均最高，分别是74.4%和79.8%，同时，HNS-CBOW的整体准确率也最高，分别为64.8%和72.3%。但是，在语法准确率上，HNS方法表现地不够理想。

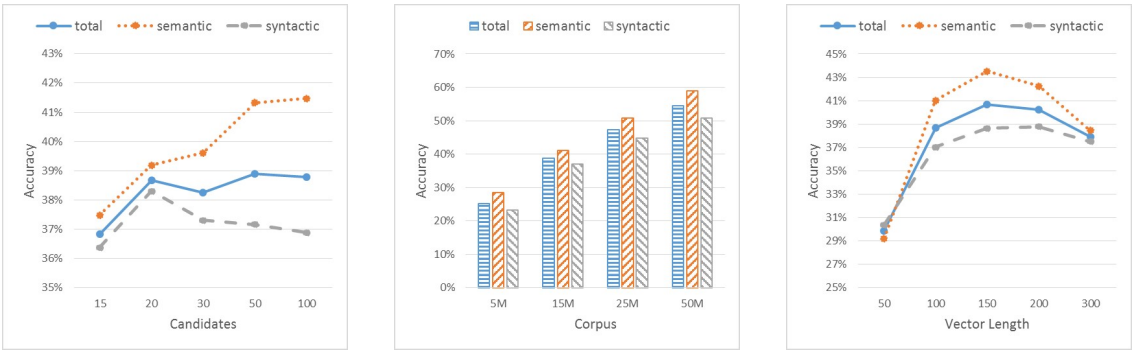
[0] https://nlp.stanford.edu/projects/glove/

Table 1: Accuracy of various word embeddings on the Semantic-Syntactic test set
表 1: 不同词嵌入在语义-语法测试集上的准确率

Model	Dim	Size	Semantic	Syntactic	Total
GloVe	100	6B	65.3%	**61.3%**	63.1%
HNS-CBOW	100	2.6B	**74.4%**	56.8%	**64.8%**
HNS-SG	100	2.6B	62.6%	46.1%	53.6%
CBOW	300	6B	63.6%	**67.4%**	65.7%
SG	300	6B	73.0%	66.0%	69.1%
GloVe	300	6B	77.4%	67.0%	71.7%
HNS-CBOW	300	2.6B	**79.8%**	66.1%	**72.3%**
HNS-SG	300	2.6B	74.6%	57.1%	65.1%

4.4 模型分析：参数分析

本节将讨论候选负样本集合大小、语料库大小、词向量长度等参数对于词嵌入质量的影响，包括语义准确率、语法准确率以及整体准确率。本节实验使用的训练数据是从维基百科语料库中随机抽取的5000篇文章，约1500万个词例组成。所有结果使用HNS-CBOW模型迭代15轮计算得到。



(a) (b) (c)

Figure 1: Semantic-syntactic accuracy on different params
图 1: 不同参数对语义-语法准确率的影响

在图1(a)中，我们展示了从不同大小的候选负样本集合中抽取10个强负样本对于词嵌入性能的影响。从图中可以看出，对于10个强负样本的采样要求，当候选集合从15增大到20时，总体、语义、语法3个准确率都有比较明显的提升；当候选集合从20增到100时，总体准确率就不再有明显变化，但此时语义准确率的提升还是比较明显的，不过其代价是降低了语法准确率。

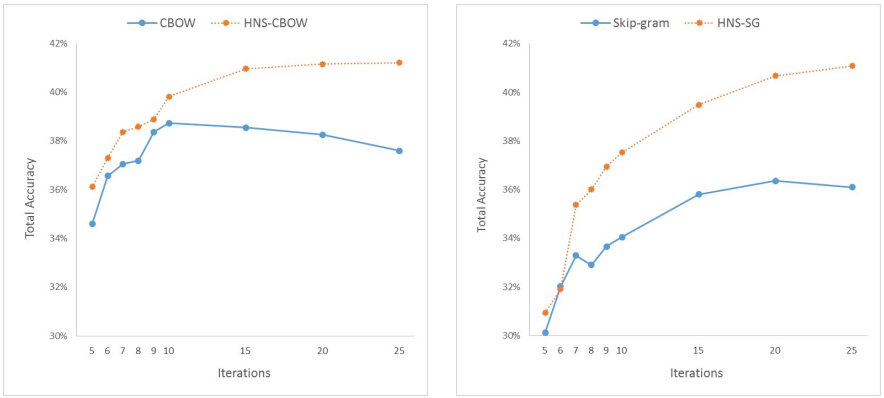
在图1(b)中，我们分别对拥有500万、1500万、2500万、5000万个词例的语料库进行了训练。意料之中的是，语料库越丰富，词嵌入的效果也越好。使用5000万词训练出的词向量比500万词的准确率提升了一倍以上。不过，随着语料库词例数量的增加，增大语料库对于词嵌入质量的提升效果是在逐渐降低的。

在图1(c)中，我们探索了在50到300的向量长度上，15轮迭代后能够达到的准确率。实验中使用1500万词例的语料库，当向量长度从50扩大到150时，准确率随着向量长度的增加而增大。但是当向量长度达到300时，准确率反而发生了下滑，这是因为，高维向量想要达到更高的准确率需要更多的迭代次数和更大的语料库来支撑，因此，在训练时间和语料库来源都受到限制的情况下，合理选择向量长度是必要的。

4.5 模型分析：与word2vec比较

为了严格比较HNS与word2vec原方法的区别，我们按照4.4的参数设置对上下文窗口、初始学习率、亚采样概率进行了控制，以使这些参数对最终准确率的影响降到最低。同时，设置候

选负样本个数为100，强负样本个数为15。作为对比，在使用word2vec工具时，选择通过负采样的方式进行训练，每次选取负样本的个数也设置为15个。另外，本次实验每一种方法都对相同的语料库进行了25轮迭代学习，语料库包含1500万个词例，并在每轮迭代后记录下词嵌入的整体准确率。



(a) (b)

Figure 2: Total accuracy on CBOW and Skip-gram
图 2: CBOW和Skip-gram方法的整体准确率

图2展示了不同迭代次数下基于HNS的词嵌入与word2vec工具在语义-语法测试集上的整体准确率变化。我们发现，在相同的参数下，不论CBOW还是Skip-gram框架，HNS方法几乎在每轮迭代后都表现出了更高的质量。同时，HNS方法也没有出现准确率下降的情况，如果不考虑训练时间的限制，HNS方法可以取得更好的结果。

5 结论

本文提出了一种在词嵌入训练过程中使用余弦距离衡量样本重要性的采样方法。该方法吸收了强采样的思想，结合word2vec的负采样方法，改善了词嵌入学习的的质量。本文使用公开的语义-语法测试数据，证明了HNS的有效性，改进后的HNS方法在CBOW和Skip-gram框架下比原方法都取得了更好的结果，相比于部分公开的预训练词向量也拥有更高的准确率。此外，本文探索了在HNS方法下不同参数对于词向量质量的影响，对比了不同候选负样本集合大小、语料库大小、词向量长度下语义-语法准确率的变化。结果表明，向量空间需要结合实际目标，针对特定的训练任务选择不同的参数组合。后续工作主要集中在两点，一是更高效的距离计算及排序方法，二是探索优化后的词向量对NLP下游任务的影响。

参考文献

Bansal M, Gimpel K, and Livescu K. *Tailoring Continuous Word Representations for Dependency Parsing* [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers). Baltimore, Maryland, USA: Association for Computational Linguistics, 2014: 809-815

Goldberg Y and Levy O. *word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method* [J]. arXiv preprint arXiv:1402.3722, 2014

Habibi M, Weber L, Neves M, et al. *Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition* [J]. Bioinformatics, 2017, 33(14): i37-i48

Ji Shihao, Satish N, Li Sheng, et al. *Parallelizing Word2Vec in Multi-Core and Many-Core Architectures* [J]. arXiv preprint arXiv:1611.06172, 2016

Katharina S. *Adapting word2vec to Named Entity Recognition* [C] // Proceedings of the 20th Nordic Conference of Computational Linguistics. Vilnius, Lithuania: Linköping University Electronic Press, 2015: 239-243

Lilleberg J, Zhu Yun, and Zhang Yanqing. *Support Vector Machines and Word2vec for Text Classification with Semantic Features* [C] // 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). Beijing, China: IEEE, 2015

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

Mikolov T, Sutskever I, Chen Kai, et al. *Distributed Representations of Words and Phrases and their Compositionality* [J]. arXiv preprint arXiv:1301.4546, 2013

Miyato T, Dai A, and Goodfellow I. *Adversarial Training Methods for Semi-Supervised Text Classification* [J]. arXiv preprint arXiv:1605.07725, 2017.

Pennington J, Socher R, and Manning C. *GloVe: Global Vectors for Word Representation* [C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1523-1543

Schroff F, Kalenichenko D, and Philbin J. *FaceNet: A Unified Embedding for Face Recognition and Clustering* [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 815-823

Wang Ling, Dyer C, Black A, et al. *Two/Too Simple Adaptations of Word2Vecfor Syntax Problems* [C] // Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. Denver, Colorado, USA: Association for Computational Linguistics, 2015: 1299-1304

Ye Xin, Shen Hui, Ma Xiao, et al. *From Word Embeddings to Document Similarities for Improved Information Retrieval in Software Engineering* [C] // ICSE '16 Proceedings of the 38th International Conference on Software Engineering. New York, NY, USA: ACM, 2016: 404-415