文章编号：2095-2163(2021)03-0217-03 中图分类号：TP391.1 文献标志码：A

基于替换方法的无监督双语词典抽取

郭晋鹏，曹海龙
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：双语词典抽取任务是自然语言处理一个重要课题。本文基于替换方法重新训练词向量，使得词向量具有跨语言特性。本文主要研究了训练词典的获取方法，以及词向量共训练模型，在中英维基百科语料上进行实验。实验结果表明，按照确信度的方法选取训练词典，基于替换的方法得到的词向量跨语言性质较好，最终抽取的词典具有较高的准确率。

关键词：双语词典抽取；无监督；替换方法

Unsupervised bilingual lexicon induction based on word substitution

GUO Jinpeng, CAO Hailong
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Bilingual lexicon induction is an important task in natural language processing. This paper retrains the word vector based on the substitution method, so that the word embedding gets cross-language characteristics. This paper mainly studies the acquisition of training dictionary and the co-training model of word vector, and carries out experiments on the corpus of Chinese and English Wikipedia. The experimental results show that using the selected training dictionary according to the method of confidence, the word vector obtained by the method of substitution has a good cross-language property, and the dictionary extracted finally has a high accuracy.

【Key words】 bilingual lexicon induction; unsupervised learning; substitution method

0 引 言

在各种跨语言任务中，双语词典抽取是目前备受各方关注的研究课题。在多数跨语言自然语言处理任务，如机器翻译[1]、跨语言文本分类[2]、跨语言情感分析[3]中，跨语言词典都起着至关重要的作用。但是，进行跨语言词典抽取往往需要人工标注的跨语言知识，如平行语料或者人工标注的翻译词典等。但世界上大多数语言对之间的平行语料或者种子词典是十分匮乏的。因此，近年来学者们开始研究无监督跨语言词典抽取，旨在使得计算机能够在不借助跨语言知识的前提下即可得到跨语言信息，从而高效、自动地获取跨语言知识。无监督跨语言词典抽取都基于如下的一个基本假设：对于不同语言的基于分布式表示的词向量空间，都存在某种映射关系，可以使其投影到相同的空间中，并且具有相同语义的单词在这个空间中的距离会彼此接近。

目前，无监督跨语言词典抽取方法已经取得了很大突破，典型工作有：Zhang 等人[4]提出了基于生成对抗网络的跨语言词典抽取方法；Hoshen 等人[5]提出了基于迭代最近点(ICP)算法的无监督翻译词典获取方法；Aldarmaki 等人[6]提出了一种不需要线性变换的映射方法来获得初始化词典。然而现有工作大都先在单语语料上获得词向量，再将词向量空间对齐。本文提出了加入反馈机制重新训练词向量的新思路：先利用无监督方法得到双语词典，再借助词典利用单词替换的方式重新训练词向量。这种方法使得词向量在保持单语特性的同时具有更好的跨语言特性。

1 具有反馈机制的无监督跨语言词典抽取模型

本课题按照 Conneau 等人基于自学习的模式(Vecmap)[7]来进行研究。其过程主要分为：初始词典的选取、迭代的自学习过程。其中，自学习过程是映射矩阵的求解和双语词典的更新两步骤反复迭代直至收敛。在此基础上，本文加入反馈机制，用得到的词典重新训练词向量，整个模型框架如图 1 所示。

vecmap 认为 2 种语言的向量空间严格满足同构性假设，使用正交变换来对齐 2 种语言的词向量空间。但单独训练得到的词向量并不能完全使正交变换来进行对齐。为使词向量具有更好的几何相似性，

作者简介：郭晋鹏(1996-)，男，硕士研究生，主要研究方向：自然语言处理、机器翻译；曹海龙(1976-)，男，博士，副教授，主要研究方向：自然语言处理、机器翻译、机器学习与人工智能。

收稿日期：2020-06-03

项目加入反馈机制,利用得到的翻译词典再重新训练具有更好跨语言特性的词向量,从而提高准确率。

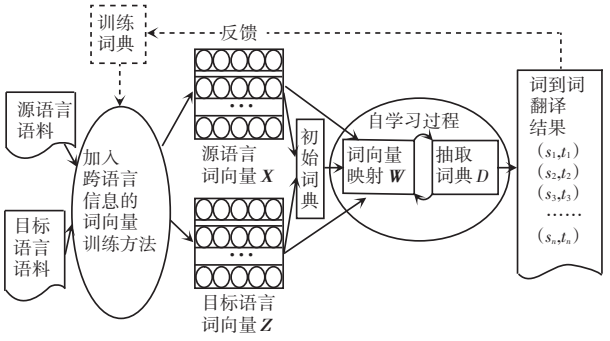

图 1 基于反馈的无监督双语词典抽取模型图

Fig. 1 Unsupervised bilingual lexicon induction based on feedback system

2 基于替换的共训练方法

共训练模型的输入为 2 种语言的单语语料和无监督反馈得到的训练词典,输出为 2 种语言的具有跨语言特性的词向量。由于无监督方法得到的翻译结果并不是完全正确的,则要从中筛选出可能作为训练指导的翻译词对作为训练词典。具体地,需要确定翻译词表中选取哪些词作为词条以及每个词条的候选翻译个数。若只取最可能的一个作为翻译,反馈过程就没有意义;若候选词太多,会使训练变得困难,也会增加时间复杂度。本文评估了经自学习过程映射后词向量翻译的 $top-k$ 准确率来确定候选词表的大小,并且比较了按照频率和置信度两种标准来筛选词条对结果的影响,经过筛选得到的词条加入训练词典指导下一轮词向量的共训练过程。

本文的共训练方法在 word2vec 中的 CBOW 模型[8]基础上加入跨语言信息。在训练词典的指导下,模型得到的词向量保持单语特性的同时要有很好的跨语言特性,即互为翻译的词所对应的词向量在空间中应该彼此接近。对于单语词向量而言,近义词或相关词由于上下文相似,训练后在空间中彼此接近。因此,本文提出基于替换的共训练方法:在语料中将训练词典中互为翻译的词按照一定概率进行替换,如此使两者就有了相同的上下文,便可以得到较为接近的词向量。例如,在翻译词典中"吃"对应的翻译为 eat ,在训练语料中句子"你喜欢吃苹果吗"时,中文单词"吃"和英文单词 eat 基于二者在词表中互为翻译的确信度以一定概率用同样的上下文进行训练。为了进一步融合双语语料,在训练过程中按照翻译的确信度以一定概率替换上下文。如图 2 所示。

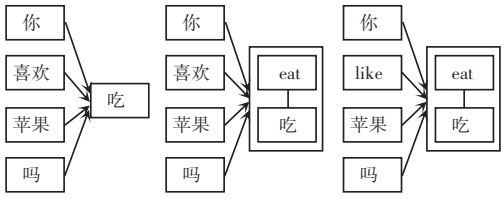

图 2 CBOW、双语 CBOW 和上下文 CBOW

Fig. 2 CBOW, BI-CBOW and Context-CBOW

由于筛选出的词典不能保证其中的词条一一对应,即一个源语言的词可能有若干个目标语言的词成为其候选翻译。本次研究在训练过程中根据词向量当前值为每一个词选出一个最可能的候选翻译,这些候选翻译实际上就组成了一个一对一的翻译集合。再利用这个确定的翻译来指导词向量的更新,该过程其实是一个 EM 算法:要求得 word2vec 的参数 θ(包括词向量 $\boldsymbol{U}$ 和上下文向量 $\boldsymbol{V}$),随机初始化后,利用当前词向量得到确定的词典,再利用词典更新词向量,如此迭代直至收敛。EM 算法具体如下:

1:随机初始化 $\boldsymbol{V}$, $\boldsymbol{U}$

2: for $i < iter$ do

3: for $i \in D_s \cup D_t$ do

4: $s \leftarrow U_{w_i} + c_i$

5: $\bar{w}_i = argmax_{w \in dic(w_i)} \cos(s, w)$

6: $\theta \leftarrow \theta + \eta \dfrac{\partial L(t(w_i), w_i, c_i)}{\partial \theta}$

7: end for

8: end for

其中, $\theta = (\boldsymbol{U}, \boldsymbol{V})$, D_s、D_t 表示词典中出现的源语言和目标语言单词集合, c_i 表示单词 w_i 的上下文词向量, $dic(w_i)$ 表示单词 w_i 的候选词表, $t(w_i)$ 为 w_i 在 D 中最可能的翻译。

3 实验

关于候选词大小的实验,本文在中英维基百科语料上用 CBOW 模型分别训练 2 种语言,再利用 vecmap 将 2 组词向量映射到同一空间,对于 vecmap 得到的映射后的词向量进行 $top-k$ 准确率评估。分别采用最近邻(Nearest Neighbor, NN)和 CSLS (Cross Domain Similarity Local Scaling)两种距离度量方式计算准确率。结果如图 3 所示。可以看出,随着词表数目的增加,准确率的增长越来越缓慢,本文后续实验使用准确率曲线拐点附近的值(5~10)作为候选词表大小设置。

利用替换方法进行无监督双语词典抽取的结果见表 1。vecmap 给出的实验结果在中英双语词典抽