文章编号：2095-2163(2021)08-0164-03 中图分类号：TP391.2 文献标志码：A

基于混合语料的无监督双语词典抽取

韩梦凡，曹海龙

（哈尔滨工业大学 计算学部 机器智能与翻译实验室，哈尔滨 150001）

摘 要：双语词典抽取作为机器翻译的基础是自然语言处理领域的重要任务。由于不需要任何监督信息，无监督双语词典抽取方法逐渐成为研究热点。无监督方法依赖于不同语言词向量之间的同构性，但是目前却少有提升词向量同构性的方法。本文提出了一种基于混合语料的同构性增强方法来提升不同语言词向量之间同构性，进而提升双语词典性能。该方法在中英维基百科上的抽取词典的性能有明显的提升。

关键词：双语词典抽取；混合语料；同构性增强

Unsupervised bilingual dictionary induction based on mixed corpus

HAN Mengfan, CAO Hailong

(Machine Intelligence and Translation Lab, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Bilingual dictionary induction as the basis of machine translation is an important task in the field of natural language processing. Unsupervised bilingual dictionary induction (UBDI) has gradually become a research hotspot because they don't require any supervision information. UBDI relies on the isomorphism between word embedding in different languages, but there are currently few methods to improve the isomorphism of word embedding. This paper proposes an isomorphism enhancement method based on mixed corpus to improve the isomorphism between word embedding in different languages, and then improve the performance of UBDI. The performance of proposed method in the induction of dictionaries on English-Chinese Wikipedia has been significantly improved.

【Key words】 bilingual dictionary induction; mixed corpus; isomorphism enhancement

0 引 言

双语词典抽取能够抽取出不同语言中含义相同的单词。作为机器翻译的基础，双语词典也被应用到跨语言自然语言处理任务中。在跨语言任务中，双语词典作为共享的跨语言特征将在一个语言上训练得到的模型应用到其它语言上。跨语言命名实体识别、跨语言信息检索以及跨语言文档分类等都利用该思想进行跨语言任务学习。

双语词典的抽取需要大规模高质量的平行语料，例如 Mikolov 和 Xing 等人的工作都采用了规模较大的词典作为监督方式学习跨语言词向量，进而抽取双语词典。由于高质量大规模的平行语料难以获取，不需要任何监督信息的无监督方法逐步成为研究热点[1-2]；Barone 等人首次提出采用生成对抗训练进行无监督学习[3]；zhang 等人在此基础上提升生成对抗训练方法的性能[4]；Artetxe 等人利用无监督初始化和迭代自学习进行无监督跨语言词向量表示学习来抽取词典[5]；Lample 等人将生成对抗训练与迭代学习过程进行结合，利用对抗训练获取初始化词典之后进行迭代增强[6]。尽管无监督方法在部分语言上（如英语-西班牙语）的性能与有监督方法不相上下，但是无监督方法隐含了不同语言的词向量是同构的假设。Søgaard 等人的研究表明词向量之间的同构性受到多种因素的影响，不同语言的同构程度是不同的[7]。基于以上原因，本文提出了一种同构性增强的方法，来提升无监督方法在双语词典抽取上的性能，该方法首先利用基线模型抽取双语词典，根据双语词典替换且合并单语语料，对混合语料进行训练，提升不同语言词向量的同构性，进而提升双语词典性能。在维基百科语料英文-中文实验上，本文提出的方法有明显的提升。

1 基于混合语料的无监督双语词典构建模型

本文在 Artetxe 等人提出的无监督双语词典抽取模型（vecmap）的基础上，提出了一个基于混合语料的无监督双语词典构建模型，模型的示意图如图 1 所示。

作者简介：韩梦凡（1997-），女，硕士研究生，主要研究方向：机器翻译、跨语言词向量学习、词典抽取等；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、跨语言词向量学习等。

收稿日期：2021-06-08

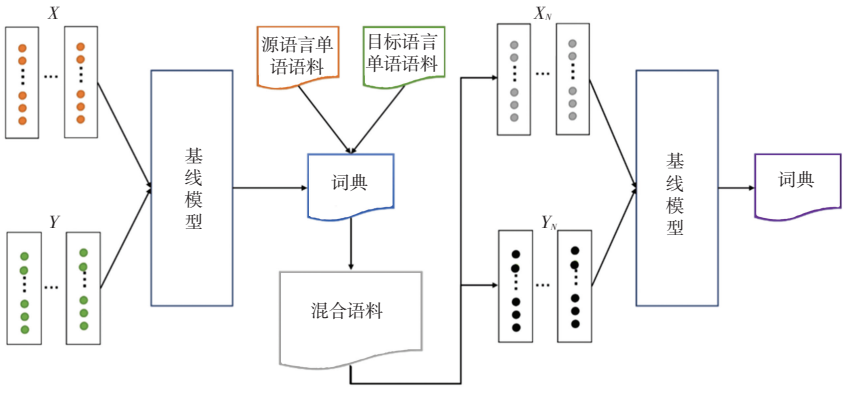

图 1　基于混合语料的无监督双语词典抽取模型图

Fig. 1　Unsupervised bilingual dictionary induction based on mixed corpus

基于混合语料的无监督双语词典构建模型包含 3 部分，第一部分利用基线模型将单语词向量映射至同一个空间并抽取词典；第二部分利用抽取的词典将源语言及目标语言单语语料中的单词替换并合并为混合语料，训练混合语料得到新的单语词向量 X_N 与 Y_N；第三部分利用基线模型映射词向量 X_N 与 Y_N 至同一空间并抽取词典。

训练混合语料过程中被词典替换后的单词能够影响上下文单词，使对应上下文单词更加接近，从而可以增强不同语言之间单语词向量的同构性。

2　基于混合语料的无监督双语词典模型细节

本文提出的模型结构中，基线模型 vemcap 采用了无监督初始化词典以及迭代自学习，不断更新词典和映射矩阵，最终利用映射矩阵 $\boldsymbol{W}$ 把源语言词向量映射到同一个空间。本文采用 Lample 等人提出的跨域相似性局部缩放方法（cross-domain similarity local scaling，CSLS）[6] 替代最近邻方法抽取词典。

在抽取词典时采用 CSLS 方法寻找源语言到目标语言的翻译，得到对应的翻译对，根据翻译对抽取词典。本文提出了两种抽取词典方式：

（1）基于频率进行词典抽取。在抽取词典的过程中，根据源语言单词出现的频率作为选取准则，源语言单词出现的频率越高，该源语言单词对应的翻译对越优先被抽取；源语言单词出现的频率越低，该源语言单词对应的翻译对越靠后被抽取；

（2）基于 *CSLS* 值进行词典抽取。该方式在抽取词典的过程中，根据已有翻译对对应的 *CSLS* 值进行词典抽取，翻译对对应的 *CSLS* 值越大，对应翻译对越容易被抽取；翻译对对应的 *CSLS* 值越小，对应翻译对越难以被抽取。

利用抽取得到词典替换合并语料：首先将词典中的词对联结成为一个特殊的联结对，接下来将单语语料中出现在词典中的单词替换成对应的联结对，具体例子见表 1。

表 1　替换合并语料例子

Tab. 1　Example of replacing and merging corpus

词典与替换翻译对	数学 math -> 数学＊＊＊math
源语言单语语料	你喜欢学习数学吗 我不喜欢和他玩
目标语言单语语料	I don't know much about China. He likes learning math very math.
混合语料	你喜欢学习数学＊＊＊math 吗 我不喜欢和他玩 I don't know much about china. He likes learning 数学＊＊＊math very much.

在训练混合语料过程中，本文采用了 word2vec 方法进行混词向量的训练。混合语料训练词向量中，根据上下文预测中心词的过程如图 2 所示。

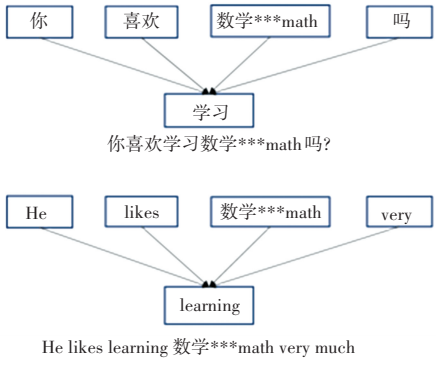

图 2　混合语料预测中心词

Fig. 2　Example of predicting center word from Mixed Corpus

根据图 2 可以发现，数学＊＊＊math 的翻译联结对能够影响“学习”和“learning”，根据单词的语义是由上下文决定的分布假设，经过词向量训练后的“学习”和“learning”会更加接近彼此。采用混合语料训练词向量的方式能够提升单语词向量的同构性。

在得到混合词向量后，将混合词向量分离为源语言单语词向量与目标语言单语词向量，具体见表 2。

表 2 分离混合词向量

Tab. 2 Example of separating mixed word embedding

混合词向量	数学 * * * math 1.332 2.265 0.638 2.747 喜欢 2.567 -0.107 -0.016 0.840 task 3.323 -2.535 0.639 -1.690
源语言词向量	数学 1.332 2.265 0.638 2.747 苹果 2.567 -0.107 -0.016 0.840
目标语言词向量	math 1.332 2.265 0.638 2.747 task 3.323 -2.535 0.639 -1.690

3 实 验

本文的实验在维基百科中文和英文单语语料进行,评价指标包括抽取双语词典的准确率以及词向量同构性的程度。词向量同构性程度的衡量采用了 Søgaard 等人提出的奇异向量相似度(Eigenvector Similarity, EVS)[7]。*EVS* 值越低,同构性越好;*EVS* 值越高,同构性越差。

本文提出的方法在双语词典抽取任务上的结果见表 3,其中 CSLS、frequency 分别表示基于 *CSLS* 值抽取词典以及基于频率抽取词典,参数 dict 表示抽取词典的规模。

表 3 基于混合语料的词典抽取结果

Tab. 3 Dictionary induction accuracy based on mixed corpus

方法	参数	覆盖率	精确度
基线模型	-	96.33	46.44
CSLS	dict = 500	96.33	49.41
	dict = 1 000	96.33	50.45
	dict = 2 000	96.33	50.52
	dict = 4 000	**96.33**	**51.14**
	dict = 6 000	96.33	44.91
frequency	dict = 500	96.33	48.37
	dict = 1 000	96.33	48.85
	dict = 2 000	**96.33**	**51.97**
	dict = 4 000	96.33	44.7
	dict = 6 000	96.33	43.18

可以发现不论是基于频率方法还是基于 *CSLS* 值方法,在词典规模合适的情况下,本方法面向词典抽取任务上的结果有明显的提升。在基于 *CSLS* 值替换的方法中最高能够达到 51.14%,在基于频率替换的方法中最高能够达到 51.97%,远远超过基线模型的 46.44%。验证了本文提出的方法在双语词典抽取任务上的有效性。

根据表 3 可以发现,随着抽取词典规模的增大,双语词典的性能并没有随着提升。一个可能的原因是由于随着抽取词典规模的增大,词典对应的质量随之降低。词典中错误翻译对上下文也产生了影响,最终导致双语词典抽取任务性能下降。

本文基于混合语料训练得到单语词向量在同构性评价指标上的结果见表 4,其中 10 k、20 k 表示抽取最常用 10 k 或者 20 k 单词衡量对应词向量之间的同构性。

表 4 词向量同构性结果

Tab. 4 Isomorphism results of word embedding

语言	单语训练	混合训练	
		CSLS	frequency
ESV(10K)	127.37	104.63	121.41
EVS(20K)	218.60	185.30	170.53

通过表 4 可以发现,本文提出方法词向量同构性相对于原始方法有明显的提升,验证了本文提出方法能够提升不同语言词向量之间的同构性。对比在 10k 与 20k 的结果可以发现,频率越高的单词对应的同构性越好。

4 结束语

本文提出了一种基于混合语料训练的无监督双语词典构建方法。该方法根据单语词向量训练方法,采用分布假设的特性,提出了将单语语料中的单词替换成抽取词典翻译联结对,并将原始单语语料合并的混合语料的方法。该方法增强了单语词向量之间的同构性,同时在双语词典抽取任务上有明显的提升。无监督双语词典抽取的同构性假设制约了无监督算法的性能,除了增强不同语言词向量之间的同构性,未来还可以探索其它不需要同构性假设的方法。

参考文献

[1] MIKOLOV T, LE Q V, SUTSKEVER I. Exploiting similarities among languages for machine translation[J]. arXiv e-prints, 2013:arXiv:1309.4168.

[2] XING C, WANG D, LIU C, et al. Normalized word embedding and orthogonal transform for bilingual word translation[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1006-1011.

[3] BARONE A V M. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders[J]. arXiv preprint arXiv:1608.02996, 2016.

[4] ZHANG M, LIU Y, LUAN H, et al. Adversarial training for unsupervised bilingual lexicon induction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1959-1970.

[5] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[J]. 2018:789-798.

[6] CONNEAU A, LAMPLE G, RANZATO M A, et al. Word translation without parallel data[J]. arXiv preprint arXiv:1710.04087, 2017.

[7] SØGAARD A, RUDER S, VULIC I. On the limitations of unsupervised bilingual dictionary induction[J]. arXiv preprint arXiv:1805.03620, 2018:778-788.