文章编号：1003-0077(2020)02-0001-15

跨语言词向量研究综述

彭晓娅，周 栋

（湖南科技大学 计算机科学与工程学院，湖南 湘潭 411201）

摘 要：随着人们对互联网多语言信息需求的日益增长，跨语言词向量已成为一项重要的基础工具，并成功应用到机器翻译、信息检索、文本情感分析等自然语言处理领域。跨语言词向量是单语词向量的一种自然扩展，词的跨语言表示通过将不同的语言映射到一个共享的低维向量空间，在不同语言间进行知识转移，从而在多语言环境下对词义进行准确捕捉。近几年跨语言词向量模型的研究成果比较丰富，研究者们提出了较多生成跨语言词向量的方法。该文通过对现有的跨语言词向量模型研究的文献回顾，综合论述了近年来跨语言词向量模型、方法、技术的发展。按照词向量训练方法的不同，将其分为有监督学习、无监督学习和半监督学习三类方法，并对各类训练方法的原理和代表性研究进行总结以及详细的比较；最后概述了跨语言词向量的评估及应用，并分析了所面临的挑战和未来的发展方向。

关键词：跨语言词向量；深度学习；有监督方法；半监督方法；无监督方法

中图分类号：TP391　　**文献标识码**：A

Survey of Cross-Lingual Word Embedding

PENG Xiaoya, ZHOU Dong

(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China)

Abstract: With the increasing demand of multilingual information on the Internet, cross-lingual word embedding has become an important basic tool, which has been successfully applied to the natural language processing fields such as machine translation, information retrieval and text sentiment analysis. Cross-lingual word embedding is a natural extension of monolingual word embedding. The cross-lingual representation of words transfers knowledge among different languages by mapping different languages into a shared low-dimensional vector space, so as to accurately capture the meaning of words in multilingual scenario. In recent years, there have been a lot of research achievements on the cross-lingual word embedding model. This paper reviews the existing literature on cross-lingual word embedding models, and comprehensively discusses the development of cross-language word vector models, methods and technologies in recent years. According to the different ways of word embedding training, it is divided into three kinds of methods: supervised learning, unsupervised learning and semi-supervised learning. Finally, we summarize the evaluation and application of the cross-lingual word embedding, and analyze the challenges and future development directions.

Keywords: cross-lingual word embedding; deep learning; supervised method; semi-supervised method; unsupervised method

0 引言

近年来，随着神经网络的快速发展，词向量(word embedding)[1]已经成功地应用于许多自然语言处理(natural language processing, NLP)的任务中，如情感分析[2]、依存分析[3]、机器翻译[4]等，并成为这些任务中解决问题的基础和主流方法之一。在 NLP 中，最直观、最简单的词表示是基于词袋(bag-of-words, BOW)的 One-Hot 表示。这种方法

收稿日期：2019-07-01　　**定稿日期**：2019-10-16

基金项目：国家自然科学基金(61876062)；湖南科技大学科学基金(KJ1746)

将每个词表示为一个很长的向量。向量的维度通常为词表的大小,其中只有一个维度的值为1,表示当前词,其余元素全部为0。然而,One-Hot表示存在两大缺点:一是向量的维度会随着语料中词的数量增大而增大;二是任意两个词之间都是孤立的,在语义层面上没有任何联系。因此,如何将语义融入词表示中成为了亟需解决的问题。

Harris[5]在1954年提出了分布假设(distributional hypothesis)。该假设指出,出现在相似上下文中的词往往具有相似的含义。随后,Firth[6]对分布假设做了进一步的阐述和明确,即词的语义由上下文决定。基于分布假说的词表示方法,根据不同的建模方式主要可以分为三类:基于矩阵的分布表示[7]、基于聚类的分布表示[8]和基于神经网络的分布表示[7]。其中,前两种属于高维表示,后一种属于低维表示,它们的核心思想是大致相同的,都由两部分组成:即选择一种方式描述上下文、选择一种模型刻画某个词与其上下文之间的关系。基于矩阵的分布表示通常又称为分布语义模型,需要构建一个"词—上下文"矩阵,该矩阵中,每行对应一个词,每列表示一种不同的上下文,矩阵中的每个元素对应相关词和上下文的共现次数。由于分布假说认为上下文相似的词其语义也相似,因此在这种表示下,两个词的语义相似度可以直接转化为两个向量的空间距离。基于矩阵分布表示的经典模型是隐式语义分析(latent semantic analysis,LSA)[9],通过"词—文档"矩阵使用奇异值分解技术(singular value decomposition,SVD)降维,获取词的低维向量表示。基于聚类的分布表示最经典的方法是布朗聚类[10],主要根据两个词的公共类别来判断这两个词的语义相似度。

基于神经网络的分布表示是分布假设的有效性最为明显的表现,一般将其称作词向量、词嵌入或分布式表示(distributed representation)[7]。词向量降低了向量的维度,将语义相近的词映射到相近的集合空间上,易于捕捉语义和句法间的关系。词向量最先由Bengio等于2003年提出,即著名的神经网络语言模型(neural network language model,NNLM)[11]。2008年,Collobert和Weston首次展示了预先训练好的词向量的实用性[12]。Mikolov等人在2013年提出了著名模型Word2Vec[13]。Word2vec是由NNLM改进而来,包括了连续词袋模型(continuous bag-of-words,CBOW)和跳字模型(skip-gram)两种方法。随后,Pennington等[14]提出了基于全局信息统计和上下文关系预测的GloVe模型,该模型经过预先训练而得到了一套完整的词向量集。

随着单语词向量在许多NLP任务中的成功应用[15],词向量在跨语言自然语言处理中的潜力引起了人们的广泛关注。跨语言词向量是单语词向量的一种自然扩展,它认为具有相似概念的不同语言的词在向量空间中的词向量非常接近[16],这就使我们能够在多语言环境下对词义进行推理。而跨语言应用对词义表达和知识转移的需求,催生了跨语言词向量模型[17],该模型在一个联合向量空间中学习词的跨语言表示。随着多语言自然语言处理研究的蓬勃发展,越来越多的学者将目光转向了跨语言词向量模型的研究。

与其他跨语言模型,如基于本体的跨语言模型[18]等相比,跨语言词向量模型受到众多研究者的青睐是因为其具有两大独特之处。第一,能够对跨语言语义信息进行建模,准确计算跨语言词语相似度等信息,是跨语言词典构建[19]、跨语言自动链接[20]、跨语言信息检索[21]等多种跨语言应用的基础。第二,借助迁移学习技术,跨语言词向量模型能够在语言间进行结构和语义上的迁移[22],即在一种语言中训练词向量模型,然后借助某些跨语言特征,如双语词典等将该语言的词向量信息迁移到其他语言,构成另一种语言的词向量模型,特别是在资源不对等的语言对之间构建跨语言词向量。

跨语言词向量模型近几年的研究成果比较丰富,研究者们提出了较多生成跨语言词向量的方法。这些方法都引入了不同对齐形式的跨语言语料资源,包括词级对齐[23-24]、句子级对齐[25-26]、文档级对齐[27]等。这些生成方法大多依赖于源语言和目标语言之间的昂贵人工标识语料,如平行语料库或种子词典等。对于资源贫乏的语言,这样的语料并不容易获得,且在不同的语言对之间,资源和基线方法的可用性是高度不平衡的。最近,越来越多的研究致力于解决这一问题,研究者们倾向于使用极小的种子词典(如使用25个词对[22])作为跨语言监督信号,这种方法可归类于半监督学习方法[28-29]。随后又提出了无监督学习跨语言词向量的方法。该类方法不需要语言对之间的任何对齐或标注语料,在NLP领域受到广泛的关注。

在本文中,我们首先通过对现有的跨语言词向量模型研究的文献回顾,综合论述了近年来跨语言词向量模型的发展。按照训练方式的不同进行分

类，并对各类训练方法的原理和代表性研究进行总结以及详细的比较。本文贡献如下：

（1）根据跨语言词向量不同的训练方法，将其分为有监督训练方法、半监督训练方法和无监督训练方法三类，详细描述了这三大类模型的不同特征。

（2）概述了跨语言词向量模型目前研究覆盖的语言以及其评估任务和应用。

（3）对比跨语言词向量三种不同训练方法，概述了跨语言词向量研究中面临的重要挑战，并提出了有待探索的研究方向。

本文的组织结构如下：第 1 节首先简要地介绍跨语言词向量模型的基本概念与训练数据需求，特别是其对齐形式以及数据集的可比较性；第 2 节对有监督训练方法中的基于词对齐平行语料、基于句对齐平行语料以及基于文档对齐可比语料的跨语言词向量模型的主要研究工作进行总结，并对各种模型进行了分析与比较；第 3 节概述半监督跨语言词向量模型的代表性研究工作及方法原理，并对其做了分析与比较；第 4 节对无监督跨语言词向量模型的主要研究工作及方法原理进行总结并对其进行分析与比较；第 5 节概述了跨语言词向量模型目前研究覆盖的语言以及其评估任务和应用；第 6 节总结了跨语言词向量三种不同的监督训练方法的差异，并讨论了跨语言词向量模型所面临的一些挑战和未来的研究方向。

1　跨语言词向量模型的基本概念与训练数据需求

跨语言词向量即跨语言的词表示，通过将不同的语言映射到一个共享的向量空间，在不同语言间进行知识转移，从而在多语言环境下对词义进行建模[16]。词向量作为 NLP 任务中的核心表征技术，能够很好地捕获语言中的规律[1]。跨语言词向量则利用此规律在不同的语言间进行知识转移。前人的工作[17]发现利用两种语言的单语词向量在向量空间中存在近似同态性，因而可以使用线性映射把这两个向量空间联系起来。此外，通过从单语设置转换到双语设置，并构建共享的双语向量空间，可以在不同语言间扩展或概括语义任务[30]，例如，语义相似性[31]计算、同义词检测或单词类比计算等[24]。

关于跨语言词向量模型的研究表明[32]，模型所需要的数据比实际的底层架构对最终模型性能更重要。特别是模型性能之间的巨大差异源于它们的训练数据量与数据本身的质量，而其他细粒度的差异则是所选体系结构、超参数以及所使用的额外技巧和调参工作。跨语言词向量模型的数据需求可基于对齐方式和数据的可比性进行分类。如图 1 所示，基于对齐方式可大致分为词级对齐、句级对齐、文档级对齐、不同对齐类型引入不同程度的双语监督信号。基于数据的可比性可将训练语料分为平行语料库和可比语料库。平行语料库是不同语言间的实际翻译，而可比语料库则是具有某些共同特性的文档的集合，仅仅在某种程度上相似，如标题或内容相似，没有明确显示出源语言和目标语言的关系。

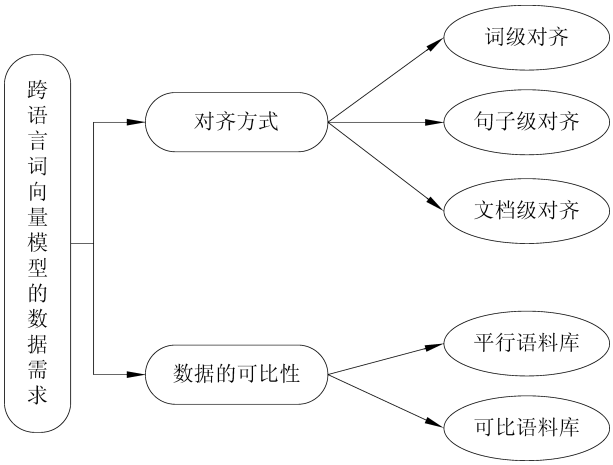

图 1　跨语言词向量模型的数据需求

大多数基于词对齐的跨语言词向量模型都需要以跨语言特征，如双语词典为基础的词对齐平行语料库作为建模的依据[17]。可比语料库虽然更为常见，但在其基础上进行建模比较困难，通常需要其他信息，如图片信息作为辅助[33]。基于句对齐的模型通常使用机器翻译中常见的标准句对齐平行语料库如 Europarl① 等作为训练语料，有些工作也借助了词对齐的信息[25]。句对齐可比语料库则与词对齐语料库一样，使用了其他模式如图像等的信息作为辅助[34]，但大多数工作集中在从平行数据中学习跨语言表示，对可比数据的研究较少。通常，不同语言的文档可能基于同一主题，如 Wikipedia② 中不同语言的文档，因此基于文档对齐的模型通常使用跨语言文档可比语料库进行建模[35-36]。在该语料库中，不同语言的文档在 Wikipedia 网站上已经自动对

① http://www.statmt.org/europarl/

② http://www.wikipedia.org

齐,或已经与主题对齐并谈论类似的话题。平行语料在这一模型中基本没有使用,这是由于平行语料通常假设句子也是对齐的。因而常用基于句对齐平行语料的模型来解决跨语言词向量生成问题,而不使用文档对齐的平行语料。为方便理解,图 2 分别给出了基于不同对齐语料的跨语言词向量模型的示例,其中图 2(a)是词级平行对齐的双语词典,图 2(b)是句级对齐的平行翻译,图 2(c)是使用类似文档的文档级可比对齐。

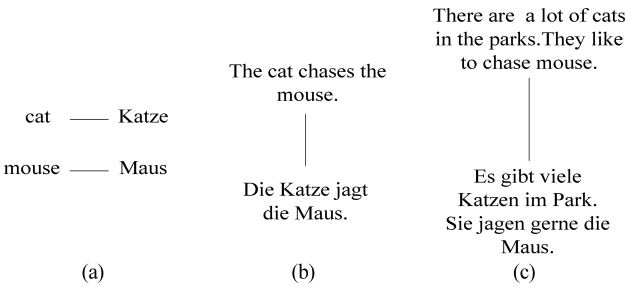

图 2 基于不同类型的对齐语料的示例

有监督学习通常需要大量人工标注的数据,而以上描述的训练数据需求主要是有监督学习的。对于资源较为贫乏的语言对,这样的数据需求很难满足。半监督学习可以缓解这一问题,使用较少的人工标注数据作为训练数据,如种子词典等。无监督学习无需任何人工标注数据,因此近年来受到了研究者们的关注。下面我们将分别从有监督学习、半监督学习以及无监督学习三方面对跨语言词向量模型的主要研究展开详述。

2 有监督的跨语言词向量模型

近几年,跨语言词向量模型在有监督学习方面相对于半监督和无监督学习的研究成果较为丰富,但有监督的学习方法往往需要大量的对齐语料。如图 3 所示,本节将在前人工作[16]基础上,基于不同类型的对齐语料对跨语言词向量模型进行分类讨论。由于基于词对齐的跨语言词向量模型与基于句对齐的跨语言词向量模型在平行语料方面的研究更为常见,基于文档对齐的跨语言词向量模型的研究主要集中在可比语料方面,所以本节的有监督学习主要围绕它们展开讨论。

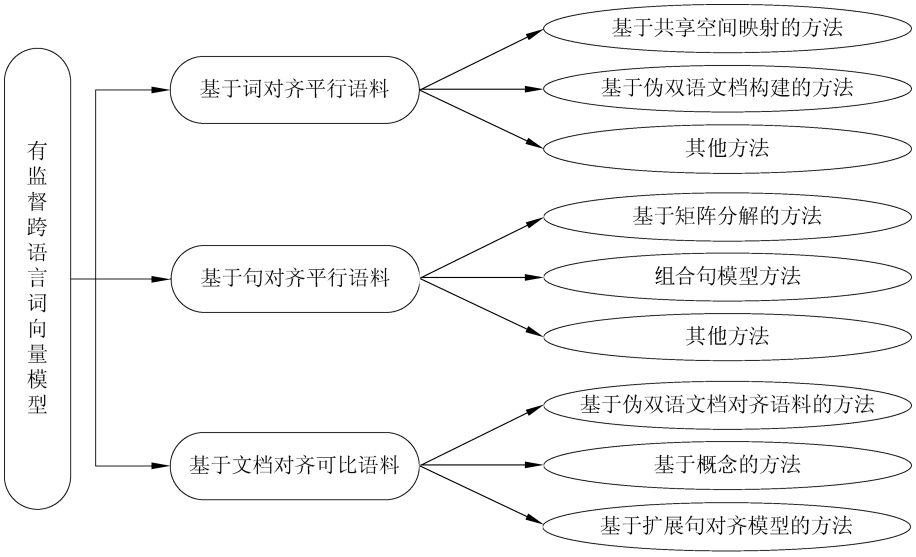

图 3 有监督跨语言词向量模型的分类

2.1 基于词对齐平行语料的跨语言词向量模型

使用词对齐平行语料库的模型可基于不同的方法来获取跨语言特征,包括基于共享空间映射的方法、基于伪双语文档构建的方法[37]以及其他方法,其中第一类方法研究较多。

2.1.1 基于共享空间映射的方法

该类方法的工作原理是独立地训练两种语言的词向量,然后使用线性转换将它们映射到跨语言向量共享空间。这些方法大部分使用数千条的双语词典来学习映射。现有的这类方法可以分为回归方法、规范方法、正交方法以及边距方法等。

回归方法通过最大化源语言的相似性,将源语言的向量映射到目标语言空间。Mikolov 等[17]利用最小二乘目标映射目标语言的向量,将余弦相似度作为距离度量,在目标语言空间中找到最接近源

语言词的表示，并使用这种方法学习基于大型单语数据的语言结构和小型双语数据之间的映射来翻译缺失的单词和短语。其缺点是使用均方差（mean squared error，MSE）导致了中心度问题（即一些单词往往作为许多其他单词的最近邻出现）[38-39]。Shigeto 等[40]在零样本学习[40]（zero-shot learning，ZSL）中，对源对象空间中的目标对象进行映射回归，并在源空间中进行最近邻搜索。实验表明，该方法降低了中心度，提高了双语词向量的准确性。Dinu 等[39]根据映射向量上潜在邻域的邻近性分布，通过调整映射后的相似矩阵，从而降低中心度。

规范方法使用典型相关分析（canonical correlation analysis，CCA）及其扩展方法将两种语言中的向量映射到一个新的共享空间，从而最大化它们的相似性。Faruqui 等[23]使用典型相关分析法将源语言和目标语言的向量映射到共享空间。如图 4 所示，不同于线性映射的是，CCA 为每种语言都学习了一个转换矩阵，其中，转换矩阵 V 用于将一种语言的词向量从向量空间 Σ 映射到新的向量空间 Σ^*，转换矩阵 W 用于将另一种语言的词向量从向量空间 Ω 映射到新的向量空间 Ω^*。注意，Σ^* 和 Ω^* 可看作同一个共享空间。随后，Lu 等[41]在映射过程中引入非线性，将双语 CCA 扩展到深层双语 CCA：即训练两个深层神经网络，最大限度地提高两种单语向量空间投影之间的相关性。

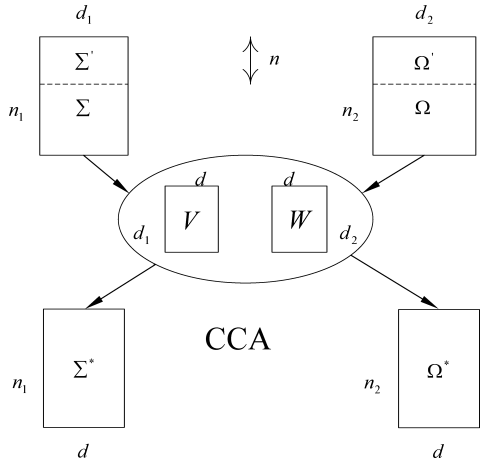

图 4　基于 CCA 映射的规范方法[23]

正交方法是在正交变换的约束下，使用一种或两种语言映射向量。Xing 等[42]为解决学习词向量的目标函数、词向量的距离度量以及学习线性变换的目标函数之间的不一致性，提出了一种基于归一化词向量的双语词翻译正交变换。他们首先对词向量进行归一化处理。然后，为了保持映射后的单位长度，将双语投影中的线性变换约束为正交变换。Artetxe 等[43]实验表明，正交性质对性能的影响比长度归一化更重要。Zhang 等[44]使用翻译对来建立单语向量之间的标准正交映射，从而学习跨语言词向量。Smith 等[45]证明了词向量空间之间的最优线性变换应该是正交的，并利用 SVD 得到了双语词向量。

边距方法将向量映射到一种语言中，最大限度地扩大正确翻译与其他候选翻译之间的边距。为解决中心度问题，Lazaridou 等[46]使用最大边距损失函数（max-margin hinge loss，MMHL）而不是 MSE。该方法在跨语言环境中极大地提高了词向量映射的准确性。

2.1.2　基于伪双语文档构建的方法

该类方法中，伪双语文档一般使用种子双语词典，通过随机地将源语言语料库中的单词替换为它们的翻译来构建伪双语语料库。Xiao 和 Guo[37]利用 Wikipedia 词典将源语言语料库中出现的所有词翻译成目标语言，过滤出多义词以及未出现在目标语言语料库中的译文，从而构建种子双语词典。Gouws 等[24]显式地创建伪双语语料库，他们连接源语言和目标语言语料库，并随机使用翻译对中的部分词来替换源语言中的所有词，然后在这个语料库上训练 CBOW。Ammar 等[47]将这种方法扩展到多种语言，使用双语词典确定不同语言中的同义词簇。他们将不同语言的单语语料库连接起来，用词簇 ID 替换同一簇中的词，然后在这个连接的语料库上训练 SGNS。Duong 等[48]也提出了类似的方法，但在 CBOW 训练过程中，并没有将语料库中的每个词都随机替换成译文，而是用即时翻译替换每个中心单词。另外，他们还明确提出一个期望最大化风险的训练算法，该算法通过合并多个字典翻译来显式处理一词多义问题。Adams 等[49]使用相同的方法对跨语言词向量进行预训练，用于低资源语言建模。

2.1.3　其他方法

Kočiský等[50]提出了一种同时学习词对齐和双语分布式词表示的概率模型。该方法充分利用了平行语料库，实现了高质量的词对齐。Shi 等[51]提出了一种用于学习跨语言词向量的矩阵因子分解框架，能够在一个统一的向量空间中捕捉不同语言的词汇相似性。如图 5 所示，总体目标分为单语部分和双语部分，通过对单语上下文词逐点互信息[52]（point-wise mutual information，PMI）矩阵 M_i^1 进

行矩阵因子分解，得到转换矩阵 W_i^l 和上下文向量矩阵 C_i^l，并引入约束条件，保证跨语言语义关系。Wick 等[53]利用人工代码转换(artificial code switching，ACS)来学习语言的不变性，ACS 转换使用翻译字典将文本中的一些单词替换为来自另一种语言的(可能是粗略的)翻译。

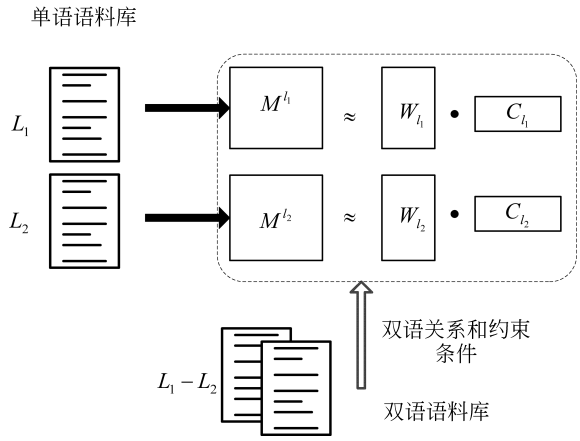

图 5 基于矩阵因子分解的有监督学习方法[54]

2.2 基于句对齐平行语料的跨语言向量模型

句对齐的训练数据是比较难获取的，因为它需要细粒度的监控。由于机器翻译(aachine translation，MT)的研究提供了大量句对齐的平行数据，因此很多研究工作都集中在从平行数据中学习跨语言词向量，可比数据方面的研究较少。利用句对齐数据的方法通常是成功的单语模型的扩展。我们将按照基于矩阵分解的方法、组合句模型方法、其他方法三类方法对基于平行语料库的句对齐模型进行概述。

2.2.1 基于矩阵分解的方法

基于矩阵分解的方法是将矩阵分解技术应用到双语设置中，通常还需要额外的词对齐信息。Zou 等[55]提出了一种从大量未标记语料库中学习双语词向量的方法，同时利用 MT 词对齐约束翻译等价性。Huang 等[56]提出了一种基于平移不变性概念的多语言词向量构造方法。该方法将平移不变性的概念形式化到矩阵分解的目标函数中，提供了一种灵活的、可伸缩的方法来获取在学习向量空间中相互平移的单词向量。Vyas 和 Carpuat[57]提出了另一种基于矩阵分解的方法来学习稀疏跨语言词向量。首先通过 GloVe 从预先训练过的源语言和目标语言的单语向量矩阵中学习单语稀疏表示，然后将单语向量矩阵进行分解，引入约束条件，使两个平行语料中的单词紧密对齐来学习跨语言词向量。Guo 等[58]为了提高映射的鲁棒性，使用了一种基于形态学的机制，将向量从词汇表内传播到词汇表外(out-of vocabulary，OOV)，并使用编辑距离作为形态相似性近似来学习跨语言词向量，即对于每个 OOV 单词，提取一个候选单词列表，这些候选单词在编辑距离上与其相似，然后将这个候选单词列表的平均向量作为 OOV 单词的向量。

2.2.2 组合句模型方法

组合句模型方法使用单词表示来构造对齐句的句子表示。Hermann 和 Blunsom[59]提出了一种基于组合语义的跨语言分布式表示方法，该方法使用句子对齐的数据成功地训练语义表示。如图 6 所示，输入语句 a 和 b(其中 a_1,a_2,a_3,a_4 和 b_1,b_2,b_3 分别表示语句 a 和 b 的单词)，通过组合向量模型(compositional vector model，CVM)生成组合句级表示 a_{root} 和 b_{root}，并最小化两个句表示之间的距离，从而使彼此更接近。随后，其又提出了一种利用平行数据学习多语言词向量的方法，并结合多语言的组合向量模型。该方法将分布假设扩展到多语言数据和联合空间表示，并将这种方法扩展到文档，递归地应用组合和目标函数将句子组合成文档[60]。Soyer和 Stenetorp[61]提出了一个基于神经网络的架构来生成跨语言词表示，该方法同时利用双语和单语数据，将词级表示形式限制为组合式。

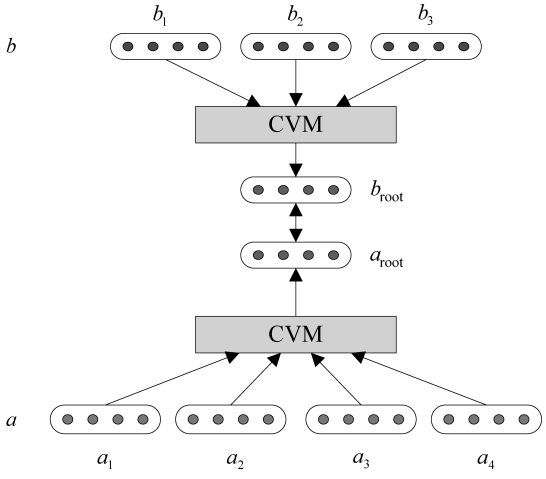

图 6 基于组合语义的跨语言分布式表示方法[59]

2.2.3 其他方法

Lauly 等[62]研究了一个用于学习多语言词表示的自动编码器模型，该模型能够学习将成对的词袋语句的隐藏表示关联起来。首先从源句中重构出目标句，将句子编码为其词向量的总和，然后用特定语

言层的编码器和解码器层以及分层 softmax 训练一个自动编码器，从每个句子中重构句子本身及其翻译。Chandar 等[63]也提出了类似的双语自动编码器模型，但使用了二进制词袋(bag-of-words，BOW)替换 softmax。他们将平行句的词袋表示编码到同一个向量中，以至于每个单词并不直接匹配到另一个单词，而是用于创建独立的句子表示。

Gouws 等[64]提出了一种基于句对齐的双语词袋模型，其本质上是基于 SGNS 训练跨语言向量的扩展。该方法同时优化了每个单词与其中介语上下文(出现在对齐的目标语言句子中的单词)及其语内上下文(原始单语模型中)的相似性。Luong 等[26]提出了双语 Skip-gram 模型联合学习双语和单语学习表征，该模型使用 SGNS 来预测每种语言的周围单词，除了在语言内部进行预测外，还基于句对齐信息进行跨语言预测。Coulmance 等[65]提出了不需要词对齐信息的 Trans-gram 方法，假设语料库已经句对齐，预测源句中的每个单词的上下文单词以及对齐的目标语句中的所有单词，反之亦然。同样，Pham 等[66]通过强制不同语言的对齐语句共享相同的向量表示将段落向量扩展到多语言环境。Rajendran 等[67]基于中枢语言的思想，提出了 Bridge Correlational Neural Networks 模型，该模型需要每种语言和中枢语言之间的平行数据，并且能够在没有任何直接对齐信号的情况下学习两种语言的共享向量空间。Oshikiri 等[68]提出了跨语言特征词方法，该方法将跨语言信息整合到单语特征词中，仅需要句子对齐来捕获跨语言关系。Levy 等[69]基于在平行语料库中所有语言都共享句子的特征空间的事实，使用句子级双语信号，大幅提高了跨语言词向量的性能。

2.3　基于文档对齐可比语料的跨语言词向量模型

2.3.1　基于伪双语文档对齐语料的方法

基于伪双语文档对齐语料的方法通过混合来自文档对齐级别的不同语言文档中的单词，自动构建包含源语言和目标语言单词的伪双语语料库。VuliVulić和 Moens[36]提出了一种合并与随机交换(merge and shuffle)的方法，如图 7 所示，将两种不同语言的对齐文档合并为一个伪双语文档，然后随机打乱伪双语文档中的单词。由于完全随机的变换步骤可能导致次优效果。作者提出了构建伪双语文档的另一种方法，即基于长度比例交换(length-ratio shuffle)的方法，如图 8 所示。即假设两个文档的结构是相似的，根据两个单语文档的长度比例和单词在单语文档中出现的顺序将源语言(浅色框)和目标语言(深色框)的单词依次插入(最初为空)伪双语文档。从而构建跨语言词向量并将其应用于跨语言检索[21]。

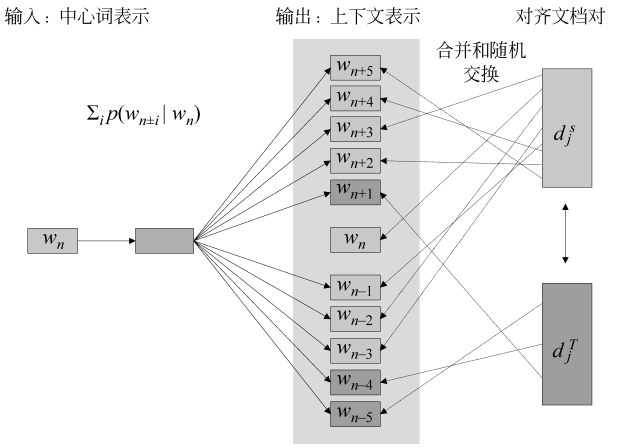

图 7　基于合并与随机交换的方法[36]

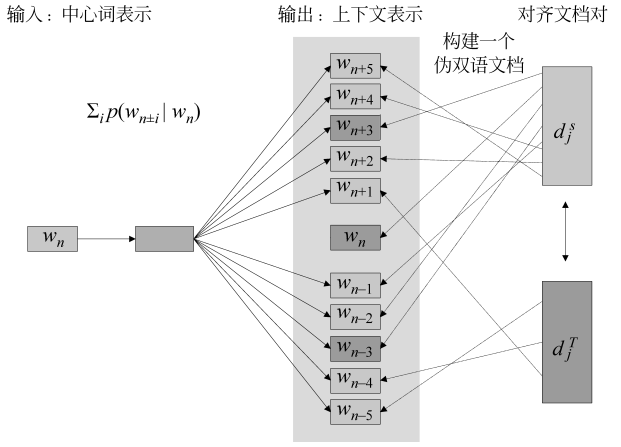

图 8　基于长度比例交换的方法[36]

2.3.2　基于概念的方法

基于概念的方法利用了不同语言中的词语在谈论同一主题或概念时具有一定的相似性这一特性来构建跨语言词向量。Vulić和 Moens[35]基于认知理论，使用多语概率模型，构建跨语言词向量。由于该技术构建的词向量比较稀疏，作者使用了矩阵分解技术来进一步构建稠密向量。Søgaard 等[70]则假设不同语言中描述同一概念的词语基本相同这一思想，借助 Wikipedia 的主题与概念组织结构，使用倒排索引技术来构建跨语言词向量。

2.3.3　基于扩展句对齐模型的方法

该方法在句对齐模型的基础上进行扩展，使用可比语料信息来构建跨语言词向量。Mogadala 和

Rettinger [71]在前人研究的基础上,使用迁移学习技术,将一种语言中学习得到的词向量借助句对齐和文档对齐信息迁移到另外一种语言中,生成跨语言词向量。

2.4 有监督的跨语言词向量模型的分析与比较

基于词对齐平行语料的跨语言词向量模型相对于句级和文档级的模型通常需要昂贵且较为严格的跨语言监督。在实践中,基于伪双语文档构建的方法更昂贵,因为文档中没有词对齐的信息,需从文档中抽取相关信息训练跨语言词向量。相比之下,基于共享空间映射的方法计算效率更高,因为它们充分利用了词对齐信息,采取映射的方法能非常有效地学习跨语言词向量。由于其概念简单易用,颇受研究者的青睐,成为了目前最突出、最流行的方法。

基于句对齐平行语料的跨语言词向量模型,对数据对齐要求介于词级和文档级之间,若没有特别的说明,该模型通常使用 Europarl 语料库,该语料库由欧洲议会议事录中的句子对齐文本组成,常用于训练统计机器翻译模型。在实践中,句子级的监督比词汇级的监督要昂贵得多,即使对于许多低资源的语言,也有双语词典的形式。由于这个原因,近期的工作主要集中在词级的有监督方法学习跨语言向量。然而,词级的监督只允许学习跨语言的单词表示,而对于更复杂的任务,我们通常对跨语言的句对更感兴趣。在文档对齐方面,可比较的文档级对齐更有吸引力,因为它通常更便宜。现有的方法通常使用 Wikipedia 文档,它们要么自动对齐 Wikipedia 文档,要么使用已经与主题对齐的 Wikipedia 文档讨论类似的主题。

近年来,有监督的跨语言词向量模型受到了研究者的广泛重视,因而其研究成果较为丰富。然而,该类方法需要昂贵且规模较大的人工标注训练数据,如双语词典依赖于专家知识和人工归纳,工作量大,代价较高,且领域移植性差。根据前文的论述,表 1 对有监督的跨语言词向量模型进行了大致的归纳总结。

表 1 有监督的跨语言词向量模型对比

对齐语料	方法类别	代表文献	工作原理
词对齐平行语料	基于共享空间映射的方法	[17,23,39-44,46]	独立训练两种语言的词向量,然后使用线性变换将它们映射到跨语言向量共享空间
	基于伪双语文档构建的方法	[24,37,47-49]	通过随机将源语言语料库中的单词替换为其对应目标语言翻译来构建伪双语语料库
	其他方法	[50-53]	
句对齐平行语料	基于矩阵分解的方法	[55-58]	将矩阵分解技术应用到双语词向量的学习中,通常还需要额外的词对齐信息
	组合句模型方法	[59-61]	使用单词表示来构造对齐句的句子表示
	其他方法	[26,62-69]	
文档对齐可比语料	基于伪双语文档对齐语料的方法	[36]	通过混合源文档和目标文档中的单词,自动构建包含源语言和目标语言单词的伪双语语料库
	基于概念的方法	[35,70]	利用文档对齐数据中潜在主题或概念的分布信息来表示单词
	基于扩展句对齐模型的方法	[71]	在句对齐模型的基础上进行扩展,使用可比语料信息来构建跨语言词向量

3 半监督的跨语言词向量模型

3.1 半监督的跨语言词向量模型

以上的有监督方法都是基于较大的双语词典或平行语料库,而对于资源不对等的语言对来说,这些双语数据较难获取。因此,为了减少双语监督需求,研究者们转向使用较少的训练数据来诱导词向量。启发式种子词典是减少双语监督需求的一种切实可行的方法,其可以促进低资源的语言对的诱导。种子词典是通过收集不同语言中相同或拼写相似的单词来构建的,它跨越了最初的双语向量空间。生成种子词典的方法的核心思想是从最初的几个种子词

开始，然后迭代地扩展。传统的基于计数的向量空间模型也有类似的想法，其依赖于共享的单词和同源词[72]。Vulić 和 Korhonen [73]深入分析了种子词典在学习跨语言词向量中的重要性和特点，认为种子双语词典在跨语言或多语言的 NLP 任务中具有重要的应用价值。并提出了一个简单而有效的混合双语词向量模型，该模型仅使用来自种子文档级向量空间的高度可靠的对称翻译对来学习两种单语向量空间之间的映射。实验表明，共享的双语词向量空间可以通过仅利用非常微弱的双语信号（文档对齐）和单语数据来构建。

Zhang 等[74]利用一个小规模的种子词典，从非平行数据出发，在一个共享的语义空间中训练双语词向量。为提高双语词向量的质量，他们通过在学习目标中引入一个匹配的词汇，最大限度地发挥其跨语言配对单词的潜力。作者使用了最少 10 个种子来生成更好的双语词典。为减少双语监督的需要，Smith 等[45]利用正交变换优越的鲁棒性，丢弃训练字典，利用两种语言中出现的相同字符串来形成一个伪词典。实验表明，在没有专家双语信号的情况下使用这个伪词典可以获得双语向量空间。Artetxe 等[22]仅使用了 25 对单词或简单的数字作为种子词典，然后利用词典学习向量映射，之后利用向量映射以自学习方式迭代地生成新的种子词典，从而在没有任何实际双语数据的情况下学习高质量的跨语言词向量。

3.2　有监督与半监督的跨语言词向量模型的分析与比较

有监督的跨语言词向量模型虽然在学习跨语言词向量时能取得较好的效果，但需要大量的人工标注语料，耗时费力。为克服对大规模语料的需要，半监督方法仅需要少量人工标注语料，因此适用于资源缺乏的语言对。然而，该方法存在一个缺点，即对初始的种子词典的质量要求较高，领域迁移时需要构建高质量的种子词典，且种子词典迭代生成的过程中易引入噪声而导致语义漂移。

4　无监督的跨语言词向量模型

4.1　无监督的跨语言词向量模型

前文介绍的跨语言词向量学习方法依赖于大型的双语词典或平行、可比语料库，但这对于资源稀缺的语言是很难获取的。近年来，无需大量标注数据的无监督的跨语言词向量模型受到了研究者们青睐。早期的无监督工作仅仅是对单词的出现信息进行分析[75-76]，后来的研究涉及更复杂的统计数据。Haghighi 等[77]首先利用上下文计数、正字法子串和典型相关分析等特征探索了无监督双语词典的归纳。Barone 和 Cao 等[78-79]基于分布信息，进一步地尝试在没有双语证据的情况下学习双语词向量。前者使用对抗性自动编码器[80]，如图 9 所示，结合一个将源语言向量映射到目标语言空间的编码器，一个从映射向量中重构原始向量的解码器，和一个辨别映射向量与真实目标语言向量的判别器来训练生成双语词向量。而后者则将单语词向量的分布假设为高斯分布，在同时训练单词向量的过程中添加两个正则项来分别约束词向量矩阵的均值和方差，使不同的语言在每个维度的均值和方差相互接近。Zhang 等[81]采用了对抗性训练，并使用对抗性网络[82]进行优化，在没有并行数据的情况下获取跨语言词向量。尽管这些方法听起来很吸引人，但由于没有使用标注语料，这类方法取得的效果都普遍低于有监督方法。

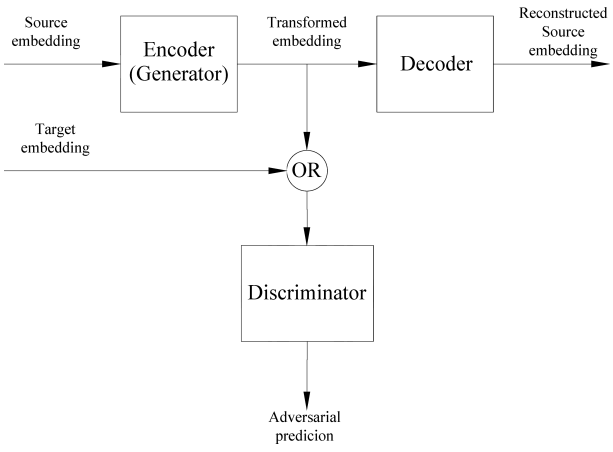

图 9　基于对抗性自动编码器的无监督学习方法[78]

Conneau 等[83]的研究表明，在没有任何跨语言监督的情况下，通过对齐单语词向量空间的方式可以构建一个双语词典。他们使用了两个单语语料库，利用对抗性训练分两步操作来学习从源空间到目标空间的线性映射。首先，训练一个判别器来区分映射的源向量和目标向量，而映射（可以看作是一个生成器）被联合训练来欺骗判别器。其次，从共享向量空间中提取一个合成字典，并使用 Schönemann[84]提出的解决方案对映射进行微调。

Lample 等[85]提出了一种从两种语言的单语语料库中提取句子并将其映射到同一潜在空间的模型。该模型通过学习从这个共享的特征空间中重构两种语言,并利用判别器对源语言和目标语言的潜在分布进行对齐,有效地实现了在不使用任何标注数据的情况下的翻译。

Artetxe 等[28]提出了一种完全无监督的跨语言映射方法,在不需要种子字典的情况下构建初始的单词配对。该方法利用向量的结构相似性以及向量空间结构的初始弱映射,并结合鲁棒自学习方法来学习跨语言词向量。上述三种方法都依赖词向量图近似同构的假设。Søgaard 等[86]针对 Conneau 等[83]的工作,指出了其使用的单语词向量不是近似同构的。为克服这一缺点,他们引入了一个基于拉普拉斯特征值的度量方法来量化词向量的相似性。他们还发现,无监督双语词典归纳的性能在很大程度上取决于这三个因素:语言对、单语语料库的可比性以及词向量算法的参数。

跨语言词向量可通过迁移学习对应词向量空间上的转换函数,建立不同语言词之间的语义映射。Xu 等[29]提出了一种基于迁移学习的无监督学习方法,在给定任意语言对的两种单语单词向量空间的情况下,同时优化两个方向的转换函数,并最小化反向翻译损失。其关键思想为:在两个方向上为每一对语言进行优化映射,将词向量从语言 A 转换为语言 B,从而匹配语言 B 中的单词分布,再将语言 B 转换为语言 A,从最大程度上找到接近原词的向量。如图 10 所示,该模型包括两个由神经网络参数化的映射函数,即 $G:X\rightarrow Y$ 和 $F:Y\rightarrow X$,X 和 Y 分别是两种不同语言对应的单语向量,并作为模型的输入;损失函数包括两个部分:用于匹配转移向量的分布与其目标向量分布的 Sinkhorn 距离[87],以及防止转换变化的反向翻译损失。

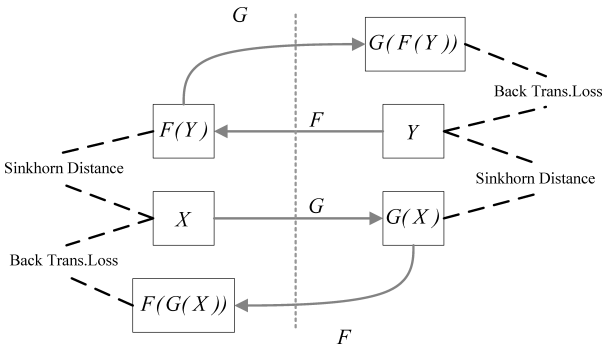

图 10 基于迁移学习的无监督学习方法[29]

4.2 无监督的跨语言词向量模型的分析与比较

对一个新的语种来说,获取大量标注数据来进行有监督训练代价太大,因此如何从已经训练好(或有大量标注数据)的源语言迁移到一个新的语言,使模型在新语言上也能有一个比较好的表现,同时还不需要人工去标注,成为了所要应对的挑战。根据前文所述,有监督和半监督的跨语言词向量模型都需要或多或少的监督信号。而无须任何监督信号的无监督方法逐渐受到了研究者们重视,且部分研究证明无监督方法能获得与有监督方法相媲美的结果。无监督的跨语言词向量模型无须依赖任何人工标注的数据,具有领域无关性,适合处理大规模开放领域数据。对于资源缺乏的语言对来说,是一种较好的处理方法。

5 跨语言词向量的评估及应用

在社会全球化快速发展的今天,人们接触的语言信息越来越多样化和复杂化。特别是在 NLP 领域中,不再局限于英语、汉语、法语、德语、意大利语、荷兰语等资源丰富的语言,资源稀缺的语言如缅甸语、斯瓦西里语等也得到了更多的关注,从而帮助人们跨越语言鸿沟。对于资源信息缺乏的语言,跨语言词向量模型是 NLP 领域中较为流行的方法,它能很好地学习跨语言词表示。由于语言的复杂多样性,所使用的语料库的复杂度及其大小、模型算法的设计等都会使跨语言词向量模型在不同语言上呈现出的效果存在差异。因此,需要可靠的评估方法来对各类跨语言词向量模型进行性能评价。

对于生成的跨语言词向量的质量,通常可以通过直接方法和间接方法进行评估。第一类评估方法通过某种可直接描述词向量质量的指标来评估;第二类方法则使用与词向量的相关 NLP 任务来间接评估词向量的质量。直接方法中较为广泛使用的是词相似度,即用词向量计算两个词对的余弦相似度,然后计算其与人工标注的相似性值的斯皮尔曼等级相关系数,虽然计算简单快速,但是仍存在以下几个缺点:①人为标注的相似性值过于主观;②数据集评估的是语义相似性而不是基于某个任务上的相似性;③没有标准的分割;④没有考虑一词多义。间接评估方法虽然复杂且计算过程缓慢,但一般评价结果较准确。典型方法包括词对齐预测以及双语词典生成评价等。其中,评价双语词典生成的准

确性是一项很有吸引力的评估工作，因为高质量、免费、覆盖面广的手工构建的词典是罕见的。因此，没有针对所有场景的最佳度量方法，最有用的方法是将其与任务的目标相结合，以达到更合理的评估效果。

跨语言词向量作为 NLP 的核心表征，在 NLP 相关的多项任务中都得到了广泛应用。跨语言词向量模型在跨语言文档分类、机器翻译、情感分析、信息检索、命名实体识别、词性标注、超语义标记、依存关系分析等任务中都作为基本模型使用。根据方法的类型和使用的数据类型，不同的方法往往能够很好地完成不同的任务。Upadhyay 等[88]在不同的任务上对不同监督形式的跨语言词向量模型进行了评估。他们发现，在词相似度数据集上，使用弱监督形式（如句子对齐和文档对齐的数据）的模型几乎与使用强监督形式（词对齐）的模型一样好。在语义任务中，使用强跨语言监督（词对齐或句子对齐）训练的单词向量表现得最好，如对跨语言分类和词典生成来说，信息含量越高的监督效果越好。而对于语法任务，较弱形式的跨语言监督的模型（如上下文无关的翻译词典）相对于需要强监督的模型更具竞争力。最后，对于跨语言依存分析，具有词级对齐的模型能够更准确地捕获语法，从而在整体上执行得更好。跨语言词向量模型已经在机器翻译、自动双语词典生成、跨语言信息检索、平行语料库提取和生成以及跨语言剽窃检测等多种任务中发挥了作用。随着跨语言词向量技术的发展，其不仅在 NLP 领域中得到广泛的应用，对其他应用程序也都很有效，比如推荐和链接预测等。

6　未来发展及挑战

随着跨语言词向量建模技术在情感分析、机器翻译、信息检索等领域的成功应用，跨语言词向量技术已逐渐成为跨语言自然语言处理的主流技术之一。但目前跨语言词向量模型大多数都还存在一些需要解决的问题，特别是在资源不对等语言对之间训练跨语言词向量还存在很大的困难。具体原因表现在：第一，单语言下训练的模型无法适用于其他语言；第二，不同语种之间存在特征空间异构；第三，就资源不对等语言对来说，词对齐和句对齐这两类语料库较难获取。而文档对齐语料库如 Wikipedia 等则相对比较容易得到，但数量极度不对等；第四，由于资源稀缺语言数据较少，相对于资源丰富语言来说较难建模。迁移学习[89]是一种可行的方法，但目前在跨语言词向量方面，迁移学习只应用于词对齐和句对齐语料，在文档对齐语料方面还未有研究。表 2 根据前文对三类方法的分析与比较进行了总结。

表 2　跨语言词向量模型对比

跨语言词向量模型	有监督方法	半监督方法	无监督方法
数据需求	基于双语词典、句子对齐语料、文档对齐语料	仅需少量人工标注数据	无需任何人工标注数据
人工干预程度	强	中	弱
领域移植性	弱	中	强
性能提升方法	改进跨语言词向量算法，改进映射方法，扩展特征	改进种子词典的生成方法	改进假设（单语词向量近似同构）

总体说来，跨语言词向量模型的应用还较少，研究也比较初步，同时伴随着新的挑战，需要进一步探讨和深入研究，特别是有如下问题亟待解决：

（1）子词级信息应用。虽然其也被用于学习词表示，但到目前为止还没有被纳入跨语言词汇表征的学习中。

（2）多词表达。其是由单词非组合方式组合起来形成的，处理多词表达仍然是一个挑战，且在跨语言环境中很少受到关注。

（3）语言功能。模型学习的跨语言表示与其他语言的向量空间模型存在一个弱点，即不能恰当地模型化语言的含义，如区分“Give me a pencil”和“Give me that pencil”。语言功能方面的建模在对话等场景中尤为重要，在这些场景中必须考虑语言的语用学。

（4）一词多义。在跨语言向量空间中，一词多义这个问题较为突出，词汇歧义仍悬而未决。

（5）语料库的获取。目前提出的大多数跨语言词向量模型都是基于平行语料库的。然而，对于资源稀缺的语言对，平行或可比语料库是很难获取的。因此，用尽可能少的语料创建健壮的跨语言单词表示是一个重要的研究途径。一个重要的相关方向是利用可比性语料库，这些语料库通常更丰富，并包含如来自多模态上下文的其他信号。

另外，在跨语言词向量算法设计方面，主要面临数据规模增大、数据表示稀疏、单词位置、参数设置以及模型训练速度等挑战。因此，为了设计更加健

壮的跨语言词向量模型,有必要开发能够有效处理大量文本数据并使词向量更具表现力的技术,以适应新的应用需求。

7 结束语

词向量作为自然语言处理任务中的核心表征技术,它能够很好地捕获语言中的规律,并利用此规律在不同的语言间进行知识转移。依赖于大量的人工标注语料有监督的跨语言词向量模型的研究成果较为丰富,其中基于词对齐平行语料、句子对齐平行语料的研究相对较多,文档对齐可比语料方面研究则比较稀缺。半监督的跨语言词向量模型只需少量的人工标注数据,但好的种子词典较难生成。最近兴起的无监督的跨语言词向量模型无需依赖标注语料,因而受到较多研究者的关注。基于前人的工作,本文对跨语言词向量的三种主流方法进行了描述与比较,并对跨语言词向量模型所面临的挑战和未来研究方向进行了研讨。

参考文献

[1] Mikolov T,Sutskever I,Chen K,et al.Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2 (NIPS),2013: 3111-3119.

[2] Socher R,Perelygin A,Wu J,et al.Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,2013: 1631-1642.

[3] Dyer C,Ballesteros M,Ling W,et al.Transition-based dependency parsing with stack long short-term memory[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),2015: 334-343.

[4] Bahdanau D,Cho K,Bengio Y.Neural machine translation by jointly learning to align and translate[J].CoRR,2015,abs/1409.0473.

[5] Harris Z S.Distributional structure[J].Word.1954,10(2-3): 146-162.

[6] Firth J R.A synopsis of linguistic theory,1930-1955[J].Studies in Linguistic Analysis (Special volume of the Philological Society).Oxford University Press,London,1957: 1-32.

[7] Turian J,Lev R,Bengio Y.Word representations: A simple and general method for semi-supervised learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL),2010: 384-394.

[8] Fernando P,Tishby N,Lee L.Distributional clustering of English words[C]//Proceedings of the 31st Annual Meeting on Association for Computational Linguistics,1993: 183-190.

[9] Deerwester S,Dumais S T,Furnas G W,et al.Indexing by latent semantic analysis[J].Journal of the Association for Information Science & Technology,1990,41(6): 391-407.

[10] Brown P F,Desouza P V,Mercer R L,et al.Class-based n-gram models of natural language[J].Computational Linguistics,1992,18(4): 467-479.

[11] Bengio Y,Ducharme R,Vincent P,et al.A neural probabilistic language model[J].Journal of Machine Learning Research,2003: 1137-1155.

[12] Collobert R,Weston J.A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine Learning,2008,ACM: 160-167.

[13] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[J].CoRR,2013,abs/1301.3781.

[14] 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.

[15] 何炎祥,孙松涛,牛菲菲,等.用于微博情感分析的一种情感语义增强的深度学习模型[J].计算机学报,2017,40(4): 773-790.

[16] Ruder S,Vulic I,Søgaard A.A survey of cross-lingual word embedding models[J].Journal of Artificial Intelligence Research,2018: 1-55.

[17] Mikolov T,Le Q V,Sutskever I.Exploiting similarities among languages for machine translation[J].CoRR,2013: abs/1309.4168.

[18] 王进,陈恩红,张振亚,等.基于本体的跨语言信息检索模型[J].中文信息学报,2004,18(3): 2-9.

[19] Heyman G,Vulić I,Moens M F.Bilingual lexicon induction by learning to combine word-level and character-level representations[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,2017: 1085-

1095.

[20] Tsai C T, Roth D. Cross-lingual wikification using multilingual embeddings [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 589-598.

[21] Vulić I, Moens M F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings [C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, ACM: 363-372.

[22] Mikel A, Gorka L, Eneko A. Learning bilingual word embeddings with (almost) no bilingual data [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 451-462.

[23] Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation [C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014: 462-471.

[24] Gouws S, Søgaard A. Simple task-specific bilingual word embeddings [C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015: 1386-1390.

[25] Zou W Y, Socher R, Cer D, et al. Bilingual word embeddings for phrase-based machine translation [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1393-1398.

[26] Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind [C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, 2015: 151-159.

[27] Vulic I, Moens M-F. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 719-725.

[28] Artetxe M, Labaka G, Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings [C]//Proceedings of the ACL, 2018: 789-798.

[29] Xu R, Yang Y, Otani N, et al. Unsupervised cross-lingual transfer of word embedding spaces [C].//Proceedings of the EMNLP, 2018: 2465-2474.

[30] 廖祥文，刘德元，桂林，等. 融合文本概念化与网络表示的观点检索 [J]. 软件学报，2018，29(10)：7-22.

[31] 庞亮，兰艳艳，徐君，等. 深度文本匹配综述 [J]. 计算机学报，2017，40(4)：985-1003.

[32] Levy O, Søgaard A, Goldberg Y. A strong baseline for learning cross-lingual word embeddings from sentence alignments [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017: 765-774.

[33] Kiela D, Vulić I, Clark S. Visual bilingual lexicon induction with transferred ConvNet features [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 148-158.

[34] Gella S, Sennrich R, Keller F, et al. Image pivoting for learning multilingual multimodal representations [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017: 2839-2845.

[35] Vulić I, Moens M F. Cross-lingual semantic similarity of words as the similarity of their semantic word responses [C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 106-116.

[36] Vulić I, Moens M F. Bilingual distributed word representations from document-aligned comparable data [J]. Journal of Artificial Intelligence Research, 2016, 55(1): 953-994.

[37] Xiao M, Guo Y. Distributed word representation learning for cross-lingual dependency parsing [C]//Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 2014: 119-129.

[38] Shigeto Y, Suzuki I, Hara K, et al. Ridge regression, hubness, and zero-shot learning [C]//Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015: 135-151.

[39] Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem [C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2015.

[40] Shigeto Y, Suzuki I, Hara K, et al. Ridge regression, hubness, and zero-shot learning [C]//Proceedings of the ECML PKDD, 2015: 135-151.

[41] Lu A, Wang W, Bansal M, et al. Deep multilingual correlation for improved word embeddings [C]//Pro-

ceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015: 250-256.

[42] 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.

[43] Aretetxe M, Labaka G, Agirre E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2289-2294.

[44] Zhang Y, Gaddy D, Barzilay R, et al. Ten pairs to tag-Multilingual POS tagging via coarse mapping between embeddings [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1307-1317.

[45] Smith S L, Turban D H, Hamblin S, et al. Offline bilingual word vectors, orthogonal transformations and the inverted softmax[C]//Proceedings of the ICLR, 2017.

[46] Lazaridou A, Dinu G, Baroni M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 270-280.

[47] Ammar Waleed, Mulcaire George, Yulia Tsvetkov, et al. Massively multilingual word embeddings [C]// Proceedings of the CoRR, 2016.

[48] Duong L, Kanayama H, Ma T, et al. Learning crosslingual word embeddings without bilingual corpora [C]//Proceedings of the EMNLP, 2016: 1285-1295.

[49] Adams O, Cohn T, Bird S, et al. Cross-lingual word embeddings for low-resource language modeling [C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017: 937-947.

[50] Kočiský T, Hermann K M, Blunsom P. Learning bilingual word representations by marginalizing alignments[C]//Proceedings of the ACL, 2014: 224-229.

[51] Tianze S, Liu Z, Liu Y, et al. Learning cross-lingual word embeddings via matrix co-factorization [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 567-572.

[52] Church K W, Hanks P. Word association norms, mutual information, and lexicography [J]. Journal of Computational Linguistics. 1990, 16(1): 22-29.

[53] Michael Wick, Pallika Kanani, Pocock A. Minimally-constrained multilingual embeddings via artificial code-switching[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 2849-2855.

[54] Shi T, Liu Z, Liu Y, et al. Learning cross-lingual word embeddings via matrix co-factorization[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 567-572.

[55] Will Y Zou, Socher R, Cer D M, et al. Bilingual word embeddings for phrase-based machine translation [C]//Proceedings of EMNLP, 2013: 1393-1398.

[56] Huang K Gardner, M, Papalexakis E, et al. Translation invariant word embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1084-1088.

[57] Vyas Y, Carpuat M. Sparse bilingual word representations for cross-lingual lexical entailment [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1187-1197.

[58] Jiang Guo, Che W, Yarowsky D, et al. Cross-lingual dependency parsing based on distributed representations[C]//Proceedings of the ACL, 2015: 1234-1244.

[59] Moritz H K, Phil B. Multilingual distributed representations without word alignment[C]//Proceedings of the ICLR (Conference Track), 2013.

[60] Hermann K M, Blunsom P. Multilingual models for compositional distributed semantics[C]//Proceedings of the ACL, 2014: 58-68.

[61] Soyer H, Stenetorp P, Aizawa A. Leveraging monolingual data for crosslingual compositional word representations[C]//Proceedings of the CoRR, 2015: abs/1412.6334.

[62] Lauly S, Boulanger A, Larochelle H. Learning multilingual word representations using a bag-of-words autoencoder[C] Proceedings of the CoRR, 2014: abs/1401.803.

[63] Chandar A S, Stanislas L, Hugo L, et al. An autoen-

coder approach to learning bilingual word representations[C]//Proceedings of the NIPS, 2014: 1853-1861.

[64] Gouws S, Bengio Y, Corrado G. Bilbowa: Fast bilingual distributed representations without word alignments[C]//Proceedings of the ICML, 2015: 748-756.

[65] Coulmance J, Marty J-M, Wenzek G, et al. Transgram, fast cross-lingual word-embeddings[C]//Proceedings of the EMNLP, 2016: 1109-1113.

[66] Pham H Q, Luong T, Manning C D. Learning distributed representations for multilingual text sequences [C]//Proceedings of the Learning Distributed Representations for Multilingual text Sequences, 2015: 88-94.

[67] Rajendran J, Khapra M M, Chandar A P S, et al. Bridge correlational neural networks for multilingual multimodal representation learning[C]//Proceedings of the NAACL-HLT, 2016: 171-181.

[68] Oshikiri T, Fukui K, Shimodaira H. Cross-lingual word representations via spectral graph embeddings [C]//Proceedings of the ACL, 2016: 493-498.

[69] Levy O, Søgaard A, Goldberg Y. A strong baseline for learning cross-lingual word embeddings from sentence alignments[C]//Proceedings of the EACL, 2017: 765-774.

[70] Søgaard A, Agić Ž, Alonso H M, et al. Inverted indexing for cross-lingual NLP[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP), 2015: 1713-1722.

[71] Mogadala A, Rettinger A. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 692-702.

[72] Peirsman Y, Padó S. Cross-lingual induction of selectional preferences with bilingual vector spaces[C]//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010: 921-929.

[73] Vulić I, Korhonen A. On the role of seed lexicons in learning bilingual word embeddings[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 247-257.

[74] Zhang M, Peng H, Liu Y, et al. Bilingual lexicon induction from non-parallel data with minimal supervision[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence 2017: 3379-3385.

[75] Rapp R. Identifying word translations in non-parallel texts[C]//Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995: 320-322.

[76] Fung P. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus[C]//Proceedings of the Third Workshop on Very Large Corpora, 1995: 173-183.

[77] Haghighi A, Liang P, Berg-Kirkpatrick T, et al. Learning bilingual lexicons from monolingual corpora [C]//Proceedings of the ACL-08: Hlt, 2008: 771-779.

[78] Barone A V M. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders[C]//Proceedings of the Rep4-NLP@ACL, 2016: 121-126.

[79] Cao H, Zhao T, Zhang S, et al. A distribution-based model to learn bilingual word embeddings[C]//Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 1818-1827.

[80] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders[J]. CoRR, 2015: abs/1511.05644.

[81] 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.

[82] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the Neural Information Processing Systems, 2014: 2672-2680.

[83] Conneau A, Lample G, Ranzato M A, et al. Word translation without parallel data[J]. CoRR, 2018: abs/1710.04087.

[84] Schönemann P H P. A generalized solution of the orthogonal procrustes problem[J]. Psychometrika, 1966: 1-10.

[85] Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only [C]//Proceedings of the ICLR, 2017.

[86] Søgaard A, Ruder S, Vulić I. On the limitations of unsupervised bilingual dictionary induction[C]//Proceedings of the ACL, 2018: 778-788.

(下转第 26 页)

于东(1982—),通信作者,博士,副教授,主要研究领域为自然语言处理。
E-mail:yudong_blcu@126.com

吴思远(1998—),硕士研究生,主要研究领域为第二语言习得、自然语言处理。
E-mail:wusiyuan2401@163.com

耿朝阳(1996—),硕士研究生,主要研究领域为自然语言处理。
E-mail:yangican@163.com

(上接第 15 页)

[87] Cuturi M.Sinkhorn distances:Lightspeed computation of optimal transport[C]//Proceedings of the Advances in Neural Information Processing Systems,2013:2292-2300.

[88] Shyam Upadhyay M F,Chris Dyer,Dan Roth.Cross-lingual models of word embeddings:An empirical comparison[J].ArXiv.2016,abs/1604.00425.

[89] Lu J,Behbood V,Hao P,et al.Transfer learning using computational intelligence:A survey[J].Knowledge-Based Systems,2015,80:14-23.

彭晓娅(1996—),硕士研究生,主要研究领域为信息检索、自然语言处理。
E-mail:302834020@qq.com

周栋(1979—),通信作者,博士,教授,主要研究领域为信息检索、自然语言处理。
E-mail:dongzhou1979@hotmail.com