中国科学: 信息科学 2018 年 第 48 卷 第 5 期: 564–573
SCIENTIA SINICA Informationis
人工智能优青专刊 · 论文

基于非平行语料的双语词典构建

张檬[1,2,3], 刘洋[1,2,3*], 孙茂松[1,2,3]

1. 清华大学计算机科学与技术系, 北京 100084
2. 清华大学智能技术与系统国家重点实验室, 北京 100084
3. 北京信息科学与技术国家研究中心, 北京 100084
* 通信作者. E-mail: liuyang2011@tsinghua.edu.cn

收稿日期: 2017–11–27; 接受日期: 2018–01–26; 网络出版日期: 2018–05–11

国家自然科学基金优秀青年项目 (批准号: 61522204) 资助

摘要 在进行跨语言自然语言处理时, 缺少双语资源是非常棘手的问题, 而这在语言资源匮乏的场景下是非常普遍的. 此时, 利用好非平行语料中蕴含的翻译知识变得更为重要. 由于语料不平行, 从中获取翻译知识意味着小数据、无监督学习, 因此极具挑战, 而取得的结果通常是双语词典的形式. 这既是人工智能领域重要的学术问题, 也在语言资源匮乏场景有着巨大的应用价值. 本文针对前人研究中存在的问题, 介绍一系列工作, 从各个角度探索如何更好地利用非平行语料构建双语词典.

关键词 双语词典构建, 非平行语料, 双语词向量, 跨语言自然语言处理, 人工智能

1 引言

伴随着全球化的进行, 国际交流日渐频繁, 人们开始需要越来越多种语言的翻译. 我国 “一带一路” 倡议的提出, 更是对沿线国家的语言处理带来了巨大的实际需求. 然而, 对于许多小语种而言, 不仅翻译人才匮乏, 自动的机器翻译性能也很差. 这主要是因为机器翻译的性能非常依赖于平行语料的规模和质量. 所谓平行语料, 是指语料中两种语言的文本是互为翻译的. 然而, 对于语言资源匮乏的小语种而言, 平行语料是非常稀缺甚至不存在的. 实际上, 对于大语种而言, 平行语料往往也只存在于部分领域之中, 对于许多专门的领域, 这种珍贵的语言资源很多时候也难以获得.

在小语种和专门领域这样的语言资源匮乏场景下, 相比而言, 非平行语料是容易获得且较为丰富的语言资源. 因此, 如何从非平行语料中获取翻译知识成为相关研究人员广泛关注的问题. 这个研究课题在学术和应用两方面均有重要意义. 从学术意义上讲, 翻译知识的获取在非平行语料的条件下意味着小数据乃至无监督的场景, 而小数据学习和无监督学习是人工智能和深度学习的重要问题, 在跨语言自然语言处理中也是如此. 从应用意义上讲, 全球化的深化、“一带一路” 的倡议, 使得语言资源匮乏场景下基于非平行数据获取翻译知识成为越发迫切的重要任务.

引用格式: 张檬, 刘洋, 孙茂松. 基于非平行语料的双语词典构建. 中国科学: 信息科学, 2018, 48: 564–573, doi: 10.1360/N112017-00256
Zhang M, Liu Y, Sun M S. Bilingual lexicon induction from non-parallel corpora (in Chinese). Sci Sin Inform, 2018, 48: 564–573, doi: 10.1360/N112017-00256

图 1 (网络版彩图) 基于非平行语料构建双语词典

Figure 1 (Color online) Bilingual lexicon induction from non-parallel corpora

不过, 由于机器翻译系统在大规模平行语料上训练得到的效果还有待提高, 在非平行语料的条件下, 直接训练机器翻译系统显得不太可行. 所以, 研究人员主要关注利用非平行语料构建双语词典, 换言之, 基于非平行语料获取词汇级别的翻译知识. 图 1 是此任务的示意图. 可见, 两种语言的文本并不互为翻译, 但其中仍然存在有价值的词汇翻译知识可供挖掘, 如图中相同颜色的色块对应的词对, 只是在语料非平行的条件下, 此任务极具挑战.

双语词典提供词汇语义的跨语言等价信息, 是一种重要的双语资源. 它不仅易于人工编辑和使用, 而且对于许多跨语言自然语言处理任务也很有帮助, 例如跨语言信息检索 [1]、机器翻译 [2]、跨语言标注投射 [3] 等. 在语言资源匮乏的场景下, 可供利用的双语资源很少, 双语词典常常是为数不多的一种选项, 而且往往存在着质量低、规模小的问题, 有时甚至连这样的双语资源也完全没有. 在这种情况下, 利用非平行语料构建、扩充、改善双语词典 (统称双语词典构建), 通常是开展后续跨语言自然语言处理的必要步骤.

2 相关工作

双语词典构建是一个研究历史较为长久的跨语言自然语言处理任务, 在 20 世纪末即有研究开展. 研究思路大体上可以分为 3 步:

- 第一步, 将两种语言中的每个词表示为向量;
- 第二步, 为两种语言的向量空间建立联系, 得到共有的双语向量空间;
- 第三步, 在双语向量空间中进行查找, 获取双语词典.

在第一步中, 早期的方法一般利用单语语料上的统计量如 PMI 作为向量表示 [4,5]. 近年来, 伴随着深度学习的研究热潮, 基于神经网络模型学习得到的词表示 (通称词向量) [6] 得到了广泛应用. 这种词向量表示也为双语词典构建的方法打开了新的思路. 具体而言, 词向量在这个任务中可有以下两种使用方式.

第一种方式是将上述第一步中的向量表示换用基于神经网络模型的词向量 [7~13]. 这种方式以

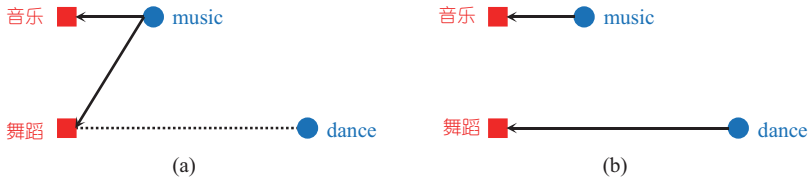


图 2 (网络版彩图) Hubness 问题

Figure 2 (Color online) The hubness problem. (a) The nearest neighbor suffers from the hubness problem and produces wrong translation; (b) the earth mover's distance is able to produce correct translation for this example

Mikolov 等 [7] 在 Google 完成的工作为代表: 他们使用 word2vec [6] 在单语语料上训练得到词向量, 完成第一步, 随后使用一个种子词典完成第二步中两种语言向量空间的建立.

第二种方式是将第一步和第二步联合起来, 直接训练神经网络模型得到共有的双语词向量空间. 训练过程中为了联系两种语言的词向量空间, 通常需要使用双语监督信号, 这种监督信号的形式可能是篇章级别对齐的 [14], 句子级别对齐的 [15~22], 或词级别对齐的 (即种子词典) [23~26]. 这种方式以 Gouws 等 [19] 在 Google 完成的工作为代表: 他们使用句子级别对齐的平行语料, 并为平行句对做均匀的词对齐假设, 以此作为联系两种语言词向量空间的监督信号.

本文将主要关注基于词向量且不使用平行语料的双语词典构建方法. 虽然新兴的基于词向量的方法极大地推进了双语词典构建的研究, 但是在构建过程的各个环节仍然存在有待解决的挑战, 包括 hubness 问题、一词多译问题、双语监督信号缺乏问题. 第 3 节具体介绍这些问题, 并针对性地介绍应对这些挑战的方案. 具体而言, 第 3.1 小节针对 hubness 问题, 第 3.2 小节针对一词多译问题, 第 3.3~3.5 小节针对双语监督信号缺乏, 乃至完全无监督的场景.

3 研究内容

3.1 基于 earth mover's distance 的词汇翻译查找

这份工作针对双语词典构建的第三步中存在的所谓 hubness 问题. 在这一步中, 我们已经获得了共有的双语词向量空间, 现在需要从中获取双语词典. 这一步通常是采用最近邻查找完成的: 对于一个源语言词向量, 在目标语言词向量空间中查找与之最近的词向量作为翻译结果. 然而, 最近邻查找常常遇到如图 2(a) 所示的问题. 图中的示例双语词向量空间展示了两个源语言中文词向量和两个目标语言英文词向量, 分别用方块和圆点表示. 当用最近邻对源语言的 "音乐" 和 "舞蹈" 进行查找时, 返回的翻译结果均为 "music", 此时 "舞蹈" 翻译错误, 而 "music" 被称为 "hub", 这就是 hubness 问题 [10,11]. 这个问题在双语词向量空间中广泛存在, 但由于最近邻查找本质上是一种局部的操作, 因此不可避免地受到这个问题的影响.

如何才能获得如图 2(b) 所示的正确翻译呢? 为了避免最近邻查找的局部性, 我们考虑进行全局的带权匹配, 其基本思想可以借由下面的比喻加以说明. 将图中的圆点视为土堆, 方块视为坑洞, 它们的大小代表土堆的体积和坑洞的容积, 或者说相应的权重. 在图 2 的示例中, 所有的权重都相等. 在这个设定下, 我们希望用最小的整体代价来移动土堆填满坑洞, 而代价是由移动土堆的距离和体积的乘积衡量的. 可以想见, 图 2(b) 中的箭头即代表了这个示例下的最优移动方案, 而这个方案正好可以视为词汇翻译的结果. 从微观看, 由于 "music" 土堆中的泥土已经全部用来填 "音乐" 坑洞, 它将不会去干涉 "舞蹈" 坑洞, 从而由 "dance" 土堆负责填满 "舞蹈" 坑洞. 从宏观看, 整体移动代价的最小化使得

我们可以考虑全局的信息, 从而克服最近邻查找的局部性, 应对 hubness 问题.

上述比喻代表的全局带权匹配思想在数学上可以用 earth mover's distance (EMD) 来实现, 它的名字正是来源于上述的比喻. 其对应如下的线性规划问题:

$$\begin{aligned} &\min \sum_{i=1}^{V_t}\sum_{j=1}^{V_s} W_{ij}C_{ij} \\ &\text{s.t. } W_{ij} \geqslant 0, \\ &\quad \sum_{j=1}^{V_s} W_{ij} \leqslant t_i,\ i \in \{1,\ldots,V_t\}, \\ &\quad \sum_{i=1}^{V_t} W_{ij} = s_j,\ j \in \{1,\ldots,V_s\}, \end{aligned} \tag{1}$$

其中, V_s 代表源语言词汇表大小, V_t 代表目标语言词汇表大小, C_{ij} 代表第 i 个土堆与第 j 个坑洞之间的距离, t_i 代表第 i 个土堆的体积, s_j 代表第 j 个坑洞的容积, W_{ij} 为优化问题的决策变量, 代表从第 i 个土堆转移到第 j 个坑洞的泥土体积, 因此, 目标函数即为最小化整体的移动代价. 求解完成后, 非零的 W_{ij} 值即代表第 j 个源语言词与第 i 个目标语言词之间存在翻译关系.

实验发现, 使用 EMD 进行词汇翻译可以取得比最近邻更好的效果 [27].

3.2 基于 earth mover's distance 正则化的一词多译模型

一词多译指的是一个源语言词在目标语言中存在着多种可能的翻译. 一词多译现象在自然语言之间是广泛存在的 [28], 实际上, 在两种语言的词汇之间, 甚至可能存在着复杂的多对多的对应关系.

尽管一词多译现象很常见, 之前的双语词典构建方法通常在建模时不予考虑, 而是将问题直接简化为一对一翻译的情形. 在查找词汇翻译阶段, 如果需要返回多个词汇翻译候选, 则使用 K 近邻查找, 但翻译候选数量 K 须事先指定并且是固定的. 可见, 一词多译现象未能得到很好的处理.

在使用 EMD 处理 hubness 问题时, 我们还发现它表现出处理一词多译的能力: 对于每个源语言词, 它能自动确定翻译候选的数目, 而不像 K 近邻总是返回 K 个翻译候选. 这个效果可以从图 3 中看出. 在这个例子中, 中文词 "汽车" 存在着 "automobile" 和 "car" 两个英文翻译, 而 "car" 又对应着中文词 "汽车" 和 "车厢". 此时, 如果我们为中文词 "汽车" 和英文词 "car" 赋予较大的权重, 如图中方块和圆点的大小所示, 则 EMD 求解之后即得图中箭头所示的正确翻译结果.

前面我们使用 EMD 时, 假定已经获得了双语词向量空间, 而将 EMD 用在双语词典构建的第三步来取代最近邻查找. 为了能够更好地发挥 EMD 处理一词多译现象的能力, 我们提出将 EMD 引入双语词向量的训练过程中. 在训练的目标函数中, EMD 作为其中一项以正则的形式参与训练, 使得训练得到的双语词向量能够更好地捕捉一词多译现象. 它的效果通过实验得到了印证 [29].

3.3 基于隐变量的双语匹配模型

一般来说, 为了获得双语词向量空间, 需要某种形式的双语监督信号把两种语言联系起来, 常用的双语监督信号以种子词典的形式存在. 尽管我们希望利用非平行语料构建双语词典, 但往往仍然需要使用双语资源, 有的时候使用量还不小. 这与我们使用非平行语料的初衷是违背的, 因为在资源匮乏的场景下, 可供利用的双语资源非常有限, 这就是双语监督信号缺乏问题.

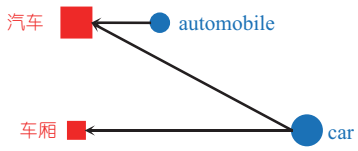


图 3 (网络版彩图) Earth mover's distance 处理一词多译现象

Figure 3 (Color online) Multiple alternative translation based on earth mover's distance

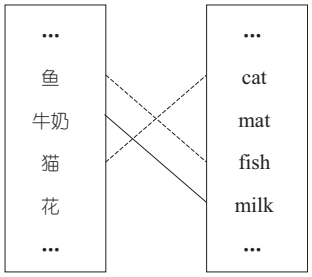


图 4 基于隐变量的双语匹配模型

Figure 4 Bilingual matching model based on latent variables

对此, 我们希望尽可能减小对于种子词典的依赖, 或者说, 给定一个规模有限的种子词典, 我们希望对其加以充分利用, 使得双语词向量空间能够更好地建立起来. 其基本思想是这样的: 首先利用初始种子词典为两种语言的词向量空间建立初步的联系, 这个过程一般可以引出一些新的较为可靠的翻译词对, 如果可以将这些新的翻译词对利用起来, 应该可以更好地联系两种语言的词向量空间, 从而可以引出更多的翻译词对, 如此迭代下去. 在图 4 中, 实线代表初始给定的种子翻译词对, 虚线代表潜在的可供发掘利用的翻译词对. 我们将上述想法中潜在的翻译词对表示为隐变量, 建立起一个隐变量模型, 其学习过程可以采用 Expectation-Maximization (EM) 式的迭代训练, 这个过程恰好对应着上述的直观想法. 我们将这个模型称为基于隐变量的双语匹配模型. 实验发现, 仅用 10 组种子翻译词对就能成功将两种语言的词向量空间联系起来 [30].

3.4 基于对抗学习的无监督双语词典构建

沿着监督信号缺乏的角度进一步深入, 我们考虑是否可能完全不使用任何双语监督信号, 仅仅利用非平行语料构建双语词典. 这不仅对完全缺乏双语资源的低资源语言打开了连接其他语言的可能, 而且暗示了人类语言表示词汇概念可能存在着普遍结构. 初看起来, 只有两种语言的单语语料, 想要学到双语的词汇关联似乎是不可能的. 然而, 这实际上是可行的. 首先, 观察图 5, 图中所示的西班牙语和英语的词向量是分别独立地在各自语言的单语语料上训练得到的, 不过可以看到, 两种语言的单语词向量空间表现出近似的同态性, 这意味着存在线性映射能够近似地连接这两个空间. 前人工作 [7] 利用种子词典来学习这个线性映射, 而现在, 我们希望完全不使用双语监督信号, 因此需要设计一种方法来学习这个映射, 而这个方法不能依赖于种子翻译词对这种级别的监督信号. 我们的灵感来自于生成对抗网络 [31], 把词向量的跨语言映射学习建模成一个对抗游戏.

图 6(a) 展示了对抗学习的基本结构. 图 6 中, 方块代表源语言词向量, 圆点代表目标语言词向量. 整个模型由互相对抗的两部分组成, 即生成器 G 和鉴别器 D. 生成器 G 负责学习联系两个语言空间的线性映射, 它的目标是将源语言词向量映射到目标语言空间后, 与目标语言词向量难以区分. 鉴别器 D 负责鉴别词向量是由生成器 G 映射得来的, 还是真正的目标语言词向量, 它的目标是希望尽可能鉴别准确. 可见, 生成器 G 和鉴别器 D 的目标是互相对抗的. 这样的对抗目标可以表示为 $\min_G \max_D V(D,G)$ 的形式, 其值函数 $V(D,G)$ 为

$$V(D,G) = \mathbb{E}_{y\sim p_y}[\log D(y)] + \mathbb{E}_{x\sim p_x}[\log(1 - D(G(x)))], \tag{2}$$

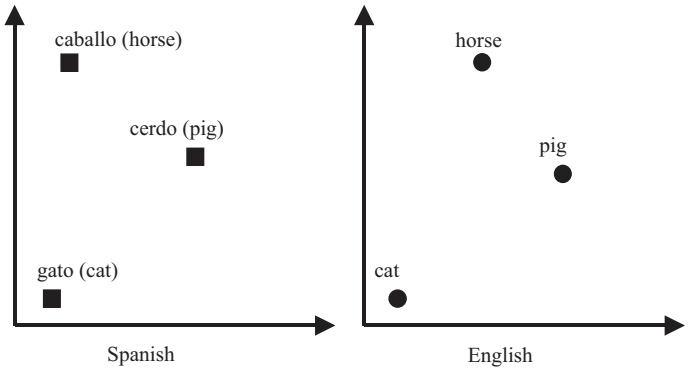


图 5 西班牙语和英语的单语词向量空间

Figure 5 Monolingual word embedding spaces of Spanish and English

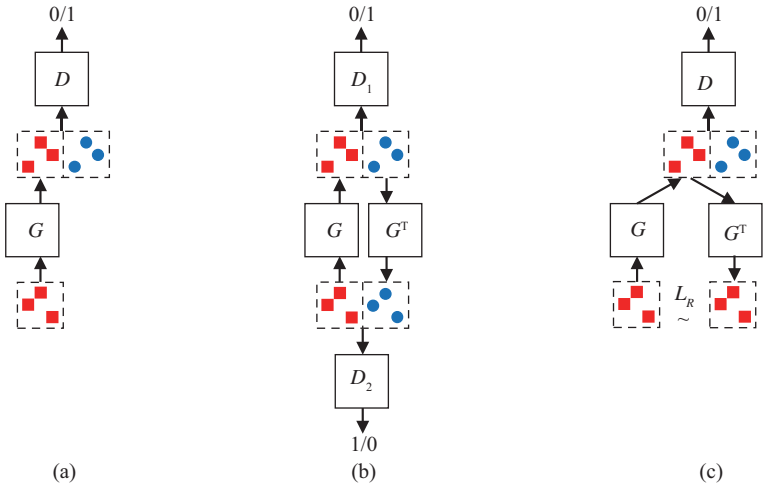


图 6 (网络版彩图) 对抗学习

Figure 6 (Color online) Adversarial training. (a) Unidirectional transformation model; (b) bidirectional transformation model; (c) adversarial autoencoder model

其中, x 为源语言词向量, p_x 为源语言词向量服从的分布, y 与 p_y 则代表目标语言.

据此, 可以写出鉴别器 D 的损失函数为

$$L_D = -\log D(y) - \log(1 - D(Gx)). \tag{3}$$

为了简化符号, 这里写出的损失函数中仅有一对源语言与目标语言词向量.

生成器 G 的损失函数为

$$L_G = -\log D(Gx). \tag{4}$$

此形式与原始形式 $\log(1 - D(Gx))$ 有所不同, 因为它相对较易训练.

如果训练成功, 生成器 G 学到的线性映射能够较好地联系两个语言空间, 而鉴别器 D 分类的准确率则较低. 然而, 与通常的生成对抗网络类似, 图 6(a) 所示的基本模型训练起来也非常困难. 因此, 我们设计了一系列改进的模型, 如图 6(b) 和 (c) 所示.

在图 6(b) 中, 我们不仅希望生成器 G 较好地将源语言词向量映射到目标语言空间, 还希望它的

转置 G^{T} 将目标语言词向量映射回源语言空间, 而在两个空间中各有一个鉴别器负责鉴别. 两个鉴别器的损失函数与前类似, 而生成器的损失函数为

$$L_G = -\log D_1(Gx) - \log D_2\left(G^{\mathrm{T}}y\right). \tag{5}$$

在图 6(c) 中, 源语言词向量经由生成器 G 映射到目标语言空间后, 又由其转置 G^{T} 映射回源语言空间, 最终的词向量应与初始的源语言词向量尽可能相似, 相似程度可由余弦相似度衡量. 在这个模型下, 生成器的损失函数为

$$L_G = -\log D(Gx) - \lambda\cos\left(x, G^{\mathrm{T}}Gx\right), \tag{6}$$

其中 λ 是平衡目标函数中两项的超参数.

此外, 我们还探索了相关的训练技术, 使得训练能够顺利进行. 最终, 实验成功实现了不使用任何双语监督信号联系两种语言的词向量空间, 使得单纯的基于非平行语料的双语词典构建成为可能 [32].

3.5 基于 earth mover's distance 最小化的无监督双语词典构建

前面, 面对无监督地为两种语言的词向量空间建立联系的问题, 我们通过建模为一个直观的对抗游戏加以解决. 接下来, 尝试从更加数学化的角度进行建模. 这不仅使我们可以从另一个角度看待对抗学习, 也为无监督联系双语词向量空间提供了新的可能.

面对两个有待联系的词向量空间, 本文的目标是在不使用种子词典的条件下, 寻找一个线性映射使这两个空间中的词向量能较好地对齐. 种子词典中包含的是词汇级别的跨语言监督信号, 为了避免使用这种级别的监督信号, 我们考虑使用词汇表级别的准则来指导线性映射的学习. 为此, 我们的想法是将词向量视为概率分布, 将分布之间的距离作为词汇表级别的准则. 这样的建模方式使我们可以考虑各种形式的分布距离. 实际上, 前面对抗学习的方法也可以放在这个框架下看待, 因为对抗学习隐式地优化了 Jensen-Shannon divergence [31]. 但是对于词汇翻译的任务来说, 可能有其他更好的分布距离供选择. 由于 EMD 也是分布之间距离的一种度量, 并且在第 3.1 和 3.2 小节已经看到, 它对词汇翻译任务非常适合, 所以考虑使用 EMD 作为词汇表级别的准则来指导线性映射的学习, 即寻找一个映射 G, 使得源语言经过映射后的词向量分布与目标语言的词向量分布之间的 EMD 最小化, 如图 7 所示. 使用数学公式可以表示成如下的形式:

$$\min_{G} \mathrm{EMD}\left(p_{G(x)}, p_y\right), \tag{7}$$

其中 $p_{G(x)}$ 代表经过 G 映射后的源语言词向量分布, p_y 代表目标语言词向量分布.

不过, 涉及 EMD 的优化问题求解起来也很有挑战, 为此, 设计了两种方案. 第一种方案利用了 WGAN [33], 它可以视为优化 EMD 的 GAN 变种; 第二种方案直接将 EMD 代入上述式 (7), 尝试进行直接优化. 最终发现, 两者的结合能够稳定而有效地最小化 EMD, 找到相应的映射 [34].

4 总结

本文关注的问题是, 如何利用两种语言的非平行语料获取词汇级别的翻译知识. 我们讨论了前人工作中存在的一些挑战, 并针对性地给出了改进方案. 由于从非平行语料构建双语词典非常困难, 这个

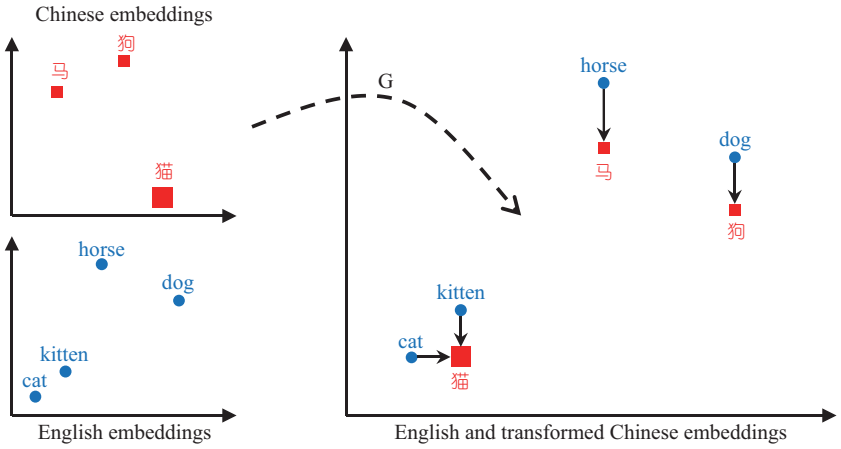


图 7 (网络版彩图) Earth mover's distance 最小化学习

Figure 7 (Color online) Learning by earth mover's distance minimization

任务仍有很大的提升空间, 而它的效果的改善将为跨语言处理所面临的资源匮乏场景带来帮助.

参考文献

1 Levow G A, Oard D W, Resnik P. Dictionary-based techniques for cross-language information retrieval. Inf Process Manage, 2005, 41: 523–547

2 Och F J, Ney H. A systematic comparison of various statistical alignment models. Comput Linguist, 2003, 29: 19–51

3 Täckström O, Das D, Petrov S, et al. Token and type constraints for cross-lingual part-of-speech tagging. Trans Assoc Comput Linguist, 2013, 1: 1–12

4 Rapp R. Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, 1999. 519–526

5 Gaussier E, Renders J M, Matveeva I, et al. A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, 2004

6 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, 2013. 3111–3119

7 Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation. ArXiv: 1309.4168

8 Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, 2014. 462–471

9 Lu A, Wang W, Bansal M, et al. Deep multilingual correlation for improved word embeddings. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL, Denver, 2015. 250–256

10 Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem. ArXiv: 1412.6568

11 Lazaridou A, Dinu G, Baroni M. Hubness and pollution: delving into cross-space mapping for zero-shot learning. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 270–280

12 Ammar W, Mulcaire G, Tsvetkov Y, et al. Massively multilingual word embeddings. ArXiv: 1602.01925

13 Smith S, Turban D, Hamblin S, et al. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. ArXiv: 1702.03859

14 Vulic I, Moens M F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 719–725

15 Zou W Y, Socher R, Cer D, et al. Bilingual word embeddings for phrase-based machine translation. In: Proceedings

of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, 2013. 1393–1398
16 Chandar A P S, Lauly S, Larochelle H, et al. An autoencoder approach to learning bilingual word representations. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, 2014. 1853–1861
17 Hermann K M, Blunsom P. Multilingual distributed representations without word alignment. ArXiv: 1312.6173
18 Kočiský T, Hermann K M, Blunsom P. Learning bilingual word representations by marginalizing alignments. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, 2014. 224–229
19 Gouws S, Bengio Y, Corrado G. Bilbowa: fast bilingual distributed representations without word alignments. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, 2015. 748–756
20 Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, 2015. 151–159
21 Coulmance J, Marty J M, Wenzek G, et al. Trans-gram, fast cross-lingual word-embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 1109–1113
22 Oshikiri T, Fukui K, Shimodaira H. Cross-lingual word representations via spectral graph embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, 2016. 493–498
23 Gouws S, Sogaard A. Simple task-specific bilingual word embeddings. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL, Denver, 2015. 1386–1390
24 Wick M, Kanani P, Pocock A. Minimally-constrained multilingual embeddings via artificial code-switching. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 2849–2855
25 Duong L, Kanayama H, Ma T, et al. Learning crosslingual word embeddings without bilingual corpora. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, 2016. 1285–1295
26 Shi T, Liu Z, Liu Y, et al. Learning cross-lingual word embeddings via matrix co-factorization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 567–572
27 Zhang M, Liu Y, Luan H, et al. Building earth mover's distance on bilingual word embeddings for machine translation. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 2870–2876
28 Resnik P, Yarowsky D. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Nat Lang Eng, 1999, 5: 113–133
29 Zhang M, Liu Y, Luan H, et al. Inducing bilingual lexica from non-parallel data with earth mover's distance regularization. In: Proceedings of the 26th International Conference on Computational Linguistics, Osaka, 2016. 3188–3198
30 Zhang M, Peng H, Liu Y, et al. Bilingual lexicon induction from non-parallel data with minimal supervision. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017. 3379–3385
31 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, 2014. 2672–2680
32 Zhang M, Liu Y, Luan H, et al. Adversarial training for unsupervised bilingual lexicon induction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 1959–1970
33 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. ArXiv: 1701.07875
34 Zhang M, Liu Y, Luan H, et al. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, 2017. 1934–1945

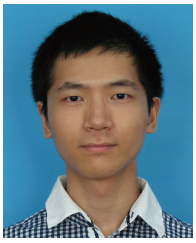
Bilingual lexicon induction from non-parallel corpora

Meng ZHANG[1,2,3], Yang LIU[1,2,3*] & Maosong SUN[1,2,3]

1. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China*;
2. *State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing* 100084*, China*;
3. *Beijing National Research Center for Information Science and Technology* (*BNRist*)*, Beijing* 100084*, China*
* Corresponding author. E-mail: liuyang2011@tsinghua.edu.cn

Abstract In cross-lingual natural language processing, the lack of parallel data is a serious problem. However, this is common in scenarios with scarce language resources. In this case, better utilizing translational equivalence encoded in non-parallel corpora becomes more important. Owing to the non-parallelism of the corpora, acquiring translational equivalence faces the challenging problem of small data or unsupervised learning, and the result usually takes the form of a bilingual lexicon. Not only is this an important research problem in the field of artificial intelligence, but it also has significant application value in scenarios with scarce language resources. This paper introduces a series of studies that address problems in prior research, exploring how to obtain better bilingual lexica with non-parallel corpora from various perspectives.

Keywords bilingual lexicon induction, non-parallel corpora, bilingual word embeddings, cross-lingual natural language processing, artificial intelligence



Meng ZHANG was born in 1991. He is a Ph.D. student in the Department of Computer Science and Technology, Tsinghua University. His current research interests include cross-lingual natural language processing and machine learning.

Yang LIU was born in 1979. He is an associate professor in the Department of Computer Science and Technology, Tsinghua University. His current research interests include natural language processing and machine translation.

Maosong SUN was born in 1962. He is a full professor in the Department of Computer Science and Technology, Tsinghua University. His current research interests include natural language processing, Chinese information processing, and computational social science.