计算机科学
COMPUTER SCIENCE

无监督句对齐综述

谷仕威, 刘静, 李丙春, 熊德意

引用本文

谷仕威, 刘静, 李丙春, 熊德意. 无监督句对齐综述[J]. 计算机科学, 2024, 51(1): 60-67.

GU Shiwei, LIU Jing, LI Bingchun, XIONG Deyi. Survey of Unsupervised Sentence Alignment[J]. Computer Science, 2024, 51(1): 60-67.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100024

无监督句对齐综述

谷仕威[1]　刘　静[2]　李丙春[2]　熊德意[1]
1 天津大学智能与计算学部　天津 300350
2 喀什大学计算机科学与技术学院　新疆 喀什 844000
(swgu98@qq.com)

摘　要　无监督句对齐在自然语言处理领域是一个重要而具有挑战性的问题。该任务旨在找到不同语言中句子的对应关系，为跨语言信息检索、机器翻译等应用提供基础支持。该综述从方法、挑战和应用3个方面概括了无监督句对齐的研究现状。在方法方面，无监督句对齐涵盖了多种方法，包括基于多语言嵌入、聚类和自监督或者生成模型等。然而，无监督句对齐面临着多样性、语言差异和领域适应等挑战。语言的多义性和差异性使得句对齐变得复杂，尤其在低资源语言中更为明显。尽管面临挑战，无监督句对齐在跨语言信息检索、机器翻译、多语言信息聚合等领域具有重要应用。通过无监督句对齐，可以将不同语言中的信息整合，提升信息检索的效果。同时，该领域的研究也在不断推动技术的创新和发展，为实现更准确和稳健的无监督句对齐提供了契机。

关键词：无监督句对齐；自然语言处理；机器翻译；自监督；低资源
中图分类号　TP391

Survey of Unsupervised Sentence Alignment

GU Shiwei[1], LIU Jing[2], LI Bingchun[2] and XIONG Deyi[1]
1 College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
2 School of Computer Science and Technology, Kashi University, Kashgar, Xinjiang 844000, China

Abstract　Unsupervised sentence alignment is an important and challenging problem in the field of natural language processing. This task aims to find corresponding sentence correspondences in different languages and provide basic support for cross-language information retrieval, machine translation and other applications. This survey summarizes the current research status of unsupervised sentence alignment from three aspects: methods, challenges and applications. In terms of methods, unsupervised sentence alignment covers a variety of methods, including based on multi-language embedding, clustering and self-supervised or generative models. However, unsupervised sentence alignment faces challenges such as diversity, language differences, and domain adaptation. The ambiguity and diversity of languages complicates sentence alignment, especially in low-resource languages. Despite the challenges, unsupervised sentence alignment has important applications in fields such as cross-lingual information retrieval, machine translation, and multilingual information aggregation. Through unsupervised sentence alignment, information in different languages can be integrated to improve the effect of information retrieval. At the same time, research in this field is also constantly promoting technological innovation and development, providing opportunities to achieve more accurate and robust unsupervised sentence alignment.

Keywords　Unsupervised sentence alignment, Natural language processing, Machine translation, Self-supervised, Low-resource

1　引言

随着大数据时代的到来，海量的多语言文本数据成为可利用的资源。然而，由于对平行句对进行标注的难度较大，传统的监督学习方法在句对齐任务中存在局限性[1]。在这种背景下，无监督句对齐方法成为一个引人瞩目的研究领域[1]。无监督句对齐方法能够在不依赖人工标注数据的情况下发现不同语言句子之间的对应关系，具有广泛的实际应用前景[2]。然而，无监督句对齐面临着多样性、语言差异和领域适应等挑战[3-4]。语言的多义性和差异性使得句对齐变得复杂，尤其在

到稿日期：2023-11-02　返修日期：2023-12-10
基金项目：新疆自治区自然科学基金重点项目(2022D01D43)；云南省重点研发计划(202203AA080004)；基于汉语-乌尔都语平行语料库的研究(KS2022084)
This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region(2022D01D43), Key Research and Development Program of Yunnan Province(202203AA080004) and Research on the Parallel Corpus of Chinese Urdu Language(KS2022084).
通信作者：熊德意(dyxiong@tju.edu.cn)

低资源语言中更为明显[5]。本文将探讨当前无监督句对齐方法的研究进展，以及面临的挑战和未来的发展方向。

在方法方面，无监督句对齐的研究需要借助词向量来完成，因此基于多语言嵌入来进行句对齐获取的方法尤为引人注目。这些方法利用先进的预训练模型将句子映射到语义空间，从而揭示句子之间的语义相似性[6-7]。此外，通过词向量，基于距离计算、聚类、自监督模型和图模型等方法也在解决句对齐问题上发挥了重要作用。首先是针对基于多语言嵌入距离计算的方法，这类方法依赖于预训练的多语言嵌入模型，如BERT，Word2Vec 等，将不同语言中的句子映射到共享的语义空间[8]。通过在这个空间中测量句子之间的相似性，可以找到句子对应关系[9]。这些方法在跨语言文本处理中表现出色，因为它们能够捕捉语义信息并处理语言差异[10-11]。其次是基于聚类的方法，聚类方法将句子组织成簇，其中每个簇包含相似的句子，通过簇内句子的相似性，可以识别句子对应关系[12]。这些方法通常不需要嵌入学习，因此适用于数据较少的语言或领域。基于自监督学习的无监督句对齐方法旨在通过自动构建句子表示并在无需平行数据或句对齐标签的情况下对齐句子，这些方法通常利用大规模的单语文本数据，使用自监督任务来训练句子表示，然后将这些表示映射到一个共享的空间中来实现句对齐[12-13]。最后是基于图模型的方法，图模型方法使用图结构来表示不同语言中的句子及其关系，节点表示句子，边表示句子之间的对应关系，通过图匹配算法，可以找到对应关系，这种方法通常适用于多对多的对应问题[14]。这些方法在无监督句对齐问题上采用了不同的策略和技术，但它们之间存在一些逻辑关系，即它们都利用了词向量和语义空间来揭示句子之间的语义相似性，并通过不同的方式来建模句子之间的关系和对应关系。这些方法的选择可以根据具体的应用场景和数据情况来确定。各种无监督句对齐方法总结如图 1 所示。

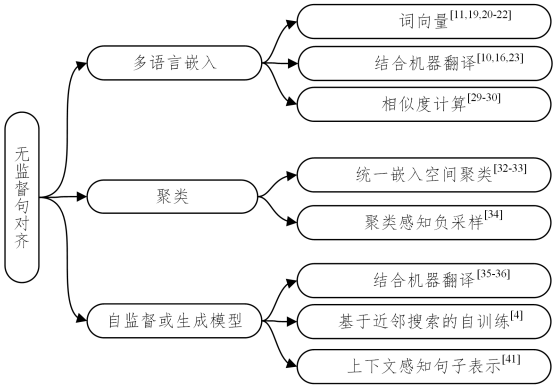

图 1 无监督句对齐方法

Fig. 1 Unsupervised sentence alignment methods

然而，无监督句对齐也面临一系列挑战，语言的多样性、多义性以及低资源语言的困难性使得句对齐问题变得复杂，而领域特异性和语言风格差异也增加了难度。主要的挑战有：1）多样性和多义性，不同语言中的句子可能有多种不同的表达方式，这增加了对应关系的复杂性[15]，同一个概念可能在不同语言中以多种方式表达，涉及多义性和多样性的处理[16-17]；2）语言差异，语言之间存在显著的语法、词汇和结构差异，这对句对齐构成了挑战，低资源语言或特殊领域的语言的句对齐尤其困难[18]；3）领域适应，句对齐模型的性能可能受到领域特异性和语言风格的影响，不同领域和文体之间的差异会增加句对齐的难度。克服上述挑战对于实现准确和稳健的句对齐至关重要。

尽管面临挑战，无监督句对齐在多领域中具有广泛应用，从跨语言信息检索到机器翻译再到多语言信息聚合，通过无监督句对齐，我们能够整合来自不同语言的信息，提高信息检索的效果。同时，这一领域的研究也在不断推动技术的创新和发展，为实现更准确和稳健的无监督句对齐提供了机遇。本文将深入研究这一领域的最新进展，以期为研究者和从业者提供有价值的见解和指导。

2 无监督词向量

无监督句对齐是自然语言处理领域的一个重要问题，它的主要目标是在没有预先标注的平行语料（即已对齐的句子对）的情况下，自动地找到不同语言中句子之间的对应关系。通常，无监督句对齐涉及两种不同语言的文本。我们将这两种语言分别称为“源语言”和“目标语言”，例如源语言可以是英语，而目标语言可以是法语。无监督句对齐的目标是找到源语言和目标语言中相对应的句子对，这些对应关系指示了句子在不同语言中表示的相似或相关的内容。下文是该任务的主要概念，以及相关的公式化描述：假设源语言的句子集合为 $S=(s_1,s_2,s_3,\cdots,s_m)$，目标语言的句子集合为 $T=(t_1,t_2,t_3,\cdots,t_n)$，将两个语言端的 S 和 T 做笛卡尔积即可得到 $m*n$ 个候选的相似句子对。若经过判断后，源端集合 S 中的第 i 个句子 s_i 与目标端集合 T 中的第 j 个句子 t_j 互为翻译，则 s_i 与 t_j 组成一对互为翻译的平行句子。无监督句对齐任务的主要目标就是寻找这样的互为翻译的句对，同时在判断时没有平行数据信息的参与，实现获取平行数据的目的。

进行无监督句对齐的基本思想就是根据已经对齐的词向量，进一步实现句子级别的对齐。国内外诸多学者对此开展了研究。针对双语语料弱监督和无监督的获取，研究者提出了一种通用的方案：首先构建多语言的跨语言词向量映射，然后通过这种跨语言词向量结合有限的双语监督信息进行双语词对和双语句对的获取。对于双语词对的获取，Ren 等[11]提出结合图神经网络和对抗学习进行跨语言词向量的映射，然后使用推导出的跨语言词向量进行双语词对的推导。首先需要讨论的问题是该方法在进行跨语言向量映射时主要依赖于一个基本的假设：不同语言语义相近的词的表示具有相似的空间分布，然而这种相似性假设并非完全成立，特别是在词源差异较大的语言间，词向量的空间非等距性更加明显，这种非等距性严重影响着跨语言词向量的映射，Ormazabal 等[19]和 Patra 等[20]都指出了该问题的存在。Zhao 等[21]借助预训练的跨语言词向量，设计了一种松弛匹配算法，用于实现无监督跨语言词向量的推导。在松弛匹配部分，引入 KL 散度来缓解最优传输问题中过于严格匹配的局限性，另外用 rcsls 距离来取代欧氏距离，测量不同语言词的语义嵌入之间的差异，产生差异的原因在于不同语言间的词容易产生聚类的现象，即同一语言语义相近的词更容易聚集到一起（又称为语义中心偏移），导致在利用跨语言词向量进行双语词对推导时，容易

产生违反直觉的错误的匹配结果。在松弛匹配以外，又引入双向优化框架，一改以往将从 S 到 T 与 T 到 S 的匹配视为两个独立问题的观点，在匹配过程中的每一次迭代都随机地选择一个方向进行。

3 基于多语言嵌入距离计算的无监督句对齐方法

早期的方法主要是通过借助词典来进行句对齐，这些方法主要强调不使用平行句对数据就可以被认为是一种无监督方法，这与目前的无监督方向有着巨大的差距，后文将不再具体讨论该类方法[1]。

Hangya 等[22]研究了在无双语监督的情况下，如何从单语语料库中挖掘出平行句子对的方法。传统方法通常依赖于双语词嵌入(BWEs)来进行平行句子提取，但需要强有力的双语信号，以训练 BWEs、用于句对提取的分类器或进行特征工程。这些方法的缺点在于，对于许多语言对来说，所需的双语信号不可用，这也是平行句子提取具有重要意义的原因之一。与这些方法相反，本文首次提出了基于词向量的无监督方法，Hangya 等使用后续映射(post-hocmapping)创建 BWEs，这使他们能够利用大量的(廉价的)单语数据来训练良好的单语词嵌入(MWEs)，然后将其映射到 BWEs。他们还使用结合了对抗训练和后续映射的方法来学习 BWEs，而无需任何双语信号。此外，动态阈值设置也在双语句对筛选中具有良好的效果，并且这种无监督的通过双语词嵌入进一步实现了平行句对挖掘的方法在多种语言中都具有普适性。Artetxe 等[10]提出了一种用于学习多语言通用句子嵌入的方法，可以处理 93 种不同的语言，具有广泛的潜在应用前景，特别是对于资源有限的语言。该方法使用单一的编码器来处理多种语言，使不同语言中的语义相似句子在嵌入空间中接近。这种方法在多种任务中表现出了很强的性能，包括跨语言自然语言推理、跨语言分类、平行语料库挖掘和多语言相似性搜索。

Kim 等[23]提出了一种新颖的模型架构和训练算法，用于从平行和单语数据的组合中学习双语句子嵌入。该方法将自动编码和神经机器翻译连接起来，强制源句子和目标句子嵌入在不借助中间语言或额外转换的情况下共享相同的空间，在句子嵌入之上训练了一个多层感知器，以从非平行或嘈杂的平行数据中提取良好的双语句子对。Bañón 等[24]介绍了通过抓取网络并使用开源软件创建最大的公开可用平行语料库的方法。Zhu 等[5]以及 Hangya 等[25]指出，Hangya 等[22]简单地使用词向量的相似度得分进行双语语料的获取，容易引入一些非完全平行的双语句子，这些句子尽管在单词层面具有较高相似度，但是放大到句子层面并非是完全平行的句子。针对该问题，他们提出通过检测候选句对中是否含有连续的互为翻译的短语来进行噪音数据的过滤。这类方法主要的问题在于：1)不同语言的句子内部的单词并非完全的一对一，多数情况下，一对多和多对多的现象更加常见，仅计算词向量的相似度确定双语句对容易引入大量的噪声；2)主要使用静态词向量，这种词向量本身有着缺陷，无法有效地表示一词多义的现象，在计算词向量的相似度时容易导致匹配错误。Hangya 等[25]通过检测候选句子对中的连续平行段来解决这个问题，他们不仅仅是对双语单词的相似性进行平均，而对候选句子对中的相似单词进行对齐，然后利用这些对齐来检测在两侧都与彼此对齐的连续子句段。为了提高系统的精确性，他们只挖掘那些检测到的平行段在整个句子对中占据很大一部分的相似句子对，从而解决了前面提到的非完全平行句子对的误判问题。Kvapilíková 等[16]介绍了一种新颖的无监督方法，用于生成多语言句子嵌入，只依赖于单语数据，而不需要大规模的平行数据资源。该方法首先通过无监督机器翻译生成一个合成的平行语料库，然后利用它来微调预训练的跨语言遮蔽语言模型(XLM)，以生成多语言句子表示。Hong 等[26]引入跨语言语义表示作为中介语义表示，以完全无监督的方式实现了跨语言迁移学习，为低资源语言中的自然语言处理任务提供了一种有效的解决方案，通过将不同语言的语义信息映射到相似的语义空间中，可以实现跨语言任务的迁移，而不依赖于传统的平行语料和机器翻译方法。Tien 等[3]提出了一种无监督的方法，利用双语对齐数据来生成多语对齐数据。首先，使用现有的双语对齐模型将双语文本进行对齐，然后利用对齐的双语数据，通过一系列迭代步骤，将其扩展到多语对齐数据。具体而言，该方法使用了一种基于词嵌入的相似度度量方法，将双语对齐数据映射到一个共享的嵌入空间中，然后通过最大化多语对齐数据的相似度，来优化嵌入空间的质量，最后使用生成的多语对齐数据来训练一个多语言翻译模型，以实现无监督的平行文本挖掘。

尽管上述方法都在多个语言对上取得了较好的效果，但是在实际双语语料获取中的应用有待验证，上述工作都没有探索出一种基于无监督方法的多语言大规模无监督双语语料获取的实际实践，Schwenk 等[27]最早使用无监督方法来获取多语言大规模双语语料。他们从维基百科单语语料库中提取双语句子，并且使用基于 BiLSTM 的多语言编码器将 93 种语言的单语词表示映射到同一向量空间，虽然其中需要大量平行语料的参与，但是也做到了训练数据极少甚至没有训练数据参与训练的语种同样可以实现对齐效果。其初衷是设计一种语言无关的词向量表示，然后通过词向量的平均池化或最大池化构建句子向量，并计算句向量的相似度进行双语语料的提取。他们在含有 300 种语言的维基百科单语语料中提取了大约 1 600 多组语言对的双语平行句对，其中数量超过 10 万级句对的语言对超过 100 组，且多数是以英语为源语言或目标语言的平行语料。该项工作有效证明了从维基百科中获取双语语料的可行性。尽管多语网站上现成的双语语料很少，但是单语语料是极为丰富的。同时 Schwenk 等[28]指出维基百科上的许多文章虽然是独立撰写的，但由于相同事件或者话题可利用不同的语言进行描述，其中仍可能包含互为翻译的句子。Schwenk 等[28]从中获取的超过千余种语言对的双语平行句对很好地证明了这一点；其次该项工作通过构建一个语言无关的 encoder，将不同语言的词向量映射到统一向量空间，进行双语句对的获取，这点表明在不考虑某种具体的语种和相关语言本身属性的情况下，仅通过深度学习技术仍然可以获取相当规模的双语平行语料。

在计算相似度方法上，Lian 等[29]提出了一种计算相似度的新方法。具体来说，使用 Wasserstein Barycenter 来对齐多个语言之间的分布。Wasserstein Barycenter 是一种用于计算

多个概率分布之间的平均分布的方法,使用自编码器将每个语言的单语语料映射到一个低维的语义空间,然后通过计算每个语言对应的语义空间中的分布与中心分布之间的 Wasserstein 距离,来衡量语言之间的差异,最后通过最小化这些距离,可以得到每个语言与中心分布之间的对齐关系。Chousa 等[30]通过预测源语言和目标语言句子中的片段,使用整数线性规划来找到最佳的句对齐。具体而言,首先使用预训练的语言模型来预测源语言和目标语言句子中的片段,然后将这些片段作为变量输入到整数线性规划模型中,通过最大化对齐的片段之间的相似性来确定最佳的句对齐。

总的来说,上述双语平行句对的弱监督方法都需要借助跨语言词向量,但是并未考虑知识迁移过程中不同语言引起的语言一致性的影响或者并未深入地分析不同粒度的语言相关性对双语句对获取过程中迁移学习的影响。Zhu 等[31]深入分析了词级、句子级不同粒度的知识迁移对低资源语言对双语句子对获取的效果,综合研究了融合词级、句子级语言一致性的知识迁移模型。该工作旨在利用不同语言中的内部语言不变性,将高资源语言对的知识转移到低资源语言对中,以解决低资源语言对中双语资源稀缺的问题。该方法的核心思想是使用多视图分类器,将句子嵌入视为分类任务,并使用两种视图来识别语义信息:1)词级表示;2)句级表示。句级表示用于捕获两个句子之间的语义相似性,而词级表示用于捕获一对平行句子中的词汇翻译,以避免语义相似但非平行的句子误对齐问题。为了获得双语平行语料库,Zhu 等[31]首先使用预训练的无监督多语言词嵌入模型将单语词表示映射到共享的词表示向量空间。然后,通过计算对 M-BERT 输出的均值池化,可以获得句子嵌入。此外,文中还引入了一种新的正则化训练策略,使用单语语料库来提高挖掘到的双语数据的质量。

基于多语言嵌入距离计算的无监督句对齐方法具有一些优点和缺点。主要优点为:1)多语言嵌入模型通常是语言无关的,这意味着它们可以应用于多种语言对,而不需要针对每种语言对进行特定的定制,这增加了方法的通用性;2)基于多语言嵌入距离计算的方法能够捕获句子之间的语义相似性,而不仅仅是基于表面文本相似性,这有助于识别在意义上相似但在形式上不同的句子对;3)与一些方法需要额外的外部资源(如词典或平行语料库)不同,多语言嵌入方法通常只需要大规模的单语语料库和一个预训练的多语言嵌入模型。主要缺点为:1)虽然无监督方法不需要平行句对作为训练数据,但它们仍然需要大规模的单语数据来训练多语言嵌入模型,这对于一些低资源语言对而言仍然是一个挑战;2)基于嵌入距离的方法通常依赖于多语言嵌入模型,这些模型可能相对复杂,并需要大量的计算资源来训练和计算距离,这可能会限制其在某些环境中的可行性;3)如果多语言嵌入模型存在错误的语义关联,这些错误可能会传播到句对齐过程中,导致不准确的对齐结果。

4 基于聚类的无监督句对齐方法

基于聚类方法的无监督句对齐是一项重要的自然语言处理任务,旨在发现不同语言中的句子对应关系,为跨语言信息检索、机器翻译等应用提供基础支持。这一领域的研究致力于解决缺乏平行句对标注的问题,通过聚类相似的句子来实现无监督的句对齐,其目标是将来自不同语言的句子分组到共享的簇中,其中每个簇内的句子被认为是对应的。研究者使用各种技术来衡量句子之间的相似性,以便将它们正确地划分到簇中。这种方法的优势在于不需要大量的平行句对数据,因此适用于低资源语言和特殊领域。

Chi 等[32]提出了一种基于聚类的无监督句对齐方法,它使用句子嵌入和聚类技术来实现句对齐。首先将句子嵌入到连续空间中,然后使用一种带有谱聚类的无监督方法来对句子进行聚类。此外还提出了一个用于评估双语上下文单词相似性的数据集。Wang 等[33]旨在将不同语言的句子表示对齐到一个统一的嵌入空间中,从而可以使用简单的点积来计算语义相似性,无论是跨语言的还是单语的。该方法通过预训练的语言模型,并进行翻译排序任务的微调来实现。为了进一步提高对齐质量,其通过 K-Means 聚类将相似的句子分为不同簇,以改进句对齐的性能。Deng 等[34]的目标是通过聚类方法来实现无监督的句对齐,并提出了一种聚类感知的负采样方法来训练句子表示。首先,使用预训练的词向量模型将每个句子表示为向量,这样可以将句子转换为连续的向量空间,从而捕捉句子之间的语义信息。接下来使用聚类算法对句子向量进行聚类,将相似的句子聚集在一起,聚类的结果就是句对齐的结果,每个聚类簇代表了一组相互对应的句子对。然而,传统的无监督句对齐方法通常会使用随机负采样来训练句子表示,这可能导致负样本与正样本之间的相似度过高,影响对齐的准确性。为了解决这个问题,Deng 等[34]还提出了一种聚类感知的负采样方法。具体而言,根据聚类结果,为每个句子选择负样本时,会倾向于选择来自不同聚类簇的句子作为负样本,以增加负样本与正样本之间的差异性,这样可以更好地训练句子表示,提高句对齐的准确性。

上述的聚类方法可以在缺乏平行句对标注的情况下进行句对齐。这些方法使用不同的技术来衡量句子之间的相似性,并将它们正确地划分到簇中,从而实现对齐的目的。重点在于它们可以应用于低资源语言和特殊领域,而不需要大量的平行句对数据。然而,这些方法也存在一些缺点。首先,基于聚类的方法可能会受到聚类算法的选择和参数设置的影响,导致对齐结果具有很大的不稳定性;其次,这些方法可能无法处理一些复杂的语言现象和语义关系,导致对齐的准确性有限。此外,这些方法可能需要大量的计算资源和时间来进行句子嵌入和聚类,特别是在处理大规模数据时这种情况尤为突出。

5 基于自监督或生成模型的句对齐方法

Paetzold 等[35]使用机器翻译技术将文档从一种语言翻译成另一种语言,以便进行跨语言对齐,然后使用文本对齐算法将翻译后的文档与原始文档进行对齐,以找到相应的句子或段落。在对齐完成后,其还提供了注释功能,可以自动标注对齐文档中的重要信息,如实体、关键词等,这些注释可以帮助用户更好地理解和分析文档内容。Leng 等[36]引入了一种迭代的训练方法,在每一轮迭代中,使用当前的翻译模型进行

翻译，并将翻译结果作为新的句对齐训练数据，再次训练模型，通过多轮迭代可以逐渐提高句对齐的质量。Chen 等[37]通过最小化互信息来增强相似样本对之间的相关性，同时减小不相似样本对之间的相关性。Chen 等提出的模型使用了一个编码器网络来将句子映射到一个低维的嵌入空间，并使用一个对比损失函数来衡量样本对之间的相似性。通过最小化对比损失函数，模型可以学习到具有良好语义表示的句子嵌入，借助对比学习方法，选择合适的负样本，通过最小化互信息，模型可以选择那些与正样本相似度较低的负样本，从而提高对比学习的效果。

Keung 等[4]提出了一种无监督方法，从未对齐的文本中创建伪平行语料库，用于机器翻译。该方法基于自训练的上下文嵌入，使用多语言 BERT(mBERT)来创建源语言和目标语言的句子嵌入，进行最近邻搜索，并通过自我训练来改进模型。其基本框架如图 2 所示。

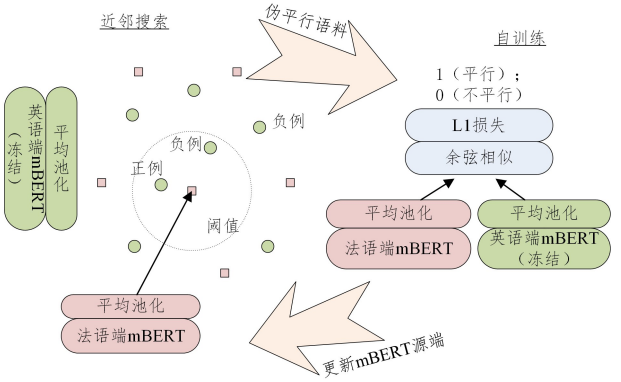

图 2 基于近邻搜索的自训练方法进行无监督句对齐的基本框架

Fig. 2 Basic framework for unsupervised sentence alignment using self-training method based on nearest neighbor search

该方法的验证结果表明，在 BUCC2017 的双语文本挖掘任务中，相比先前的无监督方法，F1 分数提高了 24.5 个百分点。此外，还在使用 Wikipedia 预训练的 XLM 模型上，通过添加从同一语料库中挖掘出的伪平行文本，提升了无监督翻译性能，特别是在 WMT'14 法英和 WMT'16 德英任务上，BLEU 值提高了 3.5，并超越了先前的最新方法。最后，还在低资源情况下展示了无监督双语文本挖掘的实际价值，通过将 IWSLT'15 英越翻译任务的英越语料库增强伪平行的 Wikipedia 句子对，BLEU 得分提高了 1.2。其关键步骤包括：

1)多语言嵌入表示：使用多语言 BERT(mBERT)[38]模型，该模型预训练于来自 104 种语言的不对齐 Wikipedia 句子，从中创建句子嵌入。

2)嵌入距离计算：通过计算 mBERT 嵌入的源语言和目标语言句子之间的相似性得分，进行最近邻搜索，以寻找候选翻译对。其中采纳了 Artetxe 等[39]提出的近邻相似度得分计算方法。

$$score=\frac{\cos(x,y)}{\sum_{z\in NN_k(x)}\frac{\cos(x,z)}{2k}+\sum_{z\in NN_k(y)}\frac{\cos(y,z)}{2k}}$$

3)自我训练：构建正例和负例句子对的数据集，并使用自我训练方法对 mBERT 进行微调，以改进句对齐质量。

4)伪平行句对：最后通过检索，获得一组伪平行句子对，其中包含实际翻译和语义相关但非翻译的句子对。

Ding 等[40]探讨了如何在双语文本中明确地整合词汇翻译作为外部知识，以增强词分布表示，从而提高句对齐的性能。在传统的句对齐方法中，通常使用手动设计的特征(例如长度比例和词对)来进行句对齐，但这些方法存在着稀疏性问题，因为语言存在歧义。随着神经网络在建模分布式表示方面的强大能力得以展现，神经网络句对齐开始受到关注。例如，一些研究依赖于句子建模，将输入的句子映射到一个固定长度的向量，然后通过这些句子向量来预测两个句子是否对齐。然而，句子级别的表示往往无法捕捉词级别的细节，这些细节对句对齐非常重要，因此一些研究提出了计算词对之间相似性的词级别方法，以获取更细粒度的词级别信息。Wu 等[41]介绍了一种用于多语言密集检索的无监督上下文感知句子表示预训练方法。在自然语言处理中，密集检索指通过计算相似度来寻找与查询相关的文本，传统的方法通常使用词袋模型或者基于语言模型的表示方法，但是这些方法忽略了句子的上下文信息。为了解决这个问题，Wu 等[41]提出了一种新的预训练方法，称为无监督上下文感知句子表示预训练，使用大规模的无标签语料库进行预训练，学习句子的上下文感知表示。具体而言，该工作提出的 UCASR 方法使用了自编码器和对比学习的思想。首先通过自编码器将输入句子编码为低维向量表示，然后使用对比学习的方法来训练模型，使得相似的句子在表示空间中更加接近，即使在不同语言中，具有相似语义的句子也能够在向量空间中靠近，而不相似的句子则更加远离。Liu 等[42]工作的核心思想是通过自动构建短语表达的统一表示来实现双语短语对齐，首先使用无监督的方法从单语语料库中提取短语，并为每个短语生成一个向量表示，然后通过最大化短语对齐的概率来学习短语对齐模型。这个模型可以将源语言短语和目标语言短语映射到同一语义空间中。

这些自监督或生成模型的方法不需要人工标注的数据，可以利用大规模的无标签语料库进行训练，从而减少了数据收集和标注的成本，特别是可以学习到将不同语言的句子映射到同一语义空间的表示，从而实现跨语言的自动学习和迁移学习。同时可以整合词汇翻译并将其作为外部知识，提高了词分布表示的性能，从而改善了句对齐的质量。但是缺点是这些方法在对齐任务中的性能可能受到限制，尤其是在存在语言歧义性和稀疏性的情况下。此外，这些方法虽然不需要人工标注的数据，但它们仍然依赖于大规模的无标签语料库进行训练，因此对于某些语言或领域可能存在数据稀缺的问题，并且需要进行多轮迭代训练或使用复杂的损失函数，这会增加训练的时间和计算成本。

6 句对齐相关任务

6.1 BUCC 共享任务

BUCC(Bilingual and Multilingual Corpus Collection)数据集是一个用于文本翻译和文本准确性评估的多语言平行语料库。该数据集由数百万语句组成，覆盖了多种语言组合，包括英语、法语、汉语、德语、俄语等，其主要目的是为语言研究人员、自然语言处理(NLP)和机器翻译(MT)算法开发者提供一个高质量和多样化的多语言语料库，以便他们利用这些

数据进行进一步的研究和应用。

其中较为经典且使用频率较高的 BUCC2017 数据集[43]是第二届共享任务提出的，专门针对于获取平行句子对。因为早期的句对齐方式主要是基于平行文档或者更多地借助句长、纯文本之外的日期等额外信息，而不是专注于文本语义内容本身，因此提出这个任务是希望将注意力转移到文本内容中。构建数据集的原始语料主要是 NewsCommentary 上的双语文本和维基百科的单语文本相混合，记录这些平行文本混入的位置，然后任务目标就是找回这些混在其中的平行文本。该数据集保证主题相同，同时也会尽力避免因单语文本中存在互为翻译的平行句子而扰乱原本的固有平行句子数量。

BUCC2017 句子检索数据集由 4 个语种组成：德语-英语(de-en)、法语-英语(fr-en)、俄语-英语(ru-en)、中文-英语(zh-en)。每个语种一个文件夹，每个语种文件夹内包含 3 类文件，一类是 training 文件，主要用于训练，包含源语言文件、目标语言文件以及 gold 文件 3 个文件，gold 文件即存储了那些平行句子插入信息的文件，可以用作判断对齐效果的工具。第二类文件是 sample，同样是包含源语言文件、目标语言文件以及 gold 文件，文件内容和 training 相同，但每个文件内的句子数量都大大减少。第三类文件就是 test 文件，用于测试，其与 training 文件的区别就是不包含 gold 文件，同时文件内句子数量比较相似。其各个文件内句子数量如表 1 所列。

表 1 BUCC2017(BUCC2018)句对齐数据集统计

Table 1 Statistics of sentence alignment dataset BUCC2017 (BUCC2018)

		de-en	fr-en	ru-en	zh-en
training	xx	413 869	271 874	460 853	94 637
	en	399 337	369 810	558 410	88 860
	gold	9 580	9 086	14 435	1 899
sample	xx	32 593	21 497	45 459	8 624
	en	40 354	38 069	72 766	13 589
	gold	1 038	929	2 374	257
test	xx	413 884	276 833	457 327	91 824
	en	396 534	373 459	566 356	90 037

BUCC2018 句子检索任务和 BUCC2017 的目的相同，都是获取平行句对，并且提供的数据也是相同的，旨在给新参与到这个任务的研究者一个机会，也给从前的参与者一个改进的机会。

此外，在无监督句对齐任务中，因为较多涉及到低资源语言，所以训练机器翻译系统进而评测翻译效果同样也是使用广泛的评判对齐效果优劣的方法。

6.2 无监督对齐方法结果介绍

本节介绍一些无监督对齐方法及其在 BUCC2018 平行句对获取数据集上的结果。

1)UPSBMWE[22]：使用无监督的双语单词嵌入来提取平行句子。此外，该方法提出了一种动态阈值方法来选择更合理的平行句对。

2)uppsd[25]：为了解决经常挖掘意义相似但不完全平行的句子对的问题，该方法选择检测候选句子对中的连续的平行片段来辅助判断是否平行。这种语言独有特征的方式在一些语言对上取得了效果，但普适性似乎有所欠缺。

3)UBMSCE[4]：使用 mBERT 在没有监督的情况下获取平行数据，并通过具有基于边界相似性分数的最近邻搜索来找到检测到的候选平行句子对。

4)UMSPCM[16]：使用无监督的机器翻译系统来产生合成的平行句子。该方法的结果证明，在跨语言迁移学习中，其内部表征并不完全是语言不可知的。

MvKD[31]：在词级和句子级两种视图下综合考虑了句子信息的相似程度，充分考虑不平行但存在一定相似度的情况，并实现了从高资源语言到低资源语言的知识迁移，充分利用语言的内部不变性来解决低资源语言平行数据的匮乏问题。

表 2 列出了各无监督句对齐方法在 4 种语言对上的精确度(precision，P)、召回率(recall，R)和 F1 分数，“—”代表未在原始论文中提及的结果。

表 2 各无监督句对齐方法在 BUCC 数据集上的结果

Table 2 Results of various unsupervised sentence alignment methods on BUCC dataset

	en-fr			en-de			en-ru			en-zh		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
UPSPSD	50.5	38.1	43.4	48.5	39.1	43.3	37.4	18.7	24.9	—	—	—
UBMSCE	—	—	73.0	—	—	74.9	—	—	69.6	—	—	60.1
UPABMWE	39.0	52.6	44.8	23.7	44.5	30.9	17.3	24.9	20.4	—	—	—
UMSPCM	—	—	78.7	—	—	80.1	—	—	77.1	—	—	67.0
MvKD	83.6	82.1	82.3	90.8	87.3	88.1	82.6	73.5	79.8	78.6	75.8	77.2

7 未来方向

无监督句对齐是自然语言处理领域的一个重要任务，旨在将两个语言中的句子进行对齐，以便进行跨语言文本分析和翻译等应用。目前，已经有一些研究工作在无监督句对齐方向上取得了一定的进展，但仍存在一些挑战。无监督句对齐一些可能的未来研究方向总结如下。

1)引入语义信息：当前的无监督句对齐方法主要基于句子的表面形式进行对齐，而忽略了句子的语义信息，或存在语义上的歧义。未来的研究可以探索如何将语义信息引入到对齐过程中，以提高对齐的准确性和鲁棒性。

2)跨语言知识迁移：在某些语言对中，可能存在一些已经对齐好的句子对，可以作为跨语言知识的迁移源。现在已经存在一些学者进行迁移学习的尝试，未来将更多地探索如何利用这些已知的对齐信息，来辅助对其他语言对的对齐，从而提高对齐的效果，这也会是未来的趋势。

3)多模态对齐：除了文本句子，还存在其他形式的语言表达，如图像、音频和视频。未来的研究可以探索如何将多模态数据进行对齐，从而实现跨模态的语义理解和交互。

4)跨领域对齐：当前的无监督句对齐方法主要关注领域内的句对齐，但在实际应用中，还存在跨领域对齐的需求，例如将新闻文本对齐到科技文本。未来的研究可以探索如何进行跨领域的无监督句对齐，以满足不同领域的应用需求。

5)对齐评估标准：目前对于无监督句对齐的评估主要

依赖于人工标注或者外部资源，缺乏一致的评估标准。今后的研究可以探索更有效的评估方法和标准，以便更好地比较不同方法的性能。

结束语 本文深入探讨了无监督句对齐在自然语言处理领域的重要性以及相关研究现状，无监督句对齐旨在找到不同语言中相对应的句子对应关系，为跨语言信息检索、机器翻译等应用提供基础支持。在方法方面，无监督句对齐采用了多种方法，包括基于多语言嵌入、聚类以及自监督或生成模型等，这些方法允许将句子映射到语义空间、聚类相似句子，或者通过大规模无标签语料库进行训练，以实现跨语言自动学习和迁移学习。然而，无监督句对齐也面临着多样性、语言差异、领域适应等多重挑战，语言的多义性和差异性增加了句对齐的复杂性，特别是在低资源语言中。此外，领域特异性和语言风格差异也增加了句对齐的难度。尽管面临挑战，由于无监督句对齐在跨语言信息检索、机器翻译、多语言信息聚合等领域具有重要应用，因此在这一方向上的探索仍未停止。通过无监督句对齐，可以整合不同语言中的信息，提升信息检索的效果，同时这一领域的研究也在不断推动技术的创新和发展，为实现更准确和稳健的无监督句对齐提供了契机。

参 考 文 献

[1] BRAUNE F, FRASER A. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora [C]// Coling 2010: Posters. 2010: 81-89.

[2] LI Z, HUANG S, ZHANG Z, et al. Dual-Alignment Pre-training for Cross-lingual Sentence Embedding[J]. arXiv: 2305. 09148, 2023.

[3] TIEN C, STEINERT-THRELKELD S. Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 8696-8706.

[4] KEUNG P, SALAZAR J, LU Y, et al. Unsupervised bitext mining and translation via self-trained contextual embeddings[J]. Transactions of the Association for Computational Linguistics, 2021, 8: 828-841.

[5] ZHU S, MI C, LI T, et al. Unsupervised parallel sentences of machine translation for Asian language pairs[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(3): 64: 1-64: 14.

[6] LAMPLE G, CONNEAU A, DENOYER L, et al. Unsupervised Machine Translation Using Monolingual Corpora Only[C]// International Conference on Learning Representations. 2018.

[7] ARTETXE M, LABAKA G, AGIRRE E, et al. Unsupervised neural machine translation[C]// 6th International Conference on Learning Representations (ICLR 2018). 2018.

[8] LAMPLE G, CONNEAU A, RANZATO M A, et al. Word translation without parallel da-ta[C]// International Conference on Learning Representations. 2018.

[9] QI Y, SACHAN D, FELIX M, et al. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 529-535.

[10] ARTETXE M, SCHWENK H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610.

[11] REN S, LIU S, ZHOU M, et al. A graph-based coarse-to-fine method for unsupervised bilingual lexicon induction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3476-3485.

[12] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embed-dings[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 789-798.

[13] GARNEAU N, GODBOUT M, BEAUCHEMIN D, et al. A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings: Making the Method Robustly Reproducible as Well[C]// Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 5546-5554.

[14] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised Cross-lingual Representation Learning at Scale[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8440-8451.

[15] LU X, QIANG J, LI Y, et al. An unsupervised method for building sentence simplification corpora in multiple languages[C]// Findings of the Association for Computational Linguistics. Punta Cana: Association for Computational Linguistics, 2021: 227-237.

[16] KVAPILÍKOVÁ I, ARTETXE M, LABAKA G, et al. Un-supervised Multilingual Sentence Embeddings for Parallel Corpus Mining[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2020: 255-262.

[17] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 789-798.

[18] HASHIMOTO K, XIONG C, TSURUOKA Y, et al. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1923-1933.

[19] ORMAZABAL A, ARTETXE M, LABAKA G, et al. Analyzing the Limitations of Cross-lingual Word Embedding Mappings [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4990-4995.

[20] PATRA B, MONIZ J R A, GARG S, et al. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 184-193.

[21] ZHAO X, WANG Z, ZHANG Y, et al. A Relaxed Matching Pro-

cedure for Unsupervised BLI[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:3036-3041.

[22] HANGYA V,BRAUNE F,KALASOUSKAYA Y,et al. Unsupervised parallel sentence extraction from comparable corpora [C]// Proceedings of the 15th International Conference on Spoken Language Translation. Brussels:International Conference on Spoken Language Translation. 2018:7-13.

[23] KIM Y,ROSENDAHL H,ROSSENBACH N,et al. Learning Bilingual Sentence Embeddings via Autoencoding and Computing Similarities with a Multilayer Perceptron[C]// Proceedings of the 4th Workshop on Representation Learning for NLP(RepL4NLP-2019). 2019:61-71.

[24] BAÑÓN M,CHEN P,HADDOW B,et al. ParaCrawl:Web-scale acquisition of parallel corpora[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:4555-4567.

[25] HANGYA V,FRASER A. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:1224-1234.

[26] HONG C,LEE J,LEE J. Unsupervised Interlingual Semantic Representations from Sentence Embeddings for Zero-Shot Cross-Lingual Trans-fer[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:7944-7951.

[27] SCHWENK H,WENZEK G,EDUNOV S,et al. CCMatrix:Mining Billions of High-Quality Parallel Sentences on the Web [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Volume 1:Long Papers). 2021:6490-6500.

[28] SCHWENK H,CHAUDHARY V,SUN S,et al. WikiMatrix:Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics:Main Volume. 2021:1351-1361.

[29] LIAN X,JAIN K,TRUSZKOWSKI J,et al. Unsupervised multilingual alignment using Wasserstein barycenter[C]// Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021:3702-3708.

[30] CHOUSA K,NAGATA M,NISHINO M. SpanAlign:Sentence alignment method based on cross-language span prediction and ILP[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:4750-4761.

[31] ZHU S,GU S,LI S,et al. Mining parallel sentences from internet with multi-view knowledge distillation for low-resource language pairs[J/OL]. Knowledge and Information Systems,2023. https://doi.org/10.1007/s10115-023-01925-3.

[32] CHI T C,CHEN Y N. CLUSE:Cross-Lingual Unsupervised Sense Embeddings[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:271-281.

[33] WANG L,ZHAO W,LIU J. Aligning Cross-lingual Sentence Representations with Dual Momentum Contrast[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021:3807-3815.

[34] DENG J,WAN F,YANG T,et al. Clustering-Aware Negative Sampling for Unsupervised Sentence Representation[J]. arXiv:2305.09892,2023.

[35] PAETZOLD G,ALVA-MANCHEGO F,SPECIA L. Massalign:Alignment and annotation of comparable docu-ments[C]// Proceedings of the IJCNLP 2017,System Demonstrations. 2017:1-4.

[36] LENG Y,TAN X,QIN T,et al. Unsupervised Pivot Translation for Distant Languages[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:175-183.

[37] CHEN S,ZHOU J,SUN Y,et al. An Information Minimization Based Contrastive Learning Model for Unsupervised Sentence Embeddings Learning[C]// Proceedings of the 29th International Conference on Computational Linguistics. 2022:4821-4831.

[38] PIRES T,SCHLINGER E,GARRETTE D. How Multilingual is Multilingual BERT? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4996-5001.

[39] ARTETXE M,SCHWENK H. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:3197-3203.

[40] DING Y,LI J,GONG Z,et al. Improving neural sentence alignment with word translation[J]. Frontiers of Computer Science,2021,15:151302.

[41] WU N,LIANG Y,REN H,et al. Unsupervised context aware sentence representation pretraining for multi-lingual dense retrieval[J]. arXiv:2206.03281,2022.

[42] LIU J,MORIN E,SALDARRIAGA S P,et al. From unified phrase representation to bilingual phrase alignment in an unsupervised manner[J]. Natural Language Engineering,2023,29(3):643-668.

[43] ZWEIGENBAUM P,SHAROFF S,RAPP R. Towards preparation of the second BUCC shared task:Detecting parallel sentences in comparable corpora[C]// Proceedings of the Ninth Workshop on Building and Using Comparable Corpora. European Language Resources Association(ELRA),Portoroz,Slovenia. 2016:38-43.

GU Shiwei,born in 1998,postgraduate. His main research interests include natural language processing and machine translation.

XIONG Deyi,born in 1979,Ph.D,professor,Ph.D supervisor. His main research interests include natural language processing and machine translation.

(责任编辑:喻藜)