计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900003

多语言问答研究综述

刘 创 熊德意
天津大学智能与计算学部 天津 300350
(liuc_09@tju.edu.cn)

摘 要 多语言问答是自然语言处理领域的研究热点之一,其目的是给定不同语种的问题和文本,模型能够返回正确的答案。随着机器翻译技术的快速发展及多语言预训练技术在自然语言处理领域中的广泛应用,多语言问答也取得了较快的发展。文中首先系统地梳理了当前多语言问答方法的相关工作,并将多语言问答方法分为基于特征的方法、基于翻译的方法、基于预训练的方法和基于双重编码的方法,分别介绍了每类方法的使用和特点;然后系统地探讨了当前多语言问答任务的相关工作,将多语言问答任务分为基于文本的多语言问答任务和基于多模态的多语言问答任务,并分别给出每个多语言问答任务的基本定义;接着总结了这些任务中的数据集统计、评价指标,以及涉及的问答方法;最后展望了多语言问答的未来发展方向。

关键词:多语言问答;机器翻译;多语言预训练技术;基于文本的多语言问答;基于多模态的多语言问答
中图法分类号 TP391

Survey of Multilingual Question Answering

LIU Chuang and XIONG De-yi
College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

Abstract Multilingual question answering is one of the research hotspots in the field of natural language processing, which aims to enable the model to return a correct answer based on understanding of the given questions and texts in different languages. With the rapid development of machine translation technology and the wide application of multilingual pre-training technology in the field of natural language processing, multilingual question answering has also achieved a relatively rapid development. This paper first systematically reviews the current work of multilingual question answering methods, and divides them into feature-based methods, translation-based methods, pre-training-based methods and dual encoding-based methods, and introduces the use and characteristics of each method respectively. Meanwhile, it also discusses the current work related to multilingual question answering tasks, and divides them into text-based and multi-modal-based tasks and gives the basic definition of each one. Moreover, this paper summarizes the dataset statistics, evaluation metrics and multilingual question answering methods involved in these tasks. Finally, it proposes the future research prospect of multilingual question answering.

Keywords Multilingual question answering, Machine translation, Multilingual pre-training techniques, Text-based multilingual question answering, Multi-modal-based multilingual question answering

1 引言

问答是自然语言处理的重要研究方向,其性能是衡量计算机模型对文本理解程度的重要标准。问答任务往往由文本或知识库作为背景信息,模型需要根据背景信息回答一系列相关的问题。随着深度学习在自然语言处理领域的快速发展,问答任务的种类也得到了快速的发展,文献[1-4]将任务集中在完型填空类型的问答上;文献[5-9]定义了多项选择类型的问答任务;文献[10-15]提出了抽取类型的问答任务;文献[16-18]则构建了生成类型的问答任务。这些数据集多由单一篇章和一系列上下文无关的独立问题组成。最近更多的数据集在原有的基础上提出了更有挑战性的任务,文献[19-23]提出了基于对话的问答任务,数据集中的问题都有上下文联系;文献[24]提出了多跳推理式问答,它需要模型在不同的篇章中整理推理线索,直到找到正确的答案。

但是上述的问答任务都是基于单语语料,在全球化快速发展的今天,一个模型对应一种语言的解决方案越来越难以应对复杂多变的实际情况,如在大规模跨境电商平台中,不同国家的用户都有可能咨询平台中商品,这也意味着模型需要同时回答不同语言提出的问题。然而不同语言的资源情况分布不均。英文和中文等高资源语言更容易构建大规模的问答数据集,而其他低资源语言的问答数据集也不断被提出,

到稿日期:2021-07-14 返修日期:2021-09-15
基金项目:国家重点研发计划(2019QY1802)
This work was supported by the National Key Research and Development Program(2019QY1802).
通信作者:熊德意(dyxiong@tju.edu.cn)

D'Hoffschmidt 等[25]利用维基百科的文章构建了一个法语问答数据集；Efimov 等[26]按照 SQuAD[11]的格式构建了俄语问答数据集；Lim 等[27]则构建了韩语问答数据集。然而，上述的问答数据集仍然属于单语言问答数据集，而且，一些更低资源的语言很难单独构建问答数据集。那么，在全球化不断发展的今天，如何使不同资源的语言之间能够相互促进，让低资源语言能从高资源的语言中受益，在不同语言中挖掘出共通的语言信息，是非常值得研究的问题。

多语言问答近年来得到了研究者的重视，一系列多语言问答数据集被提出[28-38]，研究者们希望通过这些多语言问答数据集来缓解上述相关问题。早期，研究者通过使用“回译”的手段构建多语言问答模型，即使用机器翻译技术将异构的语言转化为同构的语言，将多语言问答转化为单语言问答。然而，这种方法存在不足，首先，在翻译的过程中会产生误差，这种误差的传递会导致问答模型产生错误的结果；其次，翻译的效果仍然受到语言资源不均衡的影响，因为机器翻译的效果会受限于对齐的语料，显然高资源语言的对齐语料更容易获取，如英语到汉语，而低资源的对齐语料则更加困难，如印尼语到英语。

伴随着多语言预训练语言模型[39-49]的进步，上述问题有所缓解，基于预训练的多语言模型利用自监督的方法对自由文本建模，将多种语言映射到同一个语义空间中表示。基于预训练的好处是：首先，将所有的语言在同一语义空间表示就不会存在翻译过程中带来的误差；其次，混合不同语言的预训练将不再依赖对齐的语料，使得低资源语言也可以在整个预训练过程中受益并得到更好的表示。

本文系统地梳理了当前多语言问答的相关工作，首先介绍了多语言问答的相关方法并讨论了它们各自的特点；其次介绍了多语言问答的相关任务；接着总结并统计了这些多语言问答任务的规模、评价指标以及问答方法；最后展望了多语言问答的未来发展方向。

2 多语言问答的相关方法

本节主要介绍了多语言问答的相关方法，具体将它们分为基于特征的方法、基于翻译的方法、基于预训练的方法以及基于双重编码的方法。表 1 对比了多语言问答方法的分类以及优缺点。

表 1 多语言问答方法的分类以及优缺点分析

Table1 Methods of multilingual question answering classification and analysis of advantages and disadvantages

方法	优点	缺点
基于特征	可解释性强，可以通过专家知识编写规则	相关规则和特征的构建需要人工定义，较为烦琐；难以向低资源语言迁移
基于翻译	比较直观，可以充分利用高资源语言问答模型的优势	依赖翻译系统，不够灵活，翻译质量不好时会导致错误传播
基于预训练	迁移性强，便于在低资源语言场景下使用	在一些高资源语言上的效果并不如单语言模型的方法
基于双重编码	不同语言的适配能力更强	泛化能力不强

2.1 基于特征的方法

最早期的工作是为每种类型的问题手工制定规则或者使用一个检索函数[50]为每一个问答对打分。为了训练一个可以为候选答案排序的分类器，可以从问题、候选答案和上下文中提取各种特征。

有用的特征是机器学习算法在自然语言处理任务中成功应用的关键。研究者们通过不同的语言学特征来提高问答模型的性能。文献[51-52]将词法作为特征；文献[53-54]选择句法作为特征；文献[55-56]和文献[57]分别将语义特征和语篇特征作为特征，这些特征在回答诸如“为什么”和“如何”类型的问题时效果尤其明显[58]；文献[59]研究了概率模型在多语言问答排序中的应用。

然而，在基于特征的方法中，无论是撰写规则算法还是机器学习算法，都需要人工定义规则或者特征，正因如此，不依赖于特征工程的方法逐渐成为主流。

2.2 基于翻译的方法

基于翻译的方法是解决多语言问答任务最常见的方法，无论是早期基于特征的方式还是最近基于预训练的方式都可能使用翻译的方法作为语言之间的桥接，具体是将整个问答过程分为两个阶段，首先采用机器翻译模型或者机器翻译工具把多语言问答任务转化为单语言问答任务，再通过单语言问答模型完成建模。举例来说，问题是中文，文本是英文，先使用机器翻译方法把问题翻译为英文，再将翻译的英文问题和文本一起作为英文问答模型的输入，最后得到答案，反之亦然。

Liu 等[29]使用两种基于机器翻译系统的方法：一种方法是把训练集中的源语言翻译成目标语言，然后在翻译后的数据上训练一个标准的问答模型；另一种方法是在源语言训练集上训练问答模型，而问题和待检索的文章来自于目标语言的翻译。Jing 等[28]和 Asai 等[36]在跨语言问答和开放域检索问答任务上使用翻译工具把问题翻译成与文本相同的语言，然后在单语言问答模型上进行测试。Liu 等[30]也将多语言问答任务转换成单语言问答任务，不同的是翻译的对象是段落文本。文献[60]同样在基于翻译的方法上获得了成功。

基于翻译的方法的优势在于可以充分利用单语言问答方法，只要翻译质量尚可，那么问答模型的性能也有所保证。但是，基于翻译的方法并没有解决多语言问答任务的核心问题，即它并没有改变语言资源差异带来的影响，尤其是在语料库中语种分布不平衡时，问答在低资源语料中的迁移效果都会低于高资源语料。而且，基于翻译的方法不够灵活，可能出现翻译错误传播的现象，当多语言问答任务涉及的语种较多时，基于翻译的方法效率有所降低。

2.3 基于预训练的方法

近年来，在大规模高资源语言中训练的单语语言模型在许多基准任务上获得了巨大的成功。然而，随着全球化合作的不断深入，现实世界对多语言模型的需求与低资源语言语料匮乏之间的矛盾也日益凸显，导致训练单一语言模型的资源和成本难以解决。虽然不同的语言表达方式有区别，但是背后表达的语义却是一样。此外，有研究者们发现在同一模

型上同时训练多种语言可以得到比训练多个单一的模型更出色的效果[61]。

以 BERT 为代表的语言模型从大规模无标签的数据中通过自监督的方式学习到表示，然后在下游任务中微调。多语言预训练模型使用非并行的多语言语料库进行训练，例如，mBERT[39]使用了维基百科语料库中的 104 种语言进行训练。Pires 等[62]的研究表明，mBERT 有能力学习泛化的跨语言知识到低资源场景下。这也说明多语言模型可以学习到跨语言表示。

预训练技术的本质是获得更好的句子表示，并将这些知识迁移到更多的下游任务中。为了更好地弥补语言资源的差距，多语言预训练模型被不断提出，其中 mBERT[39]是常用的多语言预训练模型。通过使用多语言预训练模型，在大规模无标签数据训练的过程中低资源语言可以通过共享词表和语言相关性等从高资源语言中受益。多语言预训练模型通过改变模型结构中的层数和参数、训练目标以及可训练的语料库等方面来训练新的语言模型。多语言预训练模型覆盖的语言种类最少有 12 种[42]，最多有 110 种[49]。除此之外，预训练中涉及的目标函数与语料库等与单语言预训练语言模型相似，本文不再赘述。

具体到多语言问答任务中，由于多语言问答往往由人工注释编写，因此它涉及的语种规模小于多语言预训练模型的语言规模，这也使其不需要将问题与文本转化为同一语种，可以直接将多语言预训练模型作为多语言问答的编码器，在模型输出后增加适用于具体任务的分类器即可。例如，在抽取式问答任务中答案是来自文本中的一个片段，因此分类器的输出是一个开始位置和一个结束位置，那么在问题和文本经过编码器后只需要连接两个分类器，一个用来预测开始位置，一个用来预测结束位置即可。

文献[28-36]均使用了 mBERT[39]作为多语言问答实验中的预训练模型，文献[32]和文献[33]的实验中还分别包括 XLM[40]和 XLM-R[41]。需要注意的是，文献[37-38]中提到的预训练方法并非预训练语言模型，因为它们的任务是基于多模态的多语言问答任务，在文中的预训练分别指基于卷积神经网络的图像分类预训练和基于知识图谱表示的预训练。

相比基于翻译的方法，基于预训练的多语言问答方法的优势在于模型在低资源语言中的迁移能力和泛化能力较强，因为所有语言都映射到同一个语义空间，模型不需要考虑语言之间的转换。但是，在一些具有高资源平行语料的问答任务上，基于预训练的方法会弱于基于翻译的方法，因为在这些语言中它们的单语言问答模型性能已经足够强大。

2.4　基于双重编码的方法

在多语言问答任务中，基于翻译的方法和基于预训练的方法并没有消减语言资源不平衡带来的差异。因此，Cui 等[63]针对跨语言问答任务在多语言预训练模型的基础上提出了双重编码的问答模型，新的模型将高资源语言上学习到的知识通过适配注意力计算的方法迁移到低资源语言上，其结果也超过了原有基于翻译和基于预训练的方法。

目前在多语言问答任务中采用双重编码方法的工作并不多，因为随着多语言预训练模型的发展，更多的研究集中在如何通过预训练技术使语言模型具备更强大的多语言表示能力。此外，相比单语言问答任务，多语言问答任务的定义更加灵活多变，使得现阶段并没有一个统一的基准数据集。但是，基于双重编码的方法相比其他类型的多语言问答方法可以更好地适应具体的下游任务，更有效地进行知识迁移。

3　多语言问答的相关任务

多语言问答任务可以分为基于文本的多语言问答和基于多模态的多语言问答两类，表 2 列出了多语言问答任务的分类以及它们的特点。

表 2　多语言问答任务的分类以及特点

Table 2　Tasks of multilingual question answering classification and features

任务	特点
基于文本	问答围绕文本展开
基于多模态	问答围绕与文本不同模型的信息展开，如图像、视频或音频等

3.1　基于文本的多语言问答任务

给定一个英文文本和一个中文问题，多语言问答的任务要求可以给出正确的答案，答案往往来自于文本的一个片段。与单语问答不同的是，模型不仅需要理解问题与文本之间的关系，还需要对齐问题与文本之间语言的差异。近年来，随着深度学习和预训练技术的不断发展，不同语言资源之间的差异也逐渐显现，英语、汉语等资源丰富的语言更容易构建大规模的问答任务，但是一些低资源语料则十分困难，因此，多语言问答可以将高资源语言问答语料库中训练的问答模型迁移到低资源问答语料库中，从而有效缓解语言资源不均衡带来的影响。

基于文本的多语言问答任务可以细分为面向多语言的问答任务和面向跨语言的问答任务。面向多语言的问答任务在整个数据集中包含两个及以上的语种，无论是训练还是测试阶段，面向多语言的问答任务的文本、问题和答案都属于同一语种。而面向跨语言的问答任务中，问题和文本是不同的语种。基于文本的多语言问答任务的示例如图 1 所示。

英文文本： Madame ...This is the Leader's Five Dragon Disc, White Dragon Marshal,' she said. ...You are to return it when you have completed your mission.' 'Yes,' replied Trinket, ... I want you to stay here for a while,' Madame Hong ordered. The rest of you may leave.' Rootless, Black Dragon, and Yellow Dragon saluted and left.

中文文本： 洪夫人...“白龙使，这是教主的五龙令...立功之后，将令缴回。”韦小宝应道：“是。”...三人暂留，余人退去。”无根道人和黑龙使、黄龙使三人行礼退出。

Multi-lingual
英文问题/答案： Who did Madame Hong allow to leave first? / Rootless, Black Dragon, and Yellow Dragon
中文问题/答案： 韦小宝在神龙教中担任什么职位？/ 白龙使

Cross-lingual
英文问题： Who did Madame Hong allow to leave first?
中文答案： 无根道人和黑龙使、黄龙使

图 1　基于文本的多语言问答示例[28]

Fig. 1　Example of text-based multilingual question answering[28]

表3统计了已有的基于文本的多语言问答任务中覆盖的子任务，即单语言、多语言以及跨语言3类任务。

表3 多语言问答数据集细分任务

Table 3 Subtasks for multilingual question answering datasets

Dataset	Monolingual	Multilingual	Cross-Lingual
XQuAD[31]	No	No	Yes
XQA[29]	No	No	Yes
BiPaR[28]	Yes	Yes	Yes
XCMRC[30]	No	No	Yes
MLQA[32]	No	Yes	Yes
EXAMS[33]	No	Yes	Yes
LAReQA[34]	No	No	Yes
TyDi QA[35]	No	Yes	No
XOR QA[36]	No	No	Yes

从表3可以看出，单语言任务显然不再是研究的重点，而多语言任务，尤其是跨语言任务才是研究的重心。因为跨语言任务的重要应用场景就是将高资源语言上训练的模型应用到低资源语言上，这也是多语言问答任务研究的主要动机。下面将逐一介绍已有的多语言问答任务。

Jing等[28]提出了一个中英文双语问答数据集，从中英文小说中收集了3 667个文本段落，通过人工标注的方式构建了14 668个问答对。他们进一步将其拆分为7个子任务，包括单语问答任务、多语言问答任务和跨语言问答任务。实验结果表明，现阶段的问答模型离人类的水平还有很大的差距。

Liu等[29]提出了一个开放域跨语言问答数据集来解决现阶段开放域问答语料库集在英文等高资源语言上的问题。具体来说，该数据集的训练集由英文组成，而开发集和测试集由其他低资源语言组成。实验结果说明，跨语言开放域问答数据集的性能不但受不同语言之间的相似度影响，而且还受目标语言的问题难度影响。

Liu等[30]提出了一个跨语言完形填空式问答数据集，读者通过阅读一段源语言文本去补充一个由目标语言组成的句子。同样实验结果表明，现阶段的模型依然有很大的提升空间。

Artetxe等[31]在英语问答数据集SQuAD[11]的开发集的基础上通过专业的翻译将其翻译成10种语言，最终获得11种并行语言的问答数据集。数据集由240个文本和1 190个问答对组成。

Lewis等[32]提出了一个多路对齐的抽取式问答数据集，包含7种语言，即英语、阿拉伯语、德语、西班牙语、北印度语、越南语和简体中文。数据集包括12 000个英语问答对和5 000个由其他语言构成的问答对，每一个问答对平均由4种语言组成。

Hardalov等[33]提出了一个基于高中考试的多语言和跨语言问答数据集。数据集由24 000个高质量的高中考试问题组成，包含16种语言，覆盖8个语系和来自自然科学和社会科学中的24个学校科目。该数据集旨在促进学校考试问答中的知识迁移和推理方面的提升。

Roy等[34]提出了一个从多语言池中进行语言不可知论的答案检索问答任务，用来测试多语言模型在跨语言之间的校准能力，它需要语义相关的跨语言对在语义空间中更加靠近，而不是同语言但不同语义的语言对。实验结果表明，语言不可知论的检索本质上也是一种新的跨语言评测任务。

Clark等[35]提出了一个覆盖11种语言的问答数据集，包含204 000个问答对。为了更符合实际的信息搜寻场景，数据集中的问题由那些想知道答案而不知道答案的人员编写，同时每种语言的问答对都是直接编写而不是使用机器翻译。

Asai等[36]提出了一个基于开放式检索的跨语言问答数据集，即给定一个非英文的问题，需要模型在英文文本中检索并返回答案。该数据集包含来自7个非英语语言的40 000个问题。该数据集又进一步分成3个涉及跨语言文档检索的子任务。

3.2 基于多模态的多语言问答任务

人类真实的交互场景中往往混合多种模态信息，如视觉和文本、语音和视觉或者语音与文本。以问答为例，在单模态下，问题和答案会围绕一段文本或者一个知识库进行；而在多模态下，问答的主题可能围绕一张图片或者一段视频。多模态场景下的任务显然更符合现实场景，多模态模型的适用性和可迁移性也会更好。

因此，在多语言问答的研究中，研究人员同样提出了多模态场景下的多语言问答任务。如图2所示，基于多语言设定的问题和答案围绕一张给定的图片进行。

图2 基于多模态的多语言问答示例[37]

Fig. 2 Example of multi-modal-based multilingual question answering[37]

Gao等[37]提出了一个多语言的图像问答数据集，包含150 000张图片和310 000个问答对以及它们的英文翻译。通过数据集训练的问答模型将以人工方式进行评测，以此来鉴定模型生成答案的质量。

Ramnath等[38]提出了一个基于事实的视觉口语问答数据集。它需要从知识图谱中检索到相关的实体来回答关于一个图片的问题。该数据集中的问题更加口语化，并且问题的语言与知识图谱的描述语言是不一致的。

4 多语言问答的基准任务、数据集与评测

基于多语言问答的丰富性，接下来将分别从多语言数据

集问答数据集的规模、任务以及评价指标3个方面进行阐述。　表4列出了多语言问答的各项统计对比。

表4　多语言问答数据集的统计比较

Table 4　Statistical comparison of multilingual question answering datasets

Dataset	Passages or Images	Questions and Answers	Language Number	Task	Domain	Multi-model	Metric	Translate	Pre-train
FM-IQA[37]	150 000	310 000	2	Generation	Flickr	Yes	Human	No	Yes
XQuAD[31]	240	1 190	11	Extractive	Wikipedia	No	EM and F1	No	Yes
XQA[29]	906 100	90 610	9	Open-domain	Wikipedia	No	EM and F1	Yes	Yes
BiPaR[28]	3 667	14 668	2	Extractive	Novels	No	EM and F1	Yes	Yes
XCMRC[30]	113 589	113 589	2	Cloze-style	News	No	Accuracy	Yes	Yes
MLQA[32]	—	46 000	7	Extractive	Wikipedia	No	EM and F1	No	Yes
EXAMS[33]	—	24 000	16	Multiple-choice	High School Examinations	No	Accuracy	No	Yes
LAReQA[34]	—	17 289	11	Retrieval	Wikipedia	No	precision	No	Yes
TyDi QA[35]	—	204 000	11	Extractive	Wikipedia	No	F1	No	Yes
WoW[38]	2 190	5 826	3	Retrieval	Synthetic	Yes	Accuracy	No	Yes
XOR QA[36]	—	40 000	7	Retrieval	Wikipedia	No	EM and F1	Yes	Yes

4.1　数据集的规模统计

与单语言问答数据集的标注相同的是，大部分多语言问答数据集也是通过人工的方式进行标注，但是，由于多语言问答数据集的语言多样性和数据集的构建动机不同，因此不同数据集的问答对规模之间的差异巨大。举例来说，有些数据集通过双语对齐或者多语对齐等方式进行问答数据集的问答对构建；有些数据集通过在已有的单语言问答数据集上进行问答对翻译并扩充到更多的语种上。本小节分3个维度对比多语言问答数据集的规模，包括文本或图像的规模、问答对规模和语种规模。表4的第2－4列列出了不同的多语言问答数据集的规模比较。

如表4的第2－4列所列，多语言问答数据集涵盖的语种有两种语言的双语问答数据集，也有十种语言以上的多语言问答数据集。然而，并不是所有的数据集任务定义都包含背景文本或者图像，因为一些任务中是需要模型进行段落检索或者在知识图谱中进行实体检索，所以部分多语言问答数据集没有具体的文本或者图像数量。

4.2　数据集的基准任务

多语言问答数据集的任务类型与单语言问答任务比较一致，因为多语言问答数据集与单语言的主要差别是研究在多语言或者跨语言场景下研究模型的。因此多语言问答数据集的任务类型还是与单语言问答任务保持一致，而且更加单一。本小节分三个维度对比多语言问答数据集的任务类型，包括任务类型、领域类型和是否是多模态。任务类型更多的集中在几种最常见的问答任务类型，如抽取式等；领域类型中基于维基百科作为数据集语料的占多数；同样，大部分数据集都是基于单模态的。表4的第5－7列列出了各个数据集的任务对比情况。

如表4的第5－7列所列，抽取式和检索式任务是常见的多语言问答任务，多项选择式和完形填空式任务的种类很少；数据集语料最多的依然是维基百科，并且绝大多数任务都是基于单模态的。说明在多语言问答数据集构建中，当前的研究重点并不是语料覆盖的领域和模态上多样化，相对的“单一”任务更符合现阶段多语言问答的研究目标。

4.3　数据集的评测

多语言问答数据集的评价指标更加集中，除了个别任务外，EM和F1是最主要的两种指标，这也是问答任务中最常用的评价指标。除了评价指标外，本小节还从是否使用翻译技术和是否使用预训练技术这两个维度进行对比。是否使用翻译代表在数据集提供的基础方法，是否使用翻译技术将不同种语言翻译为同种语言，将多语言问答任务转换成单语言问答任务；是否使用预训练技术代表数据集中的基础方法是否使用了多语言预训练语言模型。表4的第8－10列从以上三个维度对比了各个数据集的情况。

如表4的第8－10列所列，评价指标和是否使用翻译技术都如之前所述，值得注意的是预训练技术在所有数据集的基础方法中都有使用，虽然在小部分数据集中使用的预训练技术并不是预训练语言模型，但也足以说明预训练技术目前已经是多语言问答任务的基准方法，值得更进一步的研究。

5　多语言问答研究展望

本文分别总结了多语言问答进展、问答方法以及数据集和评价指标的相关研究。结合当前多语言问答的研究进展，本文针对当前研究的不足总结出一些在未来值得深入讨论和研究的问题和方向。

(1)更丰富的多语言问答数据集。多语言问答数据集的任务设置是未来多语言问答研究的基础问题。当前的多语言问答数据集的设置还相对单一，多数任务集中在开放域问答和抽取式问答等任务上。然而，目前单语言问答数据集经过多年的发展，已经覆盖更多的场景。相比之下，多语言问答数据集的设置还应该向更多的场景延伸，如多篇章推理、多人多轮对话式以及数字逻辑计算式等方向。同时，多语言问答的答案类型也应该更加多元，目前大部分数据集的任务集中在抽取式，而完型填空式、选择题式、生成式等答案类型的数据集构建也应该加强。此外，多语言问答数据集与单语问答数据集相比，除了涵盖的语言种类更加丰富和多元，在设计的过程中还应该重视不同语言之间语系和表达习惯等方面的差异，在标注过程中，问题的设计应该更着重于语义方向的推理，减少有线索词提示的简单问题。这样，通过设计更多类型的多语言问答数据集，不但可以有效促进多语言问答模型的发展，而且也能够使多语言问答的研究更接近实际的应用场

景，从而促进其在实际场景中的应用。

(2)多语言问答中的知识融合。虽然在近年，随着预训练技术的发展，多语言模型的研究达到了新的高度。但由于语言模型中的训练语料多是来自于维基百科，缺乏特定知识的指导，因此融合知识的多语言问答将是未来问答领域研究的重点之一。首先，常识知识是人类后天成长中吸收到各类型知识的统称，如“阴天下雨不会出现太阳”“海水是咸的”。这类知识是任何种族、国家、职业的人都应该了解的，不需要进行额外的说明。如今，大部分问答任务依然属于事实型问答，即任务中问题的答案往往是由一段文本或知识库等提供，如“世界上哪座山的海拔最高?”，但如果背景文本或知识库中没有提及，问答系统将无法回答该问题。但是，通过常识知识的融合，问答系统同样可以得出正确答案，即“珠穆朗玛峰”。融入常识知识可以使问答系统更加智能化和拟人化，同时可以使问答系统处理更加多样性的问题，而不仅仅依赖于背景文本或者知识库。与常识知识区分的另一类知识是领域知识，如金融知识、法律知识和医疗知识等。这类知识往往具有较高的知识壁垒，非该专业领域的人员无法触及，也正因为如此，融合领域知识的问答任务标注成本昂贵，所以常规的问答任务缺乏对专业领域的探究。融合领域知识的问答系统受益于预训练技术的影响低于常规领域，原因如前文所述，预训练技术缺乏对领域知识的学习。因此，融入领域知识可以使问答系统更加专业化，更好地成为人类的助手。因此，融入知识对多语言问答的研究有很大的促进作用，同时也能减小不同语言不同文化背景带来的差异。

(3)新型多语言问答模型。当前多语言问答模型的主要进展来自于多语言预训练语言模型的进步。与单语言预训练语言模型相比，该模型的网络结构、训练目标以及参数量基本一样，只是在训练过程中混入更多的语种。但是，不同语言之间的表述和习惯等方面的因素并没有得到特殊的考虑，因此，针对语言差异和语言表达方面的模型设计需要被研究者们重视，通过设计更有针对性的训练目标和网络结构来改善多语言问答模型的性能，尤其是保持高资源语言中单语言问答模型的性能，提高低资源语言中问答的性能。另一方面，预训练技术本身可以得到更好的上下文表示，但是在问答任务中理解和推理也是更高阶的任务类型，因此，在模型设计中也应该考虑更多上下文内容理解和推理的目标，从而促进多语言问答模型的发展。

(4)提高多语言问答模型的可解释性。与其他自然语言处理模型一样，当前的多语言问答模型的一个很重要的缺点是可解释性不强。多语言问答数据集往往是通过标注产生的有监督数据集，模型从这些数据中学习到规模，但是人们不知道模型输出的答案的依据是什么。而且问答数据集构建的过程也会有一些与标注者自身表达习惯有关的偏差。因此，模型在训练中有可能学习到一些虚假的相关性，虽然答案正确，但推理路径是完全错误的。所以在设计多语言问答模型时重视其可解释性是非常值得研究的问题。

结束语 当前是一个全球化合作深度融合的时代，也是我国“一带一路”战略发展的关键时期。多语言问答的研究有利于改善不同国家的语言习惯差异和语言资源间的差异，提高各国合作的效率，节省人力成本，通过智能化技术推动各国之间的协作发展。基于此背景，本文对当前多语言问答的研究进行了阐述，从多语言问答研究进展、多语言问答模型以及多语言问答数据集和评价3个方面进行了总结，最后对多语言问答研究进行了展望。

参 考 文 献

[1] HERMANN K M, KOCISKY T, GREFENSTETTER E, et al. Teaching machines to read and comprehend[J]. Advances in Neural Information Processing Systems, 2015, 28: 1693-1701.

[2] HILL F, BORDES A, CHOPRA S, et al. The goldilocks principle: Reading children's books with explicit memory representations[J]. arXiv: 1511.02301, 2015.

[3] XIE Q, LAI G, DAI Z, et al. Large-scale Cloze Test Dataset Created by Teachers[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2344-2356.

[4] CUI Y, LIU T, CHEN Z, et al. Consensus Attention-based Neural Networks for Chinese Reading Comprehension[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 1777-1786.

[5] KEMBHAVI A, SEO M, SCHWENK D, et al. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4999-5007.

[6] WELBL J, LIU N F, GARDNER M. Crowdsourcing Multiple Choice Science Questions[C]// Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017: 94-106.

[7] OSTERMANN S, MODI A, ROTH M, et al. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[8] LAI G, XIE Q, LIU H, et al. RACE: Large-scale Reading Comprehension Dataset From Examinations[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 785-794.

[9] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge [J]. arXiv: 1803.05457, 2018.

[10] YANG Y, YIH W, MEEK C. Wikiqa: A challenge dataset for open-domain question answering[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 2013-2018.

[11] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100000+Questions for Machine Comprehension of Text[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.

[12] TRISCHLER A, WANG T, YUAN X, et al. NewsQA: A Machine Comprehension Dataset[C]// Proceedings of the 2nd Workshop on Representation Learning for NLP. 2017: 191-200.

[13] DUNN M, SAGUN L, HIGGINS M, et al. Searchqa: A new q&a dataset augmented with context from a search engine[J]. arXiv:

1704.05179,2017.

[14] JOSHI M,CHOI E,WELD D S,et al. TriviaQA:A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1601-1611.

[15] RAJPURKAR P,JIA R,LIANG P. Know What You Don't Know:Unanswerable Questions for SQuAD[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:784-789.

[16] HE W,LIU K,LIU J,et al. DuReader:a Chinese Machine Reading Comprehension Dataset from Real-world Applications [C]// Proceedings of the Workshop on Machine Reading for Question Answering. 2018:37-46.

[17] NGUYEN T,ROSENBERG M,SONG X,et al. MS MARCO:A Human Generated MAchine Reading COmprehension Dataset [J]. arXiv:1611.09268,2016.

[18] KOCISKY T,SCHWARZ J,BLUNSOM P,et al. The narrativeqa reading comprehension challenge[J]. Transactions of the Association for Computational Linguistics,2018,6:317-328.

[19] IYYER M,YIH W,CHANG M W. Search-based neural structured learning for sequential question answering[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:1821-1831.

[20] SAHA A,PAHUJA V,KHAPRA M,et al. Complex Sequential Question Answering:Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[21] TALMOR A,BERANT J. The Web as a Knowledge-Base for Answering Complex Questions[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2018:641-651.

[22] REDDY S,CHEN D,MANNING C D. Coqa:A conversational question answering challenge[J]. Transactions of the Association for Computational Linguistics,2019,7:249-266.

[23] CHOI E,HE H,IYYER M,et al. QuAC:Question Answering in Context[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:2174-2184.

[24] YANG Z,QI P,ZHANG S,et al. HotpotQA:A Dataset for Diverse,Explainable Multi-hop Question Answering[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:2369-2380.

[25] D'HOFFSCHMIDT M,BELBLIDIA W,HEINRICH Q,et al. FQuAD:French Question Answering Dataset[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:Findings. 2020:1193-1208.

[26] EFIMOV P,CHERTOK A,BOYTSOV L,et al. SberQuAD-Russian reading comprehension dataset:Description and analysis [C]// International Conference of the Cross-Language Evaluation Forum for European Languages. Cham:Springer,2020:3-15.

[27] LIM S,KIM M,LEE J. KorQuAD1.0:Korean QA dataset for machine reading comprehension[J]. arXiv:1909.07005,2019.

[28] JING Y,XIONG D,YAN Z. BiPaR:A Bilingual Parallel Dataset for Multilingual and Cross-lingual Reading Comprehension on Novels[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019:2452-2462.

[29] LIU J,LIN Y,LIU Z,et al. XQA:A cross-lingual open-domain question answering dataset[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2358-2368.

[30] LIU P,DENG Y,ZHU C,et al. XCMRC:Evaluating cross-lingual machine reading comprehension[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Cham:Springer,2019:552-564.

[31] ARTETXE M,RUDER S,YOGATAMA D. On the Cross-lingual Transferability of Monolingual Representations[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:4623-4637.

[32] LEWIS P,OGUZ B,RINOTT R,et al. MLQA:Evaluating Cross-lingual Extractive Question Answering[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:7315-7330.

[33] HARDALOV M,MIHAYLOV T,ZLATKOVA D,et al. EXAMS:A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:5427-5444.

[34] ROY U,CONSTANT N,AL-RFOU R,et al. LAReQA:Language-agnostic Answer Retrieval from a Multilingual Pool[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:5919-5930.

[35] CLARK J H,CHOI E,COLLINS M,et al. TyDi QA:A benchmark for information-seeking question answering in typologically diverse languages[J]. Transactions of the Association for Computational Linguistics,2020,8:454-470.

[36] ASAI A,KASAI J,CLARK J H,et al. XOR QA:Cross-lingual Open-Retrieval Question Answering[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2021:547-564.

[37] GAO H,MAO J,ZHOU J,et al. Are you talking to a machine? Dataset and methods for multilingual image question answering [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2015:2296-2304.

[38] RAMNATH K,SARI L,HASEGAWA-JOHNSON M,et al. Worldly Wise (WoW)-Cross-Lingual Knowledge Fusion for Fact-based Visual Spoken-Question Answering[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2021:1908-1919.

[39] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019:4171-4186.

[40] CONNEAU A,LAMPLE G. Cross-lingual language model pre-

training[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:7059-7069.
[41] 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.
[42] KAKWANI D, KUNCHUKUTTAN A, GOLLA S, et al. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020:4948-4961.
[43] KHANUJA S, BANSAL D, MEHTANI S, et al. Muril: Multilingual representations for indian languages[J]. arXiv: 2103.10730, 2021.
[44] LUO F, WANG W, LIU J, et al. Veco: Variable encoder-decoder pre-training for cross-lingual understanding and generation[J]. arXiv: 2010.16046, 2020.
[45] CHI Z, DONG L, WEI F, et al. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021:3576-3588.
[46] PHANG J, CALIXTO I, HTUT P M, et al. English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too[C]// Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. 2020:557-575.
[47] OUYANG X, WANG S, PANG C, et al. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora[J]. arXiv: 2012.15674, 2020.
[48] HU J, JOHNSON M, FIRAT O, et al. Explicit Alignment Objectives for Multilingual Bidirectional Encoders[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021:3633-3643.
[49] CHUNG H W, FEVRY T, TSAI H, et al. Rethinking Embedding Coupling in Pre-trained Language Models[C]// International Conference on Learning Representations. 2020.
[50] ROBERTSON S, ZARAGOZA H, TAYLOR M. Simple BM25 extension to multiple weighted fields[C]// Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. 2004:42-49.
[51] BRILL E, LIN J, BANKO M, et al. Data-intensive Question Answering[J]. NIST Special Publication, 2002(500-250):393-400.
[52] ATTARDI G, CISTERNINO A, FORMICA F, et al. PIQASso: PIsa Question Answering System[C]// Text Retrieval Conference 10. NIST, 2001:566-607.
[53] ALFONSECA E, DE BONI M, JARA-VALENCIA J L, et al. A prototype question answering system using syntactic and semantic information for answer retrieval[J]. NIST Special Publication, 2002(500-250):680-685.
[54] KATZ B, BORCHARDT G, FELSHIN S. Syntactic and semantic decomposition strategies for question answering from multiple resources[C]// Proceedings of the AAAI 2005 Workshop on Inference for Textual Question Answering. Menlo Park, CA: AAAI Press, 2005:35-41.
[55] CUI H, SUN R, LI K, et al. Question answering passage retrieval using dependency relations[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2005:400-407.
[56] HOVY E, GERBER L, HERMJAKOB U, et al. Toward semantics-based answer pinpointing[C]// Proceedings of the First International Conference on Human Language Technology Research. 2001:1-7.
[57] MORTON T S. Using coreference for question answering[C]// Proceedings of the Workshop on Coreference and its Applications. 1999:85-89.
[58] KOLOMIYETS O, MOENS M F. A survey on question answering technology from an information retrieval perspective[J]. Information Sciences, 2011, 181(24):5412-5434.
[59] KO J, SI L, NYBERG E, et al. Probabilistic models for answer-ranking in multilingual question-answering[J]. ACM Transactions on Information Systems (TOIS), 2010, 28(3):1-37.
[60] TURE F, BOSCHEE E. Learning to Translate for Multilingual Question Answering[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:573-584.
[61] NI M, HUANG H, SU L, et al. M3p: Learning universal representations via multitask multilingual multimodal pre-training [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:3977-3986.
[62] 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.
[63] CUI Y, CHE W, LIU T, et al. Cross-Lingual Machine Reading Comprehension[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019:1586-1595.

LIU Chuang, born in 1990, Ph.D. His main research interests include question answering and commonsense reasoning.

XIONG De-yi, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine translation, dialogue, and natural language generation.

(责任编辑:李亚辉)