机器音译研究综述

李卓[1,2]　王志娟[1,2,*]　赵小兵[1,2]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究少数民族语言中心
*Corresponding author: Zhijuan Wang
li_zhuo98@163.com, wangzj.muc@gmail.com, nmzxb_cn@163.com

摘要

机器音译是基于语音相似性自动将文本从一种语言转换为另一种语言的过程，它是机器翻译的一个子任务，侧重于语音信息的翻译。音译后可知道源单词在另一种语言中的发音，使不熟悉源语言的人更容易理解该语言，有益于消除语言和拼写障碍。机器音译在多语言文本处理、语料库对齐、信息抽取等自然语言应用中发挥着重要作用。本文阐述了目前机器音译任务中存在的挑战，对主要的音译方法进行了剖析、分类和整理，对音译数据集进行了罗列汇总，并列出了常用的音译效果评价指标，最后对该领域目前存在的问题进行了说明并对音译学的未来进行了展望。本文以期对进入该领域的新人提供快速的入门指南，或供其他研究者参考。

关键词： 音译；综述；语料库；评价指标

Survey on Machine Transliteration

Zhuo Li[1,2]　**Zhijuan Wang**[1,2,*]　**Xiaobing Zhao**[1,2]
[1]School of Information Engineering, Minzu University of China
[2]Natural Language Resource Monitoring and Research Center of Minority Languages
*Corresponding author: Zhijuan Wang
li_zhuo98@163.com, wangzj.muc@gmail.com, nmzxb_cn@163.com

Abstract

Machine transliteration, the process of automatically converting text from one language to another based on phonetic similarity, is a subtask of machine translation that focuses on the translation of phonetic information. After transliteration, you can know the pronunciation of the source word in another language, making it easier for people who are not familiar with the source language to understand the language, and it is beneficial to eliminate language and spelling barriers. Machine transliteration plays an important role in natural language applications such as multilingual text processing, corpus alignment, and information extraction. This paper expounds the challenges existing in the current machine transliteration tasks, analyzes, categorizes and organizes the main transliteration methods, summarizes the transliteration data sets, and lists the commonly used evaluation indicators of transliteration effects. The existing problems are explained and the future of transliteration is prospected. This article is intended to provide a quick introductory guide for newcomers to the field, or as a reference for other researchers.

Keywords: Transliteration , Research Status , Corpus , Evaluation Metrics

1 引言

机器音译是指利用计算机将源语言中的给定名称(源书写系统或拼写体系中的文本字符串)自动转换为目标语言中的名称(目标书写系统或拼写体系中的另一文本字符串)(Wei, 2004)。关于目标语言中名称表示的具体要求如下：它符合目标语言的音系，在语音上等同于源名称，并且与源语言名称的对等上符合用户的直觉。例如炒面在伦敦的中餐馆菜单里常被写作Chow Mein。机器翻译、数据挖掘以及跨语言信息检索和抽取等系统的性能极大依赖于命名实体(人名、地名、机构名、专有名词等)的音译准确性，尤其在涉及到人名、专有名称、技术术语时。因此，研究机器音译有重要的意义。

机器音译按照源语言(音译输入语言)与起源语言(来源于何种语言)是否一致可分为正向音译与反向音译(Mammadzada, 2021)。将单词从其起源语言音译为外语称之为正向音译。例如将张三(汉语)音译为Zhang San(英语)。而将用本语言拼写的外语词音译回起源语言称之为反向音译。例如将Zhang San(英语)音译回张三(汉语)。反向音译相比于正向音译来说更加困难。这是因为反向音译需要消除在正向音译中引入的噪声，正向音译的过程中往往会过滤掉不发音的音节，例如De Sciglio(意大利语)音译为德西利奥(汉语)，其中的字母g不发音。此外反向音译往往不允许有音译变体，它应该尽可能的接近原词也是反向音译更困难的一个重要的原因。比如说雷欧、李傲(汉语)音译为英语只有一个正确结果Leo。

音译与翻译和转写/转录都有所不同(Zepedda, 2020)。翻译在于使用不同语言传达语句的原始意义，其并不知道单词在原始语言中的发音。翻译与音译相反，它更注重单词的意义而不是发音。而转写是将一种字母表中的字符转换为另一种字母表中字符的过程[0]。转写字符之间是一一对应转换的，即被转换字母表中的每一个字符只能转换为另一个字母表中的一个字符，才能保证两个字母表能够完全的、无歧义的转换(冯志伟, 2012)。例如阿拉伯语单词كتب，其英语翻译为book，英语音译结果为kataba，而拉丁转写结果为ktb。

由于不同语言之间的较大差异性，音译任务存在着诸多困难与挑战。

一是源语言与目标语言使用的是不同的字母体系。例如拉丁/罗马字母源于希腊字母，它作为罗马文明的成果之一，随着征服推广到西欧地区。西里尔/斯拉夫字母是通行于斯拉夫语族部分民族中的字母书写系统。而阿拉伯/天方字母则在伊斯兰教兴盛的地区使用。音译处理的过程中需要了解不同字母体系中的字符编码。此外字母体系的书写方向也是必须要考虑的一点。例如阿拉伯字母、希伯来字母、波斯字母、乌尔都字母遵循从右到左的书写原则，而罗马字母、西里尔字母、婆罗米字母遵循从左到右的书写原则(Prabhakar and Pal, 2018)。

二是音译变体的存在。由于音译是一个基于个人认知的创造性过程，导致不同的专业音译者也有不同的观点。此外，同一种语言存在的不同方言也会导致音译变体的存在。而在音译语料的搜集过程中很难捕获到所有的变体。这种情况会让音译的质量评估变得很困难，因此很难建立起让所有人都信服的音译评估标准。

三是不同字母体系中涵盖音的范围不同，会导致发音缺失的问题。这与春秋时期创立的音阶——宫商角徵羽只能对应于现代音阶的do、re、mi、sol、la相类似(Jacques, 2017)。这将导致目标字母体系中缺少某些发音就必须使用多个字母来近似表示其发音，甚至会出现字母组合后仍无法找到类似发音的情况。因此需要让音译模型学习如何“创造”出缺失的相似发音，以保持发音的完整。

四是很难让音译模型学会“察言观色”。音译通常是对命名实体进行的。但如何让系统判断不同词采用音译还是翻译，需要模型通过从大量的训练语料或上下文中意识到这一点。例如Kunlun Mountains(英语)，第一个单词应该音译为昆仑(汉语)，而第二个单词应该翻译为山(汉语)。这对于传统的音译方法来说有着巨大的挑战，而基于深度学习的音译方法通过大量语料的学习和在注意力机制的帮助下相对来说能较容易的学习到这一点。

本文的组织方式如下。第二节描述了音译涉及到的主要语言。第三节综合阐述了具有代表性的音译方法，并对它们进行了分类整理。第四节罗列了音译的相关语料库资源。第五节介绍了音译质量/性能评估中常使用的指标。第六节对整个音译学的未来进行了展望，讨论了未来的工作方向。第七节对全文进行了总结。

[0]在实践中，如果两种语言中字母和声音之间的关系相似，则音译与转写十分接近。

第二十一届中国计算语言学大会论文集，第317页-第332页，南昌，中国，2022年10月14日至16日。

2 音译相关语言

我们统计了来自于谷歌学术的400篇与音译相关的论文，整理了其中涉及到的36种主要语言对，如表1所示。可以看到英语作为世界上使用最广泛的语言，对其相关语言对的研究占比最高。其次是语言使用人数/使用地区较多的汉语、阿拉伯语、日语、印地语等。而对于使用人数/使用地区较少的语言的研究并不多。从语系的角度来看，印欧语系(梵语、英语、旁遮普语、孟加拉语、波斯语等所属的印度-伊朗语族，法语、西班牙语所属的罗曼语族，英语、瑞典语所属的日耳曼语族，以及以俄语为代表的的斯拉夫语族)是主要的研究对象，而汉藏语系(汉语、藏语、泰语)、闪含语系(阿拉伯语、希伯来语)、南亚语系(越南语)、希腊语族(希腊语)和其他语系的研究相对较少[1]。

英语⟷汉语	英语⟷朝鲜语	英语⟷希伯来语	英语⟷阿拉伯语
英语⟷日语	英语⟷泰语	英语⟷波斯语	英语⟷孟加拉语
英语⟷俄语	英语⟷印地语	英语⟷旁遮普语	英语⟷西班牙语
英语⟷越南语	汉语⟷朝鲜语	英语⟷马拉地语	英语⟷泰卢固语
汉语⟷日语	日语⟷朝鲜语	印地语⟷乌尔都语	印地语⟷旁遮普语
英语⟶希腊语	英语⟶法语	英语⟶泰米尔语	英语⟶坎纳达语
英语⟶奥里亚语	梵文⟶英语	马拉地语⟶英语	古吉拉特语⟶英语
瑞典语⟶芬兰语	法语⟶日语	法语⟶汉语	法语⟶阿拉伯语
印地语⟶坎纳达语	藏语⟶汉语	西班牙语⟶汉语	阿拉伯语⟶印地语

Table 1: 音译研究涉及的主要语言对。日语包括日语汉字和日语片假名。朝鲜语包括朝鲜语汉字和朝鲜语谚文。朝鲜语同韩语，下同。A⟶B表示源语言A到目标语言B的音译方法。A⟷B表示A、B均可作为源语言或目标语言，下同。

3 方法

获取音译的方法主要有两类：音译生成和音译挖掘。音译生成是将一种语言中的给定单词自动生成为另一种语言表示的对应音译的过程。对于新的词语(不来源于训练语料)，生成性音译系统也可以自动生成目标语言音译词。音译挖掘是从不同资源中提取/挖掘音译对的过程，资源可以是平行语料库或可比语料库，也可以是两种语言之间的Web资源。音译对的自动提取可以用获取的新的音译对来丰富现有的音译语料库，减轻建设音译生成所需语料库的人力劳动。除这两类方法之外，还有一些方法将生成和挖掘结合起来进行音译，可以称它们为融合方法或者混合方法。

3.1 音译生成

用于音译生成的模型包括基于信道的模型、支持向量机、最大熵模型、决策树、隐马尔科夫模型、条件随机场、循环神经网络和Transformer等。**噪声信道模型(NCM)**以假设目标语言的文本T(信道意义上的输入)经过噪声信道变为源语言的文本S(信道意义上的输出)。音译模型将观察到的源语言文本S，转换成最有可能产生S的目标语言文本T'，即利用贝叶斯规则将$p(y|x)$重写为$p(x|y) * p(y)/p(x)$。由于NCM有两个组件模型($p(x|y)$和$p(y)$)意味着可以把整体的问题单独的解决，但它也受限于过高的解码成本。**信源信道模型(SCM)**是基于贝叶斯定理的混合模型，它借鉴了基于规则和统计方法的思想。其优点在于考虑了表示源语言单词的语音属性的字素，但其切分产生的错误会传播到后面的步骤中，会导致生成错误的音译结果。此外，由于两种语言都要生成可能的字素，因此时间复杂度较高。**联合信源信道模型(JSCM)**通过n-gram音译模型在不同语言之间进行直接拼写映射，JSCM与SCM相比，它在不分解联合概率的情况下估计了最优的音译结果字符，但JSCM并未使用语言学知识，且用概率模型识别源语言中的音译单元准确性仍有待提高。**支持向量机(SVM)**是一种用于二分类的机器学习算法，通过在特征空间上找到最佳的分隔超平面使得正负样本间隔最大。由于音译是一个多分类问题，因此需要先对问题进行二值化。在训练阶段，为两个不同字母体系的语言中每个字母训

[1]朝鲜语和日语所属语系现在仍存在着争议。

练一个SVM，就可在给定源语言字母序列下预测所有可能的类标签，并选择最可能的类标签。当观测样本较多时效率会明显下降。此外，核函数和参数的选择对结音译结果的影响很大。**决策树(DT)**是在已知各种情况发生概率的基础上，通过构建决策树学习如何将每个源字素转换为目标字素，从而生成音译结果。DT的优点在于它充分的考虑了广泛的上下文信息，但缺少语音信息的考虑。**最大熵模型(MEM)**是由最大熵原理推导实现的。最大熵原理可以表述为在满足约束条件的模型集合中选取熵最大的模型，即不确定性最大的模型。在选定特征作为约束且分配权重后，其余对音译有影响的特征将同等对待。MEM的优势在于它可以灵活地选择特征，鲁棒性强。但当样本量大时，对偶函数优化求解的迭代过程缓慢，开销较大。**隐马尔科夫模型(HMM)**是一种有限状态自动机，通过定义观察序列和标记序列的联合概率对生成过程进行建模。它由一个隐藏的马尔科夫链根据状态转移概率，随机生成一个状态随机序列，然后再由每个状态根据观测概率生成各自对应的一个观测，由此构成可观测的随机序列。因此输出标签的概率取决于当前输入标签和之前的输出。HMM的缺点在于它只依赖每一个状态和它对应的观察对象，忽略了观察序列的长度和上下文信息。**条件随机场(CRF)**是条件概率分布模型，定义了给定特定观察序列X(一串源语言音译单元)的标签序列Y(一串目标语言音译单元)上的条件概率$P(Y|X)$。条件随机场的优势在于避免了标签偏置的问题，所有特征能进行全局归一化，它相比HMM来说不需要独立性假设条件，因此可以容纳任意的上下文信息，但其训练代价大、复杂度高。**循环神经网络(RNN)**是处理序列数据的神经网络。初始化的字符向量表征传输到RNN中，以获取语义向量表征，并构建一个全连接网络来预测序列此时的隐藏状态，再根据隐藏状态预测出标签。但其不具备长期记忆，且会造成梯度消失的问题。而LSTM模型在此基础上加入了遗忘机制，选择性的保留或遗忘前期的某些数据，并用加法代替乘法解决了梯度爆炸的问题，但仍然难以捕获长距离的依赖关系，且RNN的序列递归结构使得难以并行化计算，效率过低。**Transformer**是一种全新架构的神经网络，它利用自注意机制解决了RNN和CNN中难以充分利用上下文信息的缺点。对齐后的双语预料的每个字符/音节被表征成向量，输入到编码器后经过自注意力层和全连接层进行残差连接和正则化，输出结果作为下一个编码器的输入，通过多次重复后又经过解码器解码得到音译结果。但它对于数据规模、计算成本要求较高。

音译生成的方法根据音译过程中信息的来源(发音或拼写)，可分为基于字素的方法(Θ_G)、基于音素的方法(Θ_P)、基于混合的方法(Θ_H)和基于组合的方法(Θ_C)，如表2所示，四类模型的示意图如图1所示。

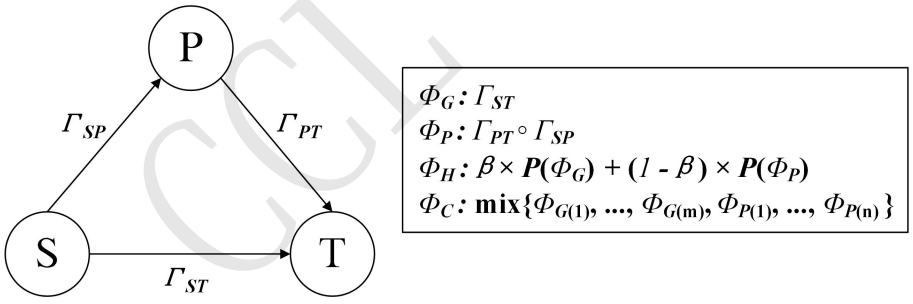

图 1. 四类音译生成模型。其中S表示源字素，P表示源音素，T表示目标字素，Γ_{ST}、Γ_{SP}、Γ_{PT}表示三个转换过程，Φ_G直接将源字素转换为目标字素，Φ_P将源字素转换为源音素，再生成目标字素，有时需要源音素到目标音素多一步的转换，Φ_H由Φ_G和Φ_P的线性插值组成，β是权重参数($0 \leq \beta \leq 1$)，Φ_C是多个字素、音素方法的混合，$(g \circ f)(x) = g(f(x))$。

基于拼写/字素的方法(Θ_G)：基于拼写的方法依赖于从单词的字符中获得的统计信息。它直接将源语言字素/字符/多个字符转换为目标语言字素/字符/多个字符。基于字素的方法旨在建立从源单词中的一组字符到目标单词中字符的直接映射模型。与基于音素的方法相比，基于字素的方法减少了音译过程中涉及的步骤，从而消除整个过程中的一些潜在错误源，实现起来也较为容易。但因缺少语音形式的信息，因此也很难获得精准的音译结果。尽管基于字素的方法总体上比优于基于音素的方法，但其在处理发音与拼写有很大差异的单词时性能较弱。Li等(2004)提出了一个音译框架，该框架通过联合信源信道模型实现了从英语到汉语的直接正字法映射。它通过n-gram模型，让正字法对齐实现自动化，能直

接从双语词典中得到对齐的音译单元。Malik(2008)提出了一个基于有限状态转换器的印地语-乌尔都语的音译模型，利用了转换器的特点，采用灵活通用的中间转录方案获得音译结果。Dhore等(2012b)使用条件随机场作为概率统计工具和n-gram作为特征集，实现了印地语-英语的命名实体音译。Wang等(2015)使用不同的字母分割方法来对英语到朝鲜语和英语到汉语的命名实体进行音译。Merhav和Ash(2018)比较了基于长短期记忆人工神经网络(LSTM)的编码器-解码器模型、Transformer和传统的加权有限状态转换器三者的性能，实验结果显示了基于Transfromer的方法表现最佳，和其余两个模型相比更适合音译。

基于音素的方法(Θ_P)：它将音译视为一个语音的过程，而不是一个正字法(字母映射)的过程。音素是声音的最小单位，该方法通常从源语言字素生成源语言音素，再从源语言音素映射到目标语言音素，最后生成目标语言字素。音素表征成为源语言与目标语言之间的中间形式/支点，所以也称为支点法。一般来说，基于音素的方法的主要优势是提升了发音在音译过程中的作用。然而，整个过程需要多个步骤，包括从源字素到源音素、源音素到目标字素、有时还需要源音素到目标音素的转换，这增加了错误传播的机会。音译系统可能在每一步生成很多个备选方案，多个步骤会导致早期的错误在后期显著增加。而这些错误直接影响候选音译目标词的最终排名，导致正确的音译排在错误的音译之后，或者可能生成完全错误的音译结果。基于音素的方法的另一个缺点是，这类方法依赖于双语发音资源，但并非所有语言都能轻易获得发音资源。Knight和Graehl(1998)利用加权有限状态变换器结合romaji(拼写日语的罗马字母)到音素、音素到对应的英语以及英语单词的概率实现了日语到英语音译建模。Jung等(2000)使用了扩展马尔可夫窗口实现了英语到朝鲜语的音译。该方法通过发音词典将英语单词转换为英语发音，再分割为音素块。每个块通过扩展马尔科夫窗口对应于手工规则定义下的朝鲜语字母。Oh和Choi(2002)提出了一个利用发音和上下文规则的英语到朝鲜语音译模型。它通过发音词典中提取英语发音单元并与对应音素对齐。对于字典中没有的单词则被拆分成两个单词继续在字典中进行搜索。若仍未找到。则用英语的希腊语起源单词。对齐后使用英语到朝鲜语标准转换规则(EKSCR)生成音译。Dhore等(2012a)提出了一个将印地语和马拉地语的命名实体音译为英语的模型。该方法基于发音统计法把音素和命名实体长度作为监督学习的特征，并将Unicode编码的印地语或马拉地语送给音节化模块，该模块将命名实体分割成若干音译单元。音译模块再通过语音图谱将梵文字母中的每个音节单元转换为英语。

基于混合的方法(Θ_H)：它结合了基于音素和基于字素的方法。通常，基于音素的方法比基于字素的方法更容易出错，它的成功率通常低于基于字素的方法。然而，基于字素的方法不能很好的处理发音与拼写有很大差异的单词。混合方法旨在结合这两类方法的优势，以提高整体精度，但实施起来较为困难。混合音译方法将基于字素和基于音素的信息结合到一个系统中，进而生成候选音译目标词。Oh和Choi(2005)提出了一种字素和音素相结合的音译模型，并且在英语到朝鲜语和英语到日语的数据集上进行了实验。AI-Onaizan和Knight(2002)基于有限状态机提出了一种结合音素和字素的混合音译模型，实现了阿拉伯语到英语的音译。

基于组合的方法(Θ_C)：能将多个基于字素或音素的方法(不局限于两者)组合起来。音译的组合方法把多个独立的系统输出汇集在一起，再将这些输出组合成一个候选音译列表。基于组合的方法可以将每个系统的独特优势相互结合，减少音译系统的错误率，给出更优的音译结果，同样它实施起来也较为复杂。Oh和Isahara(2007)使用音译系统的组合来研究英语到朝鲜语和英语到日语的音译。他们提出了一个基于支持向量机和最大熵模型的方法来对单个音译系统的输出进行重排序。每个系统生成的候选音译列表中给出不同生成目标词的置信度。Karimi等(2008)提出了基于辅音-元音的CV-Model方法。该方法在训练过程中将单词对齐后，通过替换元音和辅音生成辅音-元音的序列，序列与原字符一同分解为特定的模式。根据特定模式生成转换规则，从而形成音译模型。Najafi等(2018)的音译系统是基于多个系统的组合，他们使用了五种音译方法：DirecTL+(Jiampojamarn et al., 2010)、Sequitur(Bisani and Ney, 2008)、OpenNMT (Klein et al., 2018)、BaseNMT(Sutskever et al., 2014)和RL-NMT(Najafi et al., 2019)，取得了一定的效果。

根据使用的技术，音译生成可以分为**基于规则的方法**和**基于统计方法**(隐马尔科夫模型、条件随机场、决策树、支持向量机等)。基于规则的方法依赖大量人为规则及音译名称，系统易于实施且在缜密设计的规则下能提供良好的性能。但需要丰富的经验和大量的语言专业知识，根据一种语言设计的规则也不能转移到另一种语言，其对语言的依赖程度较高，整个工程的成本昂贵。而统计方法无需语言模型的专业知识，在大量数据和良好的算法下能让训练结果达到

满意的效果。统计学习的方法来源于机器学习和数据挖掘，都是从数据中学习。因此适应性、扩展性和维护成本都优于基于规则的方法。但其仍存在着不足，训练数据的规模对音译结果质量影响较大，有些语言甚至没有可供使用的语料库资源，且需要上下文信息来引导结果的生成。深度学习(RNN、Transformer等)最初是机器学习中的一个小分支，近年来随着算力资源的发展已经开始成为新的研究方向，它相比传统的机器学习需要更庞大的数据量和运算资源，通过一种端到端的模式解决整个问题代替传统机器学习的多个子问题逐个解决。还有一个显著差别在于深度学习提取的是单词的深层次特征，这些特征虽然目前很难解释，但从结果上看却能准确的表示出单词的某些特性。因此，也可以将基于深度学习的方法单独作为一类。

从总体上看，近年来基于规则的、基于传统机器学习的模型使用频率相对较少。这是因为基于规则的方法成本昂贵，而基于传统机器学习的模型在数据量较大的情况下，效果不如基于深度学习的模型。而基于深度学习的模型是目前机器音译研究的主流。在最近一次的命名实体研讨会(NEWS 2018)的音译评测任务中，除了新加坡国立大学和新加坡科技设计大学组成的团队提供了使用Sequitur和Moses的基线系统，其余团队则都使用了基于神经网络的模型。英国的爱丁堡大学参加了15个音译任务的评测，并在11个任务评测中取得了第一的成绩，该团队使用的是带有注意力机制的RNN编码器-解码器模型(Chen et al., 2018a)。目前为止，关于音译生成的研究涉及的语言对比较多，音译研究使用的方法也各有不同。语言的不同特点、训练语料库的大小，以及是否加入语言学家的知识等，这些都对选择适合语言的的研究方法有着影响。而且对于同一语言对之间的音译，研究者们使用的数据集也不尽相同，所以很难确切地比较出不同方法的优劣。

分类	作者及年份	语言对	模型	数据集(规模)	评价指标(分数)
基于音素	(Knight and Graehl, 1998)	日语→英语	有限状态转换器	N/A(1,549)	准确率(64%)
	(Jung et al., 2000)	英语→朝鲜语	扩展马尔科夫窗口	N/A(8,368)	召回率(87.5%)
	(Oh and Choi, 2002)	英语→朝鲜语	直接映射模型	N/A(7,185)	准确率(67.83%),字符准确率(93.49%)
	(Surana and Singh, 2008)	英语→印地语① 英语→泰卢固语②	识别适应性音译机制	N/A(2,000)	①:MRR(87%),精确率(80%); ②:MRR(82%),精确率(71%);
	(Dhore et al., 2012a)	印地语→英语 马拉地语→英语	统计模型	N/A(15,224)	准确率(97.3%),F值(94.2%),MRR(96.8%), 召回率(93.2%),精确率(95.7%)
基于字素	(Li et al., 2004)	英语↔汉语	联合信源信道模型	N/A(37,694)	正向准确率(70.10%),反向准确率(37.9%)
	(Rama and Gali, 2009)	英语→印地语	噪声信道模型	NEWS 2009 (10,949)	准确率(46.3%),F值(87.6%), MRR(57.3%),MAP(45.4%)
	(Dhore et al., 2012b)	印地语→英语	条件随机场模型	N/A(7,251)	准确率(65.01%)@1-gram
	(Grundkiewicz and Heafield, 2018)	泰语→英语① 英语→泰语② 英语→希伯来语③ 英语→孟加拉语④ 英语→坎那达语⑤ 英语→泰米尔语⑥ 英语→印地语⑦ 汉语→英语⑧ 英语→汉语⑨ 英语→越南语⑩ 希伯来语→英语⑪	带注意力机制的RNN编码器-解码器模型	NEWS 2018▲	平均F值{①(0.8454); ②(0.7760); ③(0.8042); ④(0.9006); ⑤(0.8673); ⑥(0.8405); ⑦(0.8515); ⑧(0.8300); ⑨(0.6791); ⑩(0.8893); ⑪(0.7532);}
	(Merhav and Ash, 2018)	英语→阿拉伯语① 英语→希伯来语② 英语→片假名③ 英语→俄语④ 阿拉伯语→英语⑤	Transformer I 基于LSTM的编码器-解码器模型 II	SubWikiLang▲	错误率{①I:(0.45),II(0.53); ②:I(0.44),II(0.49); ③I:(0.51),II(0.60); ④:I(0.35),II(0.40); ⑤I:(0.75),II(0.81);}
	(Chatterjee and Sarkar, 2021)	孟加拉语↔英语	支持向量机模型 I 隐马尔可夫模型 II	N/A(1,000)	准确率{正向I(53.9%),正向II(51.9%); 反向I(45.2%),反向II(43.1%);}
基于混合	(Al-Onaizan and Knight, 2002)	阿拉伯语→英语	信源信道模型 加权有限状态转换器	N/A(—)	准确率(49.08%)
	(Oh and Choi, 2005)	英语→朝鲜语① 英语→日语②	决策树模型 I 最大熵模型 II 基于记忆学习 III	EKSet(7,185) EJSet(10,398)	准确率{ ①I(62.0%),II(63.3%),III(66.9%); ②I(66.8%),II(67.0%),III(72.2%);}
基于组合	(Karimi, 2008)	英语→波斯语① 波斯语→英语②	信源信道模型 和投票法	①:N/A(1,500) ②:N/A(2,010)	①:正确率(74%) ②:正确率(53%)
	(Najafi et al., 2018)	阿拉伯语→英语① 英语→日文汉字② 英语→片假名③ 英语→朝鲜语④ 波斯语→英语⑤ 英语→波斯语⑥	DirecTL+ Sequitur OpenNMT BaseNMT RL-NMT	NEWS 2018▲	平均F值{①(0.9087); ②(0.7678); ③(0.8098); ④(0.7113); ⑤(0.9515); ⑥(0.9373);}

Table 2: 音译生成经典技术。N/A表示相关数据集并未公开，——表示数据集未注明大小，▲表示涉及到的数据集详见第四节。EKSet和EJSet数据集现已无法访问。

3.2 音译挖掘

音译对通常是从平行语料库或可比语料库或Web中挖掘出来的。平行语料库是两种或多种语言的对齐文本集合。对齐文本是一种语言到一种或多种语言的精确翻译。可比语料库也是两种或多种语言的文本集合，各种语言的文本是相似的，但不是彼此的精确翻译。

迄今为止，音译领域已经涌现出许多音译挖掘技术。这些技术可以分为三类：基于语音相似度、基于机器学习技术和基于词共现。

基于语音相似度：基于语音相似度的方法测量平行或可比语料库中词1和词2之间的相似性，并提取出词2'作为词1'最接近的候选音译词。可以用许多方法计算相似性，如莱文斯坦距离算法、最长公共子序列(LCS)算法和Jaro-Winkler距离(Jaro, 1989)算法。Udupa等(2008)提出了一种命名实体等价物/对等物挖掘方法，该方法通过跨语言文档相似度和音译相似度模型，能够有效地从可比语料库中挖掘命名实体音译的等价物。

基于机器学习：El-Kahki等(2011)提出了一种增强的音译挖掘技术，该技术使用生成图强化模型来推断源字符序列和目标字符序列之间的映射。Fukunishi等(2013)提出了一种挖掘音译对的技术，该方法在对齐过程中使用非参数化的贝叶斯方法。

基于词共现：Karimi(2011)已经对基于词共现的部分提取模型进行过讨论。Wu等(2012)为NEWS 2012提出了一种英韩音译系统。他们训练了多个音译模型，使用两种重排序方法从不同模型的预测结果中选择最佳的音译模型。其中一种重排序方法是基于网络语料库中音译对的共现。另一种是基于对齐结果特征的间接监督联合学习重排序的方法。实验结果表明使用基于网络重排序方法的音译模型可以在英语-朝鲜语音译中获得更优的结果。

3.3 融合方法

有些方法同时使用音译生成和音译挖掘技术。Zhao等(2007)提出了一种基于隐马尔可夫模型的框架来音译命名实体，此外，通过与从网络搜索引擎收集的统计数据生成的自动拼写检查器相结合，进一步提高了音译准确性。Chinnakotla等(2008)开发了一个印地语和马拉地语到英语的跨语言信息检索系统，将音译生成、音译提取等多种技术有效地融合在了一起。

4 音译语料库资源

我们对可用于音译任务的数据集进行了搜集和整理，如表3所示。列出的数据集仅包含目前可以获取(免费/收费)的，对于现无法访问的数据集未列出，数据集的详细介绍见附录A部分。

- NEWS 2018[2]：继NEWS 2009(ACL-IJCNLP 2009)、NEWS 2010(ACL 2010)、NEWS 2011(IJCNLP 2011)、NEWS 2012(ACL 2012)、NEWS 2015(ACL 2015)和NEWS 2016(ACL 2016)之后举办的连同音译共享任务的命名实体研讨会，旨在提供一个通用平台，用于对跨多种语言的不同音译方法和系统进行基准测试。

- 中日朝鲜词典研究所(CJKI)为汉语、日语、朝鲜语、阿拉伯语、西班牙语等语言编制了一系列综合字典数据集[3]，这些数据集包含大量通用词汇、专有名词和技术术语的语法、语音和语义属性。其中可用于音译任务的包括汉语拼音-汉语数据集(CHD)、汉英人名数据集(CEN)、中英地名数据集(CEP)、中日人名数据集(CJN)、中日地名数据集(CJP)、日语-多语言地名数据集(JMP)、日本公司数据集(JCD)、日英人名数据集(JEN)、朝鲜英人名数据集(KEN)、朝鲜英地名数据集(KEP)、朝鲜日人名数据集(KJN)、朝鲜日地名数据集(KJP)、朝鲜汉人名数据集(KCN)、朝鲜中地名数据集(KCP)、阿拉伯语外国名字数据集(DAFNA)、阿拉伯地名数据集(DAPNA)、美国财政部外国资产控制办公室名单扩展(XOFAC)和中越人名数据集(CVP)。

- 英语-乌克兰语数据集[4]：从维基百科中提取的英语-乌克兰语命名实体数据集，该项目由欧盟委员会赞助。

[2] http://workshop.colips.org/news2018/dataset.html
[3] https://www.cjk.org/data/all
[4] http://catalog.elra.info/en-us/repository/browse/ELRA-M0104

数据集	语言对/数据量	总规模	数据集	语言对/数据量	总规模
NEWS 2018	英语→泰语	32,781	CHD$	汉语拼音→汉语	超过60万
	泰语→英语	29,273	CEN$	汉语→英语	超过200万人名
	英语→波斯语	15,386	CEP$	汉语→英语	超过百万地
	波斯语→英语	17,677	CJN$	汉语→日语	超过200万人名
	英语→汉语	43,318	CJP$	汉语→日语	超过10万地名
	汉语→英语	34,002	JMP$	日语→13种	超过310万地名
	英语→越南语	5,256	JCD$	日语→英语	超过60万公司名
	英语→印地语	14,937	JEN$	日语→-英语	超过55万人名
	英语→泰米尔语	12,957	KEN$	朝鲜语→英语	超过200万人名
	英语→坎那达语	12,955	KEP$	朝鲜语→英语	约9万地名
	英语→孟加拉语	15,623	KJN$	朝鲜语→日语	超过200万人名
	英语→希伯来语	12,501	KJP$	朝鲜语→日语	约9万地名
	希伯来语→英语	11,447	KCN$	朝鲜语→汉语	超过200万人名
	英语→片假名	30,828	KCP$	朝鲜语→汉语	约9万地名
	英语→日文汉字	12,514	DAFNA$	英语→阿拉伯语	超过24万人名
	英语→朝鲜语谚文	9,387	DAPNA$	阿拉伯语→英语	超过1万地名
	阿拉伯语→英语	33,354	XOFAC$	英语→阿拉伯语	超过2500万人名
	波斯语→英语	8,000	CVP$	汉语-越南语	超过4.3万人名
	英语→波斯语	13,204	英语-乌克兰语数据集	英语→乌克兰语	624,168
SubWikiLang	英语→俄语	164,640	中英命名	汉语→英文	797,452
	英语→片假名	98,820	实体列表$	英文→汉语	869,136
	英语→阿拉伯语	74,973	TRANSLIT	超过180种	3,008,239实体
	英语→希伯来语	50,049	ParaNames	超过400种	14,017,168实体
	阿拉伯语→英语	15,898	Trabina	591种	1,129
	英语→Arpabet	126,191	BanglaNLP⋆	孟加拉语→英语	13,214
Xlit-Transliteration	印地语→英语	32,508	Xlit-Crowd	印地语→英语	14,919
	迈蒂利语→英语	28,88	Xlit-IITB-Par	印地语→英语	68,922
	孔卡尼语→英语	21,342	TfNSW	英语→12种	1538个车站名
Bittlingmayer⋆	英语→亚美尼亚语	39,707	Aksharantar	21种印度语言→英语	2600万
	英语→希腊语	37,505	FIRE 2013	印地语→英语	1,462
	英语→波斯语	78,663	Praneeth⋆	英语→泰卢固语	38,568
	英语→俄语	179,853	ANETAC	英语→阿拉伯语	79,924
EnToFrNE$	英语→法语	1,167,263	EnToSSLNE$	英语→6种南斯拉夫语言	26,155

Table 3: 音译数据集资源。注意：每届NEWS研究会的数据集在上一届的基础上进行扩展，在此只列出了最近举办的一届情况。词的发音模拟为表示音子和语段的符号串，音子是言语的发音，用语音符号表示。音子包含三套不同的字母符号——IPA，Arpabet，SAMPA，其中Arpabet 是高级研究计划署(ARPA)开发的语音转录代码。⋆表示数据集的作者未对数据集命名，这里用作者(团队)名称代替数据集名称。$表示数据集需要付费才可获取。

- SubWikiLang[5](Merhav and Ash, 2018)：它包括从维基数据(Wikidata)搜集后过滤的以英语为源语言的4组人名数据集、Rosca和Breuel(2016)从维基百科(Wikipedia)标题中提取的阿拉伯语到英语的数据集和CMU发音字典共同组成。维基数据是维基媒体基金会主持的一个自由的协作式多语言辅助知识库，旨在为维基百科、维基共享资源以及其他的维基媒体(Wikimedia)项目提供支持(是它们的超集)，其中的每个文档都有一个主题或一个管理页面，且被唯一的数字标识。CMU发音字典是由卡内基梅隆大学创建的一个开源发音词典，它为北美发音中的英语单词提供映射拼写/语音。

- 中英命名实体列表(LDC2005T34)[6]：由语言数据联盟(Linguistic Data Consortium, LDC)提供，包含九对从新华社通讯社文本中汇编的汉语-英文双向名称实体列表。LDC是由大学、图书馆、公司和政府研究实验室组成的语言公开联盟，隶属于宾夕法尼亚大学文理学院。

- TRANSLIT[7]：由数据集JRCNames(Ehrmann et al., 2017)、SubWikiLang[5]、Geonames[8]、

[5] https://github.com/steveash/NETransliteration-COLING2018
[6] https://catalog.ldc.upenn.edu/LDC2005T34
[7] https://github.com/fbenites/TRANSLIT
[8] https://download.geonames.org/export/dump/alternateNamesV2.zip

谷歌英语-阿拉伯语音译数据集[9]及其维基百科中所有语言的命名实体转储数据共同统一格式后合并而成。产生的数据集包含180多种语言的约160万条词条，以及约300万个名称变体(Benites et al., 2020)。

- ParaNames[10](Sälevä and Lignos, 2022)：它采用与SubWikiLang[5]中维基数据基于相同的获取方式。所不同的是，SubWikiLang只包含几种人名数据，而ParaNames包含了维基数据中的所有语言对，包含人名、地名、组织机构名三种实体。
- Trabina[11]：Wu等(2018)利用《圣经》的广泛传播性，把其中1129个英文姓名翻译成了591种语言，平均一个姓名有其他语言的52%覆盖率，超越了维基百科的覆盖率。
- BanglaNLP团队和Bengali.AI社区成员发布了一个从维基百科中爬取整理后的孟加拉语-英语的音译数据集[12]。
- Xlit-Transliteration[13]：作者在母语使用者的帮助下创建了三种印度语言(印地语、迈蒂利语、孔卡尼语)-英语的数据集。
- Xlit-Crowd[14]：由众包/外包网站Amazon Mechanical Turk获得，包含14,919对印地语-英语的音译对(Khapra et al., 2014)。
- Xlit-IITB-Par[15]：由Moses音译模块从印度理工学院和印度语言技术中心提供的英语-印地语平行语料库中自动挖掘出来，包含68,922对音译对(Kunchukuttan et al., 2018)。
- TfNSW[16]：由澳大利亚新南威尔士州交通网络中每个车站和码头的英语名称到12种语言(阿拉伯语、法语、德语、希腊语、印地语、意大利语、日语、朝鲜语、简体中文、繁体中文、西班牙语、越南语)的音译结果组成。
- Bittlingmayer从维基百科中下载了英语到亚美尼亚语、希腊语、波斯语、俄语四种语言的文章并把单词转换它们为单一字符，从而生成了四个音译数据集[17]。
- Aksharantar[18]：由21种印度语言-英语的音译对组成，是目前印度语言最大的公开音译数据集(Madhani et al., 2022)。
- FIRE 2013[19]：第五届信息检索论坛会议(FIRE 2013)提供了一个较小的印地语-英语的音译数据集，该数据集是通过使用宝莱坞歌词对齐后收集得到的(Roy et al., 2013)。
- Praneeth从维基百科平行语料的文章标题中，以及从美国社交新闻聚合、网络内容讨论网站Reddit的社区将讨论文本音译创建了英语-泰卢固语的数据集[20]。
- ANETAC[21]：一个英语到阿拉伯语的命名实体音译和分类数据集，该数据集由公开获得的平行翻译语料库建立(Ameur et al., 2019)。
- EnToFrNE[22]：由1,167,263个英语和法语平行命名实体组成，实体包括任务、组织、地点、产品和杂项五大类组成。

[9] https://github.com/google/transliteration
[10] https://github.com/bltlab/paranames
[11] https://github.com/wswu/trabina
[12] https://github.com/arijitx/BanglaNLP
[13] http://transliteration.ai4bharat.org/#/resources
[14] https://github.com/anoopkunchukuttan/crowd-indic-transliteration-data
[15] https://www.cfilt.iitb.ac.in/iitb_parallel
[16] https://opendata.transport.nsw.gov.au/dataset/tfnsw-station-names-other-languages
[17] https://github.com/deepchar/deepchar
[18] https://huggingface.co/datasets/ai4bharat/Aksharantar
[19] http://cse.iitkgp.ac.in/resgrp/cnerg/qa/fire13translit/index.html
[20] https://github.com/notAI-tech/Datasets/tree/master/En-Te_Transliteration
[21] https://github.com/MohamedHadjAmeur/ANETAC
[22] http://catalog.elra.info/en-us/repository/browse/ELRA-M0052

第二十一届中国计算语言学大会论文集，第317页-第332页，南昌，中国，2022年10月14日至16日。

- EnToSSLNE[23]：由七种语言的26,155个平行命名实体组成，包括英语和六种南斯拉夫语：波斯尼亚语、保加利亚语、克罗地亚语、马其顿语、塞尔维亚语和斯洛文尼亚语。

5 音译质量评价指标

在介绍具体指标之前，我们首先给出相关符号的定义：

- N：测试集中实体(样本/名称)的数量。
- J：测试集中实体的参考音译结果(标签/正确音译结果)的数量。
- K：音译系统输出的候选音译(音译预测/输出结果)的数量。
- n_i $(1 \leq i \leq N)$：测试集中第i个实体的参考音译结果数量，音译任务中存在一个名称有多个对应的正确音译结果的情况。
- $r_{i,j}$ $(1 \leq i \leq N,\ 1 \leq j \leq J)$：测试集中第$i$个实体的第$j$个参考音译。
- $c_{i,l}$ $(1 \leq i \leq N,\ 1 \leq l \leq K)$：测试集中第$i$个实体的第$l$个候选音译。

我们根据现有论文对音译性能的主要评价指标进行了整理，如表4所示。

定义	公式
前k预测准确率($Top-k\ ACC$，简称ACC)表示所有预测结果中概率最大的前k个结果中包含正确音译的比例，是最常用的音译评价指标。值为1表明每个单词的k个结果中至少有1个是正确的，值为0表明所有单词的音译结果都与候选词不匹配。通常k取1。有时候也用单词错误率(WER)来替代，准确率和单词错误率之和为1(Chen et al., 2018b)。	$Top-k\ ACC$ $= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{l=1}^{k(\leq K)} \begin{cases} 1, & r_{i,j}=c_{i,l}\ (\exists\ r_{i,j}) \\ 0, & \text{otherwise} \end{cases}$
召回率(R)、精确率(P)是两个衡量第i个单词音译输出结果性能的指标。它们根据候选词(c)与最佳匹配参考词(r)的最长公共子序列(LCS)长度计算得到，LCS由c、r和它们之间的编辑距离(ED)共同决定，最佳匹配由二者的最小编辑距离决定，$\|x\|$表示x的长度(Chen et al., 2018b)。	$LCS\left(c,r\right)=\frac{1}{2}\left(\|c\|+\|r\|-ED\left(c,r\right)\right)$ $r_{i,m}=\arg\min_j\left(ED\left(c_{i,k},\ r_{i,j}\right)\right)$ $R_i=\frac{LCS\left(c_{i,l},\ r_{i,m}\right)}{\|r_{i,m}\|}$ $P_i=\frac{LCS\left(c_{i,l},\ r_{i,m}\right)}{\|c_{i,l}\|}$
平均F值(简称F值)衡量了音译候选词与最接近的参考词之间的差异，它同时考虑了R和P。对于每个源词，当候选词与任意参考词的LCS为0时，F值为0。当第一个候选词匹配其中一个参考词时，F值为1(Chen et al., 2018b)。	$F_i=2\frac{R_i\times P_i}{R_i+P_i}$
平均倒数排名(MRR)是给定一组实体，其所有音译结果排序列表中第1个正确音译的排名倒数的平均值，$1/MRR$表示样本对应的正确音译结果的平均排名。MRR接近1表明正确音译结果靠近n-best列表顶部(Chen et al., 2018b)。	$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{Rank}_i}$ $=\frac{1}{N}\sum_{i=1}^{N}\begin{cases}\min_j\ \frac{1}{j}, r_{i,j}=c_{i,l}(\exists\ r_{i,j},c_{i,l}) \\ 0,\ \text{otherwise}\end{cases}$
平均精度均值(MAP)与MRR所不同，MAP考虑了所有参考音译，它衡量了第i个源名称的n-best的精准度，$num(i,k)$表示k-best列表中第i个源词的正确候选数量(Chen et al., 2018b)。	$MAP=\frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i}num\left(i,k\right)\right)$
字符准确率(CA)衡量了音译对中匹配字符的比例。有时候也用字符错误率(CER)替代，它们相加和为1(Nießen et al., 2000) 。	$CA=\frac{\|r\|-ED\left(c,r\right)}{\|r\|}$

Table 4: 音译任务常用的评价指标。

[23] http://catalog.elra.info/en-us/repository/browse/ELRA-M0051

6 未来研究方法和开放问题

6.1 构建公开、权威的音译数据集

目前虽然一些会议已经组织过音译相关的评测任务，并提供了一些公开的音译数据集。但是整体上来说仍然存在着语言对不够丰富、数据集规模不够大、数据集质量不高等问题，并没有专门针对不同语言对的统一音译任务数据集。音译数据集不同于翻译数据集，其更难获取。目前的音译数据集往往来源于各国的政要、名人、地名、组织以及外来词。若直接从翻译数据集中筛选出用于音译任务的数据集则比较困难(其在翻译数据集中占比很低、通晓不同语言的专家很少)。在不违背各国法律的基础上，急需各国各自贡献一部分标准实体名称用于全球音译数据库的构建，这对于音译的研究将会起很大的推进作用。后续的研究者也将会有统一的参考依据为方法之间的性能比较。但也要注意，这必将花费大量时间、金钱和人力资源成本。

6.2 低资源语言音译研究

从表1中看到目前音译的研究集中于世界上7000种语言中的30多种，而绝大多数语言还没有被研究。这些语言的使用者较少，其数据集资源量也较低，常称它们为低资源语言(Magueresse et al., 2020)。低资源语言缺乏有用的训练属性，如受监督的数据、母语使用者以及专家的数量等，但不应忽略其存在。仅非洲和印度就大约有2000种低资源语言，有超过25亿的人使用它们(Tsvetkov, 2017)，为这些语言开发音译技术具有相当可观的经济前景。此外，研究音译可以支持某种语言，防止其灭绝并提升其影响力。中国的“一带一路”政策涉及全球数十个国家、数十亿人口和上百种民族语言，利用技术手段突破语言障碍问题的重要性日益凸显。针对低资源语言匮乏的特点，一些学者也开始尝试了一些新的方法。Le等(2019)提出了一种利用基于RNN的模型建立一个低资源机器音译系统的方法。Upadhyay等(2018)提出了一种使用约束发现来挖掘名称对以重新训练生成模型的Bootstrapping算法。但这些方法的效果还离实际落地使用差得较远。未来针对低资源语言，我们建议从两个方面入手。一是采用多种数据增强的方式扩大其数据量并提高数据的质量。二是利用语言模式/相似性、设计更健壮的学习模型以提高音译的准确性和鲁棒性。

6.3 深度学习模型的可解释性

现有的基于深度学习的音译方法研究还较少，但它的出现可以说重振了整个音译学(Santos et al., 2018; Shillingford and Parker Jones, 2018; Khare et al., 2021)。其在音译准确性、泛化性、鲁棒性方面的表现均优于传统方法，有时甚至优于人类基线(Alam and ul Hussain, 2021)，展现出了令人难以置信的性能。传统方法是基于人工认知所驱动的，需要语言学家针对语言及其结构(语法、句法、语音学等)进行研究，设计出音译模型，人工干预性很强。但是无论是基于统计的方法、决策树、支持向量机或更复杂的模型，仍然可以理解模型的内部结构。但是基于数据驱动的深度学习的音译方法存在着乃至整个深度学习社区目前都难以解答的问题，也就是说深度学习模型是从中怎么学习到不同语言各自的特性以及他们之间的对应关系。深度学习的模型不需要手动设计的特征，但由于它的黑盒性质导致我们不能对其进行分析和理论上的改进。而透明度是模型可解释性的关键所在，有助于我们理解模型中的每一层的学习内容以及它们是如何交互的，甚至能让我们能够找到模型所存在的漏洞，防止恶意攻击和非法侵入。虽然说现在已经有一些像注意力机制(Alam and ul Hussain, 2021)、知识图谱(Moussallem et al., 2020)、模拟模型方法(Hou and Zhou, 2020)等来对其作出了一定的解释，但是其解释并不能得到整个深度学习社区的一致同意，仍然存在着分歧(Cremer, 2021)。为了让基于深度学习的音译模型不超出人类设计它时的范围(道德伦理、预设规则、潜在偏见、种族歧视等)(Muscat, 2011)，在未来将音译模型应用于现实世界迫切的需要完备的可解释性，以使系统更加透明、更具解释性、可控性、安全性和健壮性，这是接下来需要思考和研究的方向。

6.4 迁移学习方法的使用

近年来，基于深度学习的音译取得显著的进展，这是由于如NEWS研讨会所提供的大规模音译平行预料。但这些预料仅关注了高资源语言，整个音译社区极度缺少低资源语言(Khare et al., 2021)。尽管基于深度学习的音译模型已经取得了较好的效果，但支持其训练需要足够大的双语平行语料库。除英语与其他语言外，即使对于母语使用人数众多的语言，如汉语、印度语、俄语、日语，它们之间也缺少足够的音译数据集。除了我们在6.1节所介绍的扩充数

据集的方法外，迁移学习也是音译数据集稀缺的一种解决方案(Maimaiti et al., 2019; Wu et al., 2022)。迁移学习的主要思想是将父模型(源域)的部分参数和共享知识传递到子模型(目标域)中，子模型仅需较小的数据量和训练时间就可以取得较不错的性能。英语作为全世界最广泛使用的语言，与英语相关的音译研究是最多的，可以把英语作为解决稀缺资源的桥梁。在源语言-英语或英语-目标语言数据集上训练一个音译模型(父模型)，根据这个模型来微调源语言-目标语言的模型(子模型)。但这其中存在着一个问题，语族(域)是迁移学习父模型的选择的一大关键。来自同一语系的语言在句法、语义、语言特征中存在着一些相似性，这些相似性有助于提升父模型的泛化能力，对迁移学习是有利的。而英语属于印欧语系中的日耳曼语族语支，很多语言却并不属于这个语支。如果忽视语系之间的相似性，也许会对迁移学习起到反作用，这也是未来我们要研究的方向。

6.5 多种音译方法相结合

单一的机器音译方法不能完全解决音译中存在的问题，将多种音译方法相结合起来，发挥不同方法的优势，能给出更好的音译结果。不同音译方法侧重点所不同，无论是音义兼顾(Usunier and Shaner, 2002)、音同义，还是谐音音译(Chen, 2013)都存在各自的优势。深度学习模型由于黑盒性质，需要吸收传统音译方法中语言学家设置的原则性音译规则，包括消极意义、敏感政治、恐怖主义、分裂主义、极端主义、殖民文化、歧视、黄赌毒等各国法律禁止的内容都不应该在音译结果中所出现。近年来一些学者也开始尝试多音译方法组合，Karimi等(2011)将不同的传统方法进行了组合实验，Nicolai等(2015)将三种音译模型进行了组合实验，Najafi等(2018)进行了神经网络与传统方法相结合的音译实验，他们的结果共同都表明了音译方法相结合优于单独使用任一方法。

6.6 不同地区的音译差异

我们仍然需要注意一种语言的音译结果在不同国家或地区的差异性问题。比如说汉语在中国大陆、香港、澳门、台湾、新加坡等地区存在着不同的音译差异，同一模式下存在着变体问题。就人名汉语音译英语而言，中国大陆人名使用的是汉语拼音，台湾地区很多人名的英文使用的则是威妥玛拼音(Xing and Feng, 2016)，而在香港、澳门地区人名的英文往往是粤式拼音、上海话拼音、英文或它们的组合(Man, 2012)，但在新加坡地区的人名英文则是粤语、潮州话、福建话/闽南语或它们的组合(Su, 2022)。这样的问题也存在于英译汉当中，比如说Reagan在大陆译为里根，台湾译为雷根，而香港则译为列根；Bush在大陆译为布什，台湾译为布希，而香港译为布殊。这是由于各地在不同的制度、社会背景、生活习惯、文化习俗、翻译准则等多方面的因素影响下共同造成的。是根据使用的不同地区训练不同的音译模型还是让模型都吸收来自不同的地区的音译结果，这也是未来值得探讨和研究的。音译技术在一定程度上能反应出这个词的历史背景和地区，能为历史和语言研究提供不小的帮助。

7 结论

在本文的工作中，我们回顾了音译的相关经典模型并对它们进行了分类整理，同时对音译数据集和评价指标进行了汇总，最后指出了整个音译社区目前待解决的问题和对未来的研究方向进行了展望。

在这个信息洪流时代，随着新词语的不断出现，将外来词融入本国语言变得越发普遍。相信音译作为解决这一问题的支持工具必将会受到更多的关注。

参考文献

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in Arabic texts. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Mehreen Alam and Sibt ul Hussain. 2021. Deep learning-based roman-urdu to urdu transliteration. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(04):2152001.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2019. ANETAC: arabic named entity transliteration and classification dataset. *CoRR*, abs/1907.03110.

Fernando Benites, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak. 2020. TRANSLIT: A large-scale name transliteration resource. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3265–3271, Marseille, France, May. European Language Resources Association.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Soma Chatterjee and Kamal Sarkar. 2021. Machine transliteration using svm and hmm. *Int. J. Adv. Intell. Paradigms*, 19(1):3–27, jan.

Nancy Chen, Rafael E. Banchs, Min Zhang, Xiangyu Duan, and Haizhou Li. 2018a. Report of NEWS 2018 named entity transliteration shared task. In *Proceedings of the Seventh Named Entities Workshop*, pages 55–73, Melbourne, Australia, July. Association for Computational Linguistics.

Nancy Chen, Xiangyu Duan, Min Zhang, Rafael E. Banchs, and Haizhou Li. 2018b. NEWS 2018 whitepaper. In *Proceedings of the Seventh Named Entities Workshop*, pages 47–54, Melbourne, Australia, July. Association for Computational Linguistics.

Yan Chen. 2013. On lexical borrowing from english into chinese via transliteration. *English Language and Literature Studies*, 3(4):1.

Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya. 2008. Hindi to english and marathi to english cross language information retrieval evaluation. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 111–118, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carla Zoe Cremer. 2021. Deep limitations? examining expert disagreement over deep learning. *Progress in Artificial Intelligence*, 10(4):449–464.

Manikrao Dhore, Shantanu Dixit, and Ruchi Dhore. 2012a. Optimizing transliteration for Hindi/Marathi to English using only two weights. In *Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology*, pages 31–48, Mumbai, India, December. The COLING 2012 Organizing Committee.

Manikrao L Dhore, Shantanu K Dixit, and Tushar D Sonwalkar. 2012b. Hindi to english machine transliteration of named entities using conditional random fields. *International Journal of Computer Applications*, 48(23):31–37.

Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295.

Ali El Kahki, Kareem Darwish, Ahmed Saad El Din, Mohamed Abd El-Wahab, Ahmed Hefny, and Waleed Ammar. 2011. Improved transliteration mining using graph reinforcement. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1393, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Takaaki Fukunishi, Andrew Finch, Seiichi Yamamoto, and Eiichiro Sumita. 2013. A bayesian alignment approach to transliteration mining. *ACM Transactions on Asian Language Information Processing*, 12(3), aug.

Roman Grundkiewicz and Kenneth Heafield. 2018. Neural machine translation techniques for named entity transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 89–94, Melbourne, Australia, July. Association for Computational Linguistics.

Bo-Jian Hou and Zhi-Hua Zhou. 2020. Learning with interpretable structure from gated rnn. *IEEE transactions on neural networks and learning systems*, 31(7):2267–2279.

Guillaume Jacques. 2017. Traditional chinese phonology. *webpage, http://www. academia. edu/2261629/Traditional_Chinese_Phonology.*

Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating joint n-gram features into a discriminative training framework. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 697–700, Los Angeles, California, June. Association for Computational Linguistics.

Sung Young Jung, SungLim Hong, and Eunok Paek. 2000. An English to Korean transliteration model of extended Markov window. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Comput. Surv.*, 43(3), apr.

Sarvnaz Karimi. 2008. Machine transliteration of proper names between english and persian. *RMIT University, Melbourne*.

Mitesh M. Khapra, Ananthakrishnan Ramanathan, Anoop Kunchukuttan, Karthik Visweswariah, and Pushpak Bhattacharyya. 2014. When transliteration met crowdsourcing : An empirical study of transliteration via crowdsourcing using efficient, non-redundant and fair quality control. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Shreya Khare, Ashish R. Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA, March. Association for Machine Translation in the Americas.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ngoc Tan Le, Fatiha Sadat, Lucie Menard, and Dien Dinh. 2019. Low-resource machine transliteration using recurrent neural networks. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2), jan.

Haizhou Li, Min Zhang, and Jian Su. 2004. A joint source-channel model for machine transliteration. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 159–166, Barcelona, Spain, July.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Ruchi Khapra, Vivek Seshadri, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Aksharantar: Towards building open transliteration tools for the next billion users.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4), may.

M G Abbas Malik, Christian Boitet, and Pushpak Bhattacharyya. 2008. Hindi Urdu machine transliteration using finite-state transducers. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 537–544, Manchester, UK, August. Coling 2008 Organizing Committee.

Sabina Mammadzada. 2021. A review of existing transliteration approaches and methods. *International Journal of Multilingualism*, 0(0):1–15.

Joyce Man. 2012. Hong kong loves weird english names. https://www.theatlantic.com/international/archive/2012/10/hong-kong-loves-weird-english-names/263103/.

Yuval Merhav and Stephen Ash. 2018. Design challenges in named entity transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630–640, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Diego Moussallem, René Speck, and Axel-Cyrille Ngonga Ngomo. 2020. Generating explanations in natural language from knowledge graphs. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, 47:213.

Oman Muscat. 2011. The english transliteration of place names in oman. *Journal of Academic and Applied Studies*, 1(3):1–27.

Saeed Najafi, Bradley Hauer, Rashed Rubby Riyadh, Leyuan Yu, and Grzegorz Kondrak. 2018. Comparison of assorted models for transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 84–88, Melbourne, Australia, July. Association for Computational Linguistics.

Saeed Najafi, Colin Cherry, and Grzegorz Kondrak. 2019. Efficient sequence labeling with actor-critic training. In *Canadian Conference on Artificial Intelligence*, pages 466–471. Springer.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Adam St Arnaud, Ying Xu, Lei Yao, and Grzegorz Kondrak. 2015. Multiple system combination for transliteration. In *Proceedings of the Fifth Named Entity Workshop*, pages 72–77, Beijing, China, July. Association for Computational Linguistics.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Jong-Hoon Oh and Key-Sun Choi. 2002. An English-Korean transliteration model using pronunciation and contextual rules. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jong-Hoon Oh and Key-Sun Choi. 2005. An ensemble of grapheme and phoneme for machine transliteration. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Jong-Hoon Oh and Hitoshi Isahara. 2007. Machine transliteration using multiple transliteration engines and hypothesis re-ranking. In *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark, September 10-14.

Dinesh Kumar Prabhakar and Sukomal Pal. 2018. Machine transliteration and transliterated text retrieval: a survey. *Sadhana (Bangalore)*, 43(6):1–25.

Taraka Rama and Karthik Gali. 2009. Modeling machine transliteration as a phrase based statistical machine translation problem. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 124–127, Suntec, Singapore, August. Association for Computational Linguistics.

Mihaela Rosca and Thomas Breuel. 2016. Sequence-to-sequence neural network models for transliteration.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview of the fire 2013 track on transliterated search. In *Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, New York, NY, USA. Association for Computing Machinery.

Rui Santos, Patricia Murrieta-Flores, Pável Calado, and Bruno Martins. 2018. Toponym matching through deep neural networks. *International Journal of Geographical Information Science*, 32(2):324–348.

Brendan Shillingford and Oiwi Parker Jones. 2018. Recovering missing characters in old Hawaiian writing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4929–4934, Brussels, Belgium, October-November. Association for Computational Linguistics.

Yingying Su, 2022. *The Influence of Ancient Chinese Cultural Classics in Southeast Asia*, pages 37–58. Springer Singapore, Singapore.

Harshit Surana and Anil Kumar Singh. 2008. A more discerning and adaptable multilingual transliteration mechanism for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Jonne Sälevä and Constantine Lignos. 2022. Paranames: A massively multilingual entity name corpus.

Yulia Tsvetkov. 2017. Opportunities and challenges in working with low-resource languages. https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1423–1424, New York, NY, USA. Association for Computing Machinery.

Shyam Upadhyay, Jordan Kodner, and Dan Roth. 2018. Bootstrapping transliteration with constrained discovery for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 501–511, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jean-Claude Usunier and Janet Shaner. 2002. Using linguistics for creating better international brand names. *Journal of Marketing Communications*, 8(4):211–228.

Yu-Chun Wang, Chun-Kai Wu, and Richard Tzong-Han Tsai. 2015. NCU IISR English-Korean and English-Chinese named entity transliteration using different grapheme segmentation approaches. In *Proceedings of the Fifth Named Entity Workshop*, pages 83–87, Beijing, China, July. Association for Computational Linguistics.

GAO Wei. 2004. Phoneme-based statistical transliteration of foreign names for oov problem. *Master's Thesis, The Chinese University of Hong Kong*.

Chun-Kai Wu, Yu-Chun Wang, and Richard Tzong-Han Tsai. 2012. English-Korean named entity transliteration using substring alignment and re-ranking methods. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 57–60, Jeju, Korea, July. Association for Computational Linguistics.

Winston Wu, Nidhi Vyas, and David Yarowsky. 2018. Creating a translation matrix of the Bible's names across 591 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chun-Kai Wu, Chao-Chuang Shih, Yu-Chun Wang, and Richard Tzong-Han Tsai. 2022. Improving low-resource machine transliteration by using 3-way transfer learning. *Computer Speech & Language*, 72:101283.

Huang Xing and Xu Feng. 2016. The romanization of chinese language. アジア太平洋研究= *Review of Asian and Pacific studies*, (41):99–111.

Andreas Endrique Perez Zepedda. 2020. Procedure of translation, transliteration and transcription. *Applied Translation*, 14(2):8–13, Jun.

Bing Zhao, Nguyen Bach, Ian Lane, and Stephan Vogel. 2007. A log-linear block transliteration model based on bi-stream HMMs. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 364–371, Rochester, New York, April. Association for Computational Linguistics.

冯志伟. 2012. 转写和译音是两个不同的概念. 中国科技术语, 14(5):32–34, 1.