中国科学报/2018 年/7 月/12 日/第 006 版
前沿

人工智能如何拯救方言

科大讯飞多语种高级研究员　祖漪清　整理　本报记者　赵广立

当前，随着经济、文化互动的全球化，主流或通用语言更加强势，弱势语言正濒临消亡。目前世界上大约有 6000~10000 多种语言，而据语言学家预测大部分将于本世纪末消失。

濒危语言保护（下简称为“ 语保 ”）已经成为一项重要而迫切的工作。在中国，普通话的优势地位已经造成一些少数民族语言、方言等弱势语言的使用人数明显减少，如不及时对弱势语言采取措施，我们将失去对人类文化遗产完整记录的机会。

当前方法手段不能满足语保进程

我国对语保工作早就有所重视，并有着深厚的方言研究基础。2005 年我国启动国家语言资源保护工程（以下简称为“ 语保工程 ”），我国学者对方言保护的主要研究方法是田野调查，研究内容包括中国语言资源有声数据库、方言词典、方言地图等。

国际语言学家也对濒危语言进行了语言资源记录。2017 年，美国科学家提出语音罗塞塔计划，旨在通过“ 未知 ”语言的语音和“ 已知 ”语言的文本的平行关系记录没有文字的“ 未知 ”语言（即濒危语言）。

归纳、确定被研究语言的基本音位是语言记录的基本工作之一，但目前这项工作很大程度依赖于调查者对语音的主观感知和“ 口耳 ”工作。由于依赖人工，分析语料局限于孤立字、词，导致研究进度受限，很难将研究内容扩大到连续语音，从音位归纳上升到句法、语义层面的分析。并且，很多中国方言，特别是南方方言中，孤立音节的声韵调在连续话语中表现多变，在复杂的连续话语中，去除语境、韵律结构、情感等诸多因素的干扰，归纳完整的语音变化单靠人力是力所不及的。

同时，随着社会发展的日新月异，每隔数年语言会发生明显变化。因此，语言记录和分析需要高效的解决方案。

利用 AI 技术实现“ 语言复制 ”迫在眉睫

利用人工智能技术系统地研究濒危语言、方言的语音结构、语言结构，实现对一种语言的完整“ 复制 ”迫在眉睫。

“ 语言复制 ”的概念是通过智能语音技术对一种语言实现完整记录。记录内容包括确定该语言的语音结构（例如音节语言的声母、韵母、声调等）、完整分析该语言的句法结构、连续语音的音变和连读变调分析、基本意义单位和主流语言的对应关系以及这个语言的任意文本或语音和主流语言之间的互译关系。

基于主流语言语音系统，完成语音复制需要建立被研究语言的语音合成系统（文语转换系统）、语音识别系统以及和主流语音之间的翻译系统。科大讯飞智能语音技术的发展和多年来的语言积累，可以助力语保工程。一些核心技术的突破和语言积累，使得不同语种之间互译成为可能。

科大讯飞人工智能（AI）研究院有着丰厚的智能语音研究基础，到目前为止实现了中文、英文以外的 30 多种语言（包含多种少数民族语言）的语音合成、语音识别、翻译，其中许多语音系统属拓荒性系统。研究院基于深度学习技术，采用全球文本、声学解决方案，在除中文普通话、英语等强势语言以外的许多语音合成系统上突破了语音合成 MOS4.0 的门槛，目前正尝试在部分濒危语言和方言上进行语言复制。

需要更多热爱母语的人参与

不同的研究目的会产生不同的语言分类。从人工智能的角度出发，我们将语言分为主流语言和非主流语言。中文普通话就是主流语言。非主流语言又分为三个类别。第一类是文字、口语都被广泛使用的语言，例如维吾尔语、藏语等。在这类语言的使用区域，虽然文字被广泛使用，但是往往缺乏正字规范。第二类是有文字但较少使用、口语仍被正常使用的语言，例如彝语、锡伯语等，语言群体内大多数成员仅限在家乡口语交流时使用，多数群体成员不能熟练使用文字或基本不识字。第三类为濒危语言及没有文字的语言，包括只有少数老人还在使用、群体内几乎所有其他的成员都已放弃使用的语言以及没有文字的语言。对这类语言进行完整记录比较困难，实现语言复制也有相当难度。

对于文字、口语都被广泛使用的语言实现语言复制是可行的；对于有文字但较少使用、口语仍被正常使用的语言，实现语言复制也是可能的。对于没有文字的语言可以收集被研究语言的语音，并在有条件的情况下转写成主流语言的文字，使用这样的平行数据，利用人工智能领域的端—端技术实现被研究语言语音到主流语言文本之间的转换，即美国科学家正在实施的“语音罗塞塔方案”，这在逻辑上是可行的。但被研究语言的采集、文本转写缺乏规范并存在许多具体困难。

在可能的情况下尽可能多地收集自然语音和文本的平行数据是十分有意义的。有了足够大的数据，即使目前处理不了，今后仍有机会可利用。利用人工智能技术进行语言记录是一个研究方法的问题，在具体工作中仍然需要采用正确的技术路线进行操作，即使使用了人工智能技术，语言数据的处理仍然脱离不了人力支持。语言是全人类的共同财富，每种语言背后都有精彩的文化。语保工程不应该只是少数人的事业，应该有更多热爱自己母语的人群参与。