重庆科技报

2020年7月9日 星期四
主编:苏伶俐 编辑:李霜 美编:况中银
投稿邮箱:cqkejibao@sina.com

支持108种语言 日均翻译1500亿个单词

谷歌翻译利用AI技术保护土著语言

■ 映 寒

有40%的语言面临消亡的危险,其中大多数是土著语言。非洲大陆的1000种土著语言都需要紧急援助以免灭绝。近日,在一系列AI技术的加持下,谷歌翻译的水平有了重大飞跃,可支持108种语言翻译,尤其是缺乏数据的语言,如约鲁巴语、马拉雅拉姆语,平均每天翻译1500亿个单词。

混合模型和数据挖掘器

谷歌表示,其翻译突破并不是由单一技术推动的,而是针对低资源语言、高资源语言、总体质量、推理速度等一系列AI技术组合的突破。在这系列技术突破中,谷歌首先提到了混合模型和数据挖掘器。

混合模型指的是由Transformer编码器和递归神经网络解码器构成的模型。在机器翻译中,编码器通常将单词和短语编码为内部表征,解码器将其生成为所需要的语言文本。递归神经网络解码器在推理时间上比Transformer中的解码器要"快得多"。

谷歌翻译团队在将递归神经网络解码器与Transformer编码器耦合之前,对递归神经网络解码器进行了优化,以创建低延迟、质量及稳定性均比此前所使用的递归神经网络机器翻译模型更胜一筹的混合模型。

除了新颖的混合模型体系结构之外,谷歌还升级了爬虫工具,爬虫工具可以从数以百万计的示例翻译中收集编译训练数据。升级后,谷歌嵌入了14种大语言对,而不是单纯基于字典数据。这使得该数据挖掘器提取到的句子数量平均增加了29%。

"嘈杂"的数据和迁移学习

谷歌翻译另一个AI技术突破是更好地处理训练数据中的"噪声"。"噪声"即嘈杂的数据,因含有大量无法正确理解或解释的信息数据,从而会损害语料资源丰富的语言翻译。因此谷歌翻译团队部署了一个系统,该系统使用经过训练的模型为翻译示例分配分数,进而筛选出"纯净"的数据。

对于资源较少的语言,谷歌在谷歌翻译中采用了一个回译机制,来强化并行训练数据,即语言中的每个句子都与其译文相配对。在该机制中,训练数据与合成的并行数据自动对齐,目标文本为自然语言,而源文本则由神经翻译模型生成。

此外,谷歌翻译团队还建了一个M4模型。M4模型由团队在2019年提出,该模型对100多种语言的250亿对句子进行训练后,提高了30多种低资源语言的翻译质量。这一模型也证明了在机器翻译过程中可以使用迁移学习技术。这也意味着收集包括法语、德语和西班牙语,这些有数十亿个并行示例的高资源语言进行训练后,可以应用于翻译诸如约鲁巴语、信德语和夏威夷语,这些仅有数万个示例的低资源语言。

增加5种土著语言翻译

"我国现有语言130多种,至少一半语言的使用者不足万人,有25种使用者在千人以下,有11种不到百人。"语保工程首席专家、北京语言大学教授曹志耘介绍。

在2019年5月到2020年5月之间,根据人工评估和BLEU(基于翻译系统翻译和人工参考翻译之间相似性的衡量标准),谷歌翻译在所有语言中平均提高了5分以上,在50种语料资源最少的语言中平均提高了7分以上。

自2010年以来,翻译质量每年都在提高,但是机器翻译绝不是翻译问题的"终结者"。谷歌承认,即使是增强后的模型也容易出错,包括将一种语言的不同方言混合在一起,产生过多的直译,以及在特定主题、非正式用语或口语上的表现不佳。

谷歌尝试用不同的方法来解决上述的问题,曾发布一项计划旨在招募志愿者,通过检查翻译单词和短语是否正确来帮助提高低资源语言的翻译性能。今年2月份,谷歌翻译与新兴的AI技术相结合后完成了技术升级,他们又增加了卢旺达语、奥里亚语、鞑靼语、土库曼语、维吾尔语等大概仅有7500万人使用的五种语言翻译,不仅让这些语言的使用者受益,也对全球的文化多样性有着重要的贡献。

近年来,河北景县抢抓现代智能物流装备需求加大的机遇,建设龙华现代智能物流装备产业园,推动物流装备产业向智能化、高端化、服务优质化方向发展。

新华社记者 李晓果 摄

高考用上高科技 AI上阵查违规

■ 樊 华

近日,2020年辽宁省普通高考使用人工智能(AI)技术进行考试疑似违规行为检测,这是辽宁省首次将AI技术应用到考试行为分析中。

人工智能系统能够在短时间内快速对所有考场的视频文件进行分析判断,检测出考生的疑似违规行为,各级考务工作人员再对系统检测出来的疑似违规行为进一步甄别判断,并依据相关规定做出处理。

在当前新冠肺炎疫情防控常态化的形势下,将AI技术应用于普通高考违规行为管理中,能有效地提高工作效率,丰富考风考纪管理手段,提升考务工作管理水平,符合国家教育考试公平、安全、科学及规范原则。AI技术未来还将在其他国家教育考试中使用,确保国家教育考试公正公平。

智能盘点机器人"上夜班"盘点图书

■ 王祖凌

在60万的书海中如何快速找到书,闭馆之后谁来上夜班盘点书目?中新友好图书馆再添黑科技,智能盘点机器人、智能分拣还书系统和室内定位导航系统三个智慧应用在中新友好图书馆"组团上岗"。

"高大帅气"的智能盘点机器人专为解决图书盘点核对问题而研发设计,通过将RFID感知、计算机视觉及智能机器人技术的有机结合,能实现对图书的自动化盘点。目前,中新友好图书馆也是国内首家使用高频RFID技术盘点机器人的图书馆。目前馆内共有3台智能盘点机器人,能在夜晚闭馆期间对全馆在架图书进行自动盘点,同时生成错架报告发送给图书馆工作人员,提高图书排架准确率。据试运行统计,平均每台智能盘点机器人每小时可盘点2万多册图书,一晚上能盘点近19万册,这些工作量需要15名馆员同时在岗、整晚不间断盘点才能完成。

机器人上"网课"模仿手术过程

最近在谷歌、英特尔、加州大学伯克利分校的合作研究中,研究人员通过用手术教学视频来对机器人进行"训练",让其能模仿手术过程。

之前,加州大学伯克利分校的教授曾用YouTube视频指导机器人学习各种动作(比如跳跃和跳舞),而谷歌则是有训练机器人理解场景中的动作的经验。

在最近发布的论文里,研究人员简单介绍了他们如何用YouTube视频来训练两臂达芬奇机器人在针刺机上进行缝合操作。机器人从模仿学习的视频演示中,获得以运动为中心的操作技能。其中算法一致性、可解释性和监督学习的负担是该项目模仿学习中的关键问题,毕竟通常很难精确地描述定义一个片段和标记的内容。

首先研究人员尝试将一小段被标记的视频进行分类,比如手术缝合任务分解成分段动作,手术的针头刺入、拔针、针头脱落等动作。然后,研究的重点是以半监督的方式从视频演示中提取动作,进行动作分割和模仿任务。

由于训练采用的是Jigsaws数据集,其中包含3个手术任务的视频演示,分别是缝合、穿针和打结。所以在这次测试中,只有机器人尝试模拟缝合运动,并没有考虑更多的技术建模或者其他的问题。

除此之外,数据集里的视频演示并不只有一位医师,而是由8位技术等级不同的外科医师组成,所以这会产生一个问题,不同的医师,习惯可能会不同。而机器人正是要学习所有医师的缝合视频。

在缝合任务中,该团队仅用了78个医学教学视频就能训练两臂达芬奇机器人的AI引擎进行相同操作,而且声称有接近85%的成功率。

(本报综合)