自动化所提出一种适用于低资源和零资源的多语言机器翻译方法

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/5469.html

自动化所提出一种适用于低资源和零资源的多语言机器翻译方法。机器翻译是利用计算机实现从一种自然语言到另一种自然语言自动转换的技术。为了实现多语言之间的相互翻译，通常需要构建多个一对一的翻译模型。一方面每个翻译模型需要大规模存储和计算资源，从而多语言翻译的存储和计算消耗非常巨大;另一方面多语言翻译在独立模型下无法实现知识共享。现有基于编码器-解码器的统一多语言翻译框架虽然可以较好地解决资源占用问题，却面临着参数共享和语言共性未被充分利用的问题，导致目前多语言翻译系统的译文质量较低。因此，如何平衡翻译知识的共享和独立，既解决资源消耗问题同时利用语言共性提升译文质量，成为多语言机器翻译的核心挑战。针对该挑战，中国科学院自动化研究所自然语言处理团队提出一种结构紧凑且语言敏感的多语言机器翻译方法，提供了有效的解决思路，相关成果将发表在ACL-2019学术会议上。

该工作主要基于编码器和解码器框架下的多语言机器翻译。首先在模型表示方面，团队提出了一种表示器模型，共享编码器和解码器的模型结构与参数，取代多语言翻译框架下的编码器和解码器，从而显著减少了模型参数的规模，更好地利用了语言之间的共性。同时，为了提升模型对不同语言的区分能力，团队提出了三种语言敏感的模块，分别是语言敏感的词向量、语言敏感的注意力机制以及语言敏感的判别器。

针对不同语言，团队设定一个语种向量，该向量称之为语言敏感词向量。如下图最底端所示，该语种向量加到输入的词向量中，并在训练过程中进行调优。下图红色虚线标明了语言敏感注意力机制模块，该模块对于不同的翻译任务，动态地选择不同的注意力机制。下图最顶端是团队新提出的语种判别器模块，该模块对表示器的最上层的隐式表示进行语种分类，增强在解码过程中对不同语言的区分能力。

团队分别在较大规模的WMT数据集和较小规模的IWSLT数据集(如表1所示)上进行了一到多和多到多的多语言翻译实验，来验证该方法的性能。相较于之前的多语言翻译模型，该方法均有一定的提升，甚至在一些语言对上面超过了独立一对一模型在双语上训练的模型。同时，该方法极大地压缩了模型参数规模，其中在一到四多语言翻译中，在仅包含20%左右的参数规模的情况下就能取得可比的翻译性能。

在多种语言到多种语言的翻译情境下，团队对提出的方法进行了测试。该方法相较于基线系统有了显著的提升，其中在语料不平衡的翻译情境下，在英-越双向翻译任务上都取得了当前最好的性能。同时，在零资源的Zero-Shot翻译情境下，该方法也比前人的工作有了一定程度的提高，说明该方法能够很好地利用语言之间的共性，适用于低资源和零资源的翻译情形。

模型结构示意图