神经网络机器翻译中的集外词处理方法

张家俊
副研究员
中科院自动化所

摘要：

基于数据驱动的机器翻译方法严重受限于双语训练数据的规模。最为直接的影响之一便是集外词翻译问题：如何处理训练语料中未出现过的词语。由于模型约束与计算复杂度的限制，最近兴起的神经网络机器翻译方法仅仅对几万高频词进行编码和翻译，所有低频词成为集外词，从而进一步加剧了集外词翻译问题。一方面，集外词无法获得正确目标译文；另一方面，大量出现的集外词将严重破坏句子结构，影响上下文及整个句子的翻译。我们提出一种“替换-翻译-恢复”的集外词解决方案。在“替换”阶段，我们旨在寻找低频词的高频词替身，通过词语替换保持句子的语义结构；替换后的数据用于神经网络翻译模型训练；在“恢复”阶段，我们提出一种基于字符的神经网络翻译方法，从而可以处理绝大多数的集外词翻译问题，最后将句子译文中的某些词重新替换为集外词的目标译文。实验表明这种集外词处理方法可以大幅度提升神经机器翻译的译文质量。

简历：

中科院自动化所模式识别国家重点实验室副研究员，研究方向为自然语言处理、机器翻译、跨语言文本信息处理、深度学习等。现任人工智能学会青年工作委员会常务委员、中文信息学会青年工作委员会委员。在国际著名期刊 IEEE/ACM TASLP、IEEE Intelligent Systems、ACM TALLIP 与国际重要会议 AAAI、IJCAI、ACL、EMNLP、COLING 等发表学术论文 20 余篇。2009 年获亚太会议 PACLIC 最佳论文奖，2012 年获 NLPCC 最佳论文奖，2014 年获 CWMT 最佳论文奖。2014 年获中国中文信息学会“钱伟长中文信息处理科学技术奖”一等奖（排名第三）。2015 年入选首届中国科协“青年人才托举工程”计划。