朱靖波：谈谈机器翻译技术发展与产业化

[深度学习自然语言处理](javascript:void(0);) 昨天

<https://mp.weixin.qq.com/s?__biz=MzI3ODgwODA2MA==&mid=2247496533&idx=2&sn=61cd41d339cfcc3ce2cfa3276a121b7d&chksm=eb53e7c6dc246ed08a9537897e8b2de984909a705528006d1e064bc15503a7ea8c5a0a646535&mpshare=1&scene=23&srcid=1112zehvEjLzinfkf60h99LW&sharer_sharetime=1605151640881&sharer_shareid=8a1847d9b672a9232c0d78cf977ed561#rd>

转载自公众号：AI科技评论



作者 | 朱靖波

编辑 | 陈彩娴

由中国中文信息学会主办，山东大学承办的第十七届自然语言处理青年学者研讨会(YSSNLP 2020)于10月24-25日在线上举行。作为国内NLP领域最重要盛会之一，大会邀请了国内外计算机领域知名专家、企业家在线上做了精彩报告。小牛翻译董事长、东北大学计算机学院人工智能系朱靖波教授做了题为《谈谈机器翻译技术发展与产业应用》的特邀报告，以下是报告整理，有部分删减。

说起YSSNLP这个会议，我对它有着很深的感情，这要从我跟它的渊源说起。时间追溯到2004年年底，我到三亚参加首届IJCNLP会议，刚好复旦大学黄萱菁老师也在那举办国内的信息检索会议，第一次与黄老师见面，我就跟她聊起了我的想法——打算做个青年学者研讨会。黄老师听后也觉得这个很好的事情，可以组织起来，于是，我们说干就干，办起了第一届自然语言处理青年学者研讨会，我给它取了个名字——YSSNLP，这个名称沿用至今。

第一届YSSNLP会议，请到了台湾的苏克毅老师做了题为《Why and How》的报告，同大家聊了聊我们为什么要做研究。当时会议宗旨包括三点：加强学术界青年学者之间的交流、通过平台加强与企业的交流、加强与政府主管部门的交流。当时会议还是民间性质的，于是哈工大刘挺老师提了一个建议，说想做一个精英俱乐部，有名额限制，全国每个实验室只派一个代表参会。于是，YSSNLP会议就这样如火如荼地开起来了。2010年是一个转折点，我们做了一个重要的决定。2010年第七届会议在沈阳召开，会上我们决定给每个实验室增加一个新名额，但是必须要是年轻人，给我们的会议增添一些新鲜血液，会议的规模也从这届起开始扩大。



渐渐地参会的人越来越多，我们把会议交给了中文信息学会，学会又专门成立了青工委主管会议，参会的年轻人越来越多，也越来越热闹了。回想起来，当初能够参与并组织起这样个会议，真是一件让人非常自豪的事。 

**机器翻译技术发展**

渊源讲完，咱们回到正题。给大家讲一讲机器翻译技术发展和产业应用。机器翻译是快速进行大数据翻译任务的唯一解决方案，其实当初提出机器翻译的时候，动机就是为了代替人工。由于机器翻译的技术在可预期的将来，不可能达到人工翻译的水平，所以我们把它当成一种技术工具。机器翻译系统包括几个维度：语种、垂直领域、应用模式。第一代机器翻译系统被董振东老师称作是“傻子”，是因为规则是写不完的，只要是机器没有“见过”的句子，它就翻译不了。80年代末90年代初，第二代统计机器翻译出现了，它最吸引人的地方，不是技术的问题，而是不用再写规则了。用机器学习的方法构建翻译模型虽然效果有了提升，但是只要句子稍微复杂，机器就翻译不好了。比如说，它对短语翻译得不错，但是一遇到调序问题就不行了。所以，董老师当时称第二代机器翻译系统为“疯子”。到了2013，提出了第三代机器翻译系统——神经机器翻译技术。它的基本思想就是不再用短语去拼接翻译，而是把句子变成一个向量去翻译，当时提出的框架就是端到端（编码器-解码器）。最初，提出来的时候主要是应用RNN循环神经网络，后来又由Facebook提出了CNN卷积神经网络，再后来又由谷歌提出了基于自注意力的Transformer。其实，神经网络这个理论在40年代就有了，但是真正运用到机器翻译领域还是比较晚的。在2013年之前，为了解决统计机器翻译语言模型的问题，大家也引入了神经网络的技术去建模，虽然效果有所改善，但问题并没有完全解决。2013年，Encoder-Decoder框架被提出，神经网络诞生。虽然理论是新的，但是不被界内看好，因为它的性能并不好。2015年Attention机制被引入到神经网络中，大大推进了机器翻译的发展，让翻译品质得到大幅度提升，神经机器翻译技术一下子火起来了。我是做机器翻译产品的，所以非常关注最新一代技术。用统计机器翻译做成产品后，用户只能接受免费试用，付费的话用户根本不会买单。所以，在2012年-2015年那段时间，小牛翻译发展的举步维艰。2015年，我参加了一个微软研究院的会议。交流的时候，我听到了一个消息：微软总部和谷歌都在研究神经机器翻译。我就想，在统计机器翻译时代小牛翻译可以说是第一梯队的，但是这条路行不通继续走死路一条。那么，就开始做神经机器翻译吧，统计机器翻译技术从理论到工业界大概用了十五年，神经机器翻译技术发展得再快，落地到工业界差不多也需要七八年，我们再赌一次。没想到，神经网络技术从提出到进入工业界只用了短短三年的时间。2016年10月，谷歌完整的神经机器翻译系统上线了，一下子引起了业内的轰动。随后微软的神经机器翻译系统11月上线了，紧跟着，同年12月，小牛翻译的神经机器翻系统就部署到了合作伙伴的服务器上。 

**如何更好构建机器翻译系统**

神经机器翻译一登场，一时风头无两，那么问题也出现了，多家机器翻译公司如雨后春笋跑出来，特别是开源出来以后，机器翻译的门槛儿也变得越来越低了，这对我们这些做统计机器翻译出身的人来说，打击还是挺大的。但是话说回来，玩一玩是可以的，但是要真正做到系统能用，其实代价还是挺高的。大家经常会问我：如何真正做好一套好用的机器翻译系统？我认为至少需要三个东西：一是需要掌握最新的机器翻译技术；二是需要大规模的双语句对库；还有很重要的一点是要针对错误进行驱动打磨。打磨一个系统是需要花时间的，这就是为什么大家用相同的数据，用相同的模型，做的系统不一样。现在机器翻译面临一个比较大的问题——系统不会反馈学习。比如，我们在给用户提供服务的时候，用户说翻译结果不理想时也会自行修改，但是系统“记不住”，下次翻译还是不对。所以，如何让机器翻译系统实现更好的人工干预，或者通过用户提供的错误纠正实例来进行反馈学习，我觉得将是一个值得挖掘的方向。另外，现在在学术界有很多研究成果、论文，包括都在讲基于知识的机器翻译，但其实这个问题是有争议的。我个人观点，语言学的知识，甚至说外部的知识，包括领域的知识，我觉得对机器翻译是有用的。毕竟，现在机器翻译用的是数据驱动的方法，它只能从现有的数据里面得到。 

**目前机器翻译面临的问题**

**1）复杂网络建模问题**更加复杂的网络结构表示能力更强，在编码阶段能够对输入句子实现更好地表示学习，能够提高解码阶段的翻译品质。所以我们就要用更深更宽的网络去建模+训练。**2）结构学习问题**重要进展往往源于网络结构的创新，但很多未知结构没有被探索，这个过程需要人的“灵感”和大量经验性实验，我们要做的就是想办法实现结构的自动学习，让它能够针对数据，适配相应的结构。**3）效率问题**传统的NMT模型对于GPU计算资源要求比较高，比如模型大小达到几个G，难以适应小设备的应用场景，这就需要提升模型效率，这就需要我们创造出更小更快的系统。**4）适应性问题**通用模型被应用到新的数据、领域效果差，不同规模、类型的数据，甚至不同表达方式对系统表现影响很大，我们可以考虑打造面向低资源场景的高适应性模型。**5）语音翻译建模问题**传统方法通常简单将语言识别技术和文本翻译技术串联起来，会导致错误蔓延，如何采用端到端进行语音翻译建模是一个现实存在的问题。从这一点上看，多模态神经机器翻译的建模是比较值得关注的。**6）可解释性问题**大家一直在研究机器翻译的可解释性问题，在深度学习中不管哪个方向都面临这个问题，我们要搞清楚一点，我们为什么要研究机器翻译的解释性问题。举个例子，比如说我们用深度学习技术做一套判案系统，代替法官来判案，那么，对用户来说，他们是关心AI是如何推理的？还是关心判决有罪的法律依据？我觉得机器翻译的可解释性研究，它可能最关心的并不是这个东西是怎么推导出来的，也不是这个结果从语言学、翻译学怎么解释（这个观点是前段时间宋老师提出来的），我们研究可解释的目的应该是为了纠错。所以，我觉得机器翻译的可解释的研究，将来应该朝如何更好地完成纠错的方向去发展，而不是从语言学上、翻译学上去解释这个译文正确与否。 

**关于新一代的机器翻译技术**

**基于知识的机器翻译技术会成为新一代技术吗？**其实这个问题是没有解的，比如说，我们该用什么样的知识，知识怎么获取，知识怎么表示，知识怎么用，这些问题都是很难解的。所以说，基于知识的机器翻译系统能不能作为下一代，我觉得很难说。**能不能通过引入新的学习范式，比如小样本学习能力，实现新一代机器翻译？**现在人们提出一个观点就是，机器翻译要求训练的数据，双语句对量太大了，像我们做小牛翻译的时候，像英中，中英要几亿个句对，能不能用更少的语料，完成机器翻译的训练？我觉得这是一个非常值得研究的方向。**机器翻译需不需要理解？**严格地说，现在的机器翻译它根本没有理解能力，所以说将来需不需要理解？我觉得适当的理解有可能是需要的，包括稀缺资源的翻译能力。**是否存在第四代机器翻译技术？**神经机器翻译在训练数据如果足够充分的情况下，翻译品质人工评价可以达到85%甚至是90%，随着技术的迭代，可能会变得更高。那么，下一代机器翻译系统性能会达到多少呢？我觉得后续的机器翻译发展方向，不是简单的提出一套新的技术去碾压神经机器翻译，它应该是针对具体的应用场景，不断地提出更加有效地机器翻译技术。 

**技术驱动激活产业化**

小牛翻译迭代了三次，1996年做的是基于规则的系统，2003年做的也是基于规则的系统，那时候技术不行，根本不赚钱。现在技术发展不错，但是市场不够大。我预测在3-5年以后，技术会发展得更好，当然，市场也会变得很大。所以我建议大家，如果想做机器翻译创业的话，现在是个好时机，但是要有耐心，要学会等待机会。机器翻译的应用场景还是非常丰富，包括文档翻译、翻译机、大数据舆情分析、口语翻译、翻译笔、会议同传等等。我对AI同传这个方向还是看好的，但是现在它的技术不够成熟，但我相信等到将来技术成熟以后，AI同传会成为国际会议室的标配，而多模态应用即语音跟翻译结合我觉得是绝配。现在人工智能芯片发展得非常好，大家都知道语音处理有语音芯片，图像处理有图像芯片，但是机器翻译芯片现在还没有，我很看好机器翻译芯片这个方向，一旦研发出来，可以应用到复印机、传声机、翻译机里面，我觉得这是一个不错的方向。 

**机器翻译市场分析**

机器翻译是一个强刚需，随着技术的发展，用户需求不断被激活，但是机器翻译产业不属于爆发型，属于稳步增长型。将来To C应用大多以免费为主 ，机器翻译目前的主战场是To G跟To B项目，将来的主战场我觉得是在To B上。我预测在机器翻译行业，不可能有行业巨头出现，机器翻译的市场是无法垄断的。主要是机器翻译的维度太高了，它有很多语种，有很多领域，还有太多的应用场景。所以，将来机器翻译行业，它一定是百花齐放的。机器翻译市场目前还不够大，根据很多调查报告来看，保守来说应该在30-50个亿左右，但是我相信机器翻译在3-5年以后，特别是机器翻译和语音翻译的多模态结合到一起之后，我觉得是可以达到300-500亿的市场。我觉得5年以后，会出现一家以机器翻译为核心的独角兽。未来，机器翻译将来会出现两类企业，第一种只专注机器翻译引擎研发，做基础设施平台的，第二种是主要做应用与服务的，我建议大家如果想在机器翻译领域创业的还是做第二类，结合具体的应用场景，直接接触客户，投入比较小。机器翻译产业已经从蓝海进入到了红海，市场洗牌已经开始，波及最大的应该是做引擎的企业，这类企业将来可能所剩无几。但是，在应用服务这块会各显神通。所以，我觉得眼下的情况是技术不错，市场不够大。但是3-5年后技术成熟，市场会迅速扩大，我还是比较乐观的。 

**关于科技成果转化**

我是非常鼓励科研成果走出象牙塔创造更大的社会价值的。转化的方式包括专利授权，技术转让，横向项目、创业转化等。科研人员创业的优点无非就是懂技术，能够正确预测技术发展，轻易不会被技术发展所淘汰掉，但是技术没有商业模式，只有产品才有商业模式，所以要怎么解决用户的需求，这是一个关键。强技术团队的优点就是学习能力比较强，但是市场营销能力比较弱。这种情况，可以选择和别人合作，专业的事情交给专业的团队去做，一定要把强技术团队转化为强产品团队或者强运营团队，才能支撑起你的商业模式。当然，一定要解决好知识产权问题，千万别因为知识产权把你的路堵死了。我的观点是：好的品牌=好人品+好产品+好服务。只有这样，你的企业才能真正实现盈利。创业转化模式可能是一条美好的不归路，一旦走上这条路，就没有退路可言，所以大家要谨慎而为之。谢谢大家！