人类语言技术展望

文 / 宗成庆

摘　要　机器翻译伴随着世界上第一台计算机的诞生而出现，随后成为人工智能领域最具挑战性的研究课题之一。70 多年来，以机器翻译、人机对话系统、文本自动分类、自动文摘和信息抽取等为代表性应用的人类语言技术所走过的曲折发展历程，从不同的侧面折射出人工智能领域的荣禄兴衰。本文在简要回顾人类语言技术发展历程的基础上，重点介绍当前该技术面临的主要挑战和研究现状，并对未来发展的趋势进行展望。

关键词　自然语言处理；自然语言理解；计算语言学；人类语言技术

宗成庆

中国科学院自动化研究所研究员，CAAI Fellow，国际计算语言学委员会 (ICCL) 委员，亚洲自然语言处理学会 (AFNLP) 主席。主要从事自然语言处理、机器翻译等研究。主持国家项目 10 余项，国家重点研发计划重点专项首席科学家，ACM TALLIP 和《自动化学报》副主编。曾任国际一流 学术会议 ACL 2015 和 COLING 2020 程序委员会主席，多次担任 IJCAI 和 AAAI 领域主席。曾获国家科技进步奖二等奖、钱伟长中文信息处理科学技术奖一等奖等。荣获北京市优秀教师、宝钢优秀教师和中科院优秀导师等荣誉。

0 回顾

自 1956 年人工智能(Artificial Intelligence, AI)概念被提出以来，自然语言理解（ Natural Language Understanding, NLU ）就一直是这一领域研究的核心问题之一。尽管上个世纪 60 年代提出的计算语言学（ Computational Linguistics, CL ）和 70 年代衍生的自然语言处理（ Natural Language Processing, NLP ）概念分别从数学建模和语言工程角度各自诠释了不同的外延，但 NLU、CL 和 NLP 这三个术语的实质内容和共同面对的科学问题并无本质的差异，其实际应用目标是完全一致的。因此，在不引起混淆的情况下人们通常以“人类语言技术”（ Human Language Technology, HLT ）泛指这一集语言学、计算机科学和认知科学等研究为一体的多学科交叉领域。

回顾人类语言技术发展的 70 多年历史，其技术方法大致可以划分为三个阶段：① 从学科萌芽期到上个世纪 80 年代后期及 90 年代初期，为采用以模板、规则方法为主的符号逻辑阶段 ，属于理性主义方法；② 从上个世纪 90 年代初期到 2013 年前后，是以统计机器学习为主流方法的经验主义方法时期；③ 从 2013 年之后，进入了以多层神经网络为主流方法的连结主义时期。图 1 给出了整个 70 年的大致走势。

在理性主义方法为主的历史阶段，主要研究工作是建立高质量

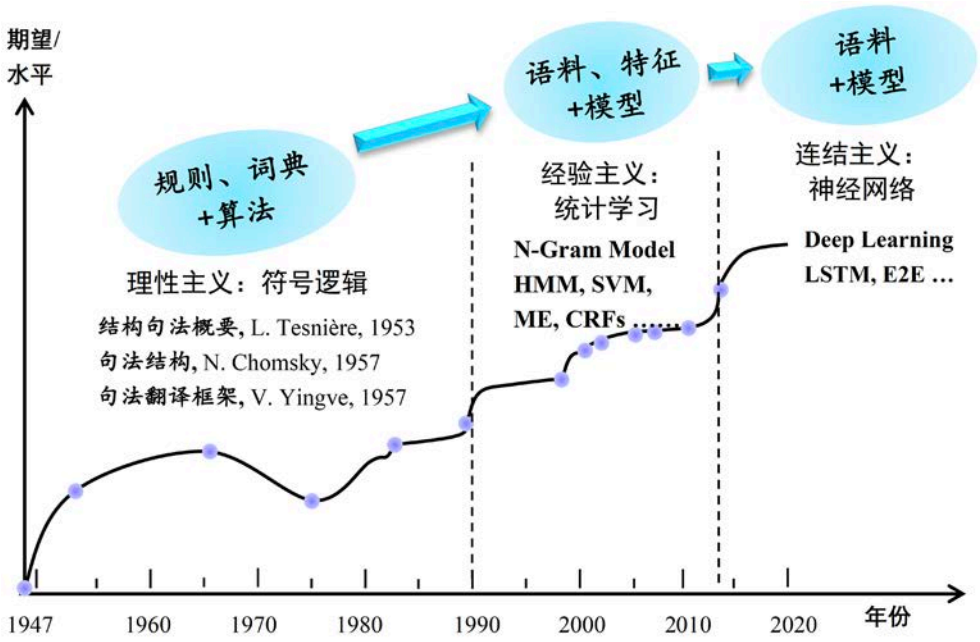

图 1 HLT 技术发展的历史阶段[①]

的词典、规则和推理算法，通过符号推理和逻辑运算实现自然语言句子的分析、转换和生成，其代表性的理论是乔姆斯基（N. Chomsky）的句法结构理论。

在经验主义方法为主流方法的历史阶段，主要研究工作是获取大规模训练样本，研究建立高质量标注体系和自动标注算法，构建基于统计方法的计算模型和算法，通过调试和优化模型参数实现面向自然语言处理任务的推断和预测，其主要理论基础是概率论和信息论。在这一阶段，*n* 元语法（*n*-gram）模型诞生，隐马尔可夫模型（Hidden Markov Model, HMM）、支持向量机（Support Vector Machine, SVM）、最大熵（Maximum Entropy, ME）和条件随机场（Conditional Random Fields, CRFs）等一系列统计学习方法，被广泛应用于自然语言处理任务。统计机器翻译（Statistical Machine Translation, SMT）系统诞生，一批开源工具公开发布，谷歌、微软和百度等公司研发的统计机器系统相继上线，推动了该技术的快速发展。

继 2006 年 G. E. Hinton（辛顿）等人将多层神经网络方法成功应用于图像识别之后，2009 年微软实现了基于多层神经网络的语音识别系统，并使识别错误率大幅度下降，深度学习方法得到大规模应用。2014 年纽约大学 Kyunghyun Cho 和加拿大蒙特利尔大学的 Yoshua Bengio 等人提出了基于注意机制的编码器 - 解码器（encoder-decoder）基本框架，对神经网络结构创新和二次开发，建立了基于神经网络的机器翻译系统，简称神经机器翻译（Neural Network based MT, NMT）系统。在此基础上谷歌于 2017 年提出了完全基于注意机制的 Transformer 模型，国内众多公司通过跟踪、完善，实现了自己的神经翻译引擎，为普通用户提供机器翻译服务。很多中小型企业利用开源平台和互联网开放数据快速搭建性能尚可的机器翻译系统，或者直接利用谷歌和百度等公司提供的翻译服务，从而使这一领域出现遍地开花、欣欣向荣的大好局面。2018 年 Google 发布的双向预训练模型 BERT（Bidirectional Encoder

① 图中曲线上的标志点表示发生在当年的标志性事件，限于篇幅，本文不一一列举。

Representation from Transformers）更将这一技术领域推向高潮。

1 现状

语言技术涉及众多领域和分支，不同的分支和方向具有相对的独立性，发展起点和速度也不一样，无论是理论基础和关键技术，还是资源建设和应用系统研发等，在不同的层面上发展状况都不一样，很难对其研究现状一概而论。以下仅对部分应用系统的性能现状进行简要的概括，希望能够达到管中窥豹的效果。

机器翻译作为自然语言处理中最具挑战性的研究课题，其译文质量的水平在很大程度上代表着自然语言处理技术的整体水平。近年来，尤其是 2014 年神经机器翻译模型提出以后，机器翻译的译文质量得到了显著提升。对于口语翻译而言，在资源较为充分的语言对上（如英汉、日汉、英法等），在说话场景不是非常复杂、口音基本标准、语速基本正常、使用词汇和句型不是非常生僻的情况下，日常口语翻译的性能基本可以满足交流的需要。对于专业领域的文本翻译而言，在训练语料较为充分时译文准确率可以达到 75% 以上。新闻领域的翻译准确率跨度较大，总体而言，新闻文本的翻译准确率基本在 70% 左右。而对于译文质量要求较高的翻译任务，如领导人的讲话稿或著作、文学名著，以及严肃场景下的演讲和对话（包括领导人的讲话、答记者问，或者有较严重口音的讲座和对话等），机器翻译系统都难以胜任。在可预见的未来看不到机器翻译系统将替代人工翻译的可能性。而对于资源稀少的小语种（如乌尔都语、波斯语等）与汉语之间的翻译，目前的机器翻译系统只能以快速获取信息为目的帮助人们大致了解原文的主题和内容。

人机对话系统一直是人们关注的热点，也是自然语言处理领域极具代表性的研究任务。对话系统通常包括面向任务的对话系统（task-oriented dialog system）和开放域的对话系统（open-domain dialog system）两大类。前者称为任务型对话系统，如机票预订系统等；后者称为闲聊式对话系统，如聊天机器人等。目前学术界研究的对话系统基本都采用数据驱动的方法，尤其端到端的神经网络模型提出之后，几乎成为类似任务实现的统一框架。这类系统的性能在很大程度上取决于训练样本的规模和质量。耐人寻味的是，目前商用的任务型对话系统基本上都采用基于规则的实现方法。对于特定领域和特定任务的对话系统而言，其任务完成的准确率可以达到 75% 以上，这对于某些特定的领域或行业，需要大量工作人员完成的重复性较大的服务任务来说，已经能够大幅度节减人力资源，提高工作效率。

总体而言，自然语言处理已经取得了丰硕成果，新的模型和方法不断被提出，并得到成功应用；很多应用系统已经被广泛使用，并直接服务于社会生活的各个方面。但是，自然语言处理仍面临若干挑战，远没有达到像人一样理解语言的程度。当前面临的主要问题可以概况为如下五点：

（1）缺乏有效的知识表示和利用手段

这里所说的知识，包括常识、领域知识、专家的经验知识和语言学知识等。对于大多数语言学知识和部分领域知识在一定程度上可以从大规模训练样本中学习到，但是很多常识和专家经验往往是“超出训练样本范围”的。例如，“Premier Li”曾经在很长的一段时间里指代李鹏总理，可是目前应该指李克强总理；“transformers”在政治领域指改革者，在电力系统指变压器，在儿童玩具中指变形金刚，而在自然语言处理领域指转换器。那么，具体指什么，需要根据上下文背景和领域确定。再如，在鸡兔同笼问题求解中，关

键常识是鸡有两条腿、兔子有 4 条腿。如果没有这种常识，这个问题就无法求解。对于人而言，这些知识都是常备的；而对于机器而言，却难以从样本中（尤其是有限的小规模样本中）归纳学习出来。

（2）缺乏未知语言现象的处理能力

对于任何一个自然语言处理系统来说，总是会遇到未知的词汇、未知的语言结构和未知的语义表达。所谓“未知”即在训练样本和词典中未曾出现过。世界上任何一种语言都在随着社会的发展而动态的变化和演化着，新的词汇、新的词义和新的句子结构都在不断出现，这些现象在微博、聊天和日常会话等非规范表述中尤为突出。例如，“李菊福”表示的意思是“有理有据使人信服”；“内牛满面”意思是“泪流满面”；等等。如果系统的前端输入是语音或者图像，语音识别或者 OCR 处理后的结果中含有大量的噪声，也是十分常见的现象。因此，一个实用的自然语言处理系统必须具有较好的未知语言现象和噪声的处理能力，即鲁棒性（robustness）。

（3）模型缺乏解释性和“举一反三”能力

尽管包括神经网络方法在内的机器学习方法已经在自然语言处理的各种应用任务和关键技术研发中发挥了重要作用，但是这些方法毕竟采用的是以概率计算为基本手段的“赌博”思维，其性能表现严重依赖于训练样本的质量和规模，当测试样本与训练样本差异较大时，模型性能急剧下降，更无从谈起“举一反三”。从纯粹的自然语言理解角度，目前的模型性能还非常有限，尤其缺乏合理的解释性。对于给定的输入，模型在“黑箱”变换过程中产生错误和丢失数据的原因是什么？每一层变换意味着什么？最终结果的可靠性有多大？目前还没有合理的解释。

（4）缺乏交互学习和自主进化的能力

自然语言处理系统在实际使用过程中会持续得到用户的反馈，包括对系统结果的修正、为系统增加新的词汇解释和补充新的标注数据等。传统的机器学习方法是将用户的反馈信息添加到训练数据中，重复进行“训练—测试”循环，以达到不断优化模型的目的。但是这种方法通常需要较长的迭代周期，难以有效利用实时的反馈信息。类比人的交互学习能力，一个智能系统应该具备在线交互学习的能力，即从用户与系统的交互过程中不断学习、补充和修正已有的知识，以达到模型自主进化的效果，而这个学习和进化过程是终生的（life-long learning）。

（5）单一模态信息处理的局限性

目前的自然语言处理研究通常指以文本为处理对象的研究领域，一般不涉及其他模型信息的处理，例如语音、图像和视频等信息，最多在某些场景下利用语音识别或 OCR 作为前端预处理，各模块之间是独立的，与语音、图像和视频等信息处理过程是相脱节的，这严重违背了“类人智能”的基本前提。对于人而言，通常是“眼观六路，耳听八方”，说出来的话，写出来的字，与看到的实际情况是一致的，而来自各个器官的信息是相互补充和验证的。试想，同样一句话借助不同的语调、重音和手势表达，意思很可能完全不同。因此，多模态信息综合利用、协调处理，势在必行。

另外，在谈论人类语言技术整体现状时，不得不对我国在该领域的迅速崛起给予充分的肯定和赞誉。近 10 年来，中国的自然语言处理研究发展迅猛，无论是在国际一流学术会议（ACL、EMNLP、COLING、AAAI、IJCAI、WWW 等）和期刊上发表的论文数量，还是我国学者在相关国际学术组织中担任重要职务的情况，都无可争

辩地标志着我国在这一领域拥有的举足轻重地位和势不可挡的发展趋势。然而，令人遗憾的是，这一领域在国内却没有得到应有的地位和话语权。

2 展望

作为人工智能领域重要的研究方向和分支，语言技术研究不仅涉及词法（形态）、句法、篇章和语义等语言学本身的特点和规律，需要解决基础性关键问题，而且需要面向实际应用构建机器翻译、自动文摘、情感分析、对话系统等特定任务的数学模型和方法。笔者认为，最终要解决人类语言理解的问题，使相关应用系统的性能达到更高的水平，满足个性化用户的需求，甚至真正做到像人一样理解语言，以下三方面将成为未来发展的重要方向。

（1）与神经科学密切结合，探索人脑理解语言的神经基础，构建更加精准、可解释、可计算的语义表征和计算方法

人脑是如何表征和处理文本语义的，这是一道难解之谜。相比于视听觉等神经系统，目前对于人脑语言系统的了解还非常初步。近年来，数据驱动的自然语言处理方法在很多方面有效地弥补了传统方法的不足，但是，正如前文所述，数据驱动的方法存在很多固有的弊端，包括性能对训练样本的依赖性、模型的可解释性和常识的表示、获取和利用等问题，而人脑在小样本数据上的归纳、抽象和举一反三的能力恰恰是目前深度学习方法所不具备的，那么如何发现和模拟人脑语言理解的机理，构建类脑语言理解模型，是摆在我们面前的一个挑战性问题。

（2）构建高质量的基础资源和技术平台

无论是以符号逻辑和规则运算为基础的理性主义方法，还是数据驱动的经验主义方法，高质量的基础资源是不可或缺的根本。这里所说的基础资源包括高质量、大规模知识库，双语对照的平行句对和词典，面向特定任务的标注样本等。尽管知识图谱已经成为目前研究的热点，而且已经构建了若干大规模的知识图谱，但是，尚没有知识图谱表示的规范，对于通用领域而言，知识图谱的规模到底应该多大？知识表示的粒度如何划分？常识如何表示和利用？对于特定领域的具体应用，知识图谱应该如何构建？等等，无数问题摆在我们面前。对于很多语言，尤其是小语种，可利用的数据资源十分可怜，甚至很多语言与汉语对应的双语词典都没有，如波斯语与汉语、乌尔都语与汉语、达利语与汉语等，更别说大规模双语平行语料。高质量的关键技术工具无论对于哪种后续的应用任务，都是不可或缺的，如命名实体识别工具、某些语言的形态分析工具等。

（3）打通不同模态信息处理的壁垒，构建多模态信息融合的处理方法和模型

如前所述，已有的语音、语言、图像和视频处理研究基本上是“井水不犯河水”，各走各的阳关道，而在真实情况下的应用任务中往往需要多模态信息的综合利用，从模拟人脑理解语言过程的角度，各类感知信息的综合利用也是情理之中的事情。

综上所述，目前的人类语言技术已经得到了广泛应用，但其性能水平基本上还是停留在“处理”层面，远没有达到“理解”的水平，未来的任务艰巨而充满挑战。同时，不得不说的是，中文有其独特的规律和热点，无论从哪个角度讲，研究和开发以中文为核心的自然语言处理技术都不应该成为被忽视的盲点。

（参考文献略）