微软周明：语言智能将迎来黄金10年|语智院成立系列

机器之心 汉语堂 6月12日

导言：

今天，AI应用已渗入人类生活的方方面面，北京语言大学顺应潮流，组建语言智能研究院。本周六（6月15日），将在北语逸夫报告厅举行语言智能研究院成立仪式与语言智能发展论坛，大咖云集，讲座缤纷，欢迎大家光临。

机器之心专栏

作者：周明、段楠、韦福如、刘树杰、张冬冬

微软亚洲研究院

> 比尔·盖茨曾说过，「语言理解是人工智能皇冠上的明珠」。自然语言处理（NLP，Natural Language Processing）的进步将会推动人工智能整体进展。
>
> NLP 的历史几乎跟计算机和人工智能（AI）的历史一样长。自计算机诞生，就开始有了对人工智能的研究，而人工智能领域最早的研究就是机器翻译以及自然语言理解。

在 1998 年微软亚洲研究院成立之初，NLP 就被确定为最重要的研究领域之一。历经二十载春华秋实，在历届院长支持下，微软亚洲研究院在促进 NLP 的普及与发展以及人才培养方面取得了非凡的成就。共计发表了 100 余篇 ACL 大会文章，出版了《机器翻译》和《智能问答》两部著作，培养了 500 名实习生、20 名博士和 20 名博士后。我们开发的 NLP 技术琳琅满目，包括输入法、分词、句法/语义分析、文摘、情感分析、问答、跨语言检索、机器翻译、知识图谱、聊天机器人、用户画像和推荐等，已经广泛应用于 Windows、Office、Bing、微软认知服务、小冰、小娜等微软产品中。我们与创新技术组合作研发的微软对联和必应词典，已经为成千上万的用户提供服务。

过去二十年，NLP 利用统计机器学习方法，基于大规模的带标注的数据进行端对端的学习，取得了长足的进步。尤其是过去三年来，深度学习给 NLP 带来了新的进步。其中在单句翻译、抽取式阅读理解、语法检查等任务上，更是达到了可比拟人类的水平。

基于如下的判断，我们认为未来十年是 NLP 发展的黄金档：

1. 来自各个行业的文本大数据将会更好地采集、加工、入库。

2. 来自搜索引擎、客服、商业智能、语音助手、翻译、教育、法律、金融等领域对 NLP 的需求会大幅度上升，对 NLP 质量也提出更高要求。
3. 文本数据和语音、图像数据的多模态融合成为未来机器人的刚需。这些因素都会进一步促进对 NLP 的投资力度，吸引更多人士加入到 NLP 的研发中来。因此我们需要审时度势、抓住重点、及时规划，面向更大的突破。

因此，NLP 研究将会向如下几个方面倾斜：

1. 将知识和常识引入目前基于数据的学习系统中。
2. 低资源的 NLP 任务的学习方法。
3. 上下文建模、多轮语义理解。
4. 基于语义分析、知识和常识的可解释 NLP。

重点知识：NLP 的技术进展

自然语言处理，有时候也称作自然语言理解，旨在利用计算机分析自然语言语句和文本，抽取重要信息，进行检索、问答、自动翻译和文本生成。人工智能的目的是使得电脑能听、会说、理解语言、会思考、解决问题，甚至会创造。它包括运算智能、感知智能、认知智能和创造智能几个层次的技术。计算机在运算智能即记忆和计算的能力方面已远超人类。而感知智能则是电脑感知环境的能力，包括听觉、视觉和触觉等等，相当于人类的耳朵、眼睛和手。目前感知智能技术已取得飞跃性的进步；而认知智能包括自然语言理解、知识和推理，目前还待深入研究；创造智能目前尚无多少研究。比尔·盖茨曾说过，「自然语言理解是人工智能皇冠上的明珠」。NLP 的进步将会推动人工智能整体进展。

NLP 在深度学习的推动下，在很多领域都取得了很大进步。下面，我们就来一起简单看看 NLP 的重要技术进展。

神经机器翻译

神经机器翻译就是模拟人脑的翻译过程。

翻译任务就是把源语言句子转换成语义相同的目标语言句子。人脑在进行翻译的时候，首先是尝试理解这句话，然后在脑海里形成对这句话的语义表示，最后再把这个语义表示转化到另一种语言。神经机器翻译就是模拟人脑的翻译过程，它包含了两个模块：一个是编码器，负责将源语言句子压缩为语

义空间中的一个向量表示，期望该向量包含源语言句子的主要语义信息；另一个是解码器，它基于编码器提供的语义向量，生成在语义上等价的目标语言句子。

神经机器翻译模型的优势在于三方面：一是端到端的训练，不再像统计机器翻译方法那样由多个子模型叠加而成，从而造成错误的传播；二是采用分布式的信息表示，能够自动学习多维度的翻译知识，避免人工特征的片面性；三是能够充分利用全局上下文信息来完成翻译，不再是局限于局部的短语信息。基于循环神经网络模型的机器翻译模型已经成为一种重要的基线系统，在此方法的基础上，从网络模型结构到模型训练方法等方面，都涌现出很多改进。

神经机器翻译系统的翻译质量在不断取得进步，人们一直在探索如何使得机器翻译达到人类的翻译水平。2018 年，微软亚洲研究院与微软翻译产品团队合作开发的中英机器翻译系统，在 WMT2017 新闻领域测试数据集上的翻译质量达到了与人类专业翻译质量相媲美的水平 (Hassan et al., 2018)。该系统融合了微软亚洲研究院提出的四种先进技术，其中包括可以高效利用大规模单语数据的联合训练和对偶学习技术，以及解决曝光偏差问题的一致性正则化技术和推敲网络技术。

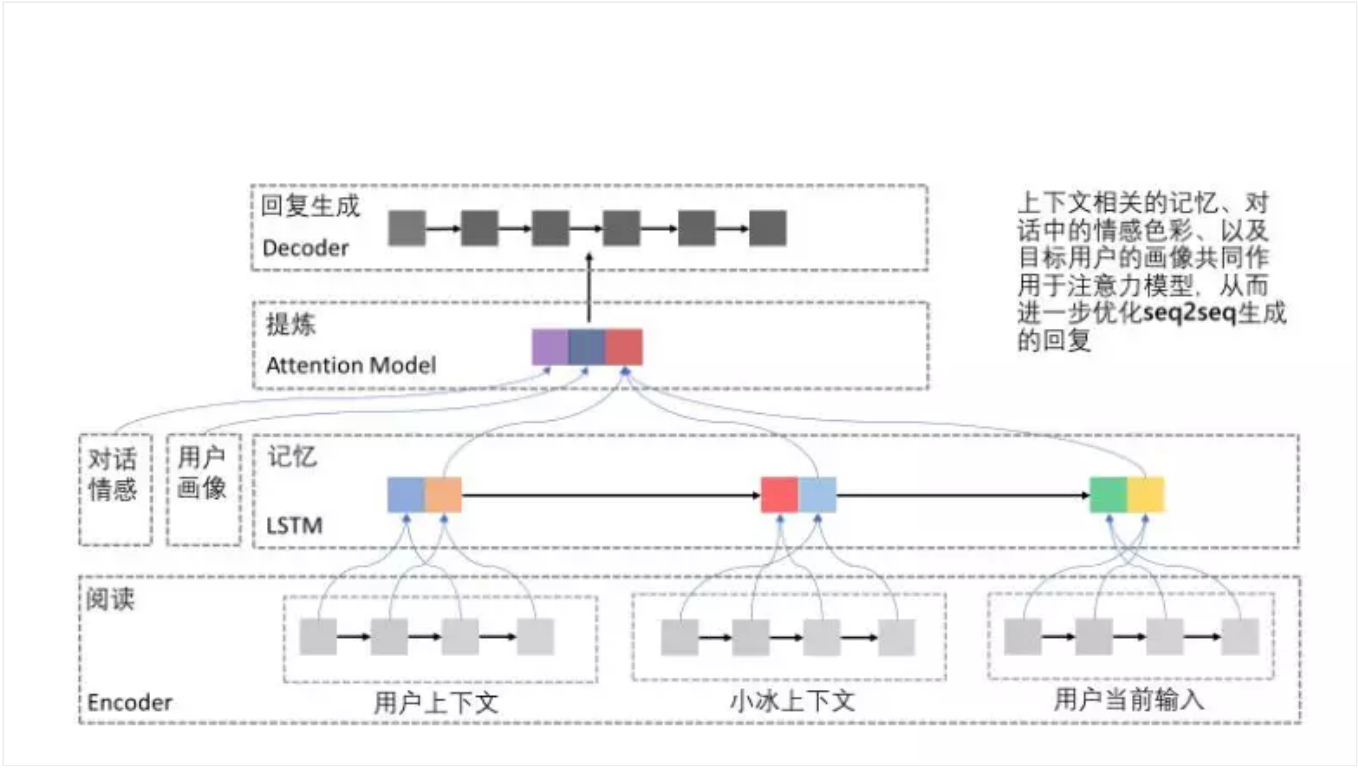
智能人机交互

智能人机交互包括利用自然语言实现人与机器的自然交流。其中一个重要的概念是「对话即平台」。

「对话即平台（CaaP，Conversation as a Platform）是微软首席执行官萨提亚·纳德拉 2016 年提出的概念，他认为图形界面的下一代就是对话，并会给整个人工智能、计算机设备带来一场新的革命。

萨提亚之所以提出这个概念是因为：首先，源于大家都已经习惯用社交手段，如微信、Facebook 与他人聊天的过程。我们希望将这种交流过程呈现在当今的人机交互中。其次，大家现在面对的设备有的屏幕很小（比如手机），有的甚至没有屏幕（比如有些物联网设备），语音交互更加自然和直观。对话式人机交互可调用 Bot 来完成一些具体的功能，比如订咖啡，买车票等等。许多公司开放了 CAAP 平台，让全世界的开发者都能开发出自己喜欢的 Bot 以便形成一个生态。

面向任务的对话系统比如微软的小娜通过手机和智能设备让人与电脑进行交流，由人发布命令，小娜理解并完成任务。同时，小娜理解你的习惯，可主动给你一些贴心提示。而聊天机器人，比如微软的小冰负责聊天。无论是小娜这种注重任务执行的技术，还是小冰这种聊天系统，其实背后单元处理引擎无外乎三层技术：第一层，通用聊天机器人；第二层，搜索和问答（Infobot）；第三层，面向特定任务对话系统（Bot）。



聊天系统的架构

机器阅读理解

自然语言理解的一个重要研究课题是阅读理解。

阅读理解就是让电脑看一遍文章，针对这些文章问一些问题，看电脑能不能回答出来。机器阅读理解技术有着广阔的应用前景。例如，在搜索引擎中，机器阅读理解技术可以用来为用户的搜索（尤其是问题型的查询）提供更为智能的答案。我们通过对整个互联网的文档进行阅读理解，从而直接为用户提供精确的答案。同时，这在移动场景的个人助理，如微软小娜（Cortana）里也有直接的应用：智能客服中可使用机器阅读文本文档（如用户手册、商品描述等）来自动或辅助客服来回答用户的问题；在办公领域可使用机器阅读理解技术处理个人的邮件或者文档，然后用自然语言查询获取相关的信息；在教育领域用来可以用来辅助出题；在法律领域可用来理解法律条款，辅助律师或者法官判案；在金融领域里从非结构化的文本（比如新闻中）抽取金融相关的信息等。机器阅读理解技术可形成一个通用能力，第三方可以基于它构建更多的应用。

斯坦福大学在 2016 年 7 月发布了一个大规模的用于评测阅读理解技术的数据集（SQuAD），包含 10 万个由人工标注的问题和答案。SQuAD 数据集中，文章片段（passage）来自维基百科的文章，每个文章片段（passage）由众包方式，标注人员提 5 个问题，并且要求问题的答案是 passage 中的一个子片段。标注的数据被分成训练集和测试集。训练集公开发布用来训练阅读理解系统，而测试集不公开。参赛者需要把开发的算法和模型提交到斯坦福由其运行后把结果报在网站上。

一开始，以 100 分为例，人的水平是 82.3 左右，机器的水平只有 74 分，机器相差甚远。后来通过不断改进，机器阅读理解性能得以逐步地提高。2018 年 1 月，微软亚洲研究院提交的 R-Net 系统首次在 SQuAD 数据集上以 82.65 的精准匹配的成绩首次超越人类在这一指标上的成绩。随后阿里巴巴、科大讯飞和哈工大的系统也在这一指标上超越人类水平。标志着阅读理解技术进入了一个新的阶段。最近微软亚洲研究院的 NL-Net 和谷歌的 BERT 系统又先后在模糊匹配指标上突破人类水平。对于阅读理解技术的推动，除了 SQuAD 数据集起到了关键作用之外，还有如下三个方的因素：首先，是端到端的深度神经网络。其次，是预训练的神经网络；最后，是系统和网络结构上的不断创新。

机器创作

机器可以做很多理性的东西，也可以做出一些创造性的东西。

早在 2005 年，微软亚洲研究院在时任院长沈向洋的提议和支持下成功研发了《微软对联》系统。用户出上联，电脑对出下联和横批，语句非常工整。

在此基础上，我们又先后开发了格律诗和猜字谜的智能系统。在字谜游戏里，用户给出谜面，让系统猜出字，或系统给出谜面让用户猜出字。2017 年微软研究院开发了电脑写自由体诗系统、作词谱曲系统。中央电视台《机智过人》节目就曾播放过微软的电脑作词谱曲与人类选手进行词曲创作比拼的内容。这件事说明如果有大数据，那么深度学习就可以模拟人类的创造智能，也可以帮助专家产生更好的想法。

就作词来说，写一首歌词首先要决定主题。比如想写一首与「秋」，「岁月」，「沧桑」，「感叹」相关的歌，利用词向量表示技术，可知「秋风」、「流年」、「岁月」、「变迁」等词语比较相关，通过扩展主题可以约束生成的结果偏向人们想要的歌词，接着在主题模型的约束下用序列到序列的神经网络，用歌词的上一句去生成下一句，如果是第一句，则用一个特殊的序列作为输入去生成第一句歌词，这样循环生成歌词的每一句。

下面也简介一下谱曲。为一首词谱曲不单要考虑旋律是否好听，也要考虑曲与词是否对应。这类似于一个翻译过程。不过这个翻译中的对应关系比自然语言翻译更为严格。它需严格规定每一个音符对应到歌词中的每一个字。例如每一句有 N 个字，那么就需要将这句话对应的曲切分成 N 个部分，然后顺序完成对应关系。这样在「翻译」过程中要「翻译」出合理的曲谱，还要给出曲与词之间的对应关系。我们利用了一个改进的序列到序列的神经网络模型，完成从歌词「翻译」到曲谱的生成过程。

趋势热点：值得关注的 NLP 技术

从最近的 NLP 研究中，我们认为有一些技术发展趋势值得关注，这里总结了五个方面：

热点一，预训练神经网络

如何学习更好的预训练的表示，在一段时间内继续成为研究的热点。

通过类似于语言模型的方式来学习词的表示，其用于具体任务的范式得到了广泛应用。这几乎成为自然语言处理的标配。这个范式的一个不足是词表示缺少上下文，对上下文进行建模依然完全依赖于有限的标注数据进行学习。实际上，基于深度神经网络的语言模型已经对文本序列进行了学习。如果把语言模型关于历史的那部分参数也拿出来应用，那么就能得到一个预训练的上下文相关的表示。这就是 Matthew Peters 等人在 2018 年 NAACL 上的论文「Deep Contextualized Word Representations」的工作，他们在大量文本上训练了一个基于 LSTM 的语言模型。最近 Jacob Delvin 等人又取得了新的进展，他们基于多层 Transformer 机制，利用所谓「MASKED」模型预测句子中被掩盖的词的损失函数和预测下一个句子的损失函数所预训练得到的模型「BERT」，在多个自然语言处理任务上取得了当前最好的水平。以上提到的所有的预训练的模型，在应用到具体任务时，先用这个语言模型的 LSTM 对输入文本得到一个上下文相关的表示，然后再基于这个表示进行具体任务相关的建模学习。结果表明，这种方法在语法分析、阅读理解、文本分类等任务都取得了显著的提升。最近一段时间，这种预训练模型的研究成为了一个研究热点。

如何学习更好的预训练的表示在一段时间内将继续成为研究的热点。在什么粒度（word，sub-word，character）上进行预训练，用什么结构的语言模型（LSTM，Transformer 等）训练，在什么样的数据上（不同体裁的文本）进行训练，以及如何将预训练的模型应用到具体任务，都是需要继续研究的问题。现在的预训练大都基于语言模型，这样的预训练模型最适合序列标注的任务，对于问答一类任务依赖于问题和答案两个序列的匹配的任务，需要探索是否有更好的预训练模型的数据和方法。将来很可能会出现多种不同结构、基于不同数据训练得到的预训练模型。针对一个具体任务，如何快速找到合适的预训练模型，自动选择最优的应用方法，也是一个可能的研究课题。

热点二，迁移学习和多任务学习

对于那些本身缺乏充足训练数据的自然语言处理任务，迁移学习有着非常重要和实际的意义。多任务学习则用于保证模型能够学到不同任务间共享的知识和信息。

不同的 NLP 任务虽然采用各自不同类型的数据进行模型训练，但在编码器（Encoder）端往往是同构的。例如，给定一个自然语言句子 who is the Microsoft founder，机器翻译模型、复述模型和问答模型都会将其转化为对应的向量表示序列，然后再使用各自的解码器完成后续翻译、改写和答案生成 (或检索) 任务。因此，可以将不同任务训练得到的编码器看作是不同任务对应的一种向量表示，并通过迁移学习（Transfer Learning）的方式将这类信息迁移到目前关注的目标任务上来。对于那些本身缺乏充足训练数据的自然语言处理任务，迁移学习有着非常重要和实际的意义。

多任务学习（Multi-task Learning）可通过端到端的方式，直接在主任务中引入其他辅助任务的监督信息，用于保证模型能够学到不同任务间共享的知识和信息。Collobert 和 Weston 早在 2008 年就最早提出了使用多任务学习在深度学习框架下处理 NLP 任务的模型。最近 Salesforce 的 McCann 等提出了利用问答框架使用多任务学习训练十项自然语言任务。每项任务的训练数据虽然有限，但是多个任务共享一个网络结构，提升对来自不同任务的训练数据的综合利用能力。多任务学习可以设计为对诸任务可共建和共享网络的核心层次，而在输出层对不同任务设计特定的网络结构。

热点三，知识和常识的引入

如何在自然语言理解模块中更好地使用知识和常识，已经成为目前自然语言处理领域中一个重要的研究课题。

随着人们对人机交互（例如智能问答和多轮对话）要求的不断提高，如何在自然语言理解模块中更好地使用领域知识，已经成为目前自然语言处理领域中一个重要的研究课题。这是由于人机交互系统通常需要具备相关的领域知识，才能更加准确地完成用户查询理解、对话管理和回复生成等任务。

最常见的领域知识包括维基百科和知识图谱两大类。机器阅读理解是基于维基百科进行自然语言理解的一个典型任务。给定一段维基百科文本和一个自然语言问题，机器阅读理解任务的目的是从该文本中找到输入问题对应的答案短语片段。语义分析是基于知识图谱进行自然语言理解的另一个典型任务。给定一个知识图谱（例如 Freebase）和一个自然语言问题，语义分析任务的目的是将该问题转化为机器能够理解和执行的语义表示。目前，机器阅读理解和语义分析可以说是最热门的自然语言理解任务，它们受到了来自全世界研究者的广泛关注和深入探索。

常识指绝大多数人都了解并接受的客观事实，例如海水是咸的、人渴了就想喝水、白糖是甜的等。常识对机器深入理解自然语言非常重要，在很多情况下，只有具备了一定程度的常识，机器才有可能对字面上的含义做出更深一层次的理解。然而获取常识却是一个巨大的挑战，一旦有所突破将是影响人工智能进程的大事情。另外，在 NLP 系统中如何应用常识尚无深入的研究，不过出现了一些值得关注的工作。

热点四，低资源的 NLP 任务

引入领域知识（词典、规则）可以增强数据能力、基于主动学习的方法增加更多的人工标注数据等，以解决数据资源贫乏的问题。

面对标注数据资源贫乏的问题，譬如小语种的机器翻译、特定领域对话系统、客服系统、多轮问答系统等，NLP 尚无良策。这类问题统称为低资源的 NLP 问题。对这类问题，除了设法引入领域知识（词典、规则）以增强数据能力之外，还可以基于主动学习的方法来增加更多的人工标注数据，以及采用无监督和半监督的方法来利用未标注数据，或者采用多任务学习的方法来使用其他任务甚至其他语言的信息，还可以使用迁移学习的方法来利用其他的模型。

以机器翻译为例，对于稀缺资源的小语种翻译任务，在没有常规双语训练数据的情况下，首先通过一个小规模的双语词典（例如仅包含 2000 左右的词对），使用跨语言词向量的方法将源语言和目标语言词映射到同一个隐含空间。在该隐含空间中，意义相近的源语言和目标语言词具有相近的词向量表示。基于该语义空间中词向量的相似程度构建词到词的翻译概率表，并结合语言模型，便可以构建基于词的机器翻译模型。使用基于词的翻译模型将源语言和目标语言单语语料进行翻译，构建出伪双语数据。于是，数据稀缺的问题通过无监督的学习方法产生伪标注数据，就转化成了一个有监督的学习问题。接下来，利用伪双语数据训练源语言到目标语言以及目标语言到源语言的翻译模型，随后再使用联合训练的方法结合源语言和目标语言的单语数据，可以进一步提高两个翻译系统的质量。

为了提高小语种语言的翻译质量，我们提出了利用通用语言之间大规模的双语数据，来联合训练四个翻译模型的期望最大化训练方法（Ren et al., 2018）。该方法将小语种（例如希伯来语）作为有着丰富语料的语种（例如中文）和（例如英语）之间的一个隐含状态，并使用通用的期望最大化训练方法来迭代地更新 X 到 Z、Z 到 X、Y 到 Z 和 Z 到 Y 之间的四个翻译模型，直至收敛。

热点五，多模态学习

视觉问答作为一种典型的多模态学习任务，在近年来受到计算机视觉和自然语言处理两个领域研究人员的重点关注。

婴儿在掌握语言功能前，首先通过视觉、听觉和触觉等感官去认识并了解外部世界。可见，语言并不是人类在幼年时期与外界进行沟通的首要手段。因此，构建通用人工智能也应该充分地考虑自然语言和其他模态之间的互动，并从中进行学习，这就是多模态学习。

视觉问答作为一种典型的多模态学习任务，在近年来受到计算机视觉和自然语言处理两个领域研究人员的重点关注。给定一张图片和用户提出的一个自然语言问题，视觉问答系统需要在理解图片和自然语言问题的基础上，进一步输入该问题对应的答案，这需要视觉问答方法在建模中能够对图像和语言之间的信息进行充分地理解和交互。

我们在今年的 CVPR 和 KDD 大会上分别提出了基于问题生成的视觉问答方法（Li et al., 2018）以及基于场景图生成的视觉问答方法（Lu et al., 2018），这两种方法均在视觉问答任务上取得了非常好的结果，实现了 state-of-the-art 的效果。除视觉问答外，视频问答是另一种最近广受关注的多模态任务。该任务除了包括带有时序的视频信息外，还包括了音频信息。目前，视频问答作为一种新型的问答功能，已经出现在搜索引擎的场景中。可以预见，该任务在接下来一定还会受到更多的关注。

未来展望：理想的 NLP 框架和发展前景

我们认为，未来理想状态下的 NLP 系统架构可能是如下一个通用的自然语言处理框架：

首先，对给定自然语言输入进行基本处理，包括分词、词性标注、依存分析、命名实体识别、意图/关系分类等。

其次，使用编码器对输入进行编码将其转化为对应的语义表示。在这个过程中，一方面使用预训练好的词嵌入和实体嵌入对输入中的单词和实体名称进行信息扩充，另一方面，可使用预训练好的多个任务编码器对输入句子进行编码并通过迁移学习对不同编码进行融合。

接下来，基于编码器输出的语义表示，使用任务相关的解码器生成对应的输出。还可引入多任务学习将其他相关任务作为辅助任务引入到对主任务的模型训练中来。如果需要多轮建模，则需要在数据库中记录当前轮的输出结果的重要信息，并应用于在后续的理解和推理中。

显然，为了实现这个理想的 NLP 框架需要做很多工作：

- 需要构建大规模常识数据库并且清晰通过有意义的评测推动相关研究；
- 研究更加有效的词、短语、句子的编码方式，以及构建更加强大的预训练的神经网络模型；
- 推进无监督学习和半监督学习，需要考虑利用少量人类知识加强学习能力以及构建跨语言的 embedding 的新方法；
- 需要更加有效地体现多任务学习和迁移学习在 NLP 任务中的效能，提升强化学习在 NLP 任务的作用，比如在自动客服的多轮对话中的应用；

- 有效的篇章级建模或者多轮会话建模和多轮语义分析；
- 要在系统设计中考虑用户的因素，实现用户建模和个性化的输出；
- 构建综合利用推理系统、任务求解和对话系统，基于领域知识和常识知识的新一代的专家系统；
- 利用语义分析和知识系统提升 NLP 系统的可解释能力。

未来十年，NLP 将会进入爆发式的发展阶段。从 NLP 基础技术到核心技术，再到 NLP+的应用，都会取得巨大的进步。比尔盖茨曾经说过人们总是高估在一年或者两年中能够做到的事情，而低估十年中能够做到的事情。

我们不妨进一步想象十年之后 NLP 的进步会给人类生活带来哪些改变？

- 十年后，机器翻译系统可以对上下文建模，具备新词处理能力。那时候的讲座、开会都可以用语音进行自动翻译。除了机器翻译普及，其他技术的进步也令人耳目一新。家里的老人和小孩可以跟机器人聊天解闷。
- 机器个人助理能够理解你的自然语言指令，完成点餐、送花、购物等下单任务。你已习惯于客服机器人来回答你的关于产品维修的问题。
- 你登临泰山发思古之幽情，或每逢佳节倍思亲，拿出手机说出感想或者上传一幅照片，一首情景交融、图文并茂的诗歌便跃然于手机屏幕上，并且可以选择格律诗词或者自由体的表示形式，亦可配上曲谱，发出大作引来点赞。
- 可能你每天看到的体育新闻、财经新闻报道是机器人写的。
- 你用手机跟机器人老师学英语，老师教你口语，纠正发音，跟你亲切对话，帮你修改论文。
- 机器人定期自动分析浩如烟海的文献，给企业提供分析报表、辅助决策并做出预测。搜索引擎的智能程度大幅度提高。很多情况下，可以直接给出答案，并且可以自动生成细致的报告。
- 利用推荐系统，你关心的新闻、书籍、课程、会议、论文、商品等可直接推送给你。
- 机器人帮助律师找出判据，挖掘相似案例，寻找合同疏漏，撰写法律报告。
- ……

未来，NLP 将跟其他人工智能技术一道深刻地改变人类的生活。当然前途光明、道路曲折是亘古不变的道理，为了实现这个美好的未来，我们需要大胆创新、严谨求实、扎实进取。讲求研究和应用并举，普及与提高同步。我们期待着与业界同仁一道努力，共同走进 NLP 下一个辉煌的十年。

✂----------