自然语言处理应用

自然语言处理(Natural Language Processing, NLP)是通过理解人类语言来解决实际问题的一门学科。自然语言处理不仅是学术界的研究热点，在工业界也有许多成果，如谷歌的文本搜索引擎、苹果的Siri、微软小冰等。

自然语言处理领域仍有许多充满挑战、亟待解决的问题。对自然语言处理问题的研究可以追溯到二十世纪三十年代，早期的处理方法往往是人工设计的规则；从二十世纪八十年代开始，利用概率与统计理论并使用数据驱动的方法才逐渐兴盛起来。近几年，随着计算机算力的提升与深度学习技术的发展，自然语言处理相关问题也迎来许多重大的创新与突破。

自然语言处理可以分为核心任务和应用两部分，核心任务代表在自然语言各个应用方向上需要解决的共同问题，包括语言模型、语言形态学、语法分析、语义分析等，而应用部分则更关注自然语言处理中的具体任务，如机器翻译、信息检索、问答系统、对话系统等。今天将介绍机器翻译、问答系统与对话系统三个应用。

机器翻译

长久以来，追求“信、达、雅”的卓越翻译官们慢慢消除着不同语言在人类信息交流和传播中带来的壁垒。随着现代科技的发展、计算机的出现，机器自动翻译逐渐走上历史的舞台。其实早在1933年，机器翻译（Machine Translation）的概念就被提出。随着双语平行语料（即同时包含源语言和与其互为译文的目标语言文本的语料）的增多，通过对语料的统计学习来进行自动翻译的统计机器翻译（Statistical Machine Translation，SMT）成为主流，但翻译的准确性和流畅性仍然和人工翻译有巨大的差距。直到深度学习兴起，神经网络翻译（Neural Machine Translation，NMT）的诞生为机器翻译领域带来了新的机遇，翻译质量也有了质的飞跃。目前Google、百度等公司都已经将线上机器翻译系统升级到神经网络翻译模型，每天为数亿用户提供服务。

神经网络翻译模型始于Google在2014年提出的基于LSTM的编码器–解码器架构。相比于主流的统计机器翻译模型，神经网络翻译模型则采用端到端的学习形式，这样可以将翻译的质量直接作为模型优化的最终目标，以对模型进行整体优化。

神经网络模型同样需要使用平行语料库作为训练数据，但和统计机器翻译模型将任务拆解成多个模块不同，神经网络模型通常是一个整体的序列到序列模型。以常见的循环神经网络为例，神经网络对翻译过程进行建模，如图1所示。通常会先使用一个循环神经网络作为编码器，将输入序列（源语言句子的词序列）编码成一个向量表示，然后再使用一个循环神经网络作为解码器，从编码器得到的向量表示里解码得到输出序列（目标语言句子的词序列）。

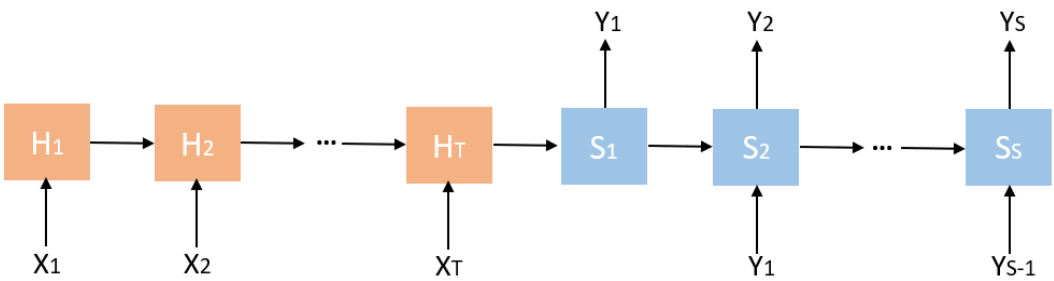

图1 循环神经网络

双语语料是机器翻译模型训练时最重要的监督信息，然而在现实应用中由于某些语言是小语种或者特定领域的语料稀缺等，经常出现双语语料不足的情况，在训练神经网络翻译模型的时候如何应对这种情况呢？这里列举几种常见的解决方案。

第一类非常直接的解决方案就是通过爬虫自动挖掘和产生更多的双语语料。

第二类比较直观的解决方案是构造伪双语平行语料。常见的构造方式有两种：一是利用目标语言端的单语语料反向翻译源语言，由于这样构造的平行语料中目标语言端为真实语料，因此有利于解码器网络的学习，提升模型的效果；二是利用数据增强的方式对原始语料进行改造。

问答系统

问答系统（Question Answering System，QA）通常是指可以根据用户的问题（question）,从一个知识库或者非结构化的自然语言文档集合中查询并返回答案（answer）的计算机软件系统。与搜索引擎系统不同，问答系统不仅能用自然语言句子提问，还能为用户直接返回所需的答案，而不是相关的网页。显然，问答系统能更好的表达用户的信息需求，同时也能更有效地满足用户的信息需求。目前已经有一些公司将问答系统的技术应用到了自己的产品之中，Google、Bing等搜索引擎就提供了根据用户的查询直接从网页结果中抽取相关答案的功能。

目前为止，问答系统没有一个明确的定义，但是一般认为问答系统的输入应该是自然语言形式的问题，输出应该是一个简洁的答案或者可能的答案列表，而不是一堆相关的文档。例如用户向问答系统提交一个问题，“电话是什么时候发明的？”，系统应该返回一个精简的答案--“1876”。

一个典型的问答系统需要完成问题分类、段落检索以及答案抽取3个任务。

a. 问题分类主要用于决定答案的类型。比如，“珠穆朗玛峰海拔有多高”这样的问题，需要根据事实给出答案；而“美国知名的互联网流媒体公司有哪些”这样的问题，则需要根据问题中的条件，返回一个符合要求的结果列表。不同的答案类型也往往意味着在系统实现上对应着不同的处理逻辑。根据期望回答方式可以将问题分类成事实型问题、列举型问题、带有假设条件的问题、询问“某某事情如何做”以及“某某东西是什么”等问题。
b. 段落检索是指根据用户的问题，在知识库以及备选段落集合中返回一个较小候选集，这是一个粗略筛选的过程。这样做的原因是知识库以及候选段落集合往往包含海量数据，以至于无法直接在这些数据上进行答案抽取，需要应用一些相对轻量级的算法筛选出一部分候选集，使得后续的答案抽取阶段可以应用一些更为复杂的算法。
c. 答案抽取是指根据用户的问题，在段落候选集的文本中抽取最终答案的过程。目前很多深度学习方法可以用来解决这个问题。

对话系统

对话系统（Dialogue System）是指可以通过文本、语音、图像等自然的沟通方式自动地与人类交流的计算机系统。对话系统有相对较长的发展历史，早期的对话系统可以追溯到二十世纪六十年代麻省理工学院人工智能实验室设计的自然语言处理程序ELIZA。经过几十年的研究发展以及数据量的增加，逐渐诞生了像苹果公司的Siri、微软的小娜（Cortana）等个人助理型的对话系统产品，以及微软小冰这样的非任务型对话系统。对话系统根据信息领域的不同（开放与闭合）以及设计目标的不同（任务型与非任务型）可以划为不同的类型：任务型对话系统需要根据用户的需求完成相应的任务，如发邮件、打电话、行程预约等；非任务型对话系统大多是根据人类的日常聊天行为而设计，对话没有明确的任务目标，只是为了与用户更好地进行沟通，例如微软小冰的设计目标之一是培养对话系统的共情能力（Empathy），更注重与用户建立长期的情感联系。

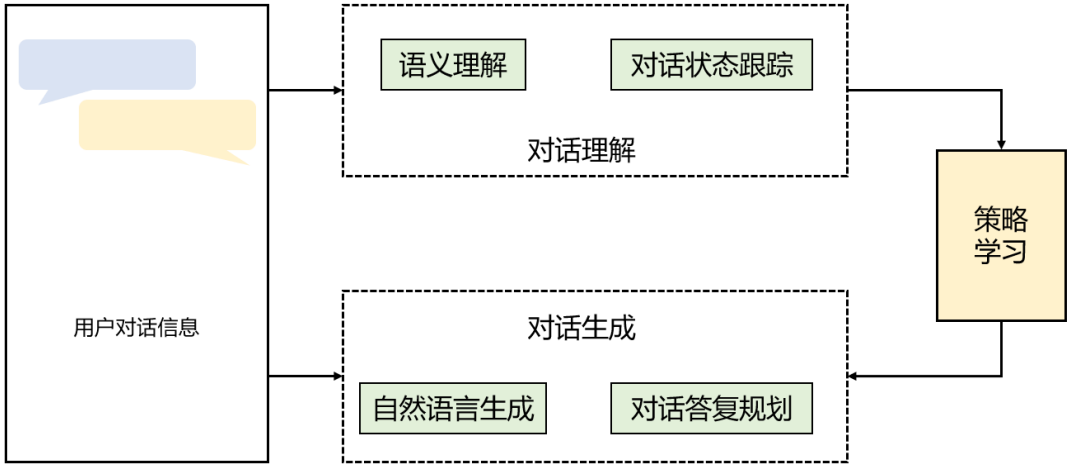

图2 任务型对话系统的结构示意图

一个典型的任务型对话系统包含图2所示的3个部分：对话理解、策略学习和对话生成。具体来说，对于用户的输入，先通过语义理解（Natural Language Understanding，NLU）单元进行编码，通过对话状态跟踪模块生成当前对话状态编码；根据当前的对话状态，系统选择需要执行的任务（由策略学习模块决定）；最后通过自然语言生成（Natural Language Generation，NLG）返回用户可以理解的表达形式（如文本、语音、图片等）。由于任务型对话系统需要完成一些特定任务，因此处理的信息领域往往是闭合的（close domain）。 对于非任务型的对话系统来说，其更注重与用户的沟通，对话的多样性以及用户的参与度比较重要，因此这类对话系统更多采用一些生成式模型（如Seq2Seq模型），或者根据当前内容从语料库中选择合适的问答语句。这类问答系统对应的信息领域往往是开放的（open domain）。

对于对话系统来说，用户的输入往往多种多样，对于不同领域的对话内容，对话系统可以采取的行为也多种多样。普通的有监督学习方法（如深度神经网络）往往无法获得充足的训练样本进行学习，而强化学习可在一定程度上解决这个问题。强化学习是深度学习领域比较热门的研究方向之一。强化学习尝试根据环境决策不同的行为（action），从而实现预期利益的最大化。当对话系统与用户的交互行为持续地从客户端传输到服务端时，强化学习方法可以对模型进行及时的更新，在线训练模型。

对于任务型对话系统，系统根据对用户的理解，采取不同的行为，这个过程可以用图2中的策略学习模块表示。由于用户的对话以及系统可采取行为的组合数量一般比较庞大，这个部分比较适合使用强化学习来解决。对于非任务型对话系统，如在微软小冰的设计中，也有类似的对话管理模块。强化学习除了可以用来为策略学习模块建模之外，还可以直接为整个对话系统进行端到端的建模，从而简化对话系统的设计。

参考文献

[1] Locke W N , Booth A D . Machine translation[J]. Journal of the lee, 1956, 2(2):109-116.
[2] Li S , Zhang J , Huang X , et al. Semantic computation in a Chinese Question-Answering system[J]. Journal of Computer Science and Technology, 2002, 17(6):933-939.
[3] Yamaguchi T , Enomoto T , Sekiyama H , et al. Dialogue system[J]. 2010.