聊天机器人的技术及展望

武 威 周 明
微软亚洲研究院

关键词：聊天机器人 检索式 生成式

婚姻破碎的作家西奥多认识了一个叫萨曼莎的姑娘。随着交流的深入，西奥多逐渐被萨曼莎所吸引，性感的嗓音、风趣的谈吐，以及善解人意……西奥多很快就爱上了她。然而，萨曼莎却不能满足西奥多的求婚要求。“放我走吧，尽管我很想留下来”，萨曼莎说。原来她只是一个机器人，是一个人工智能操作系统。这是电影《她》的内容。

《她》上映于2013年，故事的场景设定在2025年。也许当时人机自然交互仍然属于科幻。然而就在2014年夏天，一款名为“微软小冰”的聊天机器人出现了，人们惊呼，原来电影中的场景已经快触手可及了。

聊天机器人的兴起

聊天机器人是一种人工智能交互系统，其工作方式是通过语音或文字实现人机在任意开放话题上的交流。在很长一段时间内，人们建立聊天机器人的目的是模拟人类的对话行为，从而检测人工智能程序是否能够理解人类语言并且和人类进行自然交流。一个检测聊天机器人水平的重要标准就是是否能够通过图灵测试[1]。图灵测试是计算机科学的先驱阿兰·图灵(Alan Turing)于1950年提出的关于测试人工智能能力的一个标准：测试人员通过对话的方式分辨人和机器人；如果测试人员无法分辨收到的回复信息是来自于人还是机器人，则认为人工智能系统通过图灵测试。

聊天机器人的历史可以追溯到20世纪60年代。那时麻省理工学院的人工智能实验室利用大量的规则建立了一个名为ELIZA[2]的聊天机器人。很多人认为ELIZA不能通过图灵测试，因为规则系统并不能支持真正的开放域对话。尽管ELIZA在当时产生了很大的影响，但它并没有推动聊天机器人研究的进一步发展。就连ELIZA的创始人都说“机器可以表现得十分惊艳，可以蒙蔽最有经验的观察者。但如果你揭开程序的面纱，向人解释其内在原理，那么一切的魔法都荡然无存，那不过是一些工序罢了……”。正因如此，在很长一段时间内，人们认为聊天机器人除了可以证明计算机程序的“聪明”程度外并没有什么价值，聊天机器人的相关产业没有得到发展，而ELIZA也永远停留在了实验室里。此后，仍有一些基于规则系统的聊天机器人出现，比如1972年的PARRY[3]和1995年的Alicebot[4]。1997年，一款名为Jabberwacky[5]的聊天机器人出现在互联网上，其目的仍然是想通过图灵测试，但其设计原理却是通过与人的交流来让机器学习对话。相比于ELIZA，这是一个很大的进步。如今，人们仍然可以和这个机器人的升级版本Cleverbot[6]在网上交流。尽管如此，聊天机器人仍然没有真正地“火”起来。

微软小冰的出现，改变了这一局面。不同于传统的聊天机器人，小冰的定位是“情感陪护”，即通过在不断的对话中感知用户情绪的变化，实现和用户的长程交流，为用户提供“陪聊”服务。尽管有些时候，小冰也不能给出合理的回复，但多样化的回复，逗趣的语言风格，以及时不时送上的“心灵鸡汤”，却让

很多人乐此不疲地与小冰进行交流。因此，微软小冰推出短短数月，就在微博上获得了上百万的粉丝。以往的聊天机器人只能和人进行两轮左右的自然对话，而小冰可完成平均23轮的对话。截至2016年8月，小冰已经和人类产生了超过200亿次的对话。

另一方面，小冰摒弃了传统的规则系统，充分利用了存在于互联网社交媒体中的大量用户交互数据，通过数据驱动加机器学习的方式实现了人机交流。这使得小冰可以被迅速地复制到各个垂直领域以及其他语言市场。同时，这种方式实现的聊天机器人也给了已经“荷枪实弹”的大数据科学家和机器学习科学家一个可以开足马力发挥的全新战场。据不完全统计，2016年一年，出现在人工智能和自然语言顶级会议的关于聊天机器人的研究论文有13篇，而在2017年的前4个月，就已经达到了11篇，这还不包括在arXiv[1]上不断涌现的文章。可以说，小冰的出现推动了聊天机器人及其相关领域的研究，预计还将持续火热。

数据驱动的方式和市场的火爆反响也让投资人看到了聊天机器人作为通用对话平台的可能性。事实上，基于固定领域的任务导向对话系统研究以及问答系统的研究都有非常长的历史，而对话式平台的概念也早在20年前被提出。但当人们试图通过任务完成对话系统和问答系统搭建对话平台时，却发现很难在对话不中断的情况下将不同的服务连在一起，甚至在大多数情况下，单一对话系统也经常被用户的一些“题外话”所中断。而“题外话”的情况之多，根本无法通过规则系统解决。这也是为什么微软小冰之后，百度、腾讯、阿里巴巴、谷歌、Facebook等科技巨头纷纷推出自己的聊天机器人产品或者开发平台，而聊天机器人的创业公司更是如雨后春笋般涌现。微软小冰的出现带动了整个聊天机器人产业的发展。

聊天机器人技术浅析

从小冰的出现到聊天机器人的火爆已经有三年时间。这三年是聊天机器人相关研究爆发的三年，也是小冰快速成长、“勤奋学习”的三年。

聊天机器人的研究主要有三部分内容：单轮聊天、多轮聊天以及个性化聊天。单轮聊天研究的是如何针对当前输入信息给出回复。这是聊天机器人研究中最基本的问题，也是构建聊天机器人首先要解决的问题。多轮聊天是研究如何在回复过程中考虑上下文信息。这个问题不仅在聊天机器人中，在任何对话系统的研究中都是本质问题。由于对话上下文每一句都很短，而且不同上下文语境间没有明显的边界，也没有大规模的标注语料，上下文分析，特别是聊天机器人中针对开放域对话的上下文分析，一直是个难点。个性化聊天是要让聊天机器人可以根据用户的喜好以及当前的情绪等给出不同的回复。这是聊天机器人对话中独有的问题，目的是提高用户对于聊天机器人“陪伴”角色的认可度，从而增加用户黏性。

目前聊天机器人的实现方式包括检索式和生成式两类。检索式聊天机器人是利用社交网络中已有的大量对话语料来构建索引，当机器人接收到用户消息后，从索引中查找可能的候选回复，并通过对候选进行排序找到相关回复返回。而生成式聊天机器人则直接从大量的人与人的对话中学习对话模型，然后利用对话模型为收到的用户信息“创作”回复。

检索式聊天机器人

图1[7]给出了检索式聊天机器人的系统架构。研发检索式聊天机器人需要实现线上和线下两部分。线下部分由三个模块组成：索引、匹配模型以及排序模型。这三个模块分别为线上产生回复候选，信息-回复对的特征描述，以及回复候选的排序。索引中收集了大量来自社交网络上人与人的交流数据，组织成“一问一答”结构。索引是检索式聊天机器

[1] arXiv是一个收集物理学、数学、计算机科学与生物学论文预印本的网站。近年来，越来越多的学者选择先在arXiv上发布最新研究成果的论文，保护成果的发明权。

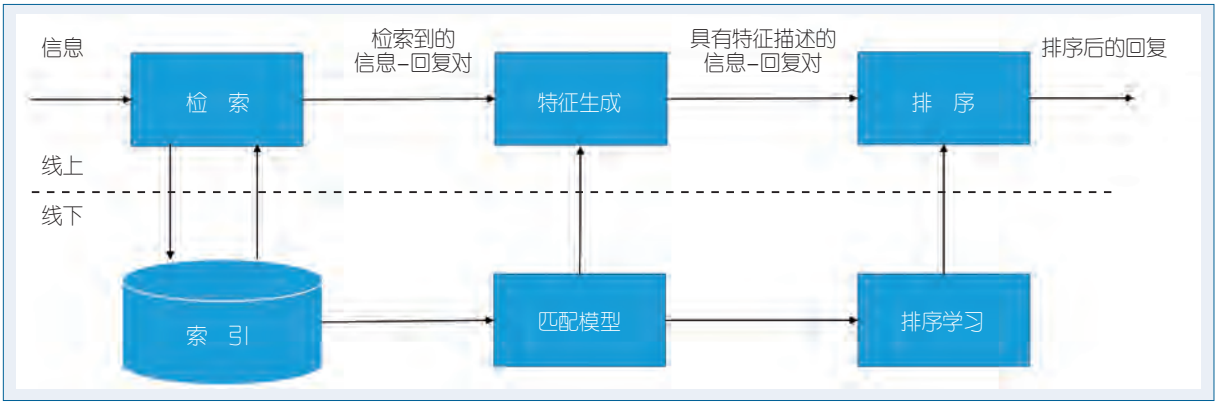

图1 检索式聊天机器人架构

人的基础，其目的是当线上来了一个用户的信息后能够快速从大量“问答对”中获得可能的回复候选。匹配模型是检索式聊天机器人的关键，其作用是实现对用户信息和回复候选的语义理解，对二者语义上构成回复关系的可能性进行打分。这些打分在线上构成了每个信息 - 回复对的特征，而这些特征最终由一个排序模型进行整合，产生最终的候选排序。机器人最后的回复从排在前面的候选中产生。

从图 1 中不难看出，检索式聊天机器人很大程度上借鉴并沿用了搜索引擎的架构。其不同点在于：(1) 索引数据没有搜索引擎那么复杂的结构，比如数据和数据之间没有链接；(2) 搜索引擎通过查询与文档的匹配实现检索结果的排序，而聊天机器人是对输入信息和回复进行匹配。相比于查询文档，信息和回复更短，语义鸿沟更大，而且由于数据结构简单，除了文字，没有其他信息帮助匹配。因此，如何实现信息回复的匹配是检索式聊天机器人面临的新挑战，也是当前学术界研究的重点。目前的方法是利用神经网络，比如卷积神经网络 [8] 或者循环神经网络 [9]，将信息和回复分别在向量空间中表示，然后通过计算向量的相似度来度量二者的匹配程度。这种方法在实践中有一定的效果，但由于训练神经网络所需的大量数据不易获得，当信息与回复的语义鸿沟很大时，这些模型仍然无法很好地识别二者之间的匹配关系。

不同于搜索引擎的网页搜索，聊天机器人在回复时还要考虑上下文，考虑当前用户的状态（比如情绪），以及聊天机器人自身的设定（比如是男是女），有什么喜好等等。检索式聊天机器人的一大优势在于，上下文信息、用户情绪理解以及自身设定等都可以通过特征的方式加入到排序中。例如，在考虑上下文的时候，一般是把一定长度的上下文通过神经网络变成向量，然后计算这些向量与回复候选的相似度，并将相似度作为特征加入到排序模型中。而在个性化聊天中，回复候选和用户画像的匹配程度也可以作为排序特征，在最终的候选排序中发挥作用。由于排序学习算法和工具在搜索引擎的发展过程中已经非常成熟，检索式聊天机器人可以利用已有的技术简单有效地解决这些问题。

检索式聊天机器人的本质是对已有的人类回复进行筛选重用来回复新的信息。一方面，这是一个优势，因为人类的回复不仅通顺流畅，而且往往还包含了网友的“智慧”，所以检索式聊天机器人只要能够找到与输入信息语义相关并且和上下文逻辑一致的回复，就可以和用户顺畅地进行对话，而且还可以时不时爆出“金句”。另一方面，这也是检索式聊天机器人的局限，回复的好坏很大程度上依赖于索引的质量和是否能够检索到合适的候选。前者很难控制，后者则在很多时候很难实现。尤其在多轮聊天中，要考虑上下文信息，如何检索到能和上下文逻辑一致的回复候选是检索系统无法控制的。而在个性化聊天中，机器人只能被动地从已有的回复候选中挑选和当前用户画像相匹配的回复。

一旦候选中没有合适的回复，后面的排序过程就无法发挥任何作用，机器人也不可能给出合适的回复。正因如此，很多时候聊天机器人的回复不好并非没有做好匹配和排序，而是根本就找不到合适的回复，而这种情况在多轮聊天和个性化聊天中表现得尤为突出。

生成式聊天机器人

也许是意识到检索式聊天机器人存在的问题，也许是不满足于停留在搜索引擎已有的技术上，研究人员又提出了利用自然语言生成技术来实现聊天机器人。事实上，早期的聊天机器人 ELIZA 就是利用生成技术实现的，只是当时的回复生成是基于大量手工制作的模板，并不具有扩展性。而生成技术作为聊天机器人的实现方法流行起来，是源于深度学习技术的发展及其在机器翻译领域的成熟运用。因此，现在说的生成式聊天机器人都是基于深度学习的回复生成。

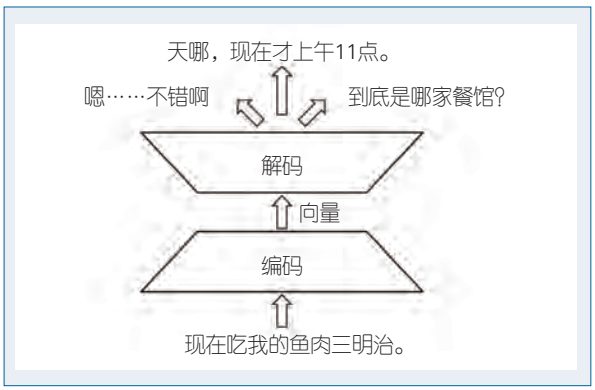

图2　生成式聊天机器人模型

图 2[10] 给出了生成式聊天机器人的基本模型。这个模型借鉴了机器翻译中的编码 - 解码模型，是该模型在回复生成中的应用。其原理比较简单：(1) 在编码阶段，利用一个神经网络（如循环神经网络）将输入信息编码成向量；(2) 在解码阶段，以编码向量为条件的语言模型生成可能的回复。这个语言模型通常也通过循环神经网络实现。一般的循环神经网络无法很好地捕捉到词与词之间长距离的依赖关系，因此，在编码 - 解码模型的实现中一般采用长短期记忆单元 (Long-Short Term Memory Unit)[11] 或者门限循环单元 (Gated Recurrent Unit)[12] 作为循环神经网络的实现。

在实际应用中，这个看似简单的模型被验证可以生成自然流畅的回复语句，而且整个过程完全由数据驱动：模型从大量的训练语料中学习到回复的模板，并存储到循环神经网络的记忆单元中，然后利用学习到的模板为新的输入信息组合出可能的回复。相比于传统的基于手工模板的回复生成，这个模型具有很好的扩展性，可以支持快速的领域迁移以及数据更新。这是生成式对话算法逐渐受到关注的最主要原因。不仅如此，基于编码 - 解码的生成式模型还能够很好地支持端到端的上下文建模以及个性化建模。在上下文建模中 [13]，一般的办法是将支持一句编码的普通编码器扩展成支持多句编码的层次编码器。层次编码器分两层：下层为多个普通编码器，将上下文中每句话变成向量；上层为一个基于句子向量的编码器，通过一个循环神经网络对句子关系进行建模，将各个句子向量编码成一个表示整个上下文的向量。这个向量作为条件，送入普通解码器中生成回复。而个性化建模 [14] 则直接将用户画像表示为向量作为解码器的输入。端到端建模使得生成模型可以自动从数据中学习到和上下文相关或者和用户画像相匹配的回复模板，而不需要多余的人工干预。

尽管生成模型可以产生自然流畅的回复，但是很多回复缺乏信息或者过于普适（比如“是啊”,“我也觉得”之类）[15]。产生这种回复的主要原因是对话的复杂性。在机器翻译中，一个源语言一般只有有限的几种翻译，而在对话中，特别是聊天机器人的开放域对话中，一条输入信息可以有上千种合适的回复。这种过于倾斜的“一对多”的对应关系使得在机器翻译中表现良好的编码 - 解码模型只能捕捉到对话中少数高频模板，从而产生普适回复。普适回复一方面会降低回复的相关性，另一方面也会使得人和机器的聊天很难进行下去。而更严重的问题是，生成式聊天机器人的性能评估需要大规模人工标注，因此，尽管已经有很多研究致力于解决普

适回复的问题，但是到目前为止，学术界对该问题的解决方案并没有一个统一的认识。

生成模型由于其特有的优势，被认为是聊天机器人未来的发展方向，同时也是学术界目前关注的重点。但是在实际应用中，需要克服“普适回复”等困难，因此除了微软小冰，大多数聊天机器人仍然停留在检索模型上。

聊天机器人未来的发展及挑战

数据的爆炸，计算设备的更新换代，以及深度学习方法的快速推进，让以微软小冰为代表的聊天机器人得到了迅猛的发展。如今的微软小冰已经集文字、图像、视频以及语音交互为一体，既支持普通的闲聊，也可以为一些企业公众号提供基于常见问题解答 (FAQ) 的商业解决方案。尽管如此，无论从技术上还是从具体场景上，当前的聊天机器人仍然处于发展早期。

在技术上，深度学习方法确实可以捕捉到输入信息和回复候选之间的一些语义关系，甚至可以合成高质量回复，但是整个过程仍然是一个黑盒子。机器只是记住了一些“模板”，却全然不知这些“模板”是什么意思。换句话说，现在的聊天机器人记忆功能已经非常强大了，但是理解能力还处在初期。而这种只记忆不理解的问题在多轮对话中体现得尤为突出。在多轮对话时，机器人往往抓不住上下文的要点，经常给出一些与上下文无关或者是前后矛盾的回复。因此，聊天机器人的下一个里程碑首先应该是能理解。这种理解，狭义上是指对当前对话的理解，比如谁说了什么，怎么说的，上下文逻辑是什么，有什么意图等等；而广义上则是对整个世界的理解，比如对话中各种实体的理解，概念的内涵、外延的理解，以及能够把握它们之间的联系，具有和一般人相当的对于常识的认识等等。在这一方面，包括微软在内的很多机构已经开始进行研究。比如有人考虑如何将符号知识引入到深度学习模型中以增强模型的理解力。尽管如此，目前的工作还没有取得大的突破，特别是在对聊天上下文的理解和知识理解上。

不同于问答以及任务导向的对话系统，如何评价聊天机器人的好坏一直是一大难题。目前学术界虽然有一些相关研究，但通用做法还是依赖于人工标注。这种方法不但费时费力，而且有很大的主观性，导致不同人的方法由于标注员的不同而不可比较。在工业界，考察的是聊天机器人与人的平均交互轮数（比如微软小冰）。这种做法的最大问题是，轮数的多少无法直接转化成对聊天机制的监督，从而无法帮助人改善现有的聊天模型。目前关于聊天机器人的研究工作有很多，但对于谁好谁坏、有没有突破却很难下结论，缺乏统一评价标准是造成这一局面的主要原因。因此，除了理解，如何评估聊天机器人也是未来的一个突破口。

在具体的应用场景上，聊天机器人除了“陪聊”，目前常见的是作为各种助手和智能硬件背后的交互引擎。尽管市场上已经出现了如京东小冰（导购助手），苹果 Siri（语音助手），以及亚马逊 Echo（智能硬件）等各式各样的助手类产品，人们发现这些产品的体验还远远达不到他们的预期。比如 Echo 对超过 70% 的问题只能回答“不知道”。究其原因是语音系统、聊天系统、问答系统等模块在准确率和召回率上还达不到实用的程度。这些产品可以看作是聊天机器人作为通用对话平台的“试水”。而聊天机器人真正落地，需要的不仅是商业场景上的创新，更是技术上的突破。这里的技术除了前述对话理解以及知识体系，还包括问答能力，语音的识别以及合成能力等等。而只有各方面的技术都达到了人的一般水平，聊天机器人才可能真的被“实用”。

过去的三年是聊天机器人从实验室走向大众生活的三年，而未来的三到五年，也许是聊天机器人走向成熟并服务于人类生活方方面面的几年。未来真正是什么样子，也许我们和小冰一样并不知道，因为它可能会远远超出我们的想象，但它一定是值得我们期待并为之努力的。 ■

武 威

微软亚洲研究院主管研究员。主要研究方向为智能对话、机器学习和信息检索。wuwei@microsoft.com

周 明

CCF高级会员，杰出演讲者，计算机术语审定工委主任，*CCCF*前动态栏目主编。微软亚洲研究院首席研究员。主要研究方向为自然语言处理、机器翻译、文本挖掘、信息检索等。
mingzhou@microsoft.com

参考文献

[1] https://en.wikipedia.org/wiki/Turing_test.

[2] https://en.wikipedia.org/wiki/ELIZA.

[3] https://en.wikipedia.org/wiki/PARRY.

[4] https://en.wikipedia.org/wiki/Artificial_Linguistic_Internet_Computer_Entity.

[5] https://en.wikipedia.org/wiki/Jabberwacky.

[6] https://en.wikipedia.org/wiki/Cleverbot.

[7] Ji Z, Lu Z, Li H. An Information Retrieval Approach to Short Text Conversation[OL]. (2014).arXiv:1408.6988.

[8] Hu B, Lu Z, Li H, et al. Convolutional Neural Network Architectures for Matching Neural Language Sentences[C]//*Advances in Neural Information Processing Systems 27* (NIPS 2014).

[9] Lowe R, Pow N, Serban I, et al. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems[C]//*Proceedings of the SIGDIAL 2015 Conference*.2015:285-294.

[10]Shang L, Lu Z, Li H. Neural Responding Machine for Short Text Conversation[C]//*Proceedings of the 53td Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL Press, 2015: 1577-1586.

[11]https://en.wikipedia.org/wiki/Long_short-term_memory

[12]https://en.wikipedia.org/wiki/Gated_recurrent_unit

[13]Serban I V, Sordoni A, Bengio Y, et al. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models[C]//*Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016: 3776-3783.

[14]Li J, Galley M, Brockett C, et al. A Persona-Based Neural Conversation Model[C]//*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL Press, 2016: 994-1003.

[15]Xing C, Wu Q, Wu Y, et al. Topic Aware Neural Response Generation[C]//*Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence(AAAI-17)*. AAAI Press, 2017: 3351-3357.

39个学生分会线上热议：

如何为学生会员提供“门前”服务？

CCF 目前已有 39 个学会分会。如何发挥 39 个学生分会的地域作用，让学生会员在“门前”享受优质的 CCF 服务？ 2017 年 8 月 28 日，CCF 会员部组织召开了学生分会在线工作会议，旨在通过学生分会主席之间的交流与互动，为 CCF 学生分会更好地服务学生会员提供新思路。会议由 CCF 会员部部长**戴丽霞**主持。

会议邀请 CCF 会员与分部工委主任**罗训**，CCF YOCSEF 主席**苗启广**，CCF 优秀学生分会指导老师**王瑞锦**，分别作了“如何在 CCF 平台上成长”，“如何设计和实施活动”，“担任 CCF 志愿者的收获”的报告。吉林大学分会主席**杜鹏**，电子科技大学分会主席**高强**，合肥工业大学分会主席**谭鑫凯**与大家分享了经验，他们谈到了“如何组织有影响力的活动”，“如何发展学生会员”，“如何留住学生会员”等问题。各个学生分会主席都积极投入其中，踊跃发言，讨论十分热烈。