计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900250

检索式聊天机器人技术综述

吴 侃[1] 李舟军[2]
1 微软亚洲研究院自然语言计算组 北京 100080
2 北京航空航天大学计算机学院 北京 100191
(yuwu1@microsoft.com)

摘 要 随着自然语言处理技术的飞速发展以及互联网上对话语料的不断积累,闲聊导向对话系统(简称聊天机器人)取得了令人瞩目的进展,受到了学术界的广泛关注,并在产业界进行了初步的尝试。当前,聊天机器人分为检索式聊天机器人和生成式聊天机器人,而检索式聊天机器人由于其生成的回复流畅且计算资源消耗小,仍然是目前工业界聊天机器人的主要实现手段。文中首先简要介绍了检索式聊天机器人的研究背景、基本架构以及组成模块,重点阐述了回复选择模块的约束要求和相关数据集;然后,针对检索式聊天机器人中最为核心的回复选择技术,进行了深入分析与详细梳理。文中将近年来经典的回复选择技术归纳为如下4类:基于统计模型的方法、基于表示的神经网络模型的方法、基于交互的神经网络模型的方法以及基于预训练技术的方法,并指出了这4类方法的优点和不足。在此基础上,分析了目前检索式聊天机器人技术研究所面临的问题,并对其未来的发展趋势进行了展望。

关键词:自然语言处理;聊天机器人;文本匹配;回复选择;预训练技术
中图法分类号 TP391

Survey on Retrieval-based Chatbots

WU Yu[1] and LI Zhou-jun[2]
1 Natural Language Computing Group, Microsoft Research Asia, Beijing 100080, China
2 School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract With the rapid progress of natural language processing techniques and the massive accessible conversational data on Internet, non-tasked oriented dialogue systems, also referred to as Chatbots, have achieved great success, and drawn attention from both academia and industry. Currently, there are two lines in chatbots research, retrieval-based chatbots and generation-based chatbots. Due to the fluent responses and low latency, retrieval-based chatbots is a common method in practice. This paper first briefly introduces the research background, basic structure and component modules of retrieval-based chatbots, and then illustrates the constraints of the response selection module and related data set in details. Subsequently, we summarize recent popular techniques for response selection problem, including: statistic method, representation-based neural network method, interaction-based neural network method, and pre-training-based method. Finally, we pose the challenges of chatbots and outline promising directions as future work.

Keywords Natural language processing, Chatbot, Text matching, Response selection, Pre-training technology

1 引言

聊天机器人研究历史悠久,早在20世纪60年代,MIT实验室的研究人员便研制出了著名的聊天机器人Eliza[1]。该聊天机器人主要依赖于模板,如果用户所说的话与已经编写好的模板匹配,便可得到很好的回复,否则会得到一些万能回复,例如:“我不知道”等。随后,在20世纪末期,基于规则的聊天机器人层出不穷,并设立了两个年赛,分别为:“罗布能奖”(The Loebner Prize)和“话匣子挑战赛”(The Chatterbox Challenge)。为了在比赛中夺冠,较为智能的聊天机器人层出不穷,其中具有代表性的有Parry和Alicebot。然而,人类语言变化多样,单纯依赖模板技术并不能枚举所有情况,因

到稿日期:2020-03-20 返修日期:2020-12-20
基金项目:国家自然科学基金(U1636211,61672081);软件开发环境国家重点实验室课题(SKLSDE-2021ZX-18)
This work was supported by the National Natural Science Foundation of China(U1636211,61672081) and Fund of the State Key Laboratory of Software Development Environment(SKLSDE-2021ZX-18).
通信作者:李舟军(lizj@buaa.edu.cn)

此聊天机器人的发展一度处于瓶颈期。

进入21世纪之后，随着机器学习技术的发展，以及可得到的互联网对话语料越来越多，数据驱动的聊天机器人技术愈发成熟[2]。其中，最有代表性有基于检索的聊天机器人[3]和基于生成的聊天机器人[4]。基于检索的聊天机器人指利用信息检索技术，为用户的会话请求匹配一个已经事先储存好的对话语料作为回复。基于生成的聊天机器人指利用自然语言生成技术自动回复用户的会话请求。目前主流的算法是借鉴在机器翻译领域取得成功的序列到序列的模型[5]。基于检索的聊天机器人和基于生成的聊天机器人各有优劣。检索式聊天机器人的优点主要包括：1)易于搭建，可依靠成熟的搜索引擎技术；2)计算开销小，可以快速给出回复；3)回复流畅且信息量较大，原因是检索式聊天机器人回复的候选集是通过抓取人类之前在网上的对话语料得到的。基于以上3个特点，检索式聊天机器人也是目前工业界的主流方案[6]。

本文聚焦于检索式聊天机器人，首先介绍了检索式聊天机器人的基本架构以及组成模块，并重点阐述了回复选择模块的约束要求和相关数据集。然后，针对检索式聊天机器人中最为核心的回复选择任务，详细介绍了近年来4种经典的回复选择技术，并探讨了其各自的优点和不足。最后，对检索式聊天机器人的未来发展趋势进行了展望。

2　检索式聊天机器人任务概述

本节描述了检索式聊天机器人的架构，以及不同模块所解决的问题，并具体阐释了回复选择模块时所遇到的挑战，以及研究领域针对这些挑战所构建的公开数据集。

2.1　概述

检索式聊天机器人通过文本匹配和排序学习技术[7]，从对话语料库中寻找最适合当前输入的回复，其架构如图1所示[3]。一个完整的检索式聊天机器人主要由3个模块构成：1)候选索引检索模块；2)相似度特征计算模块；3)排序学习模块。各个模块均包含线上和线下两部分。

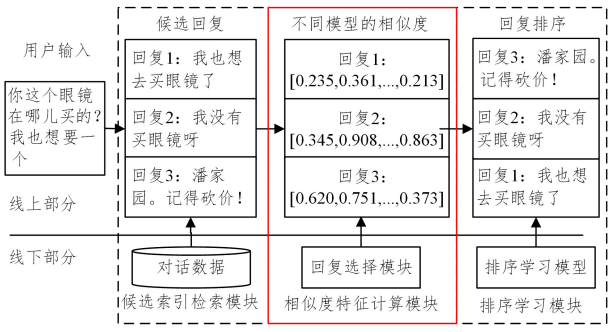

图1　检索式聊天机器人架构

Fig. 1　Structure of retrieval-based chatbot

对于线下部分，候选索引检索模块事先搜集大规模人类对话语料，并把语料整理成一问一答的形式，之后利用信息检索中的索引技术将这些对话进行索引，以便线上模块进行快速检索。相似度特征计算模块利用大量未标注或少量已标注的数据，事先训练多个不同的回复选择模型。排序学习模块则利用标注数据的标签以及相似度特征计算模块所给出的特征，训练排序学习模型。

而针对线上模块，图1给出了一个例子，对工作的具体流程进行更详尽的说明。给定一个用户输入，首先从大规模对话数据中快速检索数个候选回复，此处回复的召回率和模型效率至关重要，是检索式聊天机器人的第一步粗粒度筛选。由于对话数据往往上千万，因此候选索引检索模块使用倒排表或者向量检索的方式进行快速地回复召回。然后，这些回复选择模型为每个候选回复生成一个得分，并将这些模型所生成的得分拼接成一个向量。最后，排序学习模块将这些特征进行组合，从而得到最终的候选回复排序。相似度特征计算和排序学习两个模块对精度的要求极高，希望可以将最相关的候选回复放在排序结果的前列，因此往往采用一些比较复杂的模型。这里值得注意的是，相似度特征计算模块需要考虑当前的上下文和候选回复在多个维度的相似度，如相关性、逻辑一致性和风格一致性。

由于候选索引检索模块和排序学习模块与传统信息检索十分相似，并已得到较为充分的研究，因此往往可以采用已有的信息检索技术来实现相关模块。而相似度特征计算模块由于输入的是多轮对话，且具有复杂的句间依赖关系，与传统信息检索中的候选文档迥然不同，因此目前的主要研究热点和难点为相似度特征计算模块(也称为回复选择模型)。下文将给出相似度特征计算模块的经典任务和研究中所使用的数据集。

2.2　回复选择任务和数据集

表1列出了当前较为流行的回复选择任务数据集。

表1　检索式聊天机器人数据集概览

Table 1　Overview of datasets of retrieval-based chatbots

数据集名称	是否考虑多轮	是否人工标注	数据集语言	数据集规模(训练/开发/测试)
Ubuntu Dialogue数据集[7]	是	否	英文	1 000 000/500 000/500 000
新浪微博数据集[8]	否	是	中文	1 534 876/－/12 427
Douban对话数据集[9]	是	是	中文	1 000 000/50 000/10 000
E-commerce[10]	是	是	中文	1 000 000/10 000/10 000
Persona Dialogue数据集[11]	是	是	英文	10 907/1 000/968
Dialogue NLI数据集[12]	是	是	英文	310 110/16 500/16 500

2.2.1　单轮回复选择任务

在检索式聊天机器人的研究初期，研究者将回复选择简化为单轮回复选择任务。单轮回复选择任务指，只考虑前一轮的对话历史 $c=u_n$(忽略更早的对话历史 $u_1\cdots u_{n-1}$)和候选回复集合 $R=(r_1\cdots r_m)$，从候选集中选出与对话历史相关的回复。单轮回复选择任务将对话简化为单轮问答的任务(question answering)，虽然失去了对话复杂的上下文信息，但最后一轮对话内容往往至关重要，已经可以为很多对话提供充分的信息。为了构建单轮对话选择数据集，Wang等[8]抓取了新浪微博的相关数据集。为了限制微博话题不要太过发

散，该数据集在搜集数据时主要抓取了 NLP 学者和学生的微博以及相关回复，将微博当做对话历史 c，将微博回复当做单轮对话的回复。

在测试时，通过信息检索技术，为一个微博构建了 30 个回复作为候选，并让人进行标注，为每个回复标注相关性，然后通过传统信息检索指标 MAP[13] 和 P@1 来进行自动评测。该数据集中的训练集一共有 46 345 个微博，对应有 1 534 876 条回复，测试集一共有 422 个微博，对应有 12 427 条回复，平均每个微博对应 29.33 条回复。

2.2.2 多轮回复选择任务

由于单轮回复选择任务忽视了对话复杂的多轮历史，2015 年之后，研究者们逐渐转向多轮回复选择任务的研究。多轮回复选择任务定义为：给定多轮对话历史 $c=(u_1 \cdots u_n)$ 和候选回复集合 $R=(r_1 \cdots r_m)$，从候选集中选出和多轮对话历史相关的回复。由于多轮对话的轮数较多且句间关系复杂，如何准确理解多轮对话的上下文关系是一大难题。为了解决这个问题，目前已经有多个数据集被提出，其中最具有代表性的数据集是 Ubuntu Corpus[7]，该数据集抓取了大量 Ubuntu 论坛的帖子，将一个帖子视为一段对话，其中对话主题往往是围绕 Ubuntu 系统的技术。其中，人类真实的对话数据作为正样本 $(1, c_i, r_i)$。为了构建负样本，该数据集随机采样一个其他对话中的回复，构成 $(0, c_i, r_i)$。在测试集中正负样本比例为 1∶9。该数据集包含 100 万个训练集，50 万个开发集和 50 万个测试集。Ubuntu 数据集是第一个多轮对话数据集，由于其负例都是随机采样得到的，使得任务过于简单，并且与真实场景存在显著差异。

为了解决这个问题，Wu 等提出了豆瓣数据集[9]，该数据集的测试集模拟真实的检索式对话场景，从豆瓣社区进行数据的抓取。如图 2 所示，在豆瓣社区中有很多群组会针对某一个特定话题发起讨论，将跟帖和回复视为不同轮次的对话，能更好地模拟真实的对话。抓取对话上下文和回复后，为了构建高质量的测试集，给定一段对话历史，首先从数据库中寻找与之相似的对话历史，并将这些历史的回复当做候选，作为待标注的候选回复。

图 2 豆瓣回复样例

Fig. 2 Example of Douban response

然后，邀请 3 位专家进行相关性标注，判断回复是否与上下文相关，以此提高测试集的质量和难度。最终，豆瓣数据集得到 50 万条训练数据及 1 万条经过标注的测试数据。此后，Zhang 等针对电商数据，也提出了相应的数据集 E-commerce[10]，并将其作为多轮回复选择的评测数据集。

2.2.3 回复选择的一致性

检索式聊天系统的回复选择模块长期存在的一个问题是回复的一致性。如表 2 所列，用户两次抛出你的职业是什么？但机器却给出了完全不同的回答，机器第一次回答的职业是律师，第二次却变成了医生，产生了自相矛盾的问题。类似的问题在其他情境下也经常出现，例如“你今年多大了?”以及“你妈妈是谁”这种问题都会出现自相矛盾的情况。

表 2 对话不一致的例子

Table 2 Example of an inconsistent dialogue

	对话内容
用户	你的职业是什么？
机器人	我是个律师
用户	你的职业是什么？
机器人	我是医生

为了解决这个问题，Welleck 等[12] 提出了一个对话推断的数据集，该数据集研究的问题是：给定上下文 c 和一个相关的候选回复 r，该回复与上下文 c 的关系是矛盾的、中性的还是自然的？例如，上文说到律师之后，“我是医生”这个句子与上文存在矛盾，而“我是法律从业者”与上文一致。该数据集的训练集包含大约 30 万条数据，测试集包含 16 500 条数据。

2.2.4 个性化回复选择

之前的多轮对话选择模型并没有很好地考虑用户个人的个性化信息。为了解决这个问题，Zhang 等[11] 构建了一个基于个性化的多轮对话数据集，用于研究如何把用户的信息融入回复之中。此时，模型在回复选择时，不仅需要考虑之前的上下文 c，还需要考虑对话者的个性化信息 p。由于数据集均为人工生成，因此规模不大，一共包含 10 907 段对话，对话中的分句总共有 162 064 个。Persona-Chat 是构建的第一个基于个性化的多轮对话数据集[11]，让模型在回复选择的同时考虑用户的一些基础信息，例如：如果用户喜欢猫，它的回复更可能是“猫很可爱”，而不是“猫很惹人厌烦”。

然而，该数据集自身也存在一些问题，如对话数据集中几乎每句话都与用户的自身资料强相关，这与真实场景有所不同。在真实的场景中，用户往往只在个别时刻的对话内容是个性化的，而大部分情况下的对话是没有个性化信息的。

3 检索式聊天机器人经典回复选择模型

回复选择任务可以表示为给定句对 (c, r) 的分类任务，即 $g(c, r)$。已有工作的研究大多集中在如何设计分类函数 $g(\cdot, \cdot)$，下文将给出几种具有代表性的解决方案。

3.1 统计模型

最初，为了解决对话中的相关性问题，Wang 等[8] 使用传统的统计模型对该问题进行建模。其中，最为直接的是通过 tfidf[14] 的方法，计算回复和对话历史的相关度。具体而言，首先通过 one-hot 方法，将两句话利用 tfidf 分别表示成向量 $\boldsymbol{c}$，

$\boldsymbol{r}$，之后通过 cosine 相似度计算两句话的相似度，$g(\boldsymbol{c},\boldsymbol{r})=\frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{r}}{|\boldsymbol{c}||\boldsymbol{r}|}$。然而，由于对话上下文的相似度往往并不高，基于 tfidf 的方法对于对话相关任务的表现不如信息检索有效。为了解决这个问题，翻译模型[15]被引入，翻译模型可以计算候选回复与对话上下文中同现词的频率，例如上文出现“生病”之后，回复有很大的概率出现词语“医院”。

$$P_{TransLM}=\prod_{w\in\boldsymbol{c}}P_{TransLM}(w|\boldsymbol{r})$$

$$P_{TransLM}(w|\boldsymbol{r})=(1-\alpha)P_{mx}(w|\boldsymbol{r})+\alpha P_{ml}(w|C)$$

$$P_{mx}(w|\boldsymbol{r})=(1-\beta)[(1-\gamma)P_{ml}(w|c)+\gamma\sum_{t\in c}T(w|t)P_{ml}(t|p)]$$

其中，$P_{ml}(w|\boldsymbol{c})$代表语言模型从对话历史 $\boldsymbol{c}$ 中生成词语 w 的概率；$T(w|t)$代表翻译模型中词语 w 与 t 的转移概率；$P_{TransLM}(w|\boldsymbol{r})$代表给定回复 $\boldsymbol{r}$ 时，对话上下文中词语 w 的生成概率。除了翻译模型，主题模型也是检索式聊天机器人中最常用的模型。主题模型可以把一句话投射到主题空间，我们可以在主题空间进行相似度的计算。然而，主题模型往往需要长文本才可以很好地进行主题空间的映射，为了解决这个问题，往往使用 TwitterLDA[16]或者只抽出句子中的命名实体等词汇进行主题的预测。

除了上述 3 种非常常用的模型，最长公共子序列[17]、基于交叉熵的模型[18]等均被应用到检索式聊天机器人的回复选择任务中。基于统计的模型是传统信息检索经典算法[19]在聊天机器人上的应用，其优点是简单易实现。然而，随着神经网络算法的快速发展，基于统计的模型在表现上已经与神经网络模型相距甚远。下文将介绍基于神经网络模型的代表算法。

3.2　基于表示的神经网络模型

基于神经网络的算法可以分为基于表示的方法和基于交互的方法，两种方法各有优缺点。首先介绍基于表示的神经网络模型。基于表示的神经网络将对话上下文和回复分别用神经网络表示成固定维度的向量，之后进行向量相似度的计算，主要的探索方向是如何设计神经网络模型对句子进行建模，并将其压缩成向量。

图 3 给出了基于表示的神经网络模型的大体框架。这个框架包含 4 个重要的组成部分：句子表示函数 $f(\cdot)$，回复表示函数 $f'(\cdot)$，上下文表示函数 $h(\cdot)$以及匹配度计算函数 $m(\cdot,\cdot)$。给定上下文 $\boldsymbol{c}=\{u_1,\cdots,u_n\}$和一个候选回复 $\boldsymbol{r}$，$f(\cdot)$和 $f'(\cdot)$将 $\boldsymbol{c}$ 中的每个句子 u_i和候选回复 $\boldsymbol{r}$ 投影到向量空间。之后，将对话历史中的所有句子表示 $f(u_i)_{i=1}^n$传递给函数 $h(\cdot)$，并学习上下文 $\boldsymbol{c}$ 在向量空间中的表示 $h(f(u_1),\cdots,f(u_n))$。最终，函数 $m(\cdot,\cdot)$将 $h(f(u_1),\cdots,f(u_n))$ 和 $f'(\boldsymbol{r})$ 当做输入，计算最终的匹配得分。从宏观角度来看，原有的框架首先学习对话历史的表示以及候选回复的表示，然后计算对话历史和候选回复的匹配得分。我们在这里给出已有框架的数学形式化。

$$g(\boldsymbol{c},\boldsymbol{r})=m(h(f(u_1),\cdots,f(u_n)),f'(\boldsymbol{r}))$$

而这些研究工作的不同之处在于如何设计函数 $f(\cdot)$，$h(\cdot)$，$f'(\cdot)$和 $m(\cdot,\cdot)$。在传统的自然语言处理问答任务中，基于表示的匹配模型已经被探索过，其中具有代表性的工作有 TreeLSTM[20]，AnswerNetwork[21]，DCNN[22]等。而在检索式聊天机器人领域，Lowe 等[7]简单地将词向量进行拼接并将其作为句子表示，Zhou 等[23]利用 RNN 来学习句子表示，Yan 等[24]的 DL2R 模型利用 CNN 与 RNN 的组合来学习句子表示。在建模对话全部历史时（即函数 $g(x)$），RNN 和带有注意力的 RNN 均已被尝试。针对最后的匹配度计算，MLP 模型、Bilinear 模型和基于 Tensor[25]的匹配模型均取得了良好的效果。

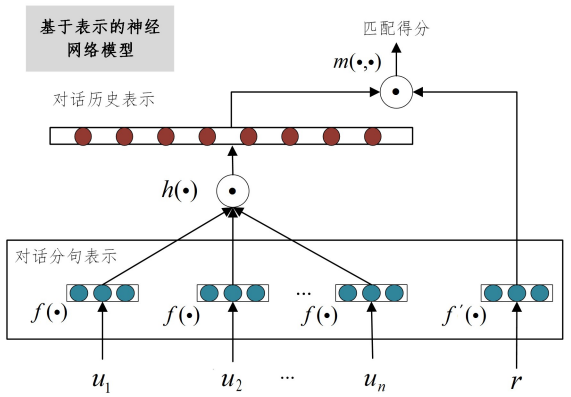

图 3　基于表示的神经网络模型结构图

Fig. 3　Representation-based neural network model

3.3　基于交互的神经网络模型

传统的基于表示的神经网络模型会丢失上下文中的重要信息，为了解决此问题，Wu 等提出了基于交互的序列匹配模型（Sequential Matching Network，SMN）[9]，并进一步将其拓展到基于交互的序列匹配框架（Sequential Matching Framework，SMF）[26]。该框架既可以捕获上下文和回复中的重要信息，也可以建模对话历史中各句的句间关系，这是对话领域中首个基于交互的序列匹配框架。

图 4 给出了 SMN 模型，该模型由 3 个模块组成，分别为：对话分句-回复匹配计算 $f(\cdot,\cdot)$（utterance-response matching）、匹配累积 $h(\cdot)$（matching accumulation）以及匹配预测 $m(\cdot)$（matching prediction）。这 3 个模块由神经网络的不同层来实现。给定对话历史 $\boldsymbol{c}=\{u_1,\cdots,u_n\}$ 和一个候选回复 $\boldsymbol{r}$，该方法首先利用 $f(\cdot,\cdot)$ 计算对话历史中每一句话和候选回复的匹配向量$\{f(u_1,\boldsymbol{r}),\cdots,f(u_n,\boldsymbol{r})\}$。Wu 等提出利用 2D CNN 模型来实现该模块[9]，之后 Zhou 等[27]提出用 Self-attention[28]的机制，并取得了更好的效果。在得到这些匹配向量之后，将向量传递给第二层进行下一步的计算，并让函数 $h(\cdot)$建模对话历史中不同句子的依赖关系，Wu 等[9]使用 GRU 进行该模块的建模，Qiu 等[25]使用 Co-attention[28]对该模块进行建模。第三个模块计算最终的匹配度得分，可以直接求各向量的算术平均，也可以对不同分句进行加权平均。除针对多轮对话任务所涉及的模型，传统自然语言处理领域也对基于交互的神经网络模型进行了探索，代表性的工作有 ABCNN[29]，PyramidMatch[30]，ESIM[31]以及 ECIM[32]。

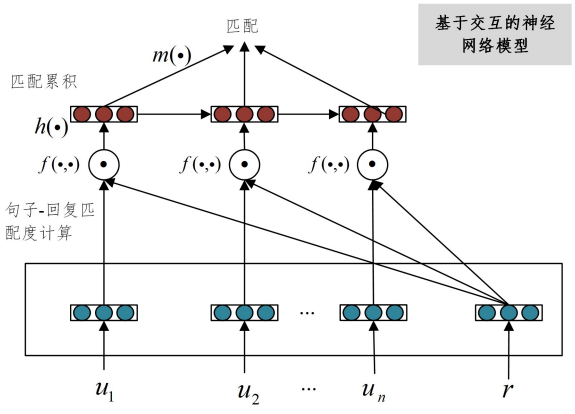

图 4 基于交互的神经网络模型结构图

Fig. 4 Interaction-based neural network model

基于交互的和基于表示的方法有两个主要的不同点。首先，基于交互的方法让对话历史中的每个句子和候选回复在一开始就进行匹配度的计算，而不是先对整个对话历史进行表示学习。通过这种方法，我们能尽可能地避免有效信息的丢失，让文本最原始的信息在匹配层体现出来。其次，基于交互的方法将对话间的依赖关系和匹配程度进行共同建模，并利用对话各句的依赖关系指导最终匹配度的计算。因为这些显著的不同，基于交互的方法有如下两个优势：1）首先计算候选回复和对话历史中各句的相似度，而不是首先将对话历史压缩成一个向量，通过该方法，可以让对话历史中的每一句话和候选回复进行充分的交互，据此可以提取对文本匹配最有用的信息并计算最终匹配得分；2）从不同的粒度对重要信息进行抽取，从而更好地识别有用的语义结构。

3.4 基于预训练的神经网络模型

近年来，预训练模型在 NLP 的各个任务上大显神威。预训练的历史要追溯到 ELMo[33] 的提出。ELMo 预训练语言模型提出了一种上下文相关的动态文本表示法，可以很好地解决已有的 Word2Vec[34] 模型在一词多义表示方面的不足。其后，GPT[35]，BERT[36]，RoBERT[37] 和 XLNET[38] 等预训练语言模型相继被提出。其中，最引人注目的是 BERT，它利用 Transformer 模型[28]，可以看到双向所有词语的优势，训练双向的语言模型，并通过完形填空任务在大规模语料上进行预训练，在多个典型下游任务上效果得到显著提升，极大地推动了自然语言处理领域的技术发展，自此开创了动态预训练技术的时代。

在对话领域中，预训练模型也展现出优秀的效果。预训练模型将对话上文 c 和回复 r 拼接起来，形成向量[***CLS***，c，***SEP***，r]并将其作为输入。与上述两种方案不同，预训练模型较为简单，直接通过一个基于 Transformer 的模型便可以算出最终得分 $g(c,r)$。如图 5 所示，其中 *CLS*，*EOU* 和 *SEP* 为预训练模型的保留字，分别为对话开始符号、上下文分隔符和历史结束符。Transformer 模型可以利用注意力机制将全局的信息融合到模型表示之中。预训练模型使检索式聊天机器人在各数据集上的表现达到了当前最高水平。例如，在 Ubuntu 数据上，之前模型在 R_{10}@1 的最高水平约为 76，而百度通过 ERNIE[39] 模型，将在该数据集上的表现提高到了 86，这是一个质的飞跃。与此同时，Whang 等[40] 探讨了如何在预训练模型的基础上更好地进行微调来提升回复选择模型的性能。

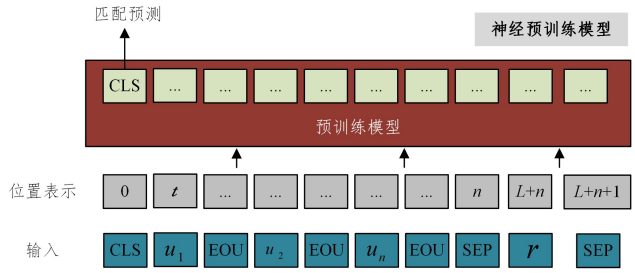

图 5 回复选择任务中的神经网络预训练模型

Fig. 5 Pre-trained model in response selection

3.5 已有方法的对比

上文介绍了 4 种经典的回复选择模型，这些模型各有优缺点，因此在不同场景下，我们应有不同的取舍。从 3 方面对比已有模型，如表 3 所列。可以看到在准确率上预训练模型是最高的，然而预训练模型需要大规模的计算资源，以及少量回复选择任务的标注数据进行训练。例如，RoBERTa 模型需要大约 1 024 块 V100 GPU 进行数天的训练才可以收敛。而简单的神经网络模型仅用数张卡就可以在一天内收敛。统计模型只需要在 CPU 上进行运算，计算开销比基于神经网络的模型更小。

表 3 已有 4 类模型优缺点的对比

Table 3 Comparison of four models

	准确率	计算速度	是否需要标注数据训练
统计模型	低	快	否
基于表示的神经网络模型	中	一般	是
基于交互的神经网络模型	中	慢	是
预训练模型	高	很慢	是

基于统计的模型往往是无监督的，即不需要大规模对话标注数据，甚至不需要对话数据就可以进行训练。在一个数据丰富并且计算资源足够的领域，我们可以使用预训练模型。然而，如果计算资源有限，则可以使用其他方法来解决问题。

4 面临的主要挑战及未来的研究方向

检索式聊天机器人是目前工业界最为主流的聊天机器人实现方式，上文已经总结了任务定义以及目前最先进的技术。虽然当前检索式聊天机器人已经取得了初步成果，但依然面临着一些严重的问题和严峻的挑战。

4.1 常识和知识

虽然检索式聊天机器人已经可以应对很多问题，但是对于客观的事实类问题，仍然无法准确地回答。例如，对于“羽绒服在冬天穿吗？”“美国总统是谁？”等问题，聊天机器人往往会给出一些与客观事实相悖的答案。造成这个问题的主要原因是，当前的检索式聊天机器人没有很好地与常识和知识相结合。在计算回复的适合程度时，仅仅考虑了语义的相关性，而没有考虑与客观世界的知识和常识进行很好的结合。

因此,未来一个很有意义的研究方向就是如何在对话中引入外部知识。为了完成这个任务,有两个问题需要重点研究。1)如何收集和对话相关的知识库?目前很多知识库,例如 ConceptNet 和 FreeBase,都是为搜索引擎设计的,但并没有考虑对话的种种特性。2)在有了知识库之后,如何将其融入检索式聊天机器人的回复选择算法之中也是需要攻克的难题。

4.2　模型推理

当前的检索式聊天机器人主要考虑上下文和候选回复的相关性,但是当其需要作出简单的模型推理时则表现得不好。例如,对于"这里比纽约晚六个小时,现在纽约是下午五点",机器往往难以推断出当前的时间,也无法针对此时间进行继续对话。此外,机器也无法很好地推理出用户对一些事物的情感极性,从而无法针对用户的态度和情感展开对话。因此,当前的机器人仍然只能进行简单的相关性回复,无法像人一样进行对话的推理。

聊天机器人模型推理的问题目前主要受限于没有很好的数据集进行评测,以致难以开展研究。当前针对推理的数据集仍然主要局限于问答领域,例如 Commonsense QA[41] 以及 DROP[42] 等数据集。因此,为了推动聊天机器人推理问题的相关研究,一个好的数据集是必不可少的,有了数据集之后相关的研究才可以开展。当前数据驱动的算法往往对推理一筹莫展,如何设计更好的推理模型不仅是检索式聊天机器人研究的难题,也是整个自然语言处理领域的一大难题。

4.3　多轮对话理解

当前,如果只考虑单轮对话,则检索式聊天机器人可以取得很好的效果,大约在 80% 的情况下回答流畅且与输入相关。然而,当输入的轮数增多,问题从单轮对话回复变为多轮对话回复时,检索式聊天机器人的表现不够理想。其主要原因是,多轮对话的句间关系十分复杂,其中可能包含着顺承、转折等多种关系。有时,由于话题的转移,多轮对话历史的内容对当前回复已经不起作用,甚至起到反作用。同样,也有可能由于顺承的关系,多轮对话历史对当前的回复选择起着不可替代的作用。

目前,为了解决这个问题,已经有一些多轮对话改写数据集和方法[43] 出现,其目标是将多轮对话改写为一句单轮对话,以此来评测模型对多轮对话的理解能力。在未来,如何更好地建模多轮对话仍然是一个极具挑战的问题。

4.4　多模态对话

除此之外,当前的聊天机器人仅仅考虑了文本信息,然而人与人之间的对话往往包含更多模态的信息,而且这些信息对于对话的理解十分关键。例如:"我哪儿有你这么聪明",如果语气是讽刺的,则说话者当时是持有负面情感;而如果语气是高兴的,则说话者当时是持正面情感。除了语气,面部表情和肢体语言也可以更好地帮助对多轮对话的理解。例如:如果面部表情十分狰狞,则代表说话者情绪十分愤怒;如果喜笑颜开,则代表说话者很开心。这时,无论说话者的语言是什么,其附带的情感都可以被更好地判断。因此,在构建聊天机器人时,引入多模态对话可以更好地帮助对多轮对话的理解,进而改善聊天机器人的用户体验。

4.5　模型迁移

目前检索式聊天机器人在数据量大的情况下已经表现较好,然而在很多情况下,某个领域只有少量数据。此时构建一个与大数据环境下的检索式聊天机器人表现类似的模型,则具有很大的挑战性,往往会因为数据量不够,导致训练出的模型不鲁棒,并且在小领域数据上表现不佳。为了解决这个问题,检索式聊天机器人的模型迁移便是一个值得研究的方向。小领域上的数据往往数据量不够,并且有很强的内容和风格的偏向,使得重新训练一个模型具有相当大的挑战。目前迁移学习的方法在问答领域已经被探索过[44],我们相信迁移学习对检索式聊天机器人领域也会有很大帮助。因此,设计迁移算法是对话系统中一个很有前途的领域。

结束语　本文首先简要概述检索式聊天机器人的基本概念以及各个组成模块,并详细介绍了检索式聊天机器人回复选择模块的约束要求和相关数据集。然后重点阐述了近年来具有代表性的 4 类模型。最后分析了目前检索式聊天机器人研究所面临的问题,并对其未来的发展趋势进行了展望。

参 考 文 献

[1] WEIZENBAUM J. ELIZA:a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM,1966,9(1):36-45.

[2] RITTER A,CHERRY C,DOLAN W B. Data-driven response generation in social media[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011:583-593.

[3] JI Z,LU Z,LI H. An Information Retrieval Approach to Short Text Conversation[J]. arXiv:1408.6988,2014.

[4] VINYALS O,LE Q. A neural conversational model[J]. arXiv:1506.05869,2015.

[5] SUTSKEVER I,VINYALS O,LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. 2014:3104-3112.

[6] ZHOU L,GAO J,LI D,et al. The design and implementation of xiaoice,an empathetic social chatbot. [J]. Computational Linguistics,2020,46(1):53-93.

[7] 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,The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Prague,Czech Republic,2015:285-294.

[8] WANG H,LU Z,LI H,et al. A Dataset for Research on Short-Text Conversations[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. EMNLP,2013:935-945.

[9] WU Y,WU W,XING C,et al. Sequential Match Network:A New Architecture for Multi-turn Response Selection in Re-

trieval-based Chatbots[C]// Proceedings of the 55th Annual Mee- ting of the Association for Computational Linguistics. 2017:496-505.

[10] ZHANG Z, LI J, ZHU P, ZHAO H, LIU G. Modeling Multi-turn Conversation with Deep Utterance Aggregation[C]// Proceedings of the 27th International Conference on Computational Linguistics 2018:3740-3752.

[11] ZHANG S, DINAN E, URBANEK J, et al. Personalizing Dialogue Agents: I have a dog, do you have pets too? [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. ACL, 2018:2204-2213.

[12] WELLECK S, WESTON J, SZLAM A, et al. Dialogue natural language inference [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019:3731-3741.

[13] BURGES C, SHAKED T, RENSHAW E, et al. Learning to rank using gradient descent[C]// Proceedings of the 22nd International Conference on Machine Learning (ICML-05):89-96.

[14] SPÄRCK J K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval[J]. Journal of Documentation, 1972:28(1):11-21.

[15] BROWN P F, DELLA S A, DELLA P V J, et al. The mathematics of statistical machine translation: Parameter estimation [J]. Computational Linguistics, 1993:19(2):263-311.

[16] ZHAO X, JIANG J, WENG J, et al. Comparing twitter and traditional media using topic models[C]// European Conference on Information Retrieval. 2011:338-349.

[17] WAGNER R A, FISCHER M J. The string-to-string correction problem[J]. Journal of the ACM (JACM), 1974, 21(1):168-173.

[18] MACKAY D J. Information theory, inference and learning algorithms [M]. Cambridge University Press, 2003.

[19] BAEZA-YATES R, RIBEIRO-NETO B, et al. Modern information retrieval[M]// volume 463. ACM press, New York, 1999.

[20] CHOI J, YOO K, LEE S. Learning to compose task-specific tree structures[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018:248-258.

[21] LIU X, KEVIN D, GAO J. Stochastic answer networks for natural language inference[J]. arXiv:1804.07888, 2018.

[22] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM. A Convolutional Neural Network for Modelling Sentences[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014:655-665.

[23] ZHOU X, DONG D, WU H, et al. Multi-view Response Selection for Human-Computer Conversation[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:372-381.

[24] YAN R, SONG Y, WU H. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System[C]// Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016:55-64.

[25] QIU X, HUANG X. Convolutional neural tensor network architecture for community-based question answering[C]// Proceedings of the 24th International Conference on Artificial Intelligence. 2015:1305-1311.

[26] WU Y, WU W, XING C, et al. A sequential matching framework for multi-turn response selection in retrieval-based chatbots [J]. Computational Linguistics, 2019, 45(1), 163-197.

[27] ZHOU X, LI L, DONG D, et al. Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018:1118-1127.

[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

[29] YIN W, SCHÜTZE H, XIANG B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016(4):259-272.

[30] PANG L, LAN Y, GUO J, et al. Text matching as image recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016:2793-2799.

[31] CHEN Q, ZHU X, LING Z H, et al. Enhanced LSTM for Natural Language Inference[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017:1657-1668.

[32] SONG S, WANG C, PU X, et al. An Enhanced Convolutional Inference Model with Distillation for Retrieval-Based QA[C]// DASFAA. 2021:511-515.

[33] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// Proceedings of NAACL-HLT. 2018:2227-2237.

[34] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013:3111-3119.

[35] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[OL]. https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf, 2018.

[36] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of NAACL-HLT. 2019:4171-4186.

[37] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized BERT pretraining approach[J]. arXiv:1907.11692, 2019.

[38] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C]// Advances in Neural Information Processing Systems. 2019:32-42.

[39] ZHANG Z, HAN X, LIU Z, et al. ERNIE: Enhanced Language Representation with Informative Entities [J]. arXiv: 1905.07129, 2019.

[40] WHANG T, LEE D, LEE C, et al. Domain Adaptive Training BERT for Response Selection[J]. arXiv:1908.04812.

[41] TALMOR A, HERZIG J, LOURIE N, et al. Commonsense QA: A Question Answering Challenge Targeting Commonsense Knowledge[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:4149-4158.

[42] DUA D, WANG Y, DASIGI P, et al. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs[C]// Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019:2368-2378.

[43] ZHOU K, ZHANG K, WU Y, et al. Unsupervised Context Rewriting for Open Domain Conversation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2017:1834-1844.

[44] YU J, QIU M, JIANG J, et al. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce[C]// Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018: 682-690.

WU Yu, born in 1992, Ph.D, senior researcher. His main research interests include natural language processing and spoken language processing.

LI Zhou-jun, born in 1963, Ph.D, professor, is a member of China Computer Federation. His main research interests include data mining, natural language processing, network and information security.