

生成式和检索式对话机器人的算法设计与实现综述

王 宇^{1,2}, 王 澈^{1,2}, 于 丹^{1,2}

(1.大连东软信息学院, 辽宁 大连 116023;

2.大连东软教育科技集团有限公司研究院, 辽宁 大连 116023)

✉wangyu@neuedu.com; wangche@neuedu.com; yudan@neuedu.com



摘 要: 依据对话机器人回复方式的不同, 可以将对话机器人分为基于生成式的对话机器人和基于检索式的对话机器人。本文主要针对上下文信息的使用, 将这两类机器人继续分类为单轮检索、多轮检索、单轮生成和多轮生成对话机器人, 对各对话机器人的构建进行了总结和阐述, 列举了应用场景和算法实现挑战, 并通过两个工业对话机器人构建中基于生成式和基于检索式的应用, 分析了目前两类对话机器人落地应用的主要难点和解决方法。最后对未来的研究趋势和方向进行了展望。

关键词: 基于生成的对话机器人; 基于检索的对话机器人; 单轮对话; 多轮对话

中图分类号: TP311.5 **文献标识码:** A

Overview of Algorithm Design and Implementation of Generative and Retrieving Chatbots

WANG Yu^{1,2}, WANG Che^{1,2}, YU Dan^{1,2}

(1. Dalian Neusoft University of Information, Dalian 116023, China;

2. Research Institute, Dalian Neusoft Education Technology Group Co. Limited, Dalian 116023, China)

✉wangyu@neuedu.com; wangche@neuedu.com; yudan@neuedu.com

Abstract: Responses of chatbots are generated by methods that are either generation-based or retrieval-based. This paper mainly focuses on the use of contextual information, to further classify these two types of chatbots into single-turn retrieval, multi-turn retrieval, single-turn generation, and multi-turn generation chatbots. And then, the paper summarizes and elaborates the construction of each type of these chatbots, and lists application scenarios and algorithm implementation challenges. Through generative and retrieval-based applications in construction of industrial chatbots, main difficulties and solutions are analyzed. Finally, the future research trends and directions are considered.

Keywords: generation-based chatbot; retrieval-based chatbot; single-turn dialogue; multi-turn dialogue

1 引言(Introduction)

通过自然语言与人进行有效交流的能力, 是机器具有智能的一个重要的体现, 同时也是一个很具有挑战性的任务。自第一个聊天机器人Eliza^[1]诞生开始, 众多研究者尝试采用了各式各样的方法, 构建智能对话机器人, 并取得了令人欣喜的结果。按照对话机器人回复方式的不同, 可以将对话机器人技术分为基于生成式的对话机器人和基于检索式的对话机

器人。

基于检索式的对话机器人是通过对话者输入的语义特征, 从成对的或者非成对的数据中, 检索出最能够匹配对话者语义特征的数据, 作为对话机器人的回复, 这是一个检索而非创造的过程。基于生成式的方法则是通过大量人类对话数据, 学习人类语言学特征, 训练可针对对话者输入而自动生成文本作为回复的模型, 这是一个生成的过程。

作为实现对话机器人有效回复的两种主要方式,学术界则一直在不断拓展二者的表现上限。而工业界则也一直将对话机器人摆在很高的位置,将其视为未来人机交互的主要形式,不断地加大研究力度,希望结合生成式和检索式的方法,将对话系统的商业应用价值实现至最大。目前以苹果Siri和微软Cortana为代表的对话系统,通过手机和电脑被广泛使用;天猫精灵和小爱则是以智能家居机器人的角色走进千家万户。微软以开放域聊天为卖点,针对多语种构建了Xiaoice、Rinna、Zo、Ruuh等聊天机器人,用户群体已达千万人。

目前已有总结对话机器人技术的论文^[2],但很少有对落地应用进行阐述的论文。本文将针对近些年检索式和生成式对话机器人的先进技术发展进行总结和综述,并对已经进行商业应用的对话机器人的回复方式进行论述。

2 问题定义(Problem definition)

根据实际需求和场景限定,可将对话系统分为单轮对话系统和多轮对话系统。如图1所示,检索式和生成式对话机器人以生成回复时是否利用上下文信息为标准,可以继续细分为单轮检索式对话机器人、多轮检索式对话机器人、单轮生成式对话机器人、多轮生成式对话机器人。

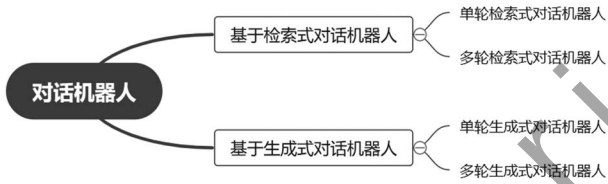


图1 基于回复的对话机器人分类

Fig.1 Classification of chatbots based on replies

单轮检索式对话可定义为:

$$r = f(q) \quad (1)$$

多轮检索式对话可定义为:

$$r = f(q|c) \quad (2)$$

单轮生成式对话可定义为:

$$r = g(q) \quad (3)$$

多轮生成式对话可定义为:

$$r = g(q|c) \quad (4)$$

其中, q : 表示查询语句;

r : 表示回复语句;

c : 表示历史对话信息;

$f(*)$: 表示从询问回复数据对库(也有称之为问答对库) q -pairs中检索最匹配数据的函数;

$g(*)$: 表示根据输入进行回复生成的函数;

u : 表示对话过程完整的一轮语境表示,既可以表示对话者的询问,也可以表示对话机器人的回复。

3 基于检索式的对话机器人(Retrieval-based chatbot)

基于检索式的对话机器人通过输入语境的特征分析,从询问回复数据对(或问答对)中选取最匹配的部分回复给对话者。可以通过是否引入上下文对话信息的情况,把检索式对话分为单轮检索对话和多轮检索对话。

3.1 单轮检索对话

单轮检索对话,仅对输入的一个询问 q 进行回复,而不涉及历史对话信息,不具有记忆功能,是很多对话机器人基本也是最为实用的一种方式。比如客服机器人,只需一轮对话,即可满足对话者当前的需求。此类对话一般为问答对话,同时也可以收集特定语料,完成一定程度的开放域闲聊。

一般来说,需要事先构造询问回复数据对库,一个询问对应了一个或多个回复。通常来说,对于输入 q ,在询问回复数据对库中,检索出跟 q 最相关的询问回复数据对,将对应回复作为 r 输出,完成一次单轮对话。根据检索对象的不同,又可以分为 q 和询问比较、 q 和回复的匹配和两者的结合使用。

(1) q 和询问比较。通过比较 q 和询问回复数据对库中所有询问的语义相似性,寻找相似度最高的询问回复数据对数据,将此数据的回复作为 r ,反馈给对话者。将问题性质转换为比较两个句子语义是否相同的语义相似性问题,可以通过语义相似性比较的方法进行实现,比如使用基于卷积神经网络的TextCNN^[3]、TextRCNN^[4]等,基于循环神经网络的TextRNN^[5]、TextAttentionRNN^[6]等;同时使用近两年大放异彩的预训练语言模型Bert^[7]、GPT^[8]提高比较的准确性,如孪生结构的SentenceBert^[9]。

(2) q 和回复的匹配。 q 和询问比较只能回答问答对里跟 r 最为相关的问题,可延伸性差。如,“大疆无人机是中国最好的无人机”,既可以回答“大疆无人机是中国的吗”,也可以回答“大疆无人机是哪国的”,但是如果问答对库里面没有“大疆无人机是哪国的”这样的问题,通过 q 和询问比较的方法是无法进行回答的。所以,借鉴了搜索引擎的一些方法,研究者们提出了 q 和回复的匹配的方法,将问题转换成两个文本关联度的语义匹配问题。可以使用常见的语义相似性匹配的方法,如DSSM^[10]、MV-LSTM^[11]、BiMPM^[12]等。

(3) q 和询问比较与 q 和回复的匹配的结合使用。另外,将 q 和询问比较与 q 和回复的匹配以某种加权的方式进行检索,或者在语义层面在模型中进行多种信息交互,进行语义匹配。前者可以认为是 q 和询问比较与 q 和回复的匹配的方法的组合使用;而后者则是在语义空间中,将 q 与询问一回复对数据的信息进行交互,计算相关性评分。SymBiMPM^[13]是对BiMPM的改进,使用了两个BiMPM模型,一个是 q 和询问比较,另一个是 q 和回复的匹配 r ,两者信息用于相似度比较。MMT-DNN^[14]借鉴MT-DNN^[15]模型模式,构建出两块Transformer Encoder^[16]层,第一块基于Bert^[7]构建 q 与询问和

回复的表征,第二块两个共享参数Transformer Encoder层,分别比较 q 和询问、 q 和回复的相似度,最后拼接后进行相似度评价。

3.2 多轮检索对话

多轮检索对话结合历史对话信息,针对对话者的query,从候选回复中选取最相关的回复。对话的上下文语境具有连续语义关系,对于query检索是必不可少的信息,这也是其与单轮检索对话最为不同的地方。常使用的方法是将上下文语境词语输入循环神经网络中,提取最后一层隐藏层作为上下文信息去匹配候选回复,但是忽略了作为一个单独的整体带来的信息。

Zhou^[17]等人通过使用两个GRU模型提取多语境下单词和语境两个维度的信息,提出Multi-view Response Selection的方法,解决多轮检索上下文信息语境信息不足的问题,但是仍然存在颗粒度信息不足的问题。Wu等人^[18]提出序列匹配网络(Sequential Matching Network, SMN),通过词语和语境两个不同的颗粒度与候选回复匹配构建匹配矩阵,卷积和池化操作提取重要的匹配信息转换为一个向量,按照时序输入循环神经网络中,通过循环神经网络的隐藏层计算匹配得分。Zhang等人^[19]对SMN模型进行了改进,提出了DUA模型,强调了语境的重要性,并进行了深度编码,认为最后一个语境很重要,将信息补充到其他语境和回复里面。DUA模型将最后一个语境的GRU Embedding与其他语境和回复连接,接入带有注意力机制的GRU去提取关键信息和去除冗余信息,再进行词语和句子匹配矩阵构建,计算匹配得分。Zhou等人^[20]放弃了对语境序列进行单独表征的方法,而是采用基于Transformer^[17]结构对文本多颗粒度信息进行提取。模型将语境和候选回复输入Transformer,Transformer每一层代表不同颗粒度的语义信息,相同颗粒度信息的语境和候选回复进行匹配生成自注意力和交叉注意力匹配矩阵,将不同语境匹配矩阵聚合3D矩阵后,使用3D卷积和池化操作进行特征信息提取,通过感知机计算匹配得分。

4 基于生成式的对话机器人(Generation-based chatbots)

生成式对话机器人通过大规模的对话语料训练和一个端到端的模型,学习人类对话的模式和特征,输入对话者的自然语言,依据语境输出生成的回复。生成式对话机器人受到机器翻译任务的启发,模型的一般范式是编码解码模式,通过编码器学习语义特征,通过解码器进行回复生成。这种方法的明显好处是具有很高的泛化性,即使不在已知对话语料中,仍可以进行普适性的回答。以是否利用上下文信息 c 为依据,生成式对话又可以分为单轮生成式对话和多轮生成式对话。

4.1 单轮生成式对话

单轮生成式对话机器人通过对话者输入的 q ,生成回复

r ,基本都是编码-解码的端到端的模型为基准进行构建。编码器将输入 q 进行语义解码,将语义编码传递给解码器用于回复生成。

Shang等人^[21]基于新浪构建语料库,采用RNN编码语义信息,并使用Attention机制,得到全局语义向量和局部语义向量,加权输入解码器机芯解码。Shao等人^[22]则是将解码器首尾相连接,将生成的部分作为Attention机制的一部分,对已有信息量进行进一步补充,用于生成具有高信息量的长回复。Wu等人^[23]对解码阶段的词语映射进行改进,提出一种基于动态解码词典的序列到序列模型(DVS2S),使得每步解码根据当前对话实际有不同的词典,以去除不相关词汇的干扰,缩小映射范围,加快解码速度。在训练时,该模型同时学习动态词典构建和回复生成。在预测中,模型使用预测模型动态地为输入分配相应的小词汇表,并且仅使用小词汇表进行解码,实验结果表明,在不影响解码质量的情况下,在线解码生成的速度提高40%。

4.2 多轮生成式对话

依据上下文 c ,并对当前输入 q 进行生成回复 r 。多轮生成式对话和单轮生成式对话的最大区别在于,多轮生成式对话需要对上下文语境信息进行建模,如何有效地利用上下文语境 u 之间的内在联系,是提升多轮生成式对话质量的关键。

Sutskever等人^[24]根据上下文语境输入LSTM进行编码,获取历史语义特征,并通过另外一个LSTM对编码器最后一个隐藏层进行解码,生成词语直到生成终止符为止,生成的语句作为多轮对话的回复。Hu等人^[25]结合对话机器人对情感分析的需求,提出了音调的重要性,将编码信息和音调信息一起输入解码器中,针对不同对话者的音调特征,进行了个性化回复。但是以上这些方法都是使用语境 u 作为上下文单位信息的来源,而没有多颗粒度下的信息提取。Serban等人^[26]使用层次神经网络(Hierarchical Neural Network, HNN),将词语输入循环神经网络提取语境信息,将语境时序序列输入另外一个循环神经网络作为上下文解码信息,输入解码器进行回复生成。Xing等人^[27]在此基础上引入了注意力机制,对词语和语境颗粒度下的重点信息进行关注和学习,提取关键信息并有针对性地回复,提高对话机器人对话质量。随着Transformers结构在自然语言处理中大放异彩,生成式对话机器人也乘着东风掀起了一次高潮,Wang^[28]等人收集并清晰整理出目前最大的中文多轮对话数据集,并在GPT模型的基础上提出CDial-GPT方法,将多轮对话进行串联后输入GPT模型,进行多轮对话训练,并在多轮生成式对话上取得了优异的成果。

5 落地应用简述(Brief description of landing application)

基于检索式的方法主要是突出了回复的稳定性上的优势,询问回复数据对库的数据是人为可控的,所有回答都不

会超过询问回复数据对库里面的内容,内容是可控可选择的,适合事实类问答功能的实现。穷举所有问题是不可能的,但是如果无法连续地对输入的 q 进行有效的回复,则会显得对话机器人“呆”和“笨”。

基于生成式的方法主要是体现在回复的多样性和泛化性上,通过学习人类对话,模拟真实对话场景,适合对不确定领域的闲聊进行构建。但基于生产式对话的实际表现远没有人们预想的那么优秀,其回答具有不可控性,前后文信息不一致、“安全性”回答一直是困扰基于生成式对话开发人员的问题。目前主流的方法是对生成的 r 的质量依据上下文和设定把控并修改。

现阶段,工业界的对话机器人并不是以单一的生成式对话和检索式对话为主,而是将二者集合在大的对话框架之下。一般来说,生成式对话用于闲聊域,检索式对话用于限定域或者基于事实回复。XiaoIce^[29]的核心对话中,则是结合检索式对话、生成式对话和基于知识图谱的检索对话方式并联,依据三种回复质量选取最优回复 r 。而阿里小蜜^[30]则是选择了一种基于检索式和基于生成式的串联的方法,首先构建了9 000万对聊天记录的客户服务询问—回复数据库,采用基于检索的方法对输入的 q 进行相似性比较,如果符合阈值,则返回最相似询问—回复数据中的回复作为输出的 r ;如果超过阈值,则对 q 进行基于生成式的方法生成 r 作为输出。

目前,从零构建一个基于检索式或基于生成式的工业对话机器人最大的阻碍可以体现在数据和硬件两个方面。尽管目前大数据和互联网的高速发展,使人类比以前任何时候都容易获得和处理海量数据,但是真正意义上的高质量数据获取起来依旧很难。对于基于检索式的方法,高质量的询问回复数据对决定了对话的上限;而对于基于生成式的方法,高质量的对话数据则是模型表现力的根本保证。XiaoIce团队花了四年收集到了300亿对询问—回复对,而阿里小蜜团队则凭借淘宝和天猫用户基数的优势,收集了9,000万对优质客服聊天记录。优质的数据是构建对话机器人不会改变的需求。

由于算法和硬件的关系一直是负相关的,优秀的算法往往要极大地消耗硬件的算力,这也是学术界的优秀成果不能第一时间在工业界应用的原因之一。权衡算法性能和硬件的关系,这是对话机器人在工业界落地必须思考的问题。

6 未来展望(Future)

现阶段工业对话机器人某种程度上来说依旧是以检索式对话为主,生成式对话为辅,将主要的对话功能中心压在检索式对话,以生成式对话作为开放域对话的补充,原因在于基于检索式对话的可控性和生成式对话的不可控性。

目前的生成式对话模型训练都是基于文本的数据,缺少了真正对话者的背景信息和对话者的世界认知,即对话者自身处于一种具体的场景下,基于自己对真实世界的理解,进行了具有意义的对话。

所以,在未来很长一段时间内,基于检索式的对话方式仍然是工业对话机器人长期使用且有效的方法,但基于对话者的背景信息和对话者的世界认知训练的生成式对话机器人,在未来极大可能作为学术界和工业界重点研究的一个方向而成为研究主流。

7 结论(Conclusion)

本文总结了基于检索式的对话机器人和基于生成式的对话机器人目前主流的算法,简述了工业落地的实践,概括了从零构建工业机器人的痛点,并对未来对话机器人的形势进行了预测。

根据是否利用上下文信息和反馈回复的方法,将对话机器人细分为单轮检索式、多轮检索式、单轮生成式和多轮生成式的对话机器人,并分别对相关的技术进行了归纳和总结。

以XiaoIce和阿里小蜜两个实际工业机器人为例,阐述了基于检索式的对话机器人和基于生成式的对话机器人的综合使用,并指出高质量数据和算法与硬件的直接矛盾是从零构建工业机器人的两个痛点。

最后,本文认为未来一段时间内,工业机器人仍将以基于检索方法为主流对话方式,而生成式为辅对话的一种手段。基于对话者的背景信息和对话者世界认知的生成式对话机器人将会在未来可能会得到大力发展。

参考文献(References)

- [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] 陈晨,朱晴晴,严睿,等.基于深度学习的开放领域对话系统研究综述[J].计算机学报,2019,42(7):1439–1466.
- [3] Kim Y. Convolutional neural networks for sentence classification[DB/OL]. [2014–09–03]. <https://arxiv.org/pdf/1408.5882.pdf>.
- [4] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015:2267–2273.
- [5] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[DB/OL]. [2016–05–17]. <https://arxiv.org/pdf/1605.05101.pdf>.
- [6] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016:207–212.
- [7] Jacob D, Ming-Wei C, Kenton L, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[DB/OL]. [2018–10–11]. <https://arxiv.org/pdf/1810.04805.pdf>.

- [8] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pretraining[DB/OL]. [2020-03-15]. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>.
- [9] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[DB/OL]. [2019-08-27]. <https://arxiv.org/pdf/1908.10084.pdf>.
- [10] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using click through data[C]. Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013:2333-2338.
- [11] Wan S, Lan Y, Xu J, et al. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN[DB/OL]. [2016-04-15]. <https://arxiv.org/pdf/1604.04378.pdf>.
- [12] Zhiguo Wang, Wael Hamza, Radu Florian. Bilateral Multi-Perspective Matching for Natural Language Sentences[DB/OL]. [2017-07-14]. <https://arxiv.org/pdf/1702.03814.pdf>.
- [13] Damani S, Narahari K N, Chatterjee A, et al. Optimized Transformer Models for FAQ Answering[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, 2020:235-248.
- [14] 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.
- [15] Liu X, He P, Chen W, et al. Multi-task deep neural networks for natural language understanding[DB/OL]. [2019-05-20]. <https://arxiv.org/pdf/1901.11504.pdf>.
- [16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[DB/OL]. [2019-09-16]. <https://arXiv:1706.03762v5>.
- [17] 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.
- [18] Wu Y, Wu W, Xing C, et al. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots[DB/OL]. [2017-05-15]. <https://arxiv.org/pdf/1612.01627.pdf>.
- [19] Zhang Z, Li J, Zhu P, et al. Modeling Multi-turn Conversation with Deep Utterance Aggregation[DB/OL]. [2018-11-06]. <https://arxiv.org/pdf/1806.09102.pdf>.
- [20] 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.
- [21] Shang L, Lu Z, Li H. Neural responding machine for short-text conversation[DB/OL]. [2015-04-27]. <https://arxiv.org/pdf/1503.02364.pdf>.
- [22] Shao L, Gouws S, Britz D, et al. Generating High-Quality and Informative Conversation Responses with Sequence-to-Sequence Models [DB/OL]. [2017-07-31]. <https://arxiv.org/pdf/1701.03185.pdf>.
- [23] Wu Y, Wu W, Yang D, et al. Neural response generation with dynamic vocabularies[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018:5594-5601.
- [24] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems, 2014:3104-3112.
- [25] Hu T, Xu A, Liu Z, et al. Touch your heart: A tone-aware chatbot for customer care on social media[C]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018:1-12.
- [26] Serban I, Sordoni A, Bengio Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016:3776-3783.
- [27] Xing C, Wu Y, Wu W, et al. Hierarchical recurrent attention network for response generation[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018:5610-5617.
- [28] Wang Y, Ke P, Zheng Y, et al. A large-scale chinese short-text conversation dataset[C]. CCF International Conference on Natural Language Processing and Chinese Computing, Springer, Cham, 2020:91-103.
- [29] Zhou L, Gao J, Li D, et al. The design and implementation of Xiaoice, an empathetic social chatbot[J]. Computational Linguistics, 2020:53-93.
- [30] Qiu M, Li F L, Wang S, et al. Alime chat: A sequence to sequence and rerank based chatbot engine[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017:498-503.

作者简介:

- 王 宇(1994-), 男, 硕士, 初级研究员.研究领域: 自然语言处理, 人工智能.
- 王 澈(1995-), 女, 硕士, 初级研究员.研究领域: 人工智能, 图像处理.
- 于 丹(1976-), 女, 博士, 研究员.研究领域: 数据分析与挖掘, 人工智能.