计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于单向 Transformer 和孪生网络的多轮任务型对话技术研究
作者：王涛，刘超辉，郑青青，黄嘉曦
DOI：10.19678/j.issn.1000-3428.0058557
网络首发日期：2020-07-28
引用格式：王涛，刘超辉，郑青青，黄嘉曦．基于单向 Transformer 和孪生网络的多轮任务型对话技术研究．计算机工程.
https://doi.org/10.19678/j.issn.1000-3428.0058557

网络首发时间：2020-07-28 09:15:59
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.TP.20200727.1507.003.html

计 算 机 工 程
Computer Engineering

基于单向 Transformer 和孪生网络的多轮任务型对话技术研究

王 涛[1] 刘超辉[1] 郑青青[1] 黄嘉曦[1]
[1](深圳市易马达科技 广东 深圳 518055)

摘 要 为了解决循环神经网络和 Transformer 在多轮对话系统的建模上依赖于大量的样本数据，且回复的准确率过低的问题，该研究提出了一种针对任务型对话系统的新的建模方式。首先，引入预训练模型对话过程进行深度编码，其次对 Transformer 模型进行精简，仅保留编码器部分的单向 Transformer，然后将应答部分抽象成不同的指令，采用孪生网络对指令进行基于相似度的排序，最终选择相似度最高的指令生成应答。在 MultiWOZ 数据集上的实验结果表明，与 LSTM 和当前最先进的基于 Transformer 的模型相比，该研究提出的方法预测速度更快，小数据集上的表现也更加优秀，在大数据集上也能取得和最先进模型相当的效果。

关键词 任务型 多轮对话系统 预训练模型 Transformer 孪生网络

开放科学（资源服务）标志码（OSID）：

MULTI-TURN TASK-ORIENTED DIALOGUE SYSTEM BASED ON UNIDIRECTIONAL TRANSFORMER AND SIAMESE NETWORK

Wang Tao[1] Liu Chaohui[1] Zheng Qingqing[1] Huang Jiaxi[1]
[1](Shenzhen Immotor Technology Co. Ltd, *Shenzhen 518055,Guangdong,China*)

Abstract In order to solve the problem that the recurrent neural network and Transformer rely on a large datasets in the modeling of the multi-turn dialogue system, and the accuracy of the reply is too low. this work presented a new modeling method, especially for the Task-Oriented dialogue system. We used the pre-training models and tools to encode the dialog contents deeply, then simplified the Transformer model, only kept the encoder of it. After that, we abstract the response to different commands, and sorted them by Siamese Network. At last, we choose the best result to generate response. The experimental results on the MultiWOZ datasets show that, compared to LSTM and State-of-the-art model based on Transformer, the method used in this study has faster prediction speed, performs better on small datasets, and achieves comparable results on large datasets.

Keywords Task-Oriented Multi-Turn Dialogue System Pre-training Model Transformer Siamese Network

DOI:10.19678/j.issn.1000-3428.0058557.

0 引 言

让机器可以以自然语言的方式与人类进行交流，完成人类下达的任务一直是人工智能[1][2][3]领域最具挑战的一项研究。从 1951 年图灵在《计算机与智能》一文中提出用人机对话来测试机器智能水平开始[4]，关于人机对话的研究就一直没有停止过，近年来，工业界更是将对话系统视为下一代人机交互的主要形式。2003 年 BENGIO 等[5]率先将神经网络

基金项目：中美绿色基金(MA009RX18)
作者简介：王涛，(1993-)，男，学士、工程师，主研方向为人机交互；刘超辉、郑青青，高级工程师；黄嘉曦，研究员。
E-mail：wt.china@outlook.com

应用于自然语言处理任务，并取得了不俗的效果。2010 年 MIKOLOV 等[6]提出的 RNNLM 更是显著提高了语言模型的准确性，之后的 RNN 及其各种变体如 LSTM[7]、GRU[8]，开始逐渐成为自然语言处理领域的常用技术。Google 在 2017 年提出了一种新的序列建模模型 Transformer[9]，该模型一经提出，就在 NLP 领域引起了极大的反响，而后 BERT[10]的发布更是将自然语言处理技术推上了一个新的台阶。

任务型对话系统[11][12]，即接受人类指令完成特定任务的对话系统是被工业界广泛使用的对话系统之一。相比于闲聊型对话系统[11][12]，任务型对话系统存在以下难点。第一，可供使用的数据集相对较小。面向任务的对话系统因为其任务的特殊性，很难像闲聊系统一样，项目启动之初即拥有大量的闲聊对话数据可以使用，针对不同的任务，通常只能生成或取得非常少量的数据。第二，任务型对话系统对应答的准确性要求较高。闲聊型对话系统应答出错一般情况下不会引起使用者的不适，然而任务型对话系统应答出错会直接导致用户下达的指令或任务无法被完成。

为了解决上述问题，本文构建了一种面向小数据集的任务型多轮对话控制模型。首先，引入多个预训练模型[13]与工具，借助外部知识对句子语意和对话过程进行深度编码。然后，对 Transformer 模型进行进一步精简，仅保留编码器[14]部分的单向 Transformer，不仅充分利用了多头自注意力机[9]优秀的特征提取能力，同时精简后的单向模型可以支持并行计算，大大提升了计算效率。最后，将应答部分抽象成指令，充分利用孪生神经网络[15]在小数据集上的优势对指令进行基于相似度的排序，最终选取相似度最高的指令生成应答。

1 相关工作

无论是学术界还是工业界，对对话机器人的研究一直都没有停止过。Zhou[16]等人提出了基于卷积神经网络[17]和循环神经网络的多轮对话检索模型，该模型将对话上下文信息作为输入，并从词序列和句子序列两个视角来计算匹配分数，最终结合两个分数来选择回复。其中，基于词序列的视角将文本中所有词按顺序输入到一个 GRU 中，将其隐藏向量作为文本的语义表示；句子序列的视角则基于卷积神经网络，先通过卷积和池化得到每个话语的表示，再输入到另一个 GRU 中输出文本的表示。

随着 Transformer 的流行，越来越多的人开始试着用 Transformer 构建多轮对话模型。Henderson[18]等人利用 Transformer 在 Reddit 的数据集上构建了一个大型的多轮对话模型，其中在对话控制和回复生成上，全都采用了 Transformer 结构，取得了非常好的效果，证明了 Transformer 在多轮对话系统建模上的优秀性能。Dinan[19]等人采用了一个类似的结构使用 Transformer 对多轮对话进行建模，只是在回复生成部分，他们的设计提供了两种方式，一种是检索式的，即 Transformer 模型用于对回复部分进行排序，另一种是生成式的，即使用 Transformer 直接生成 token-by-token 的回复。

2 多轮对话控制模型

我们提出的基于单向 Transformer 和孪生网络的多轮对话控制技术，引入了多个预训练模型来弥补数据样本集较小，信息不足的问题，借助外部知识对模型输入和对话过程进行深度编码，同时我们对 Transformer 模型进行进一步精简，仅保留编码器部分的单向 Transformer，最后的应答部分，我们没有采用传统的分类模型，而是采用孪生神经网络，通过最大化对话之间的相似度来为当前的对话状态和每个回复指令进行建模。在预测阶段，将当前的对话状态与所有可能的回复指令进行比较，并选择具有最高相似度的指令生成回复。具体的模型结构如图 1 所示：

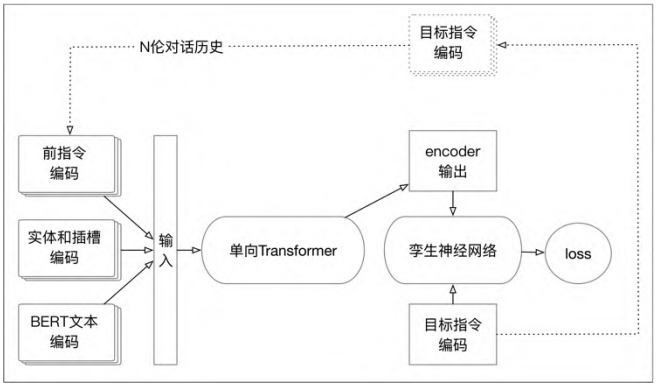

图 1 模型结构图

fig.1 model Structure

2.1 预训练模型

为了解决样本数量较小的问题，我们引入多个预训练模型和工具对句子语意和对话过程进行深度编码。首先，充分利用预训练模型 BERT 的先天优势将用户输入的文本编码成特征向量，同时，利用斯坦福大学的 StanfordNLP[20]工具对用户输入的文本进行进一步处理，提取出文本中包含的实体、预定义插槽等深度语义特征，并将上述抽取的特征统一进行 one-hot 编码。除此以外，为了尽可能的保存对话状态，我们将上一轮对话输出的目标指令同样进行 one-hot 编码，一起加入到本轮对话的输入中。最后，将上述 3 种编码后的向量进行拼接，作为单向 Transformer 的输入。

2.2 单向 Transformer

单向 Transformer 的输入包含了系统的历史指令和文本的深度语义特征，如实体、插槽、预训练特征向量。这样，我们就可以充分利用 Transformer 的自注意力机制，让其自发的选择突出一些重要的特征，同时忽略一些对对话过程影响不大的非重要特征，这一点在复杂多变的多轮对话中尤其重要。

2.3 孪生神经网络

我们将单向 Transformer 的输出作为孪生神经网络的其中一个输入，再将目标指令的 one-hot 编码作为另一个输入。输出部分，我们将正确的样本标记为 1，错误的样本标记为 0，同时由于某些指令要比其它指令多很多，负样本的数量也要比正样本多，所以我们采用随机采样算法处理样本均衡问题，最后，通过优化孪生网络的损失函数训练模型。在预测阶段，我们选用相似度最高的指令生成本轮对话中系统的回复。孪生神经网络的结构图如图 2 所示：

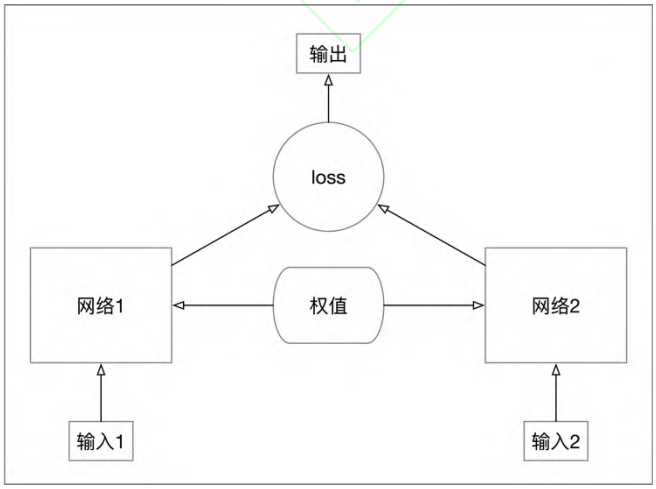

图 2 孪生神经网络

fig2. Siamese network structure

3 实验验证

本次实验中，我们使用了 2 个基线模型作为对比模型，第一个是传统的基于 LSTM 的 seq2seq 模型[21]，该模型是现阶段最稳定也是工业界应用最广泛的模型之一。第二个是 Henderson 等人于 2019 年提出的基于 Transformer 的模型，该模型是现阶段在任务型对话系统中表现最出色的模型之一。同时，我们使用 MultiWOZ 2.1 数据集[22]分别进行了三组实验，第一组实验对比了预训练模型对最终结果的影响，第二组实验通过缩减数据集规模，对比在小数据集下本文模型的表现效果，第三组实验对比了本文模型对比另外两个模型在时间效率上的差别。

3.1 MultiWOZ 数据集

在任务型对话系统中，我们需要对下一步的指令进行预测，因此类似 WikiQA[23]或 DailyDialog[24]这样的数据集无法满足需求，因为例如“ok”、“copy that”等回复实际对应的是同一个指令“YES”。因此我们选用 MultiWOZ 2.1 数据集作为我们的实验数据集。

MultiWOZ 2.1 数据集包含了酒店，饭馆，火车站，出租车，旅游景点，医院和警察局 7 个不同情境的对话数据集，共包含 10438 条数据。所有的对话都发生在用户和接待员之间。用户会问接待员相关问题，要求接待员完成相关任务，接待员会响应用户请求或要求用户补充相关信息，如要求用户提供姓名等。

本次任务中，我们将数据集按 7:3 的比例划分成训练集和测试集，其中训练集 7307 条数据，测试集 3131 条数据。

3.2 深度编码实验结果

第一轮实验中，我们采用全量的数据对上述基于 LSTM，基于 Transformer 和本文模型进行有无深度编码的对比实验。在无深度编码的分组，我们使用常用的词向量[25]技术对用户输入进行编码。在深度编码分组，我们采用本文提出的，使用预训练的 BERT 对用户输入进行编码，同时融入了实体，插槽等深度特征。最终的实验结果如表 1 所示：

表 1 深度编码实验数据

Table1. Deep-encoding experimental data

模型	历史轮数	Accuracy	F1 Score
LSTM(词向量)	5	0.58	0.21
Transformer(词向量)	5	0.64	0.23
本文模型(词向量)	5	0.62	0.20
LSTM(深度编码)	5	0.63	0.53
Transformer(深度编码)	5	0.73	0.64
本文模型(深度编码)	5	0.74	0.61

通过对比表 1 的数据我们可以发现，在任务型对话系统中，由于机器的每一轮回复都是非常明确的指令，所以传统的基于词向量的编码方式由于缺少任务中的关键信息而难以取得好的效果。分别对比 3 个模型的词向量编码方式和深度编码方式，我们发现本文提出的深度编码方式总能取得更好的效果，特别是本文提出的模型相比于传统的 LSTM 基于词向量的模型在 F1 score 上取得了近 3 倍的提升。

3.3 小数据集实验结果

第二轮实验中，为了验证本文模型在小数据集上的表现效果，我们仅使用第一轮实验五分之一的数据量，采用上述同样的深度编码的方式进行试验。实验的结果如表 2 所示：

表 2 小数据集实验数据

table2. Small dataset experimental data

模型	历史轮数	Accuracy	F1 Score
LSTM(深度编码)	5	0.59	0.42
Transformer(深度编码)	5	0.56	0.41
本文模型(深度编码)	5	0.69	0.55

通过对比表 2 和表 1 的数据我们可以发现，当训练数据缩减为原来的五分之一后，三个模型的 F1 score 都有不同程度的下降，但本文提出的模型的下降幅度远小于另外两种模型，仅下降了 9.8%，而另外两种模型分别下降了 35.9%和 20.8%。特别值得一提的，Henderson 等人提出的基于 Transformer 的模型在数据集缩减后，分类的准确率甚至不如传统的基于 LSTM 的模型。而本文提出的精简后的单向 Transformer 模型融合孪生神经网络在小数据集上取得了比传统 LSTM 和 Henderson 等人提出的基于 Transformer 模型都要好的效果。

3.4 预测时间对比实验结果

第三轮实验中，为了验证本文模型在时间效率上的表现效果，我们随机取出 1000 条数据，然后分别使用三种模型进行预测，从而对比三种模型在计算性能上的表现效果。实验的结果如表 3 所示：

表 3 预测时间实验结果

table3. Prediction time experimental data

模型	数据量（条））	预测时间（毫秒）
LSTM	1000	897
Transformer	1000	1029
本文模型	1000	780

通过对比表 3 数据我们可以发现，本文提出的模型的预测速度比 Henderson 等人提出的基于 Transformer 的模型要快 24.1%，与传统的基于 LSTM 的模型的预测速度相近。

4 结 语

本文主要研究了一种面向任务型对话系统的多轮对话控制技术。模型通过引入预训练模型和工具，借助外部知识，对模型输入和对话过程进行深度编码，同时模型还对 Transformer 模型进行了进一步精简，仅保留编码器部分的单向 Transformer，应答部分，本模型采用孪生网络对对话过程进行基于相似度的建模，最终选取相似度最高的指令生成回复。实验结果表明，在任务型对话系统中，当数据集比较大时，本文提出的模型效果远优于传统的基于 LSTM 的模型，和现阶段最先进的基于 Transformer 的模型的表现效果相当，且本文提出的深度编码方式更加适合任务型对话系统。当数据集规模减小时，在小数据集上，本文提出的模型准确率损失幅度远小于传统的基于 LSTM 的模型和目前最先进的基于 Transformer 的模型，且总体表现效果比另外两种模型都更加优秀，本文提出的模型在计算效率上也有

一定幅度的提升，说明本模型相比另外另种模型更快且更加适用于小型数据集。下一步我们将继续致力于任务型对话系统的性能提升。

参 考 文 献

[1] Zhang Weinan, liu Ting. The research progress of chatbot technology[J]. Chinese Artificial Intelligence Society Newsletter, 2016, 6(1):17-21. 张伟男，刘挺. 聊天机器人技术研究进展[J].中国人工智能学会通讯，2016，6(1): 17-21.

[2] Chen H S, Liu X R, Yin D W, et al. A survey on dialogue systems: Recent advances and new frontiers [J]. ACM Sigkdd Explorations Newsletter, 2017, 19(2): 25–35.

[3] Huang Yafang, Li Zuchao, Zhang Zhuosheng, et al. Moon IME: Neural-based Chinese pinyin aided input method with customizable association[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations. 2018: 140–145.

[4] TURING A M. Computing Machinery and Intelligence[M]. UK: Mind, 1950: 433-460.

[5] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(2): 1137-1155.

[6] Mikolov T, Karafi á t M, Burget L, et al. Recurrent neural network based language model[C]//Proceedings of INTERSPEECH 2010 Computer Science, 2010: 1045-1048.

[7] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computer, 1997, 9(8): 1735–1780.

[8] Hosseini S A, Bazrafkan S, Vatandoost H, et al. The insecticidal effect of diatomaceous earth against adults and nymphs of Blattella germanica[J]. Asian Pac J Trop Biomed, 2014, 4: 228-232.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [EB/OL]. [2017-12-06]. https://arxiv.org/abs/1706.03762.

[10] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [EB/OL]. [2019-05-24]. https://arxiv.org/abs/1810.04805.

[11] Yan R, Zhao D Y. Coupled Context Modeling for Deep Chit-Chat: Towards Conversations between Human and Computer[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 2574–2583.

[12] Chen H, Liu X R, Yin D W, et al. A Survey on Dialogue Systems: Recent Advances and New Frontiers. arXiv:1711.01731, 2017. https://arxiv.org/abs/1711.01731

[13] Qiu X P, Sun T X, Xu Y G, et al. Pre-trained Models for Natural Language Processing: A Survey. [EB/OL]. [2018-01-11] https://arxiv.org/abs/2003.08271

[14] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. [EB/OL]. [2014-09-03] https://arxiv.org/abs/1406.1078.

[15] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 1063-6919.

[16] Zhou X Y, Dong D X, Wu H, et al. Multi-view Response Selection for Human-Computer Conversation[C]//Conference on Empirical Methods in Natural Language Processing. 2016: 372-381.

[17] Lipton Z C, Berkowitz J, Elkan C. A Critical Review of Recurrent Neural Networks for Sequence Learning. [EB/OL]. [2015-10-17]. https://arxiv.org/abs/1506.00019

[18] Henderson M, Vuli ć I, Gerz D, Casanueva I, et al. Training neural response selection for

task-oriented dialogue systems. [EB/OL]. [2019-06-07]. https://arxiv.org/abs/1906.01543.

[19] Dinan E, Roller S, Shuster K, et al. Wizard of wikipedia: Knowledge-powered conversational agents. [EB/OL]. [2019-02-21]. https://arxiv.org/abs/1811.01241.

[20] Christopher D M, Surdeanu M, Bauer J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014: 55-60.

[21] Park S H, Kim B D, Kang C M, et al. Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture. [EB/OL]. [2018-10-22]. https://arxiv.org/abs/1802.06338 .

[22] Eric M, Goel R, Paul S, et al. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. [EB/OL]. [2019-12-03]. https://arxiv.org/abs/1907.01669.

[23] Yang Y, Yih S W, Meek C. WikiQA: A Challenge Dataset for Open-Domain Question Answer-ing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 2013–2018.

[24] Li Y R, Su H, Shen X Y, et al. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset[C]//Proceedings of the The 8th International Joint Conference on Natural Language Processing. 2017: 986-995.

[25] Mikolov T, Sutskever I, Sutskever I, et al. Distributed Representations of Words and Phrases and their Compositionality. [EB/OL]. [2013-10-16]. https://arxiv.org/abs/1310.4546