CCKS2021：蕴含实体的中文医疗对话生成评测报告

李宾[1]，陈恩城[2]，刘鸿儒[3]，翁诣轩[4]，孙斌[1]，李树涛[1]，
柏勇平[5]，胡美玲[6]

[1]湖南大学电气信息与工程学院
{libincn,shutao_li,sunbin611}@hnu.edu.cn
[2]中山大学数学学院
1510605163@qq.com
[3]京东科技
liuhongru3@jd.com
[5]中南大学湘雅医院
baiyongping@csu.edu.cn
[6]中南大学湘雅医院临床护理学教研室
humeiling0704@126.com

Abstract. 本文介绍了由2021年全国知识图谱与语义计算大会（CCKS2021）组织的蕴含实体的中文医疗对话生成竞赛第一名的方法，本届竞赛要求生成与历史条件一致且有医疗意义的回复。因此我们提出了一个由实体预测和实体感知对话生成模块组成的流水线系统，通过融合机制将预测实体与对话上下文添加到生成模型中，以利用来自不同来源的信息。在解码阶段，我们使用多样性集束搜索，以提高回复的多样性。在比赛后期，我们利用两种不同的融合方法，提升了最终回复的长度和质量。

Keywords: 医疗实体预测，医疗感知融合对话生成，模型融合

1 引言

1.1 背景

在新冠疫情期间，中国存在着医疗资源短缺、医生负担沉重、患者等待时间长等问题。因此，建立一个可以自动回复的医疗对话系统有利于提高临床会诊的效率，减轻医生的负担。因此，第十五届中国知识图谱与语义计算会议（CCKS2021）设置了蕴含实体的中文医疗对话生成的任务，参与者需要根据胃肠病学中的医患对话语料库内容构建对话生成模型。近年来，医学对话生成技术越来越受到研究者的关注，落地需求也逐渐加大[1-5]。为了实现生成模型能够模仿真实的医生回复并实现真正的落地，有两个亟需解决的问题：一是模型需要能够给出合理的回答，这通常涉及到正确的医疗实体信息[1,2]。二是模

[4]Email：wengsyx@qq.com.

型需要模仿人思考习惯，生成流畅且长度较长的回复[3,4]。为此，我们提出了一个包含实体预测和实体感知融合对话生成的两阶段的系统。我们在这项工作可以总结如下：

1. 我们构建了一个医学对话生成框架，采用具有很强的灵活性的流水线系统，以获得高质量的回复；
2. 我们开发了一种编码融合模块，充分利用对话历史的实体编码信息与预测编码信息；
3. 通过两种不同的模型融合方法，最后提升了最终回复的长度和质量

接下来，本文分别从模型与方法、实验结果、以及结论进行介绍。

1.2 数据描述

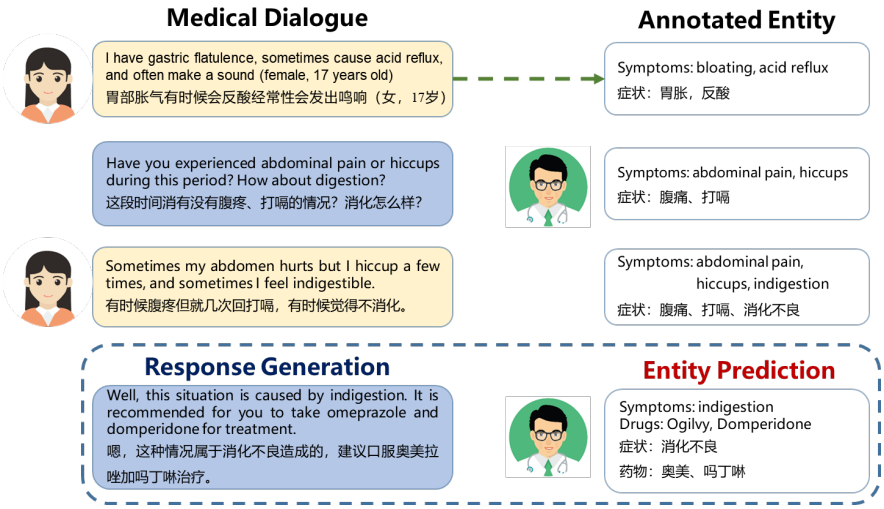

图1. 医疗对话数据示例

医疗对话数据示例如图1所示，本次比赛提供的训练集共有17864段对话，训练集的数据参数参考文献[4]，验证集共有452段对话，该验证集为测试集中抽出来的452段对话数据。验证集中对话中回复每段对话平均10句，每段对话平均198个字，每句话平均19.8个字，最大长度为300，最小长度为11，平均长度为57.4，共有1124个实体，平均每句话有2.5个实体。测试集A一共2747段对话，每段对话平均205.5个字，每段对话平均10.7句，每句话平均19个字。测试集B一共1600段对话，每段对话平均194.5个字，每段对话平均9.8句，每句话平均19.76个字。

2 模型及方法介绍

2.1 系统构架

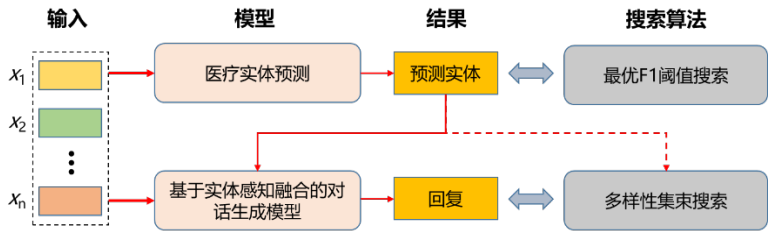

图2. 医疗对话数据示例

我们方法的框架如图2所示，其中我们采用了两阶段式的管道方法，采用Seq2Seq架构[6]。在上游推理预测出最佳的预测实体集合，并通过最佳F1阈值搜索进行结果优化。在下游，将预测的实体被与对话上下文输入到实体感知融合对话生成模型中，通过多样性集束搜索得到最终的响应[7-8]。

2.2 实体预测模型

实体预测模块，用于对下一句（也就是医生回复）中可能出现的实体进行预测，本质上还是一个分类模型，本次比赛中我们用到的方法有：

A. 多种预训练模型：

我们的医疗实体预测模型如图 2 所示，我们选择不同的预训练模型进行实体预测，包括利用 BERT [9]、RoBERTa [10]、PCL-MedBERT[1]、RoBERTa-wwm-ext [11] 等模型作为主干。至于一般领域的预训练模型，我们使用 MacBert 的预训练方法 [12] 利用在线医学数据作为继续的预训练[13]，并且利用医学领域语料继续微调，以提高模型泛化能力：

在网络上查询相关的开源语料，这次比赛微调利用的语料有：

1. MLPCP[2] a 榜数据（17864 段对话）（同时也是 CCKS 的训练样本集）；
2. MLPCP[2] b 榜数据，因为没有标识对话人（Speaker）的信息，因此训练了一个 chr2bert 的模型，将预测的对话人信息用作伪标签，加入训练；
3. CCKS b 榜的数据，因为直接标注了说话人，因此用模板去标注实体就可以加入训练中；
4. Kamed[3] 数据集，其中包括了整理之后的 Meddialogue 数据集。

[1]https://code.ihub.org.cn/projects/1775

[2]https://competitions.codalab.org/competitions/29703

[3]https://github.com/lddsdu/VRBot

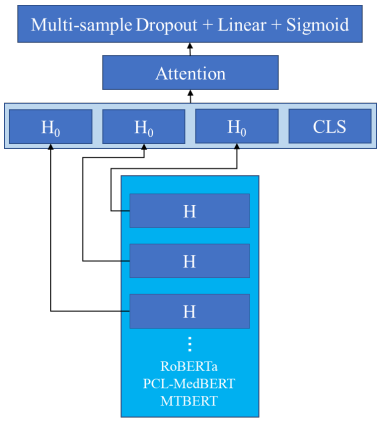

图2. 医疗实体预测模型

在实体预测模型部分，我们采用了以下的Trick：

1. FGM对抗训练[15]、混合精度训练、移动滑动平均策略；

2. 分层学习率与学习率衰减策略。具体为BERT上层学习率设置较大值，BERT内部学习率较小，并且越接近底层学习率越小；

3. F1指标优化技术优化类别不平衡。在分类问题中，当类别不平衡时，使用交叉熵损失优化得到的结果，F1并不是全局最优的。考虑每一个实体为二分类问题，可以使用阈值搜索，得到一个合理阈值。

2.3 实体感知融合对话生成模型

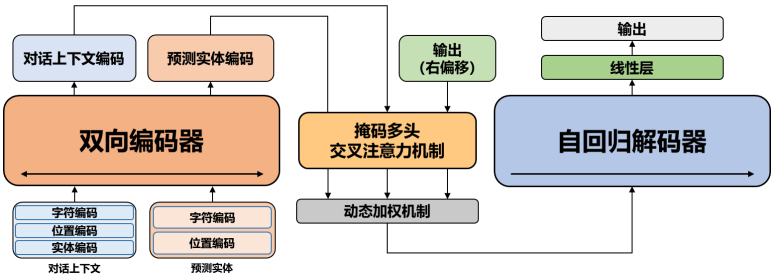

图3. 实体感知融合对话生成模型

通过encoder-decoder[9]的方式将对话上下文信息与实体信息通过Masked Muti-head Cross Attention的机制进行融合，使得最后的回复具有预测的实体信息。

2.3.1 上下文编码模块

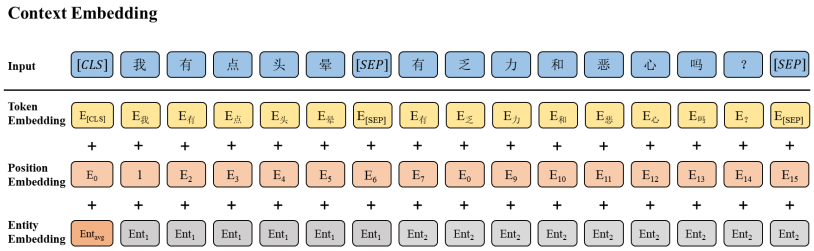

图4. 上下文编码模块示意图

设计上下文编码模块，将输入进行tokenize得到的token与位置编码、一段句子出现的实体编码进行相加，注意E_{avg}为整个句子的平均编码（除以句子长度），其余每一部分都是对应句子的编码，三者相加得到输入encoder端的上下文编码。

2.3.2 实体编码部分

将实体进行拼接，通过[SEP]进行隔开，通过tokenizer进行映射成token的形式，通过与位置编码进行相加得到实体编码。

2.3.3 编码融合机制

具体计算方式：

query key value

$$O_{ent} = \boldsymbol{MHCA}[E_{prev}, E_{ent}, E_{ent}],$$
$$O_{prev} = \boldsymbol{MHCA}[E_{prev}, E_{prev}, E_{prev}],$$
$$O_{C} = \boldsymbol{MHCA}[E_{prev}, E_{C}, E_{C}],$$

$$O_{enc} = (O_C + O_{ent} + O_{prev})/3.$$

图5. 编码融合机制示意图

MHCA 为 Masked multi-head Cross Attention 方法。进行MHCA后进行编码平均，使得输入解码端的信息包括了历史实体信息，预测编码信息，上下文等信息。

2.4 模型融合

我们使用了5-fold交叉验证训练策略，bagging融合机制，两类融合机制能够提高最终的得分，并且一定程度上缓解了模型学习偏执的问题，增加了最后回复的多样性缓解exposure bias问题，进一步提升BLEU[14]与F1值。

3 实验

3.1 数据预处理

样本构造 EDA：在准备 a 榜的过程中，筛选最终的回答是数据清洗后的医生说的样本、数据清洗后的医生说的带实体的样本，数据清洗后的医生说的长文本的样本，为之后的课程学习微调准备数据。

另外，在准备 b 榜的过程中，通过观察公开的测试集合样本，最短长度为 11 字，一共出现 1124 个实体，平均每句话 2.5 个实体，并且绝大部分回复都包含实体且长度比较长。因此筛掉过短回复对应的样本 (少于 11 字的)，删掉不包含实体的样本。

3.2 模型实现

3.2.1 训练策略

继续预训练 对面向医疗文本的PCL-MedBERT上使用Mac预训练的方法继续预训练，并在比赛数据集上进行微调。在语句中，使用正则匹配相关医疗领域词语，以总体15%的比例替换医疗词语，而15%中则是80%替换为相似词，10%替换随机词，10%不替换，让模型预测原句。

3.2.2 微调策略

课程学习&5-Fold 采用几个阶段，主要分为三步：

第一、 利用训练好的 Seq2Seq 模型(BertGPT/T5)初始化 encoder 和 decoder 的参数，在所有经过数据清洗的数据上进行微调，采用 boost 的方式训练 4 个 epoch 共 5 折；

第二、 利用所有医生的带有实体的对话进行训练，使得回复具备医生的共性特征，采用 boost 的方式训练 4 个 epoch 共 5 折。

第三、 用长度大于 11 的医生的带有实体的对话语料训练模型，因为这些对话有较强的实体回复特征，并且长度也很长，所以模型更容易学习如何使用实体进行回复，训练 2 个 epoch 共 5 折。

3.3 实验结果

表1. 不同主干预测结果

Model	F1
BERT-base-chinese [9]	31.23
RoBERTa-wwm-ext [7] [11]	31.68
RoBERTa-large [10]	33.23
Mac-BERT-large [12]	33.64
PCL-BERT-wwm	34.68
PCL-BERT-wwm-Post	35.71

表2. 不同结构预测结果

Model	F1	Rec.	Acc.
RNN_CNN	33.29	36.62	30.53
Last_MaxPool	33.17	37.11	29.98
Last3_Embedding	34.43	38.22	31.32
Last3_Attention	34.79	38.82	31.51
Last3_MulDropout	35.30	37.74	33.16
Last3_Atten_MulDrop	36.39	41.23	32.56

不同主干网络预测结果如表1所示，符号-Post代表在收集的数据中使用Mac-BERT预训练方法进行继续预训练。可以观察到，BERT-base-chinese和RoBERTa-wwm-ext具有相似的预测结果。当主干网络被 RoBERTa-large取代后，F1改进约为1.62。我们最后选择 PCL-BERT-wwm 作为我们的基线主干网络，继续预训练后F1提升1.03。我们还尝试了具有最高主干的不同生成模型结构，如表3所示。结果表明利用最后三层embedding融合并加入注意力机制和multidropout方法时，模型表现具备相当的竞争优势。。

表3. 对话生成模型结果

	Test A (Dev)			Test B		
Model	Avg.	F1	BLEU	Avg.	F1	BLEU
Transformer [16]	15.40	24.71	6.09	-	-	-
GPT2 [17]	16.56	25.75	7.37	-	-	-
BertGPT [6]	16.80	26.57	7.03	-	-	-
+ Curriculum 5-fold	19.42	28.72	10.12	19.16	28.34	9.99
T5-pegasus-small [18]	16.55	23.76	9.34	-	-	-
T5-pegasus-base [18]	17.42	25.41	9.43	-	-	-
T5-pegasus-base-Post [8] [18]	17.58	25.55	9.61	-	-	-
CPM2-prompt [19]	18.21	26.38	10.04	18.94	27.10	10.78
Logit-ensemble	19.83	28.92	10.74	20.73	29.54	11.92
+ Context Embedding	20.02	28.74	11.30	21.03	29.52	12.54
+ Encoding Fusion	20.43	29.36	11.50	21.30	29.87	12.72
Final Results (Para. Tuning)	21.24	30.12	12.36	21.83	30.57	13.09

不同对话生成模型的结果如表3所示，原始Transformer[16]的性能比较差，这可能是编码长度限制了GPT2[17]的能力（用小词汇表）。我们利用 BertGPT [6] 进行课程5-fold训练，平均得分在原有基础上的高出2.62。随着预训练模型

规模的增加，各项指标成绩呈上升趋势。微调后的 CPM2-prompt[19]取得了最高平均分的成绩，达到18.21分。对比我们提出的不同结构，所提出的上下文编码模块有利于提高 BLEU分数，因为历史实体对于响应生成同样重要。编码融合模块也有效地提高了0.62的F1分数和0.2的BLEU分数，最后经过模型融合与参数精调，我们在A、B榜我们取得了第一的成绩。

4 结论

在本文中，我们提出了一个管道式的医疗对话生成框架，包括两部分：医疗实体预测和实体感知融合对话生成。在我们的框架中，我们首先使用F1阈值搜索优化实体预测模型。同时基于Seq2Seq架构提出了医疗感知对话生成模型，在编码端融入了上下文与历史实体编码，同时在解码端将不同的来源的信息进行融合，采用多样性集束搜索提升回复的多样性，最后采用模型集成的方式提高了最终的结果。我们在CCKS2021蕴含实体的中文医疗对话生成比赛中获得了第一名，这证明了我们提出的方案的有效性和实用性方法。在未来，我们将考虑使用知识图谱来推断预测实体，并在生成时尝试不同的融合策略，以进一步提高生成响应的正确性和质量。

参考文献

1. Wei, Zhongyu, et al. Task-oriented dialogue system for automatic diagnosis, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 2, 2018, pp. 201 - 207.

2. Xu, Lin, et al. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33. no. 01. 2019.

3. Zeng, Yang, et al. Meddialog: A large-scale medical dialogue dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9241-9250. 2020.

4. Liu, Tang, et al. MedDG: A Large-scale Medical Consultation Dataset for Building Medical Dialogue System. arXiv preprint arXiv:2010.07497 (2020).

5. Lin, Shuai, et al. Graph-Evolving Meta-Learning for Low-Resource Medical Dialogue Generation. arXiv preprint arXiv:2012.11988 (2020).

6. Lewis, Mike, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. (pp. 7871 - 7880).

7. Cohen, Eldan, et al. Empirical analysis of beam search performance degradation in neural sequence models. International Conference on Machine Learning. (pp. 1290-1299). PMLR, 2019.

8. Vijayakumar, Ashwin K., et al. Diverse beam search: Decoding diverse solutions from neural sequence models. arXiv preprint arXiv:1610.02424 (2016).

9. Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

10. Liu, Yinhan, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).

11. Cui, Yiming, et al. Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101 (2019).

12. Cui, Yiming, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. 2020. (pp. 657-668).

13. Gururangan, Suchin, et al. Don't Stop Pretraining: Adapt Language Models to Domains

and Tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. (pp. 8342–8360).

14. Chen, Boxing, and Colin Cherry. A systematic comparison of smoothing techniques for sentence–level bleu. Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014. (pp. 362–367).

15. Miyato, Takeru, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi–supervised text classification. arXiv preprint arXiv:1605.07725 (2016).

16. Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems. 2017. (pp. 5998–6008).

17. Radford, Alec, et al. Language models are unsupervised multitask learners. OpenAI blog 1.8 (2019): 9.

18. Raffel, Colin, et al. Exploring the Limits of Transfer Learning with a Unified Text to Text Transformer. Journal of Machine Learning Research 21.140 (2020): 1–67.

19. Zhang, Zhengyan, et al. CPM–2: Large–scale Cost–effective Pre–trained Language Models. arXiv preprint arXiv:2106.10715 (2021).