arXiv:2303.01341v1 [cs.CL] 2 Mar 2023

面向口语化患者查询理解的基于匹配的术语语义预训练

Zefa Hu[1,2]*,Xiuyi Chen* [1, 2], *Haoran Wu*[1,2]*,Minglun Han*[1,2]*,Ziyi Ni*[1,2]*Jing Shi*[2]*,Shuang Xu*[2]*,Bo Xu*[1,2]

[1]中国科学院大学人工智能学院，北京，中国[2]中国科学院自动化研究所，北京，中国

{huzefa2018, chenxiuyi2017, wuhaoran2018, hanminglun2018, niziyi2021}@ia.ac.cn{shijing2014, shuang.xu, xubo}@ia.ac.cn

摘要

医学槽位填充（MSF）任务旨在将医学查询转化为结构化信息，在诊断对话系统中发挥着重要作用。然而，由于缺乏充分的术语语义学习，现有方法难以捕捉医疗对话中语义相同但表达口语化的术语。在这项工作中，我们将 MSF 形式化为一个匹配问题，并提出术语语义预训练匹配网络（TSPMN），将术语和查询同时作为输入，以建模二者之间的语义交互。为了更好地学习术语语义，我们进一步设计了两个自监督目标：对比术语判别（CTD）和基于匹配的掩码术语建模（MMTM）。对于每个给定术语，CTD 判断它是否是对话中的掩码术语，而 MMTM 则直接预测被掩码的术语。在两个中文基准数据集上的实验结果表明，TSPMN 优于强基线方法，尤其是在少样本设置下[1]。

索引词— 医学对话、口语语言理解、槽位填充、低资源、预训练

1. 引言

医学槽位填充（MSF）旨在通过检测医学术语，将医学查询自动转换为结构化信息，近年来受到越来越多的关注[1–3]。它在诊疗对话系统中发挥着至关重要的作用 [4,5] 。与自然语言处理领域的传统槽位填充任务不同，后者会标注给定话语中的显式词语，并基于这些标注词语提取结构化信息（即槽值对） [6–9], , MSF 中患者查询与相应提供的医学术语槽位之间存在非对齐问题[2, 3, 10–12]。具体而言，患者查询中术语的口语化表达与正式表达有所不同。如表 1 所示，槽值“症状：腹痛”并未

患者查询
这些天我的胃感觉不舒服，该部位疼痛肚脐上方，每天排便两次，腹部鼓胀，不成形，目前正在服用头孢克肟，这是怎么回事?
我这几天肚子感觉难受，肚脐眼上面的位置疼痛，一天大便两次，肚子胀，不成形，目前在吃头孢克肟，这是怎么回事呢?
槽位-值对标签
症状：腹痛（腹痛）
症状：稀便（稀便）
症状：腹胀（腹胀）
药物：头孢克肟（头孢克肟）

表 1。患者查询及由槽位-值对组成的标签示例（例如，症状：稀便）。

在任何特定片段中明确出现，而是在查询中被隐含提及。因此，MSF 需要借助医学知识，更深入地理解术语语义。此外，医学数据更加依赖专家标注，而实际获取标注的成本很高，因此标注数据尤为匮乏。

近期已有研究 [2, 3, 13] 提出解决上述问题的方法，大致可分为两类：多标签分类和生成式方法。第一类方法 [2, 3] 将预定义的槽位-值对视为不同类别。它们可以利用带有医生回复的未标注患者查询生成伪标签，以进行数据增强 [14]。然而，这要求未标注数据必须映射到有限的预定义术语，因此难以利用规模更大的未标注医疗对话语料库。第二类方法 [13] 通常通过对话提示将 MSF 建模为响应生成任务。这样，MSF 便能受益于利用大型未标注医学对话语料库进行的对话式预训练。然而，MSF 与响应生成任务之间的差异不可避免地会削弱性能。此外，由于模型按顺序生成术语，累积的错误会传播到后续步骤 [15]。

与这些方法不同，我们提出了一种术语语义预训练匹配网络（TSPMN），它同时将

本研究得到中国科学院重点项目（编号：ZDBS-SSW-JSC006-2）和国家自然科学基金（编号：62206294）资助。[1]我们的代码可在 https://github.com/FlyingCat-fa/TSPMN 查阅。

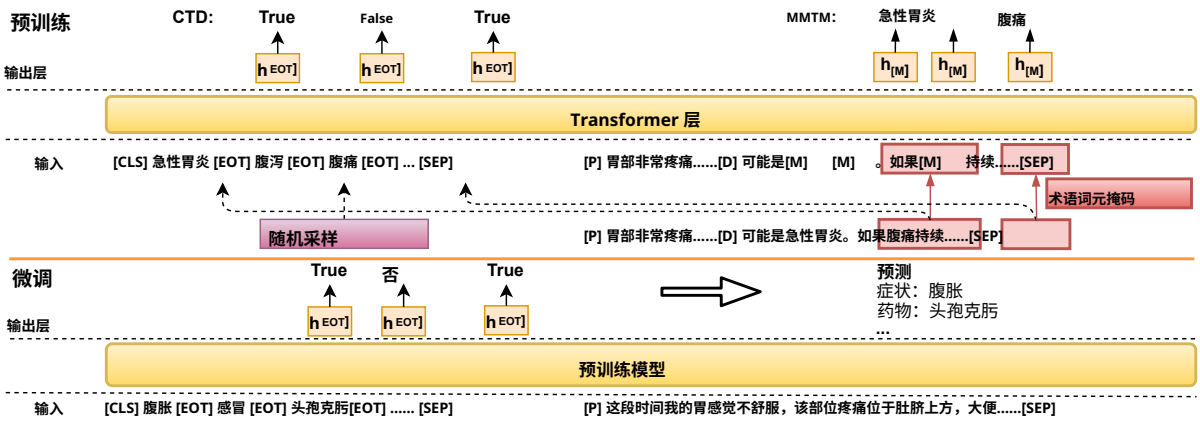

图 1。术语语义预训练匹配网络（TSPMN）示意图。输入由术语序列和医疗对话/患者查询组成。［EOT］是前一个术语的分隔符。［P］和［D］分别表示患者和医生。TSPMN首先通过我们的自监督任务学习术语语义。随后，对预训练的TSPMN进行微调，以匹配候选术语和患者查询，从而实现 MSF。请注意，所有示例均由中文翻译而来，微调示例取自表 1。

术语和查询作为输入，并将术语作为与任务相关的知识［16, 17］。因此，模型只需学习如何匹配查询与给定术语，从而降低了对数据的要求。此外，我们为 TSPMN 提出了两种自监督任务，以更好地学习术语语义，包括对比术语判别（CTD）和基于匹配的掩码术语建模（MMTM）。CTD 是一种与 MSF 相近的匹配任务，而 MMTM 则是一种自适应的掩码语言建模（MLM）任务，通过与金标准词元进行匹配，更好地预测被掩码的术语词元。通过这种方式，TSPMN 不仅能够利用大规模医疗对话语料进行预训练，还能缩小预训练阶段与微调阶段之间的差异。在两个中文基准数据集上的实验结果表明，TSPMN 的表现优于强基线方法，尤其是在少样本设置下。

2. 问题陈述

给定一个包含口语化表达的患者查询 q，医学槽位填充（MSF）任务旨在将查询 q 转换为具有离散逻辑形式（槽位：值）的、基于实际内容的形式化表示。槽位和值的候选项根据医学知识图谱预先定义，其中值为医学术语，槽位表示该术语的类别（例如，（症状：腹痛））。我们将 MSF 形式化为一个匹配问题，即将每个术语候选项与患者查询 q 进行匹配，以确定该候选项是否出现在 q中。

3. 方法

本节介绍术语语义预训练匹配网络（TSPMN）如何通过匹配术语与患者查询来对医学槽位填充（MSF）进行建模。随后将介绍 TSPMN 的两项术语语义预训练任务。

3.1. MSF 的匹配

为实现高效匹配，同时考虑模型输入的长度限制，我们首先通过拼接术语集合 T 中的术语来构建多个术语序列。每个术语序列的形式如下：

$$S_t = (T_1, \texttt{[EOT]}, \ldots, T_i, \texttt{[EOT]}, \ldots, T_n, \texttt{[EOT]}), \quad (1)$$

其中 n 和 ［EOT］分别表示术语编号和每个术语后面的分隔符，T_i 表示第 i个术语的词元。给定一个患者查询，我们将每个术语序列与查询拼接成一个完整序列 x，并使用预训练语言模型 $\mathcal{H}$（例如 BERT ［18］）对 x进行编码，以充分捕获语义信息：

$$(\mathbf{h}_{\texttt{[CLS]}}, \ldots, \mathbf{h}_{\texttt{[EOT]}}, \ldots, \mathbf{h}_{\texttt{[SEP]}}) = \mathcal{H}(x). \quad (2)$$

我们将每个隐藏状态$\mathbf{h}_{\texttt{[EOT]}}$ 作为 ［EOT］之前术语的隐藏状态。然后，按如下方式预测 T_i是否出现在查询中的概率：

$$p_i(x;\theta) = \text{Softmax}(\text{FFN}(\mathbf{h}_{T_i})), \quad (3)$$

其中，$\mathbf{h}_{T_i}$ 表示术语 i 的隐藏状态，θ 是模型参数。我们通过前馈神经网络将$\mathbf{h}_{T_i}$ 分别映射为 True 和 False 的分数，这分别表示术语 i 是否在查询中被提及（True 表示提及，False 表示未提及）。$p_i(x;\theta) \in \mathbb{R}^2$ 表示经 softmax函数归一化后的分数。如果 True 的归一化分数大于 False，我们就选择术语 i 来填充相应的槽位。

MSF 损失函数定义如下：

$$\mathcal{L}_{MSF} = -\sum_{i=1}^{n}\sum_{k} y_{i,k}\log p_{i,k}, \quad k \in \{0,1\}, \quad (4)$$

其中， n 和 k 分别表示术语数量和真或假的索引，$y_i \in \{[1,0],[0,1]\}$ 是一个标签，用于指示 T_i 是否出现在患者查询中。

3.2. 术语语义预训练

预训练语言模型（PrLM）在许多任务中表现出色 [18, 19]。此外，也有许多用于对话表示 [20–23] 或医疗领域适配 [24–26]的 PrLM。受这些工作的启发，我们聚焦于医疗对话中的口语理解，并提出两项自监督任务，以更好地建模术语语义。这些任务不仅可以利用大规模医疗对话语料库，还能尽可能缩小这些任务与 MSF 之间的差距。具体细节如下。

3.2.1. 用于预训练的匹配

我们使用三个公开的无标注医疗对话数据集 MedDialog [27], 、KaMed [28], 和 ReMeDi-large [29], 作为预训练语料库，其中包含超过 350 万条对话，涉及 100 多个医学科室。我们将公开的搜狗拼音医学词典[2] 和 THUOCL 医学词典 [30]合并，构建了一个大型医学术语词典 T_{large}。预训练语料库知识库中的术语也被加入其中 T_{large}。基于 T_{large}，我们通过字符串匹配检索每条医疗对话中的术语，并构建对话-术语对用于预训练。与第 3.1 节中针对 MSF 的匹配方式类似，我们通过拼接术语和对话作为输入，构建多个术语序列。具体而言，每个术语序列由从当前对话中采样的正术语以及不在当前对话中的负术语组成。与 MSF 的不同之处在于，我们只对采样的正术语进行掩码。这样，模型便能从对话上下文中学习术语语义，而非仅仅关注字符串匹配。我们将输入记为 x_{mask}，并按照公式 2 的相同方式对其进行编码：

$$(\mathbf{h}_{[\mathrm{CLS}]}, ..., \mathbf{h}_{[\mathrm{EOT}]}, ..., \mathbf{h}_{[\mathrm{M}]}, ..., \mathbf{h}_{[\mathrm{SEP}]}) = \mathcal{H}(x_{mask}), \quad (5)$$

其中， M 表示术语中的一个掩码词元。隐藏状态用于我们的两个自监督任务。这两个任务共享相同的输入，并通过多任务学习鼓励预训练模型捕捉语义的不同方面。我们将预训练损失定义为前述两项损失之和：

$$\mathcal{L}_{pretrain} = \lambda \mathcal{L}_{CTD} + (1-\lambda)\mathcal{L}_{MMTM}, \quad (6)$$

其中， λ 是一个可调权重，用于调整不同损失的贡献，$\mathcal{L}_{CTD}$和 $\mathcal{L}_{MMTM}$分别是 CTD 和 MMTM 的损失。这两个任务的详细信息将在下一小节中介绍。

[2]https://pinyin.sogou.com/dict/detail/index/ 15125，更新于 2017 年 10 月 13 日。

3.2.2. 自监督任务

对比术语判别。 对于术语序列中的每个术语，CTD 旨在判断其是否属于当前对话。与第 3.1 节中的 MSF 类似，我们使用 [EOT]（$\mathbf{h}_{T_i} = \mathbf{h}_{[\mathrm{EOT}]_i}$）的隐藏状态来表示与患者查询匹配后的前置术语。操作与公式 3 和 4 相同。

基于匹配的掩码术语建模。 该任务的灵感来自掩码语言建模（MLM） [18] ，并进行了两项改进以匹配 MSF：1）MMTM 仅对医学术语进行掩码；2）掩码始终出现在 T_{large} 中。因此，模型不仅能从对话上下文中学习语义，还能更充分地对术语与对话之间的语义交互进行建模。$\mathbf{h}_{[\mathrm{M}]}$ 用于预测掩码。我们计算与 MLM 相同的交叉熵损失。

4. 实验

4.1. 数据集与评估指标

数据集	训练集	Dev	Test	Slot	值（术语）
MSL	1152	500	1000	1	29
MedDG	50965	6956	3645	4	155

表 2。MSL 和 MedDG 数据集的统计信息。

我们在两个中文医疗数据集上评估了我们的方法：MSL [2]和 MedDG [1]。MedDG 最初是为医疗对话系统构建的，并标注了医学槽位，可用于医学槽位填充（MSF）。数据集的统计信息如表 2 所示。评估时，我们遵循 MSL 指南 [2] ，采用以下各项指标：**精确率**、**召回率**、**微平均F1**、**宏平均F1**和**准确率**。

4.2. 实现细节

我们使用中文 BERT-base [18]初始化模型。预训练阶段，批量大小为 48，并使用 1 位 Adam [31] 作为优化器。我们分别将学习率和预训练轮次设为 3×10^{-5} 和 5。并将λ 设为 0.9。微调阶段，我们使用 AdamW [32]作为优化器，初始学习率为 1×10^{-5}。MedDG 的批量大小为 8，MSL 的批量大小为 32。我们将每个术语序列的术语数量 n 分别设为 20、15 和 20，对应预训练、在 MSL 上微调和在 MedDG 上微调。

4.3. 主要结果

完整训练评估。从表 3 可以看出，我们的模型取得了新的最先进结果。与基线方法相比，所有指标的提升均具有统计显著性， $p < 0.05$ 来自显著性检验。与在医学语料库上增强或预训练的分类方法 BERT+TST 和基于生成的方法 PromptGen 相比，

模型	MSL					MedDG				
	P	R	微平均 F1	宏平均 F1	Acc	P	R	微平均 F1	宏平均 F1	Acc
深度循环神经网络 [2]†	83.43	67.85	74.83	65.17	52.5	96.55	97.34	96.95	82.8	73.39
深度循环神经网络+A [2]†	82.11	70.86	76.07	67.42	51.9	98.53	96.69	97.6	83.62	75.25
深度循环神经网络+A+WS [2]‡	82.94	79.44	81.15	76.95	58.3	-	-	-	-	-
文本卷积神经网络-原始 [3]†	90.37	64.31	75.14	64.28	51.6	97.5	95.65	96.57	80.64	72.57
BERT-原始 [3]†	89.78	88.63	89.2	87.03	70.9	99.3	99.38	99.34	84.82	74.15
BERT+TST [3]‡	90.95	90.81	90.88	89.28	72.9	-	-	-	-	-
PromptGen [13]§	89.11	87.57	88.34	87.75	79.6	-	-	-	-	-
TSPMN-MedBERT	90.91	**91.11**	91.01	90.45	81.4	99.38	**99.5**	99.44	86.51	98.44
TSPMN	**92.33**	90.66	**91.49**	**90.62**	**83.4**	**99.61**	99.43	**99.52**	**86.55**	**98.77**
TSPMN（无 MMTM）	91.14	90.66	90.90	89.74	82.80	99.45	99.43	99.44	86.34	98.44
无预训练	85.76	88.40	87.06	86.36	74.50	99.21	99.28	99.25	86.00	97.86

表 3。在 MSL 和 MedDG 数据集上的完整训练评估。 †：我们引用了这些模型在 MSL 上的原始论文结果 [2,3], ，并基于其发布的代码获得了它们在 MedDG 上的结果。 ‡：这些模型需要同源的未标注数据。§：由于作者未发布代码，我们引用了 PromptGen 在 MSL 上的原始论文结果 [13]。

模型	1 样本			2 样本			5 样本		
	微平均 F1	宏平均 F1	Acc	微平均 F1	宏平均 F1	Acc	微平均 F1	宏平均 F1	Acc
深度循环神经网络+A	20.66±3.32	11.63±3.15	7.16±1.88	33.69±3.38	27.96±4.89	11.74±1.15	52.83±4.84	50.24±5.3	23.74±3.56
深度循环神经网络+A+WS	70.55±1.62	64.18±4.06	42.08±1.51	71.14±1.09	64.84±0.67	42.96±2.06	75.11±0.87	70.56±1.36	46.74±0.87
BERT-原始	14.74±4.44	6.93±3.78	4.56±1.99	41.76±13.71	31.25±17.21	17.84±6.29	73.62±1.8	68.3±2.93	46.8±2.14
BERT+TST	71.68±3.19	62.5±5.68	49.76±2.56	72.76±1.06	62.55±3.51	51.46±1.9	77.55±0.68	71.08±1.16	55.66±1.63
TSPMN-MedBERT	78.19±0.86	77.81±0.25	53.32±1.6	79.34±0.93	79.1±1.43	55.76±2.14	83.34±1.04	83.27±1.13	64.08±1.96
TSPMN	**78.52±2.65**	**79.32±1.19**	**55.78±4.61**	**81.87±1.21**	**81.93±1.56**	60.5±2.18	**84.74±0.83**	83.96±0.59	**67.28±1.79**
不含 MMTM	77.82±2.32	78.24±1.47	55.68±3.63	81.37±1.2	81.49±0.59	**61.2±1.86**	84.22±0.83	**84±0.91**	66.08±1.84
不使用预训练	75.12±1.3	74.69±0.77	48.78±2.45	77.59±0.45	76.87±0.48	53.86±0.98	81.61±1.64	81.64±1.46	60.78±3.13

表 4。MSL 上的少样本评估。报告了五次运行的均值和标准差。

我们的 TSPMN 始终取得更好的表现，我们认为这是因为 TSPMN 利用了大规模无标注医学对话语料库，并缩小了预训练阶段与微调阶段之间的差异。为进一步分析，我们移除 MMTM 目标，再同时移除 MMTM 和 CTD，分别记为 TSPMN（无 MMTM）和 TSPMN（无预训练）。此外，我们用 BERT 原始目标替换自监督目标，并将其记为 TSPMN-MedBERT。从预训练语料库的角度来看，TSPMN 和 TSPMN-MedBERT 的表现均优于 TSPMN（无预训练），这说明在大规模医疗对话语料库上进行持续预训练至关重要。从预训练任务的角度来看，对比 TSPMN 和 TSPMN-MedBERT 可知，预训练任务越接近目标任务，性能提升越明显。消融实验也验证了 CTD 和 MMTM 的有效性。

少样本评估。 我们进一步在更具挑战性的少样本设置下评估了我们的方法。在 k-shot 设置下，我们从原始训练集中为每个术语选取 k 个训练样本，以构成训练数据集。少样本评估仍使用初始验证数据和测试数据。如表 4 所示，在少样本设置下，TSPMN 相较于基线方法取得了更大的性能提升。此外，我们发现即使不进行预训练，TSPMN 在大多数指标上的表现仍优于基线方法，这验证了 TSPMN 的新颖匹配范式在低资源场景中的优越性。如表 3 和表 4 所示，与 TSPMN-MedBERT 相比，TSPMN 在少样本设置下取得的相对提升大于全量训练设置下的提升，这表明在低资源场景中，预训练阶段与微调阶段之间较小的差异更为重要。

5. 结论

患者与专业医疗服务提供者所使用术语的复杂程度各不相同，因此需要更深入、更丰富的语义理解，这一直是医学槽位填充（MSF）任务中的棘手问题。为全面学习术语语义，本文提出了术语语义预训练匹配网络（TSPMN），并采用两种自监督目标：对比术语判别（CTD）和基于匹配的掩码术语建模（MMTM）。我们通过详细实验揭示了 TSPMN 及所提出训练目标的优越性。本文的局限在于，仅将医学槽位的值视为术语，忽略了在更复杂场景中与术语相对应的可能状态；这将留待未来工作探讨。

6. 参考文献

[1] Wenge Liu 等人，“Meddg：用于构建医疗对话系统的大规模医疗咨询数据集”，*arXiv* 预印本，2020 年。[2] Xiaoming Shi 等人，“利用分散关键词注意力以及来自回复的弱监督理解医疗对话”，载于*AAAI* 论文集，2020 年。[3] Xiaoming Shi 等人，“利用来自医生回复的弱监督理解患者查询”，*IEEE* 生物医学与健康信息学汇刊，2021 年。[4] Zhongyu Wei 等人，“用于自动诊断的任务导向对话系统”，载于*ACL* 论文集，2018 年。[5] Mina Valizadeh 和 Natalie Parde，“AI 医生在线：面向医疗保健应用的任务导向对话系统综述”，载于*ACL* 论文集，2022 年。[6] Gr´egoire Mesnil 等人，“利用循环神经网络进行口语语言理解中的槽位填充”，*IEEE/ACM* 音频、语音与语言处理汇刊，2015 年。[7] Ngoc Thang Vu 等人，“采用排序损失的双向循环神经网络用于口语语言理解”，载于*ICASSP* 论文集，2016 年。[8] Qian Chen 等人，“用于联合意图分类和槽位填充的 BERT”，*CoRR*，2019 年。[9] Libo Qin 等人，“用于联合槽位填充和意图检测的协同交互式 Transformer”，载于*ICASSP* 论文集，2021 年。[10] Deyu Zhou 和 Yulan He，“从未对齐数据中学习用于自然语言理解的条件随机场”，载于*ECIR* 论文集，2011 年。[11] Lina Maria Rojas-Barahona 等人，“在口语语言理解的深度神经模型中利用句子和上下文表示”，载于*COLING* 论文集，2016 年。

[12] Lin Zhao 和 Zhe Feng，“通过联合指针和注意力机制改进口语语言理解中的槽位填充”，载于*ACL* 会议论文集，2018 年。

[13] Jun Liu 等人，《面向多层级医疗对话状态跟踪的基于提示的生成方法》，*CoRR*，2022。[14] Linghui Meng 等人，《Mixspeech：面向低资源自动语音识别的数据增强》，载于*ICASS P*，2021。[15] Xiujun Li 等人，《基于强化学习的对话系统中语言理解影响的研究》，*CoRR*，2017。

[16] Minglun Han 等人，《用于端到端上下文语音识别的基于 CIF 的协同解码》，载于 *ICASS P*，2021。[17] Minglun Han 等人，《通过细粒度上下文知识选择改进端到端上下文语音识别》，载于*ICASS P*，2022。[18] Jacob Devlin 等人，《BERT：用于语言理解的深度双向 Transformer 预训练》，载于*AACL*，2019。[19] Linghui Meng 等人，《离线预训练多智能体决策 Transformer：一个大型序列模型征服所有《星际争霸 II》任务》，*arXiv* 预印本，2021。[20] Shikib Mehri 等人，《对话上下文表示学习的预训练方法》，载于*ACL*，2019。[21] Zhuosheng Zhang 等人，《对话理解的结构化预训练》，载于*ACL*，2021。[22] Yi Xu 和 Hai Zhao，《面向对话的预训练》，载于*ACL Findings*，2021。[23] Wanwei He 等人，《Galaxy：一种结合半监督学习和显式策略注入的面向任务对话生成式预训练模型》，载于*AAAI*，2022。[24] Ningyu Zhang 等人，《面向中文生物医学文本挖掘的概念化表示学习》，*arXiv* 预印本，2020。[25] Benyou Wang 等人，《生物医学领域的预训练语言模型：系统综述》，*arXiv* 预印本，2021。[26] Taolin Zhang 等人，《Smedbert：一种用于医疗文本挖掘、具有结构化语义的知识增强预训练语言模型》，载于*ACL*，2021。[27] Guangtao Zeng 等人，《MedDialog：大型医疗对话数据集》，载于*EMNLP*，2020。[28] Dongdong Li 等人，《用于医疗对话生成的半监督变分推理》，载于*SI-GIR*，2021。[29] Guojun Yan 等人，《Remedi：多领域、多服务医疗对话资源》，载于*SIGIR*，2022。[30] Shiyi Han 等人，《Thuocl：清华开放中文词库》，清华大学，2016。[31]Hanlin Tang 等人，《1 位 Adam：以 Adam 的收敛速度实现高效通信的大规模训练》，载于*ICML*，2021。[32] Ilya Loshchilov 和 Frank Hutter，《解耦权重衰减正则化》，载于*ICLR*，2019。