arXiv:2305.01195v1 [cs.CL] 2 May 2023

中文对话中的话题转移检测：语料库与基准

Jiangyi Lin[1],Yaxin Fan[1],Feng Jiang[2],Xiaomin Chu[1],和 Peifeng Li*[1]

[1] 苏州大学计算机科学与技术学院，中国苏州 {jylin, yxfa nsuda}@stu.suda.edu.cn，{xmchu, pfli}@suda.edu.cn[2] 香港中文大学（深圳）数据科学学院，中国深圳 jeffreyjiang@cuhk.edu.cn

摘要。 对话主题转移检测旨在判断对话中正在进行的话题是否已经转移或是否应该转移，可分为两类，即已知回复任务和未知回复任务。目前，只有少数研究探讨了后者，因为在没有响应信息的情况下预测话题转移仍然是一项挑战。本文首先标注了一个由 1308 段对话组成的中文自然话题对话（CNTD）语料库，以填补中文自然对话话题语料库方面的空白。随后，我们聚焦于未知回复任务，提出一种基于层次对比学习的教师-学生框架，以在没有回复的情况下预测话题转移。具体而言，在高层教师-学生框架中引入回复，以构建回复与上下文之间的对比学习；在低层学生模型中则构建标签对比学习。在中文 CNTD 和英文 TIAGE 上的实验结果表明，我们提出的模型有效。

关键词： 对话主题转移检测 · 层次对比学习 · 中文自然话题对话语料库。

1 引言

对话主题转移检测旨在检测对话中的话语是否发生了话题转换，这有助于对话系统切换话题并主动引导对话。尽管对话主题转移检测是一项新任务，但它因能显著助益许多下游任务而成为热点，例如回复生成 [1] 和阅读理解 [2,3], ，还能帮助这些实时应用生成切题或话题转换后的回复，从而在对话场景中表现出色 [4,5,6]。

对话话题转移检测任务可分为两类，即已知回复任务和未知回复任务，如图 1 所示。前者可以获取响应信息，从而取得更好的结果；后者则相反。此外，这两类任务都无法获取未来信息。这是它们与文本话题分割任务的最大区别，后者中

* 通讯作者。

所有基本话语都彼此可见。也就是说，由于现有的主题分割模型高度依赖回复及其后续话语，因此无法应用于对话主题转移检测。因此，在对话主题转移检测任务中，更难辨别话语之间的差异。由于缺少未来话语，对话主题转移检测仍是一项具有挑战性的任务。

本文聚焦于中文对话中话题转移检测的未知回复任务。中文对话中话题转移检测的未知回复任务存在两个问题：缺乏中文标注语料库，以及如何预测回复。

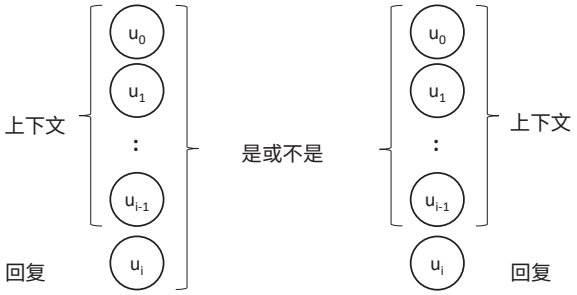


图1. 两条对话展示了话题转移检测任务：判断话语 u_{i-1} 和 u_i之间是否发生了话题转换。其中，已知回复任务（左）可以利用回复 u_i，而未知回复任务（右）可视为不依赖回复 u_i的话题转移预测。

目前可公开获取的对话主题转移语料库寥寥无几，而且大多数语料库都是为分割任务提供的，无法满足自然对话的需求。Xie 等人 [7] 对对话主题转移检测任务作出了详细定义，并标注了英语对话主题语料库 TIAGE。尽管该语料库填补了英语对话主题语料库方面的空白，但其规模仍然太小。在中文领域，Xu 等人 [8] 标注了一个中文对话主题语料库。然而，由于规模较小且质量较差，这不利于中文对话主题转移任务的进一步研究与发展。为填补中文自然对话主题语料库方面的空白，我们首次标注了一个高质量的中文自然主题对话语料库（CNTD），其中包含 1308 段对话。

Xie 等人 [7] 还基于 T5 模型 [9] 为这一未知回复任务建立了一个基准。该基准仅使用上下文来预测话题转换，由于缺乏响应信息，表现不佳。因此，在缺少有用响应信息的情况下预测自然对话中的话题转换更具挑战性。

教师-学生框架已被广泛用于获取模型 [1]无法获得的信息。为解决缺少回复信息的问题，我们提出了一种教师-学生框架来引入回复。

信息。教师可以获取响应信息，学生可以通过知识蒸馏从教师那里学习响应信息。为了促进知识迁移，学生会模仿教师的每一层，而不只是顶层，从而利用教师 [10]中的层次语义信息缓解监督信号延迟问题。

此外，我们构建了层次对比学习，将教师-学生视为高层，将学生视为低层。在高层，我们在上下文与回复之间构建信息模拟损失，以利用更可靠的预测信息提升学生模型的语义信息。在低层，我们设计了语义连贯性感知损失，以更好地区分不同的话题转换情形，并生成更可靠的预测结果。

最后，在我们的中文 CNTD 和英文 TIAGE 上进行的实验结果表明，我们提出的模型优于各基线模型。本文的贡献如下。

– 我们基于 NaturalConv 手动标注了一个包含 1308 段对话的语料库，以填补中文自然对话话题语料库的空白。– 我们提出了一个教师-学生框架，以学习话题转移检测任务中的响应信息。– 我们引入层次对比学习，进一步提升性能。– 在 CNTD 和 TIAGE 数据集上的实验结果均表明，我们的模型优于各基线模型。

2 相关工作

2.1 语料库

先前的研究探讨了对话话题相关任务，并发布了经过标注的话题对话语料库。对于英语，Xie 等人 [7]基于 PersonaChat [11]标注了 TIAGE，该语料库包含 500 段对话和 7861 个轮次。Xu 等人 [8] 通过整合现有多轮对话数据集中的对话，构建了一个包含 711 段对话的数据集：MultiWOZ 语料库 [12],和 Stanford Dialog 数据集 [13]。这两个语料库要么规模较小，要么仅限于特定领域，且均不适用于自然对话领域的研究。

对于中文，Xu 等人 [8] 标注了一个数据集，其中包含 505 条银行业务咨询客服电话录音。然而，该语料库同样局限于少数专业领域，而自然对话则更加复杂。自然对话具有多种话题转换场景、不受限制的话题，以及话语中更加自由的口语表达。上述语料库不足以弥补中文自然对话话题语料库方面的空白。

2.2 对话中的话题转换检测

对话主题转移检测任务也处于起步阶段，只有少数研究关注这一任务。正如我们在引言中所提到的，话题分段是一项类似的任务。因此，我们首先介绍对话中话题分段的相关工作

。由于缺乏对话训练数据，早期的对话主题分段方法通常采用无监督方法，利用词共现统计 [14] 或句子主题分布[15] 来衡量不同轮次之间的句子相似度，从而检测主题或语义变化。近年来，随着从维基百科采样的大规模语料库问世，单体文本主题分段的监督方法发展迅速。这类方法使用部分词元作为真实分段边界，尤其是基于神经网络的方法 [16,17,18]。由于性能更稳健、效率更高，这些监督方法受到研究人员的青睐。

对话话题转移检测与对话话题分割有很大不同。对于对话主题检测任务，Xie 等人 [7] 提出了详细的定义，将任务分为两类：同时考虑上下文和回复的已知回复任务，以及仅考虑上下文的未知回复任务。然而，仅基于上下文的方法仍然鲜见。只有 Xie 等人[7] 基于 T5 模型预测是否发生话题转换。Sun 等人 [19] 引入结构和语义信息，帮助模型检测在线讨论中的话题转移，这与已知回复任务类似。解决对话主题转移检测问题至关重要。总体而言，对话主题转移检测任务仍然具有挑战性，因为它只能依赖对话的上下文信息。本文通过利用知识蒸馏和层次对比学习，弥补了响应信息的缺失。

3 语料库

现有的中文对话主题检测语料库 [8] 规模较小，也不符合自然对话的特点。尽管可以通过机器翻译将英语对话话题语料库转换为中文，但这类语料库缺乏自然对话的口语特征，且规模较小。因此，我们基于 NaturalConv数据集 [20]标注了一个中文对话主题检测语料库 CNTD。

本节将介绍我们的标注指南，说明选择语料来源的原因，以及人工标注流程和数据统计。我们还将分析 CNTD 中的话题转移分布。

3.1 优势

我们语料库中的每段对话都有一篇新闻作为基础文档，而其他语料库并不具备这一特点；这篇新闻可作为额外信息，用于进一步研究和拓展。新闻来自六个领域，使我们的对话更贴近自然对话。此外，我们语料库中的说话者不受任何限制，这也使其更接近自然对话。另外，我们还对对话主题进行了细粒度标注，详见第 3.2 节。细粒度标签有助于推动对话主题的进一步研究。

与 Xu 等人标注的现有中文主题语料库相比， [8],我们的语料库中的对话没有无意义和重复的轮次。此外，该语料库的规模是另一语料库的两倍多。另外，语料库中的新闻也可作为对话的补充信息进行研究。

3.2 标注指南

遵循 TIAGE 中的标注指南，[7],我们判断每个对话轮次相较于上下文是否发生了话题变化。在对话中，说话人对对话上下文的回复通常属于以下情况之一，相关示例见表 1。

T表 1. 对话中不同的回复场景 。

B	“最近在玩掌上游戏《灌篮高手》， 回想起那段看动漫的时光。”
A	“我看到很多平台都在推这款游戏啊，不过 我还没玩过。”
B	“是啊，这游戏挺有意思的，等有空了你可以玩玩 再说。还有动漫剧集。”→ 不是话题转换
	（a）评论先前语境。
A	“你的孩子上几年级了？”
B	“他是大学一年级学生。” → 不是话题转换
	(b) 问答。
A	“拉瓦尔杯即将开始，欧洲队 有两位国王，Federer 和 Nadal。”
B	“我知道 Federer ，他是网球界最优秀的选手之一 world."→ 不是话题转换
	(c) 将对话拓展至子话题
B	“哈哈，那你平时喜欢做什么体育运动呀？ Do 你通常会出去跑步吗?
A	“很少，通常是躺在家里看电视、跑步和 等等，是想看看他们的胖子能不能假装看起来。
B	我也是，那你最近在看什么电视剧？你看过 《精英律师》吗？"→ 话题转换
	(d) 引入相关但不同的话题。
B	“我看这部电影时哭了，钢铁侠死了，真的很感人。”
A	“是啊，这部电影的特效非常棒。”
B	“你最喜欢哪位演员？”→话题转换
	（e）完全改变话题。

– 评论先前语境：回复是对说话人先前所说内容的评论；– 问答：回复是对说话人先前提出的问题的回答；– 将对话拓展至子话题：与上下文相比，回复拓展到了一个子话题；– 引入相关但不同的话题：与上下文相比，回复引入了一个相关但不同的话题；

– 完全改变话题：与上下文相比，回复完全改变了话题。

其中，我们统一将 CNTD 特有的问候和告别这两种情况视为话题转换。

3.3 数据来源

我们选择 NaturalConv 数据集 [20] 作为源语料库，该数据集包含约 40 万条话语和 1.99 万段多领域对话。它旨在收集一个多轮文档依据对话数据集，体现对话的场景性和自然性。

我们认为 NaturalConv 是一个很有潜力的对话主题检测数据集，原因如下：1) NaturalConv 具有自然属性，更接近人类对话，包含完整且自然的设置，例如场景假设、自由主题扩展、问候等；2) NaturalConv 包含约 40 万条话语和 1.99 万段多领域对话；3) 该语料库的平均轮次为 20，较长的对话上下文往往呈现出涉及更多话题的演变过程；4) 该语料库几乎不对说话者设置限制或假设，例如没有提出明确的目标 [21]。

3.4 标注流程

我们安排了三名标注员进行粗粒度标注，两名标注员进行细粒度标注。两类标注均分为以下三个阶段。

共同标注阶段首先，在粗粒度标注中，我们从 NaturalConv数据集的各个领域按比例抽取共 100 段对话，总计 2014 个对话轮次。在这一阶段，三名标注员每标注完 20 段对话，就需要讨论这些对话；讨论期间，每位标注员都要说明标注理由。最终，该阶段所有标注员在粗粒度标注中的 Kappa值为 0.7426。此外，我们根据完整粗粒度标注的结果对细粒度信息进行了标注。两名标注员对相同的 150 段对话进行了标注，并多次讨论以确保一致性。最终，该阶段所有标注员在细粒度标注中的 kappa 值为 0.9032。这些 kappa 值表明，我们的标注员已具备独立标注所需的充分标注能力，也证明了我们的语料库质量很高。

独立标注阶段 在开始第二阶段标注之前，我们确保了每位标注员的标注质量及判定标准。对于两种粒度的标注，我们从各个领域抽取对话，并随机分配给各位标注员进行独立标注。在此阶段，我们对 1208 段对话进行了粗粒度标注，对 1158 段对话进行了细粒度标注。

半自动复核阶段 最后，我们采用半自动复核流程，确保语料库仍保持高质量。一方面，我们会自动整理带有标注的对话，以检测格式问题

表2. 语料库的类别和比例。

类别	训练集	Val.	Test	Sum.
健康(8%)	85	11	11	107
教育(16%)	167	22	21	210
技术(17%)	176	22	22	220
体育(33%)	347	45	46	438
游戏(8%)	86	11	11	108
娱乐(17%)	180	23	22	225
总计	1041	134	133	1308

表 3. CNTD 详情。

	Min.	Max.	Avg.
对话轮次	20	26	20.1
话语词数	1	141	21.0
对话词数	194	888	421.7
对话主题	2	9	5.2
主题轮次	1	17	4.2

这些问题是由人工标注造成的。另一方面，我们会自动将相关新闻与每段对话进行匹配，并检查主题属性是否与对话一致，以排除任何可能的错误。

3.5 标注结果

由于时间有限，我们从 Natural-Conv 数据集中随机选取了 1308 段对话，并由四名标注员对其进行标注。最终，我们构建了一个包含 2.6 万个对话轮次的中文自然话题对话语料库。

如表2所示，我们按照不同类别的占比，将这些对话随机划分为训练集（1041 段）、验证集（134 段）和测试集（133 段）。此外，我们在表3中展示了 CNTD 的详细信息。数据显示，该语料库包含足够多的话题和较长的轮次，适用于对话主题检测。最后，表4展示了细粒度标签的统计数据。

我们统计了不同话题数量的对话数，如图2所示。另一方面，我们统计了对话中话题转移信号的分布，如图3所示。可以看到，对话总共有21轮，话题转移信号有三个峰值，分别出现在第 2^{nd}、 4^{th}和 18^{th} 轮。原因是，我们语料库中的对话通常以问候开场，并

表4。细粒度标签统计。

细粒度标签	数量
评论先前语境	15091
问答	3505
将对话拓展至子话题	857
引入相关但不同的话题	3106
完全改变话题	2439

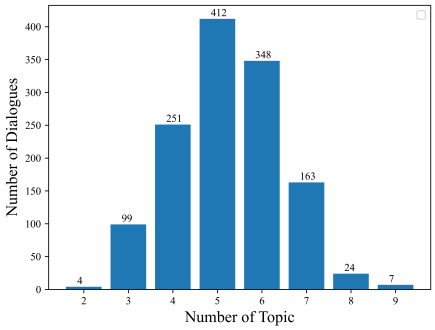


图2. 具有不同话题数量的对话数量。

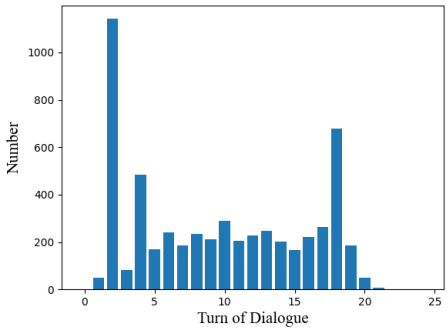


图3. CNTD 的话题转移分布。

以告别结束，因此对话开头和结尾的话题转移更多。此外，NaturalConv语料库会提供一则新闻作为对话的基础文档，因此从新闻转向派生话题的情况更为频繁，导致第 4^{th} 轮出现第三高的峰值。然而，我们认为这符合自然对话场景，因为人们在日常问候之后常常会聊起近期新闻。

4 模型

我们的模型框架如图4所示。我们提出了一种基于层次对比学习的教师-学生框架，包含两部分：知识蒸馏，以及由两种不同对比学习组成的层次对比学习。

4.1 知识蒸馏

现有研究由于缺乏未来信息，无法有效预测话题转移。为解决这一问题，我们提出了一种教师-学生框

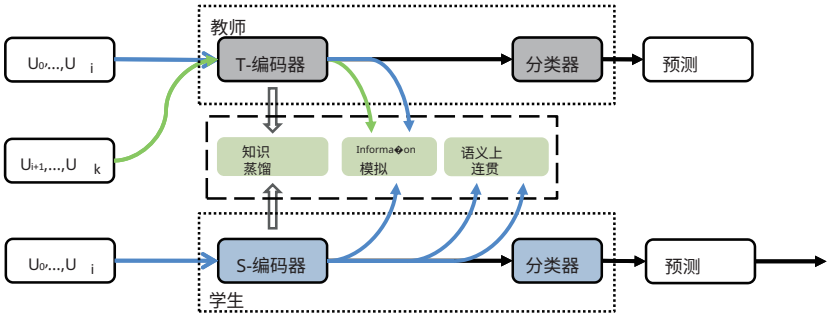


图4。 模型架构。该图包含一个教师编码器，可获取上下文和回复；以及一个学生编码器，其输入仅包含上下文，其中 u_{i+1} 表示当前对话样本中的回复， k 表示对话长度。此外，通过知识蒸馏和层次对比学习引导学生向教师学习，其中层次对比学习包含信息模拟损失和语义连贯性感知损失。

架，用于对话主题转移检测。学生端通过知识蒸馏，从教师端学习一种隐式的话题预测方式。

在该框架中，我们在两侧均采用预训练模型作为编码器，以获取对话的语义表示。此外，在编码上下文和响应信息时，我们对教师端共享编码器权重。例如，将 $[CLS]$ 最后一个隐藏层状态的表示作为对话级表示 H ，如下所示。

$$H_{C,i}^{T} = E_T\left(X_{C,i}\right) \tag{1}$$

$$H_{P,i}^{S} = E_S\left(X_{P,i}\right) \tag{2}$$

其中 $H \in R^{N\times d}$， $i \in I = \{1, 2, ..., N\}$ 表示批次中样本的索引，$C \in \{P, F, A\}$ 其中 P表示上下文， F表示回复， A 表示完整对话。 E 表示编码器， T 表示教师， S 表示学生。

在教师端，我们将对话级表示连接起来，然后将其输入线性层以获得最终检测结果；学生端则如下所示。

$$Z_i^T = W_T[H_{P,i}; H_{F,i}; H_{A,i}] \tag{3}$$

$$Z_i^S = W_S H_{P,i} \tag{4}$$

其中， W_T 和 W_S 分别表示教师端和学生端的线性层。

除了计算最终检测结果的交叉熵损失外，我们还在教师编码器和学生编码器的每个隐藏层之间建立了均方误差损失。

4.2 层次对比学习

我们认为，仅通过知识蒸馏学习响应信息是不充分的。此外，对话语料库中发生转移与否的比例不平衡，导致模型难以区分不同的转移情形。因此，我们提出了层次对比学习，分别由高层的信息模拟损失和低层的语义连贯性感知损失组成。这两种损失都基于对比学习，但前者旨在强化特征学习，后者旨在缓解标签不平衡问题。

对于信息模拟损失，该损失使每个上下文表示都能够进行主动学习。多项研究已经证明了这一有效性 [1,22,23], 融入全局信息可以使局部信息表示包含一些预测信息。这有助于我们的编码器获得具有更可靠预测信息的表示。

为缓解对话中标签比例不平衡的问题，我们基于监督式对比学习（SCL）提出了一种语义连贯性感知损失。SCL 的主要概念[24,25] 是将不同类别的样本视为彼此的正样本和负样本，以解决数据集[26]中显著的数量不平衡问题。该损失有效缓解了转移情形的不平衡，并帮助模型进一步区分不同的转移情形。

高层信息模拟损失 我们在两侧构建信息模拟损失，使上下文表示能够映射到同一个高维空间。因此，更容易学习回复信息，从而改善最终的检测结果。以下公式可用于描述该损失。

$$L_{ISL}^{M}=\sum_{i\in I}\log\frac{exp\left(H_{P,i}^{M}\cdot H_{A,i}^{T}\right)}{\sum_{j\in A(i)}exp\left(H_{P,i}^{M}\cdot H_{A,j}^{T}\right)} \tag{5}$$

其中， $A\left(i\right)=I-\{i\}$ 表示当前批次中除其自身以外的样本， P表示上下文， A 表示完整对话， $M\in\{T,S\}$ 其中 T 表示教师， S 表示学生。

低层语义连贯性感知损失 对于包含 N 个训练样本的批次，将对话最后一个隐藏状态的副本 H 复制出来，得到 H ，并将其视为正样本，同时分离其梯度。由此得到 $2N$个样本，随后，一个批次中所有样本的语义连贯性感知损失可表示如下。

$$U=[H;\overline{H}] \tag{6}$$

$$L_{SCL}=\sum_{i\in I}\frac{-1}{|P\left(i\right)|}\sum_{p\in P(i)}\log\frac{exp\left(U_{i}\cdot U_{p}\right)}{\sum_{a\in A(i)}exp\left(U_{i}\cdot U_{a}\right)} \tag{7}$$

其中， $U\in R^{2N\times d},i\in I=\{1,2,...,2N\}$ 表示批次中样本的索引， $P\left(i\right)=I_{j=i}-\{i\}$ 表示与 i 属于相同类别但不包含其自身的样本。

4.3 模型训练

我们分两个步骤训练模型。首先训练教师模型，其损失由两部分组成：预测结果与人工标注标签之间的交叉熵损失，记为 L_{NCE}；以及用于学习回复表示的信息模拟损失，记为 L_{ISL}^{T}。总体训练损失如下所示。

$$L^{T} = L_{NCE} + L_{ISL}^{T} \tag{8}$$

$$L^{S} = L_{NCE} + L_{KD} + L_{ISL}^{S} + L_{SCL} \tag{9}$$

随后，我们训练学生模型，该模型由四部分组成：预测结果与人工标注标签之间的交叉熵损失，记为 L_{NCE}；双方各个隐藏层之间的知识蒸馏，记为 L_{KD}；针对对话上下文的信息模拟损失，记为 L_{ISL}^{S}；以及针对不同转移情形的语义连贯性感知损失，记为 L_{SCL}。上述公式表示其整体训练损失。由于为这四种损失分配权重并进行微调颇具挑战，我们最终选择对它们采用相同的权重。

5 实验与分析

5.1 实验设置

基于我们在表2中划分的 CNTD 训练集/验证集/测试集，以及 TIAGE 的先前工作 [7]，我们从每段对话中提取（上下文、回复）对作为输入，并将回复标签作为回复未知任务的目标。在我们的实验中，除对话中的第一句话外，每句话都可视为回复。评估方面，我们报告精确率（P）、召回率（R）和微平均F1分数。

我们使用 BERT 作为编码器，并在训练过程中对其进行微调。对于 TIAGE 和 CNTD 语料库，所有预训练模型参数均设为默认值。我们在 NVIDIA GeForce GTX 1080 Ti 和 NVIDIA GeForce GTX 3090 上进行实验；对于 CNTD 和 TIAGE，批量大小分别为 2 和 6，初始学习率为 2e-5。训练轮次设为 20，丢弃率设为 0.5。

在实验中，我们使用 BERT-base-Chinese 和 MT5-base 获取 CNTD 对话的语义表示，并使用 BERT-base-uncased 和 T5-base 获取 TIAGE 对话的语义表示。

5.2 实验结果

对话主题转移检测是一项新任务，目前还没有可用的复杂模型，除了一个简单的 T5 [7]，它可以被视为当前最优水平。由于我们采用 BERT 作为编码器，且 TIAGE 使用了 T5 模型，因此我们将 T5 [9] 和 BERT [27] 的预训练模型用作基线。对于 BERT，我们

表5。我们模型与 CNTD 上三个基线模型的比较。

模型	**P**	**R**	**F1**
T5	27.1	46.8	34.3
BERT	55.5	43.8	48.9
TS	52.7	47.4	49.9
Ours	**56.0**	**52.0**	**53.9**

连接上下文中的话语，并使用 $[SEP]$ 分隔最后一句话。对于 T5，我们同样连接上下文中的话语，并将无法判定的预测结果归类为“非话题转换”类别。

表5展示了我们模型与各基线模型的性能对比，其中 TS 表示未采用层次比较学习（HCL）的教师-学生模型，Ours 表示最终模型，即在教师端和学生端均加入 ISL 的基础上，在学生端加入 SCL。

可以发现，在 CNTD 上，我们的模型取得了良好的提升，与基线模型相比，精确率和召回率均有所提高。尽管 T5 的召回率表现不差，但与 BERT 相比，其精确率不足，显然 T5 在预测话题方面效果不佳。相比之下，TS 的微平均 F1比 BERT 提高了 1.0，这证实了教师-学生框架在引入响应信息方面是有效的。此外，Ours 的微平均F1比 TS 提高了 4.0，精确率和召回率也均显著提升，充分证明了我们的 HCL 能够增强模型区分不同话题情境的能力。特别是，与表现最佳的基线模型 BERT 相比，我们的模型在 CNTD 上提升了 5.0，这表明了所提出模型的有效性。

5.3 消融实验

为验证模型中各组件的有效性，我们在 CTND 上进行了消融实验，实验结果如表 6 所示。

如果移除教师端（$-ISL_S$）或学生端（$-ISL_T$）的 ISL，模型的微平均 F1值分别下降 1.5 和 1.3，其中移除学生端的 ISL 后降幅最大。尽管 $-ISL_T$ 在话题预测方面的精确率最高，且错误概率低于 $Ours$ 和 $-ISL_S$，但可以看出，在教师端和学生端同时加入 ISL 能更好地提高正确预测率。此外，如果同时移除教师端和学生端的 ISL（$-ISL_{TS}$），其微平均F1表现与 $-ISL_S$ 和 $-ISL_T$ 相近，但精确率最高（58.8%）。

如果从模型中移除 SCL（-SCL）或 HCL（-HCL），模型 -SCL 和 -HCL 的微平均 F1 值将分别从 53.9 降至 52.4（-1.5）和 49.9（-4.0）。这些结果表明，我们提出的语义连贯性感知损失（SCL）和层次对比学习（HCL）对这项任务有效，尤其是 HCL。

表 6. CNTD 上我们的模型及其变体的结果，其中 ISL、SCL 和 HCL 分别指信息模拟损失、语义连贯性感知损失和层次对比学习，T、S 和 TS 分别指教师端、学生端以及两端。

	P	**R**	**F1**
Ours	56.0	52.0	**53.9**
$-ISL_S$	53.6	50.8	52.2
$-ISL_T$	56.1	49.3	52.6
$-ISL_{TS}$	**58.8**	47.0	52.3
$-SCL$	51.6	**53.1**	52.4
$-HCL$	52.7	47.4	49.9

表 7. 话题数量不同的对话上的性能。

话题数量	**我们的模型（F1）**	**BERT（F1）**
2	100	66.7
3	49.0	42.6
4	64.2	53.7
5	58.5	46.3
6	50.3	50.7
7	44.4	47.5
8	44.4	40.0
9	76.9	54.5

5.4 从不同角度分析性能

此外，如表 7 所示，我们分析了与 BERT 相比，模型在话题数量不同的对话上的表现。可以发现，在话题较少的对话中，我们的模型表现优于 BERT。在话题数量为 2 至 5 个的对话中，我们的模型在话题转换预测方面至少提升了 6%，且表现高于平均水平。当话题数量增加至 9 个时，模型表现有所提升，这是因为对话长度仍约为 20，而话题转换更加显著。

在表 8 中，我们还研究了不同主题轮次下话题转换检测的召回率。我们的模型在不同主题轮次下均有不同程度的提升，其中第 7 至第 9 轮的提升最为显著。即使在较长的主题转换情形下，我们的模型也能取得有效提升。不过，与较短的主题转换情形相比，我们模型的性能不可避免地有所下降。主题轮次较少时，话题转换情形较为简单，因此更容易判断。随着轮次变长、情形变得复杂，较长轮次中的话题包含更多信息，因此更容易识别。

5.5 英文 TIAGE 上的结果

如表 9 所示，可以发现我们的模型在英语 TIAGE 上也取得了显著提升。尽管我们的模型在精确率方面并非最佳，

表 8. 不同轮次长度下话题转换的表现。

主题轮次	我们的模型（召回率）	BERT（召回率）
1-3	56.6	53.8
4-6	30.3	21.4
7-9	28.8	11.9
10-12	40.0	32.0
13-17	40.0	30.0

表 9。TIAGE 上的结果。

	TIAGE		
	P	**R**	**F1**
T5	**34.0**	17.0	22.0
BERT	28.1	17.9	21.7
TS	26.9	20.1	22.9
Ours	27.4	**28.3**	**27.8**

我们在召回率和微平均F1值上均取得了最佳性能，尤其是微平均F1，相较于T5提升了5.8%。这证明我们的模型在英语和中文上都取得了最佳性能。

5.6 案例研究和错误分析

我们还开展了一项案例研究。我们的模型、该实例上的 BERT 模型以及人工标注所作出的预测结果如表10所示。与 BERT 模型相比，显然，我们的模型能够准确预测与话语“Yes, that’s right.”、“It is, indeed,should pay attention to it.”等相对应的实例中的话题变化，这些实例属于问答场景。然而，如果回应“Well, the policy has been imple-mented in place this time.”和“And now we are promoting the development of children’s creative and practical skills.”等属于评论先前语境场景的话语，我们的模型或 BERT 都无法准确预测该场景中的话题转换。这表明，检测自然对话中的话题转移仍然具有挑战性。

我们进一步分析了实验中模型产生的预测错误。具体而言，我们分析了该示例，以探究该示例结果中的错误是否在其他对话中也普遍存在。从表10可以看出，我们在第 14^{th} 和第 18^{th} 轮做出了错误预测。我们将“教学设备必须更新了，对吧？”预测为“非话题转换”，将“嗯，多亏了政府！”预测为“话题转换”。

我们统计了许多错误的出现情况，这些错误主要分为两类。一类是“引入相关但不同的话题”类型的话语。由于缺少有关对话后续发展的信息，模型预测没有发生话题转换。另一类是“评论先前语境”类别。由于此类回复不会影响先前话题的完整性，因此大多被预测为发生了话题转换。

表10. 不同轮次下 BERT、Ours 和 Human 的结果，其中“1”表示发生了话题转换，“0”表示未发生。我们省略了全为 0 的行。

轮次	BERT	Ours	人工
3	0	1	1
5	0	1	1
11	0	1	1
13	1	0	0
14	0	0	1
16	1	0	0
17	0	1	1
18	1	1	0
19	0	1	1

6 结论

基于NaturalConv数据集，我们通过人工标注创建了CNTD数据集，填补了中文自然对话话题语料库的空白。我们还提出了一种基于层次对比学习的教师-学生模型，以解决响应信息不足的问题。我们通过教师-学生框架引入响应信息，并在高层教师和学生中构建信息模拟学习，在低层学生中构建语义连贯性感知学习。实验结果表明，我们的模型在话题较少的对话中表现更好。然而，检测轮次较长的话题或话题较多的对话仍然是一个复杂的问题。未来，我们将重点研究如何更好地利用响应信息和新闻信息来实时检测话题转移。

致谢

作者谨向三位匿名审稿人对本文提出的意见表示感谢。本研究得到中国国家自然科学基金（项目编号：62276177、61836007 和 62006167）以及江苏高等教育机构优势学科建设工程（PAPD）的资助。

参考文献

1. Shuyang Dai、Guoyin Wang、Sunghyun Park 和 Sungjin Lee。通过对比潜在表征学习生成对话回复。载于第三届对话式人工智能自然语言处理研讨会论文集，第 189–197 页，2021 年。2. Jiaqi Li、Ming Liu、Zihao Zheng、Heng Zhang、Bing Qin、Min-Yen Kan 和 Ting Liu。Dadgraph：一种面向多方

对话机器阅读理解的、具备话语感知能力的对话图神经网络。载于*2021* 年国际神经网络联合会议（*IJCNN*），第 1–8 页。IEEE，2021 年。3. Yiyang Li 和 Hai Zhao。用于多方对话阅读理解的说话人和关键话语自监督与伪自监督预测。载于计算语言学协会论文集：*EMNLP 2021* 发现，第 2053–2063 页，2021 年。4. Asma Ghandeharioun、Judy Hanwen Shen、Natasha Jaques、Craig Ferguson、Noah Jones、Agata Lapedriza 和 Rosalind Picard。通过自我博弈近似交互式人工评估，以评估开放领域对话系统。神经信息处理系统进展，第 32 卷，2019 年。5. Arash Einolghozati、Sonal Gupta、Mrinal Mohit 和 Rushin Shah。提升面向任务对话系统的鲁棒性。*arXiv* 预印本 *arXiv:1911.05153*，2019 年。6. Bing Liu、Gokhan Tür、Dilek Hakkani-Tür、Pararth Shah 和 Larry Heck。端到端可训练的面向任务对话系统中的人工教学与反馈式对话学习。载于*NAACL-HLT* 论文集，第 2060–2069 页，2018 年。7. Huiyuan Xie、Zhenghao Liu、Chenyan Xiong、Zhiyuan Liu 和 Ann Copestake。Tiage：主题转移感知对话建模基准。载于计算语言学协会论文集：*EMNLP 2021* 发现，第 1684–1690 页，2021 年。8. Yi Xu、Hai Zhao 和 Zhuosheng Zhang。主题感知多轮对话建模。载于*AAAI*人工智能会议论文集，第 35 卷，第 14176–14184 页，2021 年。9. Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊，21(67)：140–1，2020 年。10. Zhuohan Li、Zi Lin、Di He、Fei Tian、Tao Qin、Liwei Wang 和 Tie-Yan Liu。用于非自回归机器翻译的基于提示的训练。载于*2019* 年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议论文集（*EMNLP-IJCNLP*），第 5708–5713 页，2019 年。11. Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。个性化对话代理：我有一只狗，你也养宠物吗？载于计算语言学协会第 *56* 届年会论文集（第一卷：长篇论文），第 2204–2213 页，2018 年。12. Paweł Budzianowski、Tsung-Hsien Wen、Bo-Hsiang Tseng、Iñigo Casanueva、Stefan Ultes、Osman Ramadan 和 Milica Gasic。Multiwoz：用于面向任务对话建模的大规模多领域绿野仙踪数据集。载于*2018* 年自然语言处理实证方法会议论文集，第 5016–5026 页，2018 年。13. Mihail Eric、Lakshmi Krishnan、Francois Charette 和 Christopher D Manning。面向任务对话的键值检索网络。载于第 *18* 届 *SIGdial* 话语与对话会议论文集，第 37–49 页，2017 年。14. Jacob Eisenstein 和 Regina Barzilay。贝叶斯无监督主题分割。载于*2008* 年自然语言处理实证方法会议论文集，第 334–343 页，2008 年。15. Lan Du、Wray Buntine 和 Mark Johnson。使用结构化主题模型进行主题分割。载于*2013* 年计算语言学协会北美分会人类语言技术会议论文集，第 190–200 页，2013 年。16. Omri Koshorek、Adir Cohen、Noam Mor、Michael Rotman 和 Jonathan Berant。将文本分割视为监督学习任务。载于*2018* 年

北美计算语言学协会人类语言技术分会会议论文集，第*2*卷 473, 2018 469（短篇论文），页码 – 。17. Pinkesh Badjatiya、Litton J Kurisinkel、Manish Gupta 和 Vasudeva Varma。基于注意力机制的神经文本分段。载于欧洲信息检索会议，第180–193页。施普林格，2018。18. Sebastian Arnold、Rudolf Schneider、Philippe Cudré-Mauroux、Felix A Gers 和 Alexander Löser。Sector：一种用于连贯主题分段与分类的神经模型。计算语言学协会会刊，7 184，2019 169：–。19. Yingcheng Sun 和 Kenneth Loparo。利用结构上下文检测在线讨论中的话题转换。载于*2019 IEEE*第*43*届年度计算机软件与应用会议（*COMPSAC*），第1卷，第948–949页。IEEE，2019。20. Xiaoyang Wang、Chen Li、Jianqiao Zhao 和 Dong Yu。NaturalConv：面向多轮主题驱动对话的中文对话数据集。载于*AAAI*人工智能会议论文集，第35卷，第14006–14014页，2021。21. Wenquan Wu、Zhen Guo、Xiangyang Zhou、Hua Wu、Xiyuan Zhang、Rongzhong Lian 和 Haifeng Wang。基于显式对话目标的主动式人机对话。载于计算语言学协会第*57*届年会论文集，第3794–3804页，2019。22. Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。利用对比预测编码进行表征学习。*arXiv*预印本 *arXiv:1807.03748*，2018。23. Shaoxiong Feng、Xuancheng Ren、Hongshen Chen、Bin Sun、Kan Li 和 Xu Sun。通过模仿隐式场景来正则化对话生成。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第6592–6604页，2020。24. Shimin Li、Hang Yan 和 Xipeng Qiu。对比与生成使BART成为优秀的对话情感识别器。载于*AAAI*人工智能会议论文集，第36卷，第11002–11010页，2022。25. Beliz Gunel、Jingfei Du、Alexis Conneau 和 Veselin Stoyanov。用于预训练语言模型微调的监督式对比学习。载于国际学习表征会议。26. Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu。DailyDialog：人工标注的多轮对话数据集。载于第八届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第986–995页，2017。27. Jacob Devlin、Ming-Wei Chang、Kenton 和 Lee Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。载于*NAACL-HLT*论文集，第4171–4186页，2019。