

基于话语重写的无监督对话主题分割算法

李彤亮, 李奇峰, 侯霞, 陈小明, 李舟军

引用本文

李彤亮, 李奇峰, 侯霞, 陈小明, 李舟军. 基于话语重写的无监督对话主题分割算法[J]. 计算机科学, 2025, 52(12): 215-223.

LI Tongliang, LI Qifeng, HOU Xia, CHEN Xiaoming, LI Zhoujun. [Unsupervised Dialogue Topic Segmentation Method Based on Utterance Rewriting](#) [J]. Computer Science, 2025, 52(12): 215-223.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于联合注意力机制与多阶段特征提取的图像去雨](#)

Image Deraining Based on Union Attention Mechanism and Multi-stage Feature Extraction

计算机科学, 2025, 52(11): 206-212. <https://doi.org/10.11896/jsjcx.240900013>

[MTFuse:基于Mamba和Transformer的红外与可见光图像融合网络](#)

MTFuse:An Infrared and Visible Image Fusion Network Based on Mamba and Transformer

计算机科学, 2025, 52(8): 188-194. <https://doi.org/10.11896/jsjcx.240600106>

[基于改进SOM网络的聚类算法](#)

Clustering Algorithm Based on Improved SOM Model

计算机科学, 2025, 52(8): 162-170. <https://doi.org/10.11896/jsjcx.240700017>

[基于改进TF-IIGM算法的畜禽疫病诊断模型研究](#)

Study on Diagnosis Model of Livestock and Poultry Disease Based on Improved TF-IIGM Algorithm

计算机科学, 2025, 52(6A): 240700029-7. <https://doi.org/10.11896/jsjcx.240700029>

[基于双分支小波卷积自编码器和数据增强的深度聚类方法](#)

Deep Clustering Method Based on Dual-branch Wavelet Convolutional Autoencoder and DataAugmentation

计算机科学, 2025, 52(4): 129-137. <https://doi.org/10.11896/jsjcx.240100111>

基于话语重写的无监督对话主题分割算法

李彤亮¹ 李奇峰¹ 侯霞¹ 陈小明² 李舟军³

1 北京信息科技大学计算机学院 北京 102206

2 深圳智能思创科技有限公司 广东 深圳 518052

3 北京航空航天大学计算机学院 北京 100191

(tonyliangli@bistu.edu.cn)

摘要 对话主题分割(DTS)任务旨在将一段多轮对话自动划分为不同的主题片段,从而更精准地理解和处理对话内容,在对话建模任务中具有重要作用。传统的DTS方法主要依赖语义相似性和对话连贯性来进行无监督的对话主题划分,但这些特征难以全面捕捉对话中的复杂主题转换,且未标注的对话数据尚未被充分挖掘和利用。为此,最新的DTS方法通过相邻话语匹配和伪分割,从对话数据中学习主题感知的对话表示,进一步挖掘未标注对话中的有用线索。然而,多轮对话中常见的共指和省略现象可能影响语义相似性的计算,进而削弱相邻话语匹配的准确性。为解决这一问题并充分利用对话关系中的有用线索,提出了一种新颖的无监督对话主题分割方法,结合了话语重写(UR)技术与无监督学习算法。该方法通过重写对话中的共指和省略信息,使其恢复为完整表达,从而更好地捕捉对话中的主题线索。实验结果表明,提出的话语重写主题分割模型(UR-DTS)在主题分割的准确性上取得了显著提升,达到了目前的最好水平。在DialSeg711数据集上,错误分数 P_k 和WinDiff(WD)两个指标的性能表现均提升了约6个百分点,分别达到11.42%和12.97%。在更复杂的Doc2Dial数据集上, P_k 和WD的性能表现分别提升了3个百分点和2个百分点,达到了35.17%和38.49%。这些结果表明,UR-DTS在捕捉对话主题转换方面具有显著优势,且对未标注对话数据有更大的利用潜力。

关键词: 多轮对话;无监督学习;自然语言理解;Doc2Dial

中图分类号 TP391

Unsupervised Dialogue Topic Segmentation Method Based on Utterance Rewriting

LI Tongliang¹, LI Qifeng¹, HOU Xia¹, CHEN Xiaoming² and LI Zhoujun³

1 School of Computer Science, Beijing Information Science & Technology University, Beijing 102206, China

2 Shenzhen Intelligent Strong Technology Co., Ltd., Shenzhen, Guangdong 518052, China

3 School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract Dialogue Topic Segmentation(DTS) task aims to automatically divide a multi-turn conversation into different topic segments, enabling more precise understanding and processing of dialogue content. DTS plays an important role in dialogue modeling tasks. Traditional DTS methods primarily rely on semantic similarity and dialogue coherence to perform unsupervised topic segmentation, but these features are often insufficient to fully capture complex topic transitions in conversations, and unannotated dialogue data has not been fully explored and utilized. To address this issue, recent DTS methods employ adjacent utterance matching and pseudo-segmentation to learn topic-aware representations from dialogue data, further extracting useful cues from unannotated dialogues. However, common phenomena such as coreference and ellipsis in multi-turn dialogues may affect the calculation of semantic similarity, thereby weakening the accuracy of adjacent utterance matching. To solve this problem and fully leverage the useful cues in dialogue relationships, this study proposes a novel unsupervised DTS method that combines utterance rewriting (UR) techniques with unsupervised learning algorithms. This approach rewrites coreferential and elliptical expressions in the dialogue to restore them to their complete forms, better capturing the thematic cues in the conversation. Experimental results show that the proposed utterance rewriting topic segmentation model(UR-DTS) significantly improves topic segmentation accuracy, achieving state-of-the-art performance. On the DialSeg711 dataset, the error rate P_k and WinDiff(WD) improves by approximately 6 percentage point, reaching 11.42% and 12.97%, respectively. On the more complex Doc2Dial dataset, P_k and WD improve by

到稿日期:2024-10-18 返修日期:2025-01-08

基金项目:国家自然科学基金(62406033,62276017,U1636211,61672081);教育部产学研合作协同育人项目(231004723052336)

This work was supported by the National Natural Science Foundation of China(62406033,62276017,U1636211,61672081) and University-Industry Collaborative Education Program(231004723052336).

通信作者:陈小明(chenxiaoming@aistrong.com)

3 percentage point and 2 percentage point, reaching 35.17% and 38.49%. These results demonstrate that UR-DTS has a significant advantage in capturing topic transitions in conversations and shows greater potential for leveraging unannotated dialogue data.

Keywords Multi-turn dialogue, Unsupervised learning, Natural language understanding, Doc2Dial

1 引言

在自然语言处理(Natural Language Processing, NLP)领域,对话系统因其在客户服务、个人助理、互动娱乐等多个领域中的广泛应用,受到越来越多的关注。理解和管理多轮对话的流动性是这些系统的核心能力,这不仅能提升用户体验,促进自然交互,还能够准确地响应用户需求,从而增强系统的智能性和实用性。在文本领域中,对话主题分割(Dialogue Topic Segmentation, DTS)作为一种关键技术,近年来在推动对话系统智能化发展中发挥了重要作用。DTS的目标是通过将对话分割成主题连贯的部分,从而揭示对话的主题结构^[1],这对于对话生成^[2]、摘要^[3]、响应预测^[4]和问答^[5]等下游 NLP 任务至关重要。在计算机视觉领域中,通过文本序列生成的人体运动^[6]和细粒度运动风格迁移^[7]都需要模型对文本主题有深入理解,以此来进行 3D 动画的创作^[8]。

随着大语言模型(Large Language Models, LLMs),如 GPT 系列模型的快速发展, NLP 任务的性能得到了显著提升。大模型具有强大的语言生成和理解能力,在众多 NLP 任务中展现出卓越的表现。尽管大模型能够通过广泛的预训练捕捉丰富的语言特征,但仍然面临一些特定任务上的挑战。尤其是在多轮对话中,由于对话的长程依赖、主题转换的频繁性以及语义信息的丢失,大模型常常无法准确捕捉到对话中的主题变化。大模型的黑盒性和幻觉问题进一步限制了其在需要精细化语义理解的任务中的表现。尤其是在当前大模型广泛应用于 RAG(Retrieval-Augmented Generation)模式的背景下,对话主题分割的重要性更加凸显。在多轮对话中,主题分割能够有效地将对话划分为主题连贯的片段,从而帮助检索模型更精准地获取相关内容,提升检索阶段的效率与准确性。这一过程对于多轮对话中的 RAG 尤其重要,因为准确的主题分割可以减少检索噪声,确保生成模型基于高相关性的信息进行回答。因此,尽管大模型在生成能力上表现优异, DTS 技术依然在多轮对话场景下的 RAG 任务中扮演着至关重要的辅助角色。对话主题分割和话语重写技术依然发挥着不可替代的作用。

近年来,对话主题分割的研究主要集中在有监督学习和无监督学习两大领域。有监督学习依赖于标注数据来监督对话主题分割。在相同数据量的情况下,由于有明确的标签指导,其分割效果通常优于无监督学习。但是标注数据稀缺或不可用一直是有监督学习面临的问题。无监督学习则通过聚类和主题建模技术,减少对标注数据的依赖,从而在可用数据集上具有更广泛的适用性。这类方法通常侧重于利用语义相似性和对话连贯性来评估话题转换。对话连贯性是指话语与其先前上下文之间的响应关系。然而,大多数方法通常无法全面捕捉话题间的相似性,容易忽视对话中错综复杂的动态

变化。此外,多轮对话中通常包含共同引用或信息省略,现有无监督 DTS 方法不能有效利用未标注对话数据中的这些有效信息,例如相邻话语之间的匹配关系和伪分段等,进而减弱了话语之间的语义关联。这些限制妨碍了无监督模型对主题感知话语表征的学习能力,最终影响了主题分割的准确性。

本文提出了一种融合话语重写技术的创新型无监督对话主题分割模型,以应对上述挑战。其主要目标是增强模型理解与利用对话话语细微差别的能力,从而提高主题分割的准确性。该模型通过重写对话来恢复对话中缺失的信息,能够充分挖掘未标记对话中的有用线索。这种方法不仅弥补了语义相似性计算的不足,还能够利用对话数据中丰富但未被充分挖掘的有效信息。

本文基于 Gao 等^[9]的工作,重点研究如何有效利用未标记对话数据,做出了两方面的贡献。

1)提出了话语重写对话主题分割模型(Utterance Rewriting Dialogue Topic Segmentation, UR-DTS),这是一种利用话语重写(Utterance Rewriting, UR)技术进行对话主题分割的新型无监督方法。该模型有效解决了现有无监督 DTS 方法的局限性,尤其是在处理对话中的共同指代和信息省略方面。

2)对 UR-DTS 模型进行了全面评估,结果表明,其在主题分割准确性方面优于当前先进的无监督模型。研究显示, UR-DTS 模型在多个数据集上的 P_k 错误分数和 WinDiff (WD) 指标均有显著改善,验证了该模型在捕捉复杂对话主题方面的有效性。

2 相关工作

本章将依次介绍对话主题分割和话语重写技术的现有发展状态,这些方法都已取得良好的效果,对相关领域的发展起到了重要推动作用,对本文工作也有很大的借鉴意义。

2.1 对话主题分割

对话主题分割(DTS)是将对话分割成主题一致片段的过程,类似于文档主题分割任务。在这项任务中,许多最初为文档分割设计的方法被应用于对话文本。由于早期缺乏训练数据,同时无监督学习通过设计合适的预任务来生成伪标签,因此模型能够学习到有用的特征。无监督方法可以在没有大量标注数据的情况下,仍然取得良好的效果,所以成为主流。其主要依赖于分析单词共现统计^[10]或检查句子的主题分布^[11]来识别对话转折中主题或语义的变化。

随着大规模数据集(如维基百科中的数据集)的出现,对话主题分割的格局开始发生变化,有监督的分割技术开始发展。特别是基于神经网络的技术^[12],因其更高的准确性和效率而广受欢迎。然而,与文档分割相比,对话中语言碎片化、动态化和非正式的用语等情况给对话主题分割造成了更多的

困难。此外,为训练有监督模型而收集精确注释需要大量费用^[13],且不同领域的注释指令也不尽相同,因此无监督方法始终是研究热点。

一般来说,无监督方法分两个阶段执行。首先,采用各种技术来评估潜在片段分界线(即两个话语之间的间隔)两侧的主题相似性。随后,采用分段算法(如 TextTiling^[14])来识别段落的边界。传统的对话主题分割方法主要侧重于通过对话连贯性或通过表面特征(如词汇重叠)计算的语义相似性^[15]来评估主题相似性。2016年,Song等^[16]和 Xu等^[4]结合了词嵌入技术,改进了 TextTiling 算法。后者使用预训练语言模型^[17]对对话进行编码,例如 BERT^[18]和 SentenceBERT^[19],以更好地理解对话级的依赖关系。该方法相比传统的词袋方法可以更有效地把握语义的细微差别。此外,Xing等^[20]进一步提出了连贯性评分模型(Coherence Scoring Model,CSM),该模型采用话语对连贯性来评估主题相似性。然而,这些方法往往无法全面捕捉主题相似性,因为它们忽略了对话话语中错综复杂的动态变化。

虽然无监督对话主题分割(DTS)研究取得了重大进展,但仍有一些问题尚未解决,尤其是在主题相似性建模和利用无标记对话数据方面。传统方法依赖于一般语义相似性或对话连贯性来确定主题相似性,但由于忽视了对话中错综复杂的动态变化,这些方法往往存在不足,无法完全捕捉主题相似的同一段落。具体而言,共享同一主题中的语句不一定总是表现出语义相似性。例如,在讨论天气的对话中,“今天阳光明媚”和“我喜欢晴天”虽然语义不同,但属于同一主题。此外,语义相似的话语也可能与同一主题无关。对话连贯性是指话语与其先前上下文之间的响应关系^[21],反映相邻话语是否联系在一起。然而,同一主题片段中的两个不相邻的话语可能在主题上相似但不连贯。

另外,未标记的对话数据蕴含着丰富的对话关系线索,但其潜力仍未得到充分利用。目前的语义相似性方法使用的单词或句子嵌入是在通用文本语料库和监督自然语言推理(NLI)数据集^[22]上进行预训练的,这与无标记对话数据的特点不完全相符。另一方面,在基于连贯性的方法中,CSM从

DailyDialog^[23]数据集中学习对话连贯性,学习到了一定的对话线索,并且不需要 DTS 任务的注释操作。然而,每一个话中都只涉及一个主题。因此,CSM 依赖对话级话题标签生成训练样本,并不能完全解决多话题对话中话题分割的复杂性问题。

针对上述问题,最新研究提出了有效的解决思路。Pu等^[24]提出了一种新颖的动态评估语义连贯性的方法,该方法可以通过将评估单元从单个话语动态扩展到语义相关的上下文来提高连贯性评分的准确性。Gao等^[9]提出了一种新颖的无监督 DTS 框架,称为主题感知话语表示的无监督对话主题分割模型,该模型进一步挖掘了未标记对话数据中的有用线索,通过相邻话语匹配(NUM)和伪分割,从未标记的对话数据中学习主题感知话语表示。然后,将这些主题感知话语表示与对话连贯性结合使用,以执行无监督分割,一定程度上缓解了未标记的对话数据的有效利用问题。

然而,多轮对话通常包含共指和省略^[25]的情况,这可能会模糊话语之间的语义关系,同时限制无监督模型学习主题感知话语表征的能力,最终影响主题分割的准确性。因此,更进一步挖掘未标记的对话数据中的有用线索具有挑战性。

2.2 话语重写

话语重写任务(Utterance Rewriting)是自然语言处理(NLP)领域的一个重要任务,它指的是将一段文本重新表达,使其含义保持不变或基本不变,但表达方式有所改变,也叫句子重写。这种任务在很多应用场景中都非常有用,如自动摘要^[26]、问答系统^[27]、机器翻译的后编辑^[28]等。

表 1 列出了一段多轮对话及其话语重写的结果。话语重写任务中恢复的重要对话信息可能对 DTS 模型的准确性产生积极影响。例如“哪条路”“导航到那里”等词语的指代内容通常会被省略或模糊化,这可能会产生语义噪声,影响主题相似的相邻话语在语义相似性计算中的得分。通过话语重写模型,这些被省略和模糊的信息可以被还原,恢复同一主题下的关键对话信息。同时,重写后也会增加相邻话语中的共享词汇,从而增强 DTS 模型区分对话中主题相似性的能力。

表 1 话语重写举例

Table 1 Examples of utterance rewriting

角色	原话语	重写话语
角色 A	你能帮我找到附近的加油站吗?	你能帮我找到附近的加油站吗?
角色 B	我们离雪佛龙加油站有 6 英里远,但附近有堵车。	我们离雪佛龙加油站有 6 英里远,但附近有堵车。
角色 A	哪条路最快?	哪条通往雪佛龙加油站的路最快?
角色 B	我选择了最快的方式,并将其发送到了您的地图上,现在正在导航到那里。	我选择了前往雪佛龙加油站的最快的方式,并将其发送到了您的地图上,现在正在导航到雪佛龙加油站。

近年来,话语重写任务在多个领域引起了广泛关注。在机器翻译中,通过重写操作来改进 seq2seq 模型的输出生成^[29]。Juncys-Dowmunt等^[30]探索了多种神经架构(CG-RU,GRU,M-CGRU等),这些架构适用于机器翻译输出的自动后编辑任务。在文本摘要中,重新编辑检索到的候选词可以生成更准确和抽象的摘要。Chen等^[31]提出了一种准确、快速的总结模型,该模型首先选择突出的句子,然后抽象地重

写它们(即压缩和释义),以生成简洁的整体摘要。在对话建模中,Weston等^[32]将其应用于重写检索模型的输出。Rastogi等^[33]在英语对话中采用了类似的思想,通过重新表述原始话语来简化下游的 SLU(Spoken Language Understanding)任务。将源输入重写为易于处理的标准格式也在信息检索^[34]、语义解析^[35]或问答^[36]任务中获得了显著的改进,但大多数没有注意恢复对话中的共指和省略信息^[25],而只是采用了简

单的词典或基于模板的重写策略。对于多轮对话,由于人类语言的复杂性,设计合适的基于模板的重写规则也非常耗时。本文为了很好地适应多轮对话的复杂性,采用了新的用于一般开放域对话的英语数据集^[37],基于 seq2seq 模型对话语进行重写。

3 本文方法

3.1 问题表述

对话主题分割旨在识别多轮对话中的片段边界,其中片段指同一主题连续对话段,即话语。两个相邻片段的间隔即为一个话题分割边界。

定义 1 给定一个包含 n 个话语的多轮对话 $D, D = \{u_1, u_2, \dots, u_n\}$, 称 $u_i (1 \leq i \leq n)$ 为 D 中的话语。

定义 2 对于对话 $D = \{u_1, u_2, \dots, u_{n-1}\}$, 设其间隔集合为 $V = \{v_1, v_2, \dots, v_{n-1}\}$, 称 v_i 为第 i 个和第 $i+1$ 个话语之间的间隔。

分割算法将片段边界预测为 $B = \{b_1, b_2, \dots, b_k\}$ (其中 k 表示边界数, b_i 表示对话在第 i 个间隔处被划分), 即从上述 v_{n-1} 个间隔中预测出可以分割主题的间隔集合 $B \in V$ 。

大多数无监督 DTS 方法遵循两阶段范式。首先,对于位

于 u_i 和 u_{i+1} 之间的区间 v_i , 计算相关性得分 r_i 。得分越高, 区间两侧属于同一片段的可能性就越大。然后, 给定相关性得分 $R = \{r_1, r_2, \dots, r_{n-1}\}$, 使用分割算法(如 TextTiling^[14] 或其派生算法之一)来确定分割边界。以前的方法通常根据语义相似性或对话连贯性来评估相关性得分, 没有利用到未标记的对话数据。Gao 等^[9] 提出的方法利用从主题感知话语表示中得出的对话连贯性和主题相似性来建模该相关性得分, 学习对话数据中的主题内容信息, 但对话数据中仍有信息未被利用, 如话语中被省略的主语、代词等关键信息。因此, 本文优先对话语进行重写, 优化主题感知话语表示中主题相似性的语义计算, 通过优化主题相似性来建模相关性得分。

话语重写属于不完整话语重写的一部分。

定义 3 给定对话 $D = \{u_1, \dots, u_n\}$, 给定不完整话语 $u_t (1 \leq t \leq n)$, 将其上下文定义为 $C = \{u_1, \dots, u_{t-1}\}$ 。

不完整话语重写旨在通过上下文 C 将 u_t 重写为 u_t^* 。重写后的 u_t^* 不仅应具有与 u_t 相同的含义, 而且可以单独理解。

3.2 模型架构

如图 1 所示, 所提分割模型由重写编码器、重写解码器、主题编码器、连贯性编码器和分割算法组成, 这些模块分别负责话语重写、主题表示和连贯性计算等任务。

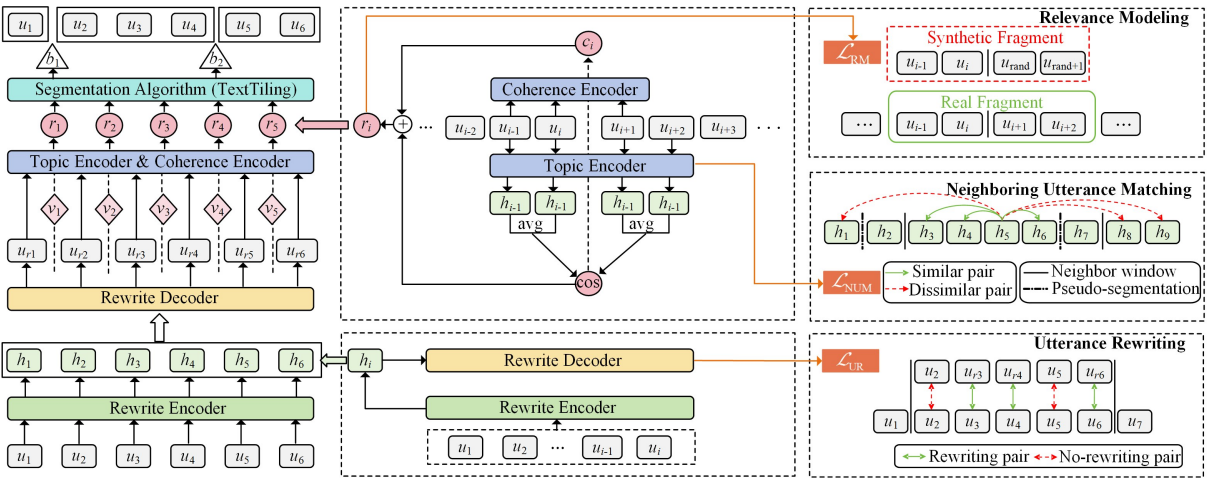


图 1 UR-DTS 框架(电子版为彩图)

Fig. 1 Framework of UR-DTS

话语重写模块包含重写编码器和重写解码器, 组合为序列到序列 (Sequence-to-Sequence, Seq2Seq) 任务, 其中编码器的第 i 个隐藏状态向量 h_i , 其输入部分的文本包含前文和当前话语的完整段落 (u_1, u_2, \dots, u_i) , 对应的训练任务是话语重写任务 \mathcal{L}_{UR} , 绿色实线箭头为重写话语对, 红色虚线箭头为原话语对。

重写之后的话语为 u_i^* , 相邻话语之间的间隔为 v_i 。为了获得更好的话语表示初始化, 选择 SimCSE^[38] 来初始化本文方法的主题编码器。SimCSE 是一个简单但有效的对比句子嵌入框架。将重写后的 u_i^* 传入主题编码器以获得每个话语的主题表示:

$$h_i^* = \text{SimCSE}(u_i^*) \quad (1)$$

其中, $h_i^* \in R^{d_h}$ 表示 SimCSE 最后一层的池化输出, d_h 是隐藏状态的维度。参照 CSM^[20] 的设计, 选择 Next Sentence Pre-

diction (NSP) BERT^[18] 作为连贯性编码器。对于每个间隔 v_i , 连贯性编码器计算得到的连贯性得分如下:

$$c_i = \text{NSP-BERT}([\text{CLS}; u_{i-1}^*; u_i^*]) \quad (2)$$

其中, u_{i+1}^* 是当前的响应, $[u_{i-1}^*; u_i^*]$ 是连接的前文。将多主题多轮对话输入主题编码器和话语连贯性编码器, 在获得主题表示 h_i^* 和连贯性得分 c_i 后, 计算相关性得分 r_i :

$$r_i = \text{sim}\left(\frac{h_{i-1}^* + h_i^*}{2}, \frac{h_{i+1}^* + h_{i+2}^*}{2}\right) + c_i \quad (3)$$

其中, $\text{sim}(\cdot, \cdot)$ 是余弦相似度, 相关度得分 r_i 被分割算法用来执行分割。

主题编码器的训练任务是邻近完整话语匹配任务 \mathcal{L}_{NCUM} , 绿色实线箭头表示定义的主题相似话语对, 红色虚线箭头表示定义的主题不相似话语对。该任务基于 Gao 等^[9] 的工作进行扩展, 改用了重写的话语, 恢复了对话中的共指和

省略等信息,旨在改进之前工作中计算主题相似性的余弦相似性计算操作,使得模型学习分辨相似主题的能力得到提升。话语连贯性编码器的训练任务是相关性建模任务 \mathcal{L}_{RM} ,绿色框内为真实的连贯片段,红色框内为构建的虚假片段,即不连贯片段。

3.3 不完整话语重写

将不完整话语重写视为序列到序列(Seq2Seq)任务,并采用两个预训练的 Seq2Seq 模型 T5^[39] 和 Pegasus^[40]。输入是串联的上下文话语和原始的最后一句话语,每句话语前插入特殊标记以指示其说话者。通过模型的编码器学习原话语的表示,捕捉重要信息。计算式如下:

$$\mathbf{h}_i = \text{Encoder}(D) \quad (4)$$

$$D = [\text{CLS}] + u_1 + [\text{SEP}] + u_2 + \dots + u_i \quad (5)$$

其中, u_i 为经过处理的包含上下文的原话语信息, \mathbf{h}_i 表示对应的隐藏状态。

重写解码器接收来自编码器的隐藏状态向量 \mathbf{h}_i ,并将其用于生成最终的输出序列。解码器以开始符号和先前生成的词为输入,逐步生成目标序列,同时利用编码器提供的源序列上下文信息指导整个生成过程。解码器的计算式如下:

$$\text{Decoder}(\mathbf{y}) = \text{FFN}(\text{SelfA}(\mathbf{y}) + \text{EDA}(\mathbf{y}, \mathbf{h}_i)) \quad (6)$$

其中,SelfA为自注意力模型;输入向量为 $\mathbf{y} = (y_1, y_2, \dots, y_n)$, y_i 是解码器的第 i 个输入的上下文话语;EDA为Encoder-DecoderAttention模型。解码器的任务是生成一个输出序列 $x = (x_1, x_2, \dots, x_m)$,其中 x_j 是重写话语输出的第 j 个字。

3.4 主题感知话语表示

为了训练具有主题感知能力的分割模型,在Gao等^[9]的工作的基础上进行了扩展,提出了一项名为邻近完整话语匹配(Neighboring Complete Utterance Matching, NCUM)的新任务。原工作基于主题变化的性质,假设话语更有可能与其邻近话语在主题上相似。为了进一步减少未标记对话中的噪声,结合了邻近话语和伪分割来获得精细的主题相似话语对和不相似话语对。然而,原工作只是针对话语间的句子级别的去噪操作,忽略了句子内部的噪声,如话语中的共指和省略等产生的噪声。为了解决这一问题,本文提出了邻近完整话语匹配(NCUM)的新任务,通过恢复信息后的对话来获得精细的主题相似话语对和不相似话语对,并将两类话语对作为边际排名损失的正样本和负样本。

首先,给定 D 中的重写后的话语 u_i^* ,将其邻近话语索引集 U_i 和非邻近话语索引集 \bar{U}_i 定义为:

$$U_i = \{j \in [1, n] \mid w \geq |i - j| \wedge j \neq i\} \quad (7)$$

$$\bar{U}_i = \{j \in [1, n] \mid w \geq |i - j|\} \quad (8)$$

其中, w 是 u_i^* 前后相邻话语的数量, n 指对话 D 的长度。

在无标签的多主题对话中,来自NCUM的监督信号容易产生相对嘈杂的声音。为了减少噪声,基于重写的话语进一步将相邻关系与伪分段相结合,以产生精细的主题相似话语对和不相似话语对。给定重写后的话语 u_i^* 及其伪分段 $\text{segment}(i)$, W_i 表示 $\text{segment}(i)$ 内的话语索引, \bar{W}_i 表示 $\text{segment}(i)$ 外的话语索引:

$$W_i = \{j \in [1, n] \mid u_j^* \in \text{segment}(u_i^*) \wedge j \neq i\} \quad (9)$$

$$\bar{W}_i = \{j \in [1, n] \mid u_j^* \in \text{segment}(u_i^*)\} \quad (10)$$

基于重写话语 u_i^* 的相邻话语和伪片段,获得了 u_i^* 的主题相似话语索引 P_i^+ 和主题不相似话语索引 P_i^- :

$$P_i^+ = U_i \cap W_i, P_i^- = \bar{U}_i \cap \bar{W}_i \quad (11)$$

其中, W_i 表示得到的主题相似话语索引, U_i 表示得到的邻近话语索引集。

4 实验

4.1 数据集

本文使用了两类数据集。

1) 话语重写数据。从DECODE^[37]数据集中抽取了10000个对话作为训练集,该数据集源自DailyDialog^[23]和BST^[41]并重新标注。此外,从DailyDialog和BST中抽取了800个对话作为测试集。

2) 对话主题分割数据。采用了DialSeg711^[4]和Doc2Dial^[42]两个广泛使用的数据集。DialSeg711包含711个真实世界的英语对话,结合了MultiWOZ^[43]和KVRET^[44]中的对话。平均每个对话包含4.9个主题段,每段包含5.6个话语。Doc2Dial包含4100多个合成英语对话,基于4个领域的450多个文档,平均每个对话包含3.7个主题段,每段包含3.5个话语。

4.2 评估指标

为了确保公平比较,采用了两个标准指标,即 P_k 错误分数^[45]和WinDiff(WD)^[46]。 P_k 通过使用一个窗口大小为 K 的滑动窗口(K 如果不指定,则为标准分割的每个块的大小的平均值的一半),判断窗口的2个边缘的节点是否属于同一个主题。WD通过直接比较预测结果和真实标注在每个文本区域内的分割边界数量是否一致来计算差异。由于它们是惩罚指标,分数越低表示性能越好。

4.3 基线

将UR-DTS模型与以下无监督基线进行比较。

Random:给定一个包含 k 条话语的对话,首先随机抽取该对话的片段边界数量 $b \in \{0, \dots, k-1\}$,然后以概率 b/k 确定某个话语是否是片段的结尾。

BayesSeg^[47]:根据与段落相关的多项式语言模型对每个主题段中的单词进行建模。通过最大化对话的观察可能性来实现词汇衔接的分段。

GraphSeg^[15]:以话语作为节点生成语义相关图,然后通过找到图中的最大团来预测句段。

GreedySeg^[4]:根据从预训练的BERT句子编码器的输出计算出的相邻话语的相似性,贪婪地确定段边界。

TextTiling^[14]:一种将文本细分为多段落单元的技术,这些单元表示段落或子主题。该方法用于识别主要子主题变化的话语线索是词汇共现和分布的模式。

TeT+Embedding^[16]:通过应用词嵌入来计算连续话语对的语义连贯性,通过GloVe词嵌入增强TextTiling。

TeT+CLS^[4]:通过预训练的BERT句子编码器增强

TextTiling, 使用 BERT 编码器的输出嵌入来计算连续话语对的语义相似度。

TeT+NSP^[9]: 通过预训练的 BERT 增强了下一句预测的 TextTiling, 利用输出概率表示连续话语对的语义一致性。

CSM^[20]: 利用话语对连贯性评分任务的监督信号, 解决无监督方法仅利用表面特征来评估话语之间的主题连贯性的限制。

CSM(unsup)^[9]: 一种改进的 CSM 变体, 不需要主题标签和动作标签。

TeT+RM+NUM^[9]: 通过相邻话语匹配和伪分割, 从未标记的对话数据中学习主题感知话语表示。

Dynamic Back^[24]: 通过将评估单元从单个话语动态扩展为语义相关的上下文来提高连贯性评分的准确性。

4.4 实验参数设置

对于话语重写, 使用了两种预训练模型 T5-Base 和 T5-Large, 其参数大小分别为 2.2×10^8 和 7.7×10^8 。实验参数如表 1 所列, 每个模型训练 4 个 epoch, 学习率为 5×10^{-5} , 并使用波束搜索(波束大小为 5)来生成。

表 1 话语重写实验参数

Table 1 Experimental parameters of discourse rewriting

实验参数	值
显卡	NVIDIA GeForce RTX 3090 * 1
数据集-UR	DECODE ^[34]
学习率	5×10^{-5}
批处理大小	12
训练轮次	4
波束	5

对于主题编码器, 从 SimCSE 的 sup-simcse-bert-base-uncased 版本开始训练。对于 DialSeg711 和 Doc2Dial, 设置相邻话语的数量 w 为 5, 主题编码器实验参数如表 2 所列。

表 2 主题编码器实验参数

Table 2 Experimental parameters of theme encoder

实验参数	值
显卡	NVIDIA GeForce RTX 3090 * 1
数据集-DTS	DialSeg711 ^[4] 和 Doc2Dial ^[39]
学习率	1×10^{-5}
批处理大小	4
训练轮次	15
滑动窗口	5

4.5 实验结果

将本文方法与两类无监督基线进行了比较。

1) 不使用 TextTiling 的基线, 包括 BayesSeg^[47], GraphSeg^[15] 和 GreedySeg^[4]。

2) 从 TextTiling 扩展而来的基线, 如 TeT^[14]、TeT+CLS^[4]、TeT+Embedding^[16]、TeT+NSP^[9]、连贯性评分模型(CSM)^[20] 和 TeT+RM+NUM^[9]。其中 CSM 和 TeT+RM+NUM 使用对话连贯性而不是语义相似性, 与其他从 TextTiling 扩展而来的无监督基线不同。

本文模型通过恢复话语中的关键信息, 使用未标记的对话进行训练, 更深层次地学习主题感知话语表示和对话连贯性。表 3 列出了本文模型和基线在两个数据集上的结果。本文模型在两个评估数据集上都实现了最佳(SOTA)性能, 相

比之前的 SOTA 有不同程度的提升。具体来说, 在 DialSeg711 数据集上, 本文模型在 P_k 错误分数和 WD 两个指标上的性能提高约 6 个百分点, 实现了 11.42% 的 P_k 和 12.97% 的 WD。Doc2Dial 是一个更复杂、规模更大的数据集, 本文模型在 P_k 错误分数和 WD 上分别提高了约 3 个百分点和 2 个百分点, 实现了 35.17% 的 P_k 和 38.49% 的 WD。这表明所提模型通过有效利用重写后的未标记对话, 恢复了对话中共指和省略等信息, 使得模型从学习主题相似性和对话连贯性中获益。

表 3 本文模型和基线的实验结果

Table 3 Experimental results of the proposed method and baseline

Method	DialSeg711		Doc2dial	
	$P_k \downarrow$	WD \downarrow	$P_k \downarrow$	WD \downarrow
Random	52.92	70.04	55.60	65.29
BayesSeg ^[47]	30.97	35.60	46.65	62.13
GraphSeg ^[15]	43.74	44.76	51.54	51.59
GreedySeg ^[4]	50.95	53.85	50.66	51.56
TextTiling ^[14]	40.44	44.63	52.02	57.42
TeT+Embedding ^[16]	39.37	41.27	53.72	55.73
TeT+CLS ^[4]	40.49	43.14	54.34	57.92
TeT+NSP ^[9]	46.84	48.50	50.79	54.86
CSM ^[20]	26.80	28.24	45.23	47.32
CSM(unsup) ^[9]	24.30	26.35	45.30	49.84
TeT+RM+NUM ^[9]	17.86	19.80	38.11	40.72
Dynamic Back ^[6]	18.42	21.15	42.01	45.06
Ours _{T5-base}	12.00	13.35	35.97	40.01
Ours _{T5-large}	11.42	12.97	35.17	38.49

4.6 消融实验

本文通过两种不同的设置, 研究了重写模块对于整体主题分割任务的影响: 1) 丢弃模型层面引入话语重写的能力, 将重写数据作为以主题编码器和连贯性编码器为主的主题分割模型的测试集; 2) 保留模型层面话语重写的能力。表 4 列出了消融研究结果。当不引入重写能力时, 观察到两个数据集上的性能均有所下降, 这表明在主题分割模型中加入重写模块对于获得良好的性能至关重要, 因为它可以提升主题编码器和连贯性编码器对于相似话语对的主题一致性判断, 防止局部对话连贯性主导主题分割。

表 4 消融实验结果

Table 4 Ablation experiment results

Method	DialSeg711		Doc2dial	
	$P_k \downarrow$	WD \downarrow	$P_k \downarrow$	WD \downarrow
TeT+RM+NUM ^[7]	17.86	19.80	38.11	40.72
Ours w/oRewrite	12.95	14.17	36.26	40.19
Ours	11.42	12.97	35.17	38.49

4.7 案例研究

表 5 列出了两组不同对话主题的多轮对话, “——”表示对话主题的分界点, 原话语中的“there, that, it”等共指或省略的信息, 通过重写模型得到了还原。本文基于 Gao 等^[9]的工作进行了扩展, 其通过邻近话语匹配和伪分割, 从未标记的对话数据中学习主题感知话语表示, 有效利用了未标记的对话。然而, 原工作仅针对话语间的句子级别进行去噪操作, 忽略了句子内部的噪声, 如共指和省略。为了进一步消除对话中的噪声, 恢复同一主题下的关键对话信息, 改进了模型计算主题相似性的余弦相似性计算操

作。重写后的相邻话语中共享词汇的增加,提高了主题相似的相邻话语的相似性计算得分,使得主题相似的正例和

负例的分数分布更明显,模型学习分辨对话中主题是否相似的能力得到提升。

表5 话语重写和主题分割示例

Table 5 Examples of utterance rewriting and dialogue topic segmentation

原话语	分割点	重写后的话语	分割点
I need to find a shopping center. 我需要找一个购物中心。	Start (True)	I need to find a shopping center. 我需要找一个购物中心。	Start (True)
The Stanford Shopping Center at 773 Alger Dr is 3 miles away. Would you like directions there? 距离位于 773Alger Dr 的斯坦福购物中心有 3 英里。您想知道那里的路线吗?		The Stanford Shopping Center at 773 Alger Dr is 3 miles away. Would you like directions to the Stanford Shopping Center? 773Alger Dr 的斯坦福购物中心距此 3 英里。您想知道前往斯坦福购物中心的路线吗?	
Yes please. 是的		Yes, I would like directions to the Stanford Shopping Center at 773 Alger Dr, please. 是的,请告诉我前往位于 773 Alger Dr 的斯坦福购物中心的路线。	
I sent all the info on the screen, please drive carefully! 我已将所有信息发送到屏幕上,请小心驾驶!	End (True)	I sent all the info on the screen, please drive carefully! 我已将所有信息发送到屏幕上,请小心驾驶!	End (True)
I am looking for a particular hotel. Its name is called autumn house. 我正在寻找一家特定的酒店。它的名字叫 autumn house。		I am looking for a particular hotel. The hotel's name is called autumn house. 我正在寻找一家特定的酒店。该酒店的名字叫 autumn house。	Start (True)
I have the Autumn House. It is at 710 Newmarket Road. Would you like the phone number? 我找到了秋屋。在新市场路 710 号。要电话号码吗?		I have the Autumn House hotel. The Autumn House hotel is at 710 Newmarket Road. Would you like the phone number of the Autumn House hotel? 我找到了秋屋酒店。Autumn House 酒店位于 Newmarket 路 710 号。您需要 Autumn House hotel 的电话号码吗?	
No thanks. Would you book the Autumn House for me starting on Monday, please. 不用了,谢谢。请您帮我预订星期一开始的秋屋酒店。	End (False)	No thanks. Would you book the Autumn House for me starting on Monday, please. 不用了,谢谢。请您帮我预订星期一开始的秋屋酒店。	
How many people would be staying and how many days will you be staying? 请问要住多少人,住几天?	Start (False)	How many people would be staying at the Autumn House and how many days will you be staying at the Autumn House? 有多少人会住在秋屋,您要在秋屋住几天?	
That's for 8 people and it's for 2 nights. 8 个人,住两晚。		The booking for the Autumn House is for 8 people and the booking for the Autumn House is for 2 nights. 秋屋的预订是 8 人,秋屋的预订是 2 晚。	
I'm sorry, your booking was unsuccessful. Would you like to book another day or a shorter stay? 很抱歉,您的预订没有成功。您想改天预订还是缩短住宿时间?		I'm sorry, your booking for the Autumn House was unsuccessful. Would you like to book another day or a shorter stay at the Autumn House? 很抱歉,您未成功预定 Autumn House。您想改天或缩短 Autumn House 的入住时间吗?	
Could you try Wednesday, instead? 您可以试试星期三吗?		Could you try Wednesday, instead of Monday? 您可以试试周三,而不是周一吗?	
Booking was successful. Reference number is : 3H0WHD4Z. 预订成功。参考编号为:3H0WHD4Z。	End (True)	Booking was successful. Reference number is : 3H0WHD4Z. 预订成功。参考编号为:3H0WHD4Z。	End (True)

结束语 本文系统地介绍了对话主题分割领域的相关工作,并提出了一种基于话语重写的无监督对话主题分割模型 UR-DTS。UR-DTS 模型进一步解决了无监督框架下未标记对话的有效利用问题。通过实验,验证了所提模型在两个评估数据集上均实现了最佳(SOTA)性能,证明了其有效性和优越性。

参考文献

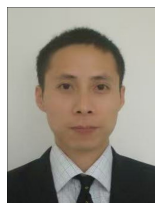
- [1] HEARST M A. Multi-paragraph segmentation of expository text[C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. 1994:9-16.
- [2] LI J, MONROE W, RITTER A, et al. Deep Reinforcement Learning for Dialogue Generation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:1192-1202.
- [3] BOKAEI M H, SAMETI H, LIU Y. Extractive summarization of multi-party meetings through discourse segmentation[J]. Natural Language Engineering, 2016, 22(1): 41-72.
- [4] XU Y, ZHAO H, ZHANG Z. Topic-aware multi-turn dialogue modeling[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021:14176-14184.
- [5] DAI Y, HE W, LI B, et al. CGoDial: A Large-Scale Benchmark for Chinese Goal-oriented Dialog Evaluation[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022:4097-4111.
- [6] LI S, ZHUANG S, SONG W, et al. Sequential texts driven cohesive motions synthesis with natural transitions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023:9498-9508.
- [7] SONG W, JIN X, LI S, et al. FineStyle: Semantic-Aware Fine-Grained Motion Style Transfer with Dual Interactive-Flow Fusion[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(11): 4361-4371.
- [8] SONG W, ZHANG X, GUO Y, et al. Automatic generation of 3d scene animation based on dynamic knowledge graphs and contextual encoding[J]. International Journal of Computer Vision, 2023, 131(11): 2816-2844.
- [9] GAO H, WANG R, LIN T E, et al. Unsupervised dialogue topic segmentation with topic-aware utterance representation[J]. arXiv, 2305. 02747, 2023.
- [10] GALLEY M, MCKEOWN K, FOSLER-LUSSIER E, et al. Discourse segmentation of multi-party conversation[C]//Proceedings of the 41st Annual Meeting of the Association for Compu-

- tational Linguistics. 2003;562-569.
- [11] RIEDL M, BIEMANN C. TOPICTILING: a text segmentation algorithm based on LDA [C] // Proceedings of ACL 2012 Student Research Workshop. 2012;37-42.
- [12] KOSHOREK O, COHEN A, MOR N, et al. Text Segmentation as a Supervised Learning Task [C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. 2018;469-473.
- [13] LO K, JIN Y, TAN W, et al. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence [C] // Empirical Methods in Natural Language Processing 2021. ACL, 2021;3334-3340.
- [14] HEARST M A. Text tiling: Segmenting text into multi-paragraph subtopic passages [J]. Computational Linguistics, 1997, 23(1):33-64.
- [15] GLAVAŠ G, NANNI F, PONZETTO S P. Unsupervised text segmentation using semantic relatedness graphs [C] // Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics. ACL, 2016;125-130.
- [16] SONG Y, MOU L, YAN R, et al. Dialogue Session Segmentation by Embedding-Enhanced Text Tiling [C] // Conference of the International Speech Communication Association (INTER-SPEECH 2016). 2016;2706-2710.
- [17] HE W, DAI Y, HUI B, et al. SPACE-2: Tree-Structured Semi-Supervised Contrastive Pre-training for Task-Oriented Dialog Understanding [C] // Proceedings of the 29th International Conference on Computational Linguistics. 2022;553-569.
- [18] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;4171-4186.
- [19] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019;3982-3992.
- [20] XING L, CARENINI G. Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring [C] // Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021;167-177.
- [21] DZIRI N, KAMALLOO E, MATHEWSON K, et al. Evaluating Coherence in Dialogue Systems using Entailment [C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;3806-3812.
- [22] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019;3982-3992.
- [23] LI Y, SU H, SHEN X, et al. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset [C] // Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017;986-995.
- [24] PU H, WANG L. Dialogue Segmentation based on Dynamic Context Coherence [C] // Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval. 2023;190-195.
- [25] SEE A, MANNING C D. Understanding and predicting user dissatisfaction in a neural generative chatbot [C] // Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021;1-12.
- [26] FANG Y, ZHANG H, CHEN H, et al. From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization [C] // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022;3859-3869.
- [27] MELE I, MUNTEAN C I, NARDINI F M, et al. Adaptive utterance rewriting for conversational search [J]. Information Processing & Management, 2021, 58(6):102682.
- [28] JIANG W, GU X, CHEN Y, et al. DuReSE: Rewriting Incomplete Utterances via Neural Sequence Editing [J]. Neural Processing Letters, 2023, 55(7):8713-8730.
- [29] NIEHUES J, CHO E, HA T L, et al. Pre-Translation for Neural Machine Translation [C] // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016;1828-1836.
- [30] JUNCZYS-DOWMUNT M, GRUNDKIEWICZ R. An Exploration of Neural Sequence-to-Sequence Architectures for Automatic Post-Editing [C] // Proceedings of the Eighth International Joint Conference on Natural Language Processing. 2017;120-129.
- [31] CHEN Y C, BANSAL M. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018;675-686.
- [32] WESTON J, DINAN E, MILLER A. Retrieve and Refine: Improved Sequence Generation Models For Dialogue [C] // Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI. 2018;87-92.
- [33] RASTOGI P, GUPTA A, CHEN T, et al. Scaling Multi-Domain Dialogue State Tracking via Query Reformulation [C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;97-105.
- [34] RIEZLER S, LIU Y. Query rewriting using monolingual statistical machine translation [J]. Computational Linguistics, 2010, 36(3):569-582.
- [35] CHEN B, SUN L, HAN X P, et al. Sentence Rewriting for Semantic Parsing [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016;766-777.
- [36] ABUJABAL A, SAHA ROY R, YAHYA M, et al. Never-ending learning for open-domain question answering over knowledge bases [C] // Proceedings of the 2018 World Wide Web Conference. 2018;1053-1062.
- [37] JIN D, LIU S, LIU Y, et al. Improving Bot Response Contradic-

- tion Detection via Utterance Rewriting[C]//23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2022). ACL, 2022:605-614.
- [38] GAO T, YAO X, CHEN D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[C]//2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). 2021.
- [39] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.
- [40] ZHANG J, ZHAO Y, SALEH M, et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization [C]//International Conference on Machine Learning. PMLR, 2020:11328-11339.
- [41] SMITH E M, WILLIAMSON M, SHUSTER K, et al. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:2021-2030.
- [42] FENG S, WAN H, GUNASEKARA C, et al. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset [C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:8118-8128.
- [43] BUDZIANOWSKI P, WEN T H, TSENG B H, et al. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:5016-5026.
- [44] ERIC M, KRISHNAN L, CHARETTE F, et al. Key-Value Retrieval Networks for Task-Oriented Dialogue[C]//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. 2017:37-49.
- [45] BEEFERMAN D, BERGER A, LAFFERTY J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34:177-210.
- [46] PEVZNER L, HEARST M A. A critique and improvement of an evaluation metric for text segmentation[J]. Computational Linguistics, 2002, 28(1):19-36.
- [47] EISENSTEIN J, BARZILAY R. Bayesian unsupervised topic segmentation[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008:334-343.



LI Tongliang, born in 1992, Ph.D, lecturer. His main research interests include artificial intelligence, natural language processing and large language model.



CHEN Xiaoming, born in 1980, master, engineer. His main research interests include artificial intelligence and document intelligent processing.

(责任编辑:何杨)