**文本生成大语言模型应用研究综述**

# 一、文本生成模型发展简述

文本生成是自然语言处理的关键任务之一。早期的统计语言模型,如n-gram通过统计语料中的词频、词组频率等特征来建模语言[1]。随着神经网络的发展,端到端的神经网络文本生成模型也在不断发展、完善和进步。

早期的序列到序列(Seq2Seq)模型采用双向LSTM等结构实现机器翻译等生成任务[2]。Transformer的提出改进了生成模型的长程依赖建模能力[3]。以Transformer为基础的GPT系列模型[4],在大规模语料预训练后,进一步提升了语言生成的逼真性和连贯性。另一类Seq2Seq模型BART[5]通过基于语料的去噪任务进行预训练,增强了模型的语言理解能力,在生成任务上也有出色表现。T5模型[6]构建了统一的文本到文本框架,使生成模型可以端到端地完成多种下游任务。

近年来,在自然语言处理领域,随着模型的参数规模不断扩大,诸如GPT-3[7]等模型问世以来,模型文本生成质量也越来越高。这些模型融合了预训练和迁移学习的思想,通过在海量数据上预训练获得通用语言能力,然后可快速迁移微调到下游生成任务[8]。它们表现出强大的语言理解和生成能力,在自然语言处理多个领域都取得了state-of-the-art的效果[9]。

预训练语言模型之所以能够表现出良好的语言处理能力,是因为它们通过大规模无监督的预训练获得了语义知识[10]。以BERT为例,它通过双向掩码语言模型在大规模文本语料上进行预训练,学习语义表示,这种半监督预训练范式使模型获得了语言的先验知识[11]。同时,预训练语言模型表现出很强的语言表示能力的迁移性,即使应用于不同的下游任务,也能取得不错的效果[12]。此外,预训练语言模型设计了transformer等统一的模型结构,这使它们在预训练后,可以灵活地微调应用于下游任务,并且通过在下游任务数据上进行迁移微调,就可以快速适配新的任务[13], [14]。所以,预训练语言模型的普适性使它们成为不同领域文本生成任务的有效基础模型,通过在特定领域进行微调,文本生成模型可以快速适配垂直领域,生成高质量结果[15]。

综上所述,从n-gram到Seq2Seq,再到预训练语言模型,文本生成模型不断发展,生成质量和逼真度持续改进,当前以Transformer为基础的预训练语言模型已成为自然语言生成的先进技术。

# 二、文本生成模型应用的相关研究

## 2.1模型与专业知识的结合

将专业知识有效地融入预训练语言模型是提升其针对不同领域应用能力的关键。目前研究主要从多个角度进行探索:

第一,构建专业领域的知识图谱,其中包含实体节点与关系边,形成领域知识体系,将知识图谱融入预训练过程,可以增强模型对专业词汇及概念的理解。例如Lee等人构建了医疗领域大规模知识图谱KGR,包含病症、疾病、药物等节点信息,并进行实体链接,他们将KGR与电子病历文本合并,以此预训练GPT-2, 实验结果显示预训练后的GPT-2生成的病历术语准确率提高14% [16]。

第二,从专业语料中抽取关键词汇,扩充预训练模型的词汇表,这使模型可以更好地对专业语言进行词向量表示和编码。例如Yuan等人从医学论文数据库中提取医学术语,用于扩充BioBERT的词表,他们在多个医学任务上验证,词汇引入确实提升了模型对专业语言的理解,除直接从语料抽取,也可综合各专业词表优化构建[17]。

第三,在大规模通用语料预训练的基础模型上,使用专业领域数据进行域自适应的进一步预训练。通过这种方法,可以使得模型可以适应专业语言环境和语义空间。例如FinBERT模型在通用BERT基础上,使用金融领域论文、新闻、报告等数据进行了domain-specific预训练,其对金融专业问答及情感分析的适应性显著提高[18]。

第四,通过多任务学习框架,将专业领域的训练任务加入预训练模型的微调过程。例如法律领域的Legal-BERT模型使用法律文书类型分类作为辅助任务,加入到BERT微调中,这帮助模型学习对法律专业术语和概念的理解[19]。

第五,建立知识存储模块,使用注意力机制来访问知识,指导模型生成时遵循专业知识。例如知识图谱可加入Transformer类文本生成模型中,当生成过程中注意到相关实体,就查询知识图谱以决定更合理的生成[20]。

综上,通过知识图谱引入、专业词汇扩充、域自适应预训练、多任务微调等方式,将专业知识与预训练语言模型有效结合,使其适配能力得到提升,这为依托预训练模型的自然语言生成技术在不同领域的应用提供了关键的技术支持。

## 2.2小样本场景的生成质量

文本生成模型在实际应用中,常面临小样本场景的挑战。小样本训练意味着模型无法获得充足的目标领域数据进行微调,这将影响其针对不同领域应用场景下的生成质量。当前研究主要从以下几个角度探索提高小样本生成效果的方法:

第一,通过元学习或迁移学习方法,提取模型在多个源任务上学习到的热启动参数或语言特征表达,快速适配新任务。Brown的研究显示,GPT-3通过在大量文本数据上预训练获得通用的语言特征,即使应用场景微调数据有限,也可以生成高质量文本[7]。

第二,使用数据增强方法扩充小样本,常见策略有同义词替换、随机删除、回译等。增强后数据量更充足,可更有效地训练模型,并提升模型的泛化能力。以医疗领域为例, Jason Wei等人通过医学术语替换等数据增加的方法使的模型拥有更多样化的电子病历数据进行训练,进一步提升模型文本生成质量[21]。

第三,多任务学习框架下进行模型训练。源领域任务可帮助模型学习本质语言特征表示,这样即使目标任务数据有限,也能得到更强的语言建模能力。在医学领域,许多学者同时使用了医学命名实体识别、医学缩写词识别等任务对模型进行多任务联合训练,实验结果也表明相比单任务而言,多任务学习具有更好的文本生成能力[22]。

第四,使用半监督学习的方法利用无标注数据对模型进行训练。Guo等人将语言模型在大规模无标注文本上进行预训练,试验结果表明模型具有更好的Embedding,并且泛化能力与性能得到显著提升[23]。

综上,元学习、数据增强、多任务学习等方法都可协同应对小样本学习问题,帮助生成模型在专业领域有限数据场景下获得更好的适配效果。

## 2.3生成内容的可解释性与安全性

当前深度学习生成模型因其“黑箱”性质,使得人类很难理解和控制其生成内容,许多学者针对这类问题展开了研究。

第一,对生成内容缺乏解释,用户难以理解模型的推理逻辑。生成模型的推理逻辑在许多专业领域尤为重要,如医疗、金融等领域。现有的研究中,许多学者通过使模型指出知识来源,或进行案例检索等方式对生成内容进行解释[24]。

第二,生成内容可能产生前后不对应或者错误的信息。大部分研究的解决方法是通过引入外部知识或监督信号约束规避该问题,例如连接知识库对文本生成内容进行评估[25],或通过标定样本判断生成内容的准确性[26]。

第三,存在生成有害内容的风险,如产生歧视性文本。目前学者解决该问题的有效方法是建立过滤机制,通过黑名单或利用对抗样本进行避让学习,以此减少不当内容生成[27]。

第四,模型可被对抗样本欺骗产生错误输出。解决该问题主要方法是增强模型的稳健性,可通过数据增强、对抗训练等方式提高模型的抗扰动能力[28]。

第五,用户难以控制生成属性如语气、长度等[29]。这需要建立生成过程的控制模块,直接对模型输出内容的形式进行直接控制,或者或通过强化学习的方法让系统适应用户喜好[30]。例如在Ouyang等人的研究中,研究者设计了一个模拟用户反馈的环境,该环境可以评价语言模型生成的文本是否遵循了指定的指令,然后基于这个反馈信号进行强化学习,以训练模型逐步适应用户的偏好,从而产生更符合指令的生成结果[31]。

综上,增强文本生成的可解释性、安全性和可控性仍面临挑战,需要生成模型与解释方法、知识库、对话管理等技术的融合,以产生更可信、可控、负责任的生成结果。

## 2.4用户交互的自然性与连贯性

改进任务导向对话系统中用户交互的自然性和连贯性,是让对话系统更完善的关键方面。具体来说,主要包括以下几个关键点:

首先,需要引入用户模拟技术,通过构建不同类型的用户模型,尽可能真实地模拟出多变的用户交互方式。目前已有研究设计出了有不同人格特点的虚拟用户,它们在交互方式上有所差异,比如有的用户表达直白,有的用户比较悠长,有的用户语气强硬,有的用户体现出明显个性。构建大规模的这类用户模拟对话数据集,有助于让对话系统在训练过程中接触到多样的语言风格,从而适应多种交互方式,提高用户体验的自然性[32]。

其次,利用预训练语言模型生成更加逼真、流畅的系统响应,取代简单的匹配模板。通过在大规模语料上进行自监督式预训练,预训练语言模型学会了语言的先验知识,没有模板的限制,可以产生更富有创造性的响应语句,这种生成方式更符合人类的表达习惯,让回复更加地道和逼真[33]。一些研究也在探索个性化的语言生成,让系统捕捉特定用户的语言风格偏好,从而产生对个别用户更富有亲和力的交互语言风格[34]。

最后,需要进行开放域的对话数据预训练,而不仅仅是针对特定垂直领域的闭域训练,以获得对常见话题和scenarios更强大的建模能力,从而更好地处理各种跳跃式的用户提问,不会因话题跃迁而无法响应[35]。一些工作采用大规模的开放域对话数据进行预训练,也通过强化学习的方式让对话系统在真实用户交互中不断优化改进,全面适应复杂多变的人类交互模式[36]。

总而言之,想要使对话系统真正达到高质量的自然语言交互,需要让对话生成模型与其他模块如对话管理、用户模拟等进行深度融合,使各部分协同支持自然、连贯的交互体验。这需要语言、心理、认知等多个学科领域的联合研究,以真正达到人机交互的自然顺畅

## 2.5多语言和多模式的统一生成

目前在多语言生成方面的研究中,一个重要途径是构建跨语言的预训练语言模型,通过在多种语言的大规模文本数据上进行知识迁移和表示学习,获得对不同语言的理解能力。这样预训练后的模型可以更好地处理多语言的下游生成任务,如机器翻译等[37]。另一种方法是在单一模型中同时使用不同语言的数据进行多任务训练,使模型在学习过程中对各语言的特征建立统一的感知,从而获得语言自适应的生成能力[38]。

关于多模式生成,一种思路是将图像、音频等不同模态信息作为文本生成的辅助输入,其可为文本提供额外的上下文和语义支持,以此增强模型的语义理解力,从而在多模式条件下指导生成更加准确合理的文本。例如,图像内容可以作为视觉先验来帮助文本更准确地描述该图像,音频也可提供语调、情绪等辅助理解的线索,不同模态间的语义对齐和理解可使多模式文本生成更加精准[39]。

另一种思路是构建统一的序列到序列生成模型,支持多模式的条件化输入与输出,实现端到端的生成训练和推理。这类模型采用统一的Encoder-Decoder结构,可同时处理图像、文本、语音等不同模态信息的编码和解码生成[40]。相比分别建模,端到端统一框架可以更简洁高效。一些研究也提出了让不同模态之间相互指导纠正的方式,进行迭代优化生成效果,以通过跨模态的互学习进行知识迁移[41]。

尽管当前多语言多模式统一生成仍面临一定的科研挑战,但已有的研究成果显示了这一方向的重要潜力。未来要进一步推进这一领域的发展,需要从多个方面进行努力:在表示学习上,强化跨语言和跨模态的表示对齐,使不同语言和模态获得语义级别的匹配;在知识迁移上,探索不同语言和模式之间知识传输的方式,实现互补的语义建模;在模型构建上,设计统一的Seq2Seq结构,简化多语言多模态生成任务的建模和训练;在训练策略上,采用多任务学习、迁移学习等方式,提升模型的跨语言跨模态的语义建模能力。只有文本生成模型与跨语言表示学习、多任务迁移技术、多模态理解等方法实现真正深入的融合,才能使语言生成系统真正具备多语言多模式环境的自然适应能力。这也将是自然语言生成领域的重要发展方向。

# 三、文本生成模型在文旅与博览行业的应用状况

在旅游领域,基于预训练语言模型构建的聊天机器人,已经成为应用自然语言生成技术的重要场景之一。这类聊天机器人可以提供路线规划、景点推荐等各类旅游咨询服务,并以流畅自然的语言进行人机交互,极大地增强了旅游信息查询的智能化和便捷性[42]。与此同时,根据用户配置文件和旅行偏好生成个性化的旅游内容也成为热点,文本生成模型可以自动化地生产符合个人兴趣的策划方案、游记等创意内容[43],这也将提升旅游体验的个性化程度。此外,多语言旅游内容转换是文本生成模型可以承担的另一关键任务,机器翻译系统可以自动实现不同语言之间的旅游资讯翻译和本地化,助力旅游企业进行多语言化的内容建设,拓展其服务范围[44]。

在博物馆、展览馆等文化场馆,语音和文字解说产品是重要的展示和讲解形式。针对特定的展览主题,使用条件文本生成模型可以高效自动化地创建大量丰富生动的讲解内容,与传统人工编撰相比,极大地提升了内容制作的效率和多样性。此外,文本生成技术也可以与图像、视频生成的技术相结合,实现多模态的历史文化遗产的宣传内容产出,使历史文化资源的传播手段更加丰富互动,也更具吸引力[45]。

除了以上典型应用场景,文本生成也可广泛支持文旅企业的营销策划、广告创意等商业化内容建设。结合用户行为数据和兴趣喜好模型,可以使营销文案更加个性化。文本生成技术同时也可辅助自动化创作相关的游记、漫游小说等创意性内容,这些应用都将丰富旅游和文化产业的创意内容建设[43]。

总而言之,文本生成技术可以极大丰富文旅和博览领域的创意内容建设和用户体验,具有非常广阔的应用前景。搭配个性化、多语言、多模态等能力,文本生成可大幅提升这些典型服务业的产品形态和运营效率。未来随着生成技术的不断成熟,其在文旅博览领域的应用空间还将持续扩大。文本生成不仅可以成为文旅企业提供服务的重要手段,也将提升用户的旅游体验和参与感。充分利用文本生成技术,将会给文旅产业带来革命性的变化与发展。

# 四、研究展望

## 4.1强化交互的对话生成

对话生成领域目前致力于探索如何实现更智能、更自然的人机交互。在这方面,强化交互能力被认为是关键的研究方向。一些最近的工作采用用户模拟技术,通过设计不同个性的用户模型并构建大规模的模拟对话数据集,来训练对话系统以适应多样的交互方式[46]。另一些研究则尝试利用强化学习框架,以“试错”的方式根据用户在交互中的反馈来调整和优化对话策略,使对话系统产生更符合用户偏好的回应[36]。

此外,语言风格的自适应也被探索用来实现个性化的对话,这需要对话生成模型更敏锐地感知语言细微差异并洞察用户喜好[34]。加入记忆模块使对话系统考虑历史信息并产生连贯一致的回应也很重要[47],外部可扩展的长期记忆被证明是增强回应质量的重要补充[48]。

强化交互需要对话生成与用户模拟、强化学习、语言适应、记忆网络等各模块及技术的融合,以产生更智能、更自然的交互体验。实现真正人性化的对话系统仍有一定挑战,还需要语言、心理、认知等多个学科的联合研究推动。

## 4.2丰富多模式理解和生成

随着人工智能的发展,文本生成任务正在朝着多模式方向演进,以实现更加沉浸式的生成效果。多模式生成需要在表示学习、知识传输、模型构建等多个层面进行创新。

在多模态表示学习方面,一些研究探索了通过联合学习文本、图像、音频等多源数据实现高质量的跨模态语义表示,为多模式生成奠定基础[49]。具体而言,可以通过让不同模态的表示互学习、让表示预测其他模态信息等方式进行表示对齐。一些工作在多任务学习框架下,共同进行多模态表示提取,利用不同模式的数据相互增强表示学习[50]。

另一方面,也有研究尝试在不同模式之间进行知识传输,以丰富单一模式的语义理解,从而改善多模式生成的质量。例如可以使用图像内容作为先验信息来增强文本表示的学习;也可以让文本描述支持更好的图像理解。跨模态知识传输充分利用了不同模态的互补性[40]。

在模型构建方面,一些最近的研究构建了统一的序列到序列生成网络,支持条件化的多模式输入与输出,实现端到端的生成[40]。这种统一框架简化了多模式组合条件下的建模。也有一些研究让不同模式之间相互指导和纠正,进行迭代优化生成结果,通过模式之间的不一致性进行知识迁移[41]。未来多模式生成还可支持用户多模式的输入,并产生同步的多模式响应,实现更沉浸的交互体验,这需要建模不同感知通道之间的协同,进行跨模式的推理[51]。

总体而言,多模式生成需要表示学习、知识传输、统一建模等多个方面技术的创新,以实现对多源信息的深度理解和建模。这也将是自然语言生成领域的重要发展方向。

# 参考文献

[1] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–394, 1999.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[3] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, and others, “Improving language understanding by generative pre-training,” 2018.

[5] M. Lewis *et al.*, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *ArXiv Prepr. ArXiv191013461*, 2019.

[6] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[7] T. Brown *et al.*, “Language models are few-shot learners,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[8] X. Liu *et al.*, “GPT understands, too,” *ArXiv Prepr. ArXiv210310385*, 2021.

[9] Y. Goldberg, “Assessing BERT’s syntactic abilities,” *ArXiv Prepr. ArXiv190105287*, 2019.

[10] J. Sarzynska-Wawer *et al.*, “Detecting formal thought disorder by deep contextualized word representations,” *Psychiatry Res.*, vol. 304, p. 114135, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv Prepr. ArXiv181004805*, 2018.

[12] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” *ArXiv Prepr. ArXiv180106146*, 2018.

[13] S. Gururangan *et al.*, “Don’t stop pretraining: Adapt language models to domains and tasks,” *ArXiv Prepr. ArXiv200410964*, 2020.

[14] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?,” in *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, Springer, 2019, pp. 194–206.

[15] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “Ctrl: A conditional transformer language model for controllable generation,” *ArXiv Prepr. ArXiv190905858*, 2019.

[16] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[17] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” *NPJ Digit. Med.*, vol. 4, no. 1, p. 86, 2021.

[18] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, “Finbert: A pre-trained financial language representation model for financial text mining,” in *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 2021, pp. 4513–4519.

[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” *ArXiv Prepr. ArXiv201002559*, 2020.

[20] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, “Commonsense knowledge aware conversation generation with graph attention.,” in *IJCAI*, 2018, pp. 4623–4629.

[21] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in *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, pp. 6382–6388.

[22] S. Zhao, T. Liu, S. Zhao, and F. Wang, “A neural multi-task learning framework to jointly model medical named entity recognition and normalization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 817–824.

[23] J. Guo, W. Che, H. Wang, and T. Liu, “Revisiting embedding features for simple semi-supervised learning,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 110–120.

[24] R. Thoppilan *et al.*, “Lamda: Language models for dialog applications,” *ArXiv Prepr. ArXiv220108239*, 2022.

[25] M. Gaur, P. He, Y. Choi, and B. Peng, “Knowledge-aware conversational semantic parsing over web tables,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[26] E. Durmus, H. He, and M. Diab, “FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[27] X. L. Li, Y. Suhara, J. R. Wang, and J. Hsu, “Mitigating gender bias amplification in distribution by posterior regularization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[28] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, “Automatic detection of generated text is easiest when humans are fooled,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[29] B. Krause, A. Kaushik, Z. Lipton, and A. Anastasopoulos, “GeDi: Generative discriminator guided sequence generation,” *ArXiv Prepr. ArXiv200906367*, 2020.

[30] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, “Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems,” *ArXiv Prepr. ArXiv180406512*, 2018.

[31] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, 2022.

[32] I. Gür, D. Hakkani-Tür, G. Tür, and P. Shah, “User modeling for task oriented dialogues,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 900–906.

[33] D. Ham, J.-G. Lee, Y. Jang, and Y. Kim, “End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 583–592.

[34] H. Song, Y. Liu, W. Bi, and X. Zhang, “Generating persona consistent dialogues by exploiting natural language inferences,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 5427–5433.

[35] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, “Recent advances and challenges in task-oriented dialog systems,” *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 2011–2027, 2020.

[36] J.-C. Gu, Z.-H. Ling, X. Zhu, and Q. Liu, “Dually interactive matching network for personalized response selection in retrieval-based chatbots,” *ArXiv Prepr. ArXiv190805859*, 2019.

[37] Y. Liu *et al.*, “Multilingual Denoising Pre-training for Neural Machine Translation,” *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 726–742, 2020.

[38] Y. Lu, K. Kann, Y. Yu, D. Pappas, and J. Henderson, “A Multi-task Self-Supervised Learning Framework for Multilingual Representations,” *ArXiv Prepr. ArXiv210903901*, 2021.

[39] A. Burns, R. Tan, K. Saenko, S. Sclaroff, and B. A. Plummer, “Language features matter: Effective language representations for vision-language tasks,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7474–7483.

[40] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, NIH Public Access, 2019, p. 6558.

[41] C. Xia, C. Xiong, P. Yu, and R. Socher, “Composed variational natural language generation for few-shot intents,” *ArXiv Prepr. ArXiv200910056*, 2020.

[42] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamdi, “AI chatbot for tourist recommendations: a case study in the city of Jeddah, Saudi Arabia,” 2020.

[43] Q. Gao, G. Trajcevski, F. Zhou, K. Zhang, T. Zhong, and F. Zhang, “DeepTrip: Adversarially understanding human mobility for trip recommendation,” in *Proceedings of the 27th ACM SIGSPATIAL international conference on advances in geographic information systems*, 2019, pp. 444–447.

[44] S. Narzary *et al.*, “Attention based English-Bodo neural machine translation system for tourism domain,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2019, pp. 335–343.

[45] X. He and L. Deng, “Deep learning for image-to-text generation: A technical overview,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 109–116, 2017.

[46] N. Crook, L. Ross, S. Pischedda, and G. Ta^no, “Simulating the crowd: An agent-based model of audience engagement with street performance,” *AI Soc.*, vol. 35, no. 3, pp. 611–628, 2020.

[47] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, “Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems,” *ArXiv Prepr. ArXiv180406512*, 2018.

[48] A. Madotto, C.-S. Wu, and P. Fung, “Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1468–1478.

[49] Q. Wang, L. Zhang, and Y. Zhang, “Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12416–12425.

[50] H. Le, D. Sahoo, N. F. Chen, and S. C. Hoi, “Multimodal transformer networks for end-to-end video-grounded dialogue systems,” *ArXiv Prepr. ArXiv190701166*, 2019.

[51] Y. Zhang, Z. Ou, and Z. Yu, “Task-oriented dialog systems that consider multiple appropriate responses under the same context,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 9604–9611.