无监督的领域自适应机器阅读理解方法

刘 皓[1] 洪 宇[1] 朱巧明[1),2)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](江苏省大数据智能工程实验室 江苏 苏州 215006)

摘 要 受益于面向大规模语言学资源的深度学习,预训练语言模型有着较强的语义表示学习能力.其能够借助特定任务场景下的迁移学习,在优化模型性能方面提供重要的支持.目前,预训练语言模型已被引入机器阅读理解研究领域,并展现了较好的优化能力.然而,针对特定领域的数据,微调后的预训练模型仍存在领域适应性问题,即无法解决未知领域中新颖的语言现象.为此,本文提出了一种融合迁移自训练和多任务学习机制的无监督领域自适应模型.具体而言,本文结合生成式阅读理解网络和掩码预测机制形成了多任务学习框架,并利用该框架实现跨领域(源领域至目标领域)的无监督模型迁移技术.此外,本文设计了文本规范化和迁移自训练模式,以此促进目标领域的数据分布适应源领域的数据分布,从而提高模型迁移学习的质量.本文将 TweetQA 作为目标领域数据集,将 SQuAD、CoQA 和 NarrativeQA 作为源领域数据集进行实验.实验证明,本文所提方法相较于基线模型有显著提升,在 BLEU-1、METEOR 和 ROUGE-L 指标上分别提升了至少 2.5、2.7 和 2.0 个百分点,验证了其优化领域适应性的能力.

关键词 无监督领域自适应;迁移自训练;多任务学习;生成式阅读理解;掩码预测

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02133

Unsupervised Domain Adaptive Machine Reading Comprehension Method

LIU Hao[1] HONG Yu[1] ZHU Qiao-Ming[1),2)]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

[2)](*BigData Intelligence Engineering Lab of Jiangsu Province, Suzhou, Jiangsu* 215006)

Abstract Benefiting from deep learning for large-scale linguistic resources, pre-trained language models have obtained strong semantic representation learning capabilities. It can use transfer learning in specific task scenarios to provide important support in optimizing model performance. Pre-trained language models such as BERT and UniLM have been widely used in natural language processing fields such as text summarization, machine translation, sentiment analysis and so on. Nowadays, pre-trained language models have been introduced into the field of machine reading comprehension, and have shown considerable optimization capabilities. However, for domain-specific data, the fine-tuned pre-trained models still suffer from weak domain adaptability. In other word, they cannot tackle novel language phenomena in unknown domains. In the social media field, it is difficult to form a standardized and normalized language representation due to the characteristics of "colloquial" and "symbolic" text. In addition, in the practical application scenarios of "innumerable" domain classes, the timeliness of manual annotation is often difficult to guarantee. The current research is mainly oriented to the field of text normalization, and the

收稿日期:2021-08-16;在线发布日期:2022-04-02.本课题得到国家重点研发计划项目(2020YFB1313601)、国家自然科学基金(62076174,61836007)资助.**刘 皓**,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为机器阅读理解.E-mail:liuh08060711@gmail.com.**洪 宇**(通信作者),博士,教授,主要研究领域为机器阅读理解、事件抽取等.E-mail:tianxianer@gmail.com.**朱巧明**,博士,教授,主要研究领域为机器阅读理解、中文语言处理等.

existing MRC models based on supervised learning necessarily require large-scale training data, while the data for the social media field is relatively scarce. Therefore, although we can fine-tune it based on large-scale pre-trained language models, existing social media data is not large enough to support a complete language system because it is different from pre-trained corpora in terms of specific linguistic phenomena. In addition, previous researches are mainly based on the methods of cloze, multiple choice and extraction models to design MRC models. Such models lack generalization ability for real MRC data, and generative models are closer to practical applications than the above models. To this end, from model level, this paper proposes an unsupervised domain adaptive model that combines transfer self-training and multi-task learning mechanisms. Specifically, this paper combines generative reading comprehension network and mask prediction mechanism to form a multi-task learning framework, and utilizes this framework to implement the cross-domain (source domain to target domain) unsupervised model migration technology. In terms of details, the training process of this model is divided into three stages. Firstly, the pre-trained language model is fine-tuned using source domain data. Secondly, the pseudo samples of the target domain are screened by confidence screening method. Finally, the source domain model is fine-tuned by using the selected pseudo-samples and multi-task mechanism. In addition, from data level, this paper designs a text normalization and transfer self-training mode to promote the data distribution of the target domain to adapt to the data distribution of the source domain, thereby improving the quality of model transfer learning. At the same time, the paper verifies the adaptability of some existing domain adaptive methods in generative reading comprehension through experiments. This article uses TweetQA as the target domain dataset, and uses SQuAD, CoQA, and NarrativeQA as the source domain datasets for experiments. Experiments have proven that the proposed method obtains a significant improvement compared to the baseline, yielding at least 2.5%, 2.7%, and 2.0% improvements for BLEU-1, METEOR, and ROUGE-L, respectively. This verifies the ability to optimize domain adaptability. In addition, experiments show that some existing domain adaptive methods can not be directly applied to the generative machine reading comprehension.

Keywords unsupervised domain adaptation; transfer self-training; multi-task learning; generative reading comprehension; mask prediction

1 引 言

机器阅读理解(Machine Reading Comprehension,简称 MRC)旨在让模型理解段落语义,并回答给定的问题(自然语言形成的问句). MRC 根据答案获取方式主要分为完型填空式、多项选择式、抽取式和生成式四种[1]. 完型填空式通过上下文语义表示预测缺失的词或实体;多项选择式凭借分类器选择给定答案候补的正确选项;抽取式借助判别器从给定的段落中自动抽取答案的字符串;生成式则依赖生成器自动生成答案的表述语言. 本文聚焦生成式的方法设计与分析.

近年来,大规模 MRC 数据集的构建(如 SQuAD[2]、CoQA[3]、NarrativeQA[4] 等)促进了阅读理解技术的发展. 尤其,借助大规模语言资源进行深度学习的预训练语言模型,能够有效提升 MRC 的性能表现,如 BERT[5]、UniLM[6] 和 UniLMv2[7].

尽管如此,现有的机器阅读理解模型依然存在领域适应性问题,即在欠缺领域知识的前提下,难以对特定领域的语言现象形成精准的特征表示(如社交媒体中口语化和符号化的表述语言). 表 1 展示了社交媒体领域中的三种语言现象.

现象 1 展示了社交媒体领域中谐音以及缩写的口语化语言现象,其段落中包含的"*u*"、"*4*"和"*conv*"等口语化文字,分别为规范化文本"*you*"、"*for*"和

表 1　TweetQA 独特的语言现象

现象 1： **段落**：*@InStyle: On KWs Cover: Beautiful statement. Thank* ***u 4*** *opening this* ***convo****. Its an important 1 that needs to be had. kerry washington (@kerrywashington) February 5, 2015* **问题**：*What is kerry thankful for?* **答案**：*Instyle opening the conversation*
段落：*@InStyle：在KWs的采访上：漂亮的声明. 谢谢开启此次对话. 这一点很重要. 克里华盛顿(@克里华盛顿) 2015 年 2 月 5 号* **问题**：*克里感谢什么？* **答案**：*InStyle开启了此次对话*
现象 2： **段落**：*Writing a bill w/* ***@MartinHeinrich*** *to prevent anyone convicted of domestic violence - be it in criminal or military court −3 from buying a gun - JeffFlake (@JeffFlake) November 7, 2017* **问题**：*With whom is jeff writing a bill?* **答案**：*Martin heinrich*
段落：*与@马丁·海因里希一起撰写法案，以防止任何被判犯有家庭暴力罪的人(无论刑事法庭或军事法庭)购买枪支 - 杰夫·弗拉克(@杰夫·弗莱克) 2017 年 11 月 7 日* **问题**：*杰夫在和谁一起撰写法案？* **答案**：*马丁·海因里希*
现象 3： **段落**：*Everything we know about love is a lie.* ***#RIPBrangelina*** *… GIPHY (@giphy) September 20, 2016* **问题**：*Who does giphy want to rest in peace?* **答案**：*Brangelina*
段落：*我们所知道的关于爱情的一切都是谎言. 安息吧·布兰吉利纳…吉菲(@吉菲) 2016 年 9 月 20 日* **问题**：*吉菲想让谁安息？* **答案**：*布兰吉利纳*

“*conversation*”的谐音或缩写.

现象 2 展示了符号化语言现象的一种形式，即“@”符号后紧跟人名或者机构名的连写形式(如“*@MartinHeinrich*”)，因此“@”符号具有标记人或机构等命名实体的作用. 此外，标准答案可能是连写词的分离形式(此例的标准答案“*martin heinrich*”就由“*@MartinHeinrich*”去掉“@”分离形成).

现象 3 展示了符号化语言现象的另一种形式，即“#”符号后紧跟多个实体的连写形式(如“*#RIPBrangelina*”). 其中，连写实体词前端的“#”符号具有标记段落主题的特殊作用. 连写实体词即为段落的中心主旨. 此外，连写实体中的部分实体也可作为标准答案(此例中的“*#RIPBrangelina*”由“*RIP*”和“*Brangelina*”两种命名实体混合组成，其段落的内容围绕“*悼念布兰吉利纳*”的主题展开，并且其中的人物实体“*Brangelina*”为最终的标准答案).

口语化和符号化的表述语言在建模过程中，适应性问题突出表现为分布式表示的跨领域欠拟合. 显然，对目标领域构建可观的标注样本，有助于解决上述问题. 但在领域类“林林总总”的实际应用场景中，人工标注的时效性往往难以得到保证. 针对这一问题，现有工作集中在无监督的领域自适应方法上开展研究，其从数据和模型两个层面分别进行应用优化：

数据层面. 相应技术通过对目标领域段落进行文本规范化操作，使其文本适应源领域文本格式，进而减小不同领域的数据分布差异. Huang 等人[8]. 针对文本不规范问题，提出了四种基于规则的规范方法，并在抽取和生成答案环节的后端，建立答案选择模块，有效发挥了两者各自的优势.

模型层面. 相应技术将已标注数据的源领域知识作为前期学习对象，形成初始的 MRC 模型. 在此基础上，相关方法尝试将无标注领域知识的表示，拟合到源领域知识的表示模式中. 上述方法具有无监督领域自适应的优势[9]. 其中，Chung 等人[10]使用了一种简单的自训练方法生成目标领域答案. Cao 等人[11]通过条件领域对抗网络[12](Conditional Domain Adversarial Networks，简称 CDAN)排除低置信度样本，从而实现对答案生成过程的降噪处理. 尽管 Cao 等人[11]提高了筛选质量，但其直接对源领域模型进行微调并不能充分提取目标领域的个性表示，模型的泛化能力仍有较大的提升空间[13].

本文继承了上述两个层面的 MRC 自适应优化思路，提出了一种融合迁移自训练与多任务学习机制的模型架构. 具体而言，本文首先对目标领域段落进行文本规范化操作，再使用源领域数据集微调预训练模型. 在此基础上，本文考察源领域模型生成目标领域样本的能力，并以此为依据，筛选置信度大于设定阈值的样本对源领域模型进行微调. 通过这种方式，源领域模型可学到目标领域的命名实体知识，以及段落问题与答案潜在的映射关系，从而降低模型在目标领域上的困惑度. 同时，本文结合多任务学习机制，以生成式阅读理解任务为主任务，掩码预测任务为辅助任务，构建编码模型. 其在生成答案的同时，预测掩码遮蔽词，借以加强模型对目标领域语义信息的理解和特征建模，从而进一步提高模型在跨领域应用过程中的泛化能力.

本文模型由模型编码层和下游任务输出层组成. 编码层架构基于 Transformer[14]进行设计，并借鉴了 UniLMv2[7]的自注意力掩码矩阵构建方法，其用于控制自注意力的作用范围. 下游任务层包括生成式阅读理解模型，以及掩码预测模型. 实验验证，该模型显著提高了 MRC 的跨领域适应性. 总体上，本文的主要贡献在于：

(1) 针对社交媒体阅读理解的数据特点，本文

结合文本规范方法，提出一种融合迁移自训练与多任务学习的模型架构；

(2)验证了部分经典领域自适应方法以及抽取式阅读理解前沿领域自适应方法在生成式方向的适用性；

(3)从困惑度的角度解释了迁移自训练能够较大幅度提升模型迁移性能的原因；

(4)实验主要以 TweetQA[15] 作为目标领域数据集，将 SQuAD[2]、CoQA[3] 和 NarrativeQA[4] 作为源领域数据集. 实验结果显示，本文方法的性能较基线均有提升. 此外，迁移自训练后的模型的跨领域性能，与利用本领域训练数据进行训练所得的性能，具有较高的可比性.

本文第 2 节回顾技术前沿；第 3 节介绍任务定义以及相关的预备知识(含生成式阅读理解和无监督领域自适应训练方法)；第 4 节详细解析本文提出的生成式 MRC 领域自适应优化方法；第 5 节给出实验及分析；第 6 节总结全文.

2 相关工作

本节首先围绕机器阅读理解介绍近年来主流的大规模数据集，以及在不同数据集上进行独立处理的前沿 MRC 框架，主要包括生成式方法和多任务架构. 在此基础上，本文介绍无监督领域自适应的主要方法，最后陈述本文方法的特点.

2.1 数据集和前沿 MRC 方法

机器阅读理解是自然语言处理中最重要的任务之一. 其旨在驱动模型理解段落语义，并在此基础上回答给定的问题. 神经机器阅读理解是近几年的主流研究方法，其将神经网络应用于 MRC 语言现象的深度学习过程中，借助表示学习、编码和解码器的自动建模，构建具有泛化能力的 MRC.

基于监督学习的 MRC 模型必然需要较大规模的训练数据. 因此，国际上不同学术组织开展了大量数据标注工作，比如 SQuAD[2] 和 MS MARCO[16]. 这些数据集接近于现实应用，对现有机器阅读理解模型构成了较高挑战. CoQA[3]、QuAC[17] 等数据集的出现促进了多轮对话式阅读理解的研究.

前人研究主要基于完形填空式(Dhingra 等人[18]、Kadlec 等人[19])、多项选择式(Zhang 等人[20]、Parikh 等人[21])以及抽取式(Zhang 等人[22]、Yu 等人[23])的模型进行方法设计. 这些方法通常难以适应真实场景. 其原因在于严格定义的填空、选择或抽取模式，对于大规模真实 MRC 数据缺少泛化能力. 而生成式阅读理解因其答案的自由形式，相较于上述模式，更贴近实际应用场景.

本文集中在生成式阅读理解方面展开讨论. Bao 等人[7] 基于 Transformer[14] 架构，提出了一种既可进行自然语言理解任务又可进行自然语言生成任务的预训练语言模型 UniLMv2[7]，其使用一种新颖的伪遮蔽语言模型. 通过仿照测试阶段的生成模式，构建注意力掩码遮蔽矩阵，UniLMv2 便可遮蔽不可观测的上下文信息. 在利用 UniLMv2 进行语义编码时，每一层的 Transformer 都将利用上述掩码遮蔽矩阵，对可观测的上下文进行动态的控制. 特别地，利用掩码矩阵，其能够在对源序列进行编码的同时，完成针对目标序列的解码.

在融入答案抽取和问题类型分类的多任务机制后，UniLMv2[7] 在多个生成式阅读理解数据集上取得了良好的性能[24]. Nishida 等人[25] 提出一种端到端的多任务机器阅读理解模型. 该模型能够同时进行文档排序、答案判断和生成式阅读理解任务，其在 MS MARCO[16] 生成阅读理解任务上取得同期最佳性能.

2.2 无监督领域自适应 MRC 研究

近期，相应研究融入了无监督领域自适应方法，借以提升机器阅读理解的跨领域鲁棒性. Wang 等人[26] 通过使用词性标注工具，标注段落中潜在答案，生成问题的同时使用梯度反转寻找不同领域的共性表示. 区别于通过词性标注选择候选答案的方法，Shakeri 等人[27] 通过端到端的模型架构，在生成问题的同时也生成答案，以此为目标领域添加标注数据，形成了一种应用数据增强的解决手段. Lee 等人[28] 通过在自训练过程中引入强化学习算法，提高了生成问题的可靠性，并以此减小生成样本与目标领域数据的分布差异. Yue 等人[29] 通过结合最大均值差异[30](Maximize Mean Discrepanc，MMD)和对比学习的对比领域自适应方法，减小源领域与目标领域数据在 MRC 模型上的特征表示差异，同时增强模型对段落中的潜在答案的判别能力. Chung 等人[10] 和 Cao 等人[11] 专注于考察目标领域数据中源领域问题分布的情况，并借助自训练模型提取目标领域的个性表示(本文的工作也基于此构建自适应模型).

值得指出的是，前人工作主要集中在抽取式阅读理解开展研究，而对于生成式阅读理解的无监督领域适应性问题却鲜有研究. 此外，前人工作在规范

文本之间进行领域迁移实验，而对于推特等具有特定语言现象的文本领域的研究却较为稀缺. 特别地，Cao 等人[11]虽然通过设定阈值实现噪声样本的过滤，并以此提高了 MRC 模型对目标领域的适应能力，但本文认为模型的泛化能力，还可以通过多任务学习进一步增强.

受上述工作启发，本文提出将多任务和迁移自训练进行结合，优化生成式阅读理解的跨领域适应性. 特别地，本文将文本规范化作为预处理，从数据层面支撑上述多任务迁移自训练过程.

3　问题形式化

本节主要介绍生成式阅读理解以及无监督领域自适应的任务定义、输入输出模式以及需要解决的关键问题.

3.1　生成式阅读理解

给定段落 $P_u=\{x_{p_u}^1,x_{p_u}^2,x_{p_u}^3,\cdots,x_{p_u}^L\}$，以及相关问题 $Q_u=\{x_{q_u}^1,x_{q_u}^2,x_{q_u}^3,\cdots,x_{q_u}^J\}$（其中 x 表示词项），生成式阅读理解旨在理解 P_u 和 Q_u 语义信息，并生成答案 $y_u=\{a_u^1,a_u^2,a_u^3,\cdots,a_u^K\}$（其中 a_u 表示答案中的词项），其中 L、J、K 分别表示段落、问题以及答案的长度. 区别于抽取式阅读理解，其生成的答案不一定是段落的连续片段.

3.2　无监督领域自适应的训练方法

假定有标签的源领域数据集和无标签的目标领域数据集，共有 N_s 个源领域样本 $\{(x_{u_j},y_{u_j})\}_{j=1}^{N_s}$（其中，$x_{u_j}=(Q_{u_j},P_{u_j})$，$y_{u_j}=\{a_u^i\}_{i=1}^K$），以及 N_t 个目标领域的无标签数据，即 $\{(x'_{u_j})\}_{j=1}^{N_t}$. 假设源领域的输入 x_{u_j} 经过模型的表示分布服从 D，目标领域的输入 x'_{u_j} 经过模型的表示分布服从 D'，且二者分布不同，即 $D\neq D'$. 无监督领域自适应旨在通过深度神经网络，减小不同数据分布中间隐状态的差异，或者通过文本规范化方法，减小数据分布差异，从而促进已有机器学习模型在目标领域上达到近似的性能，即保证跨领域的鲁棒性.

4　生成式 MRC 领域自适应优化方法

本文结合了数据层面和建模层面的自适应优化方法，并在建模过程中使用了多任务框架和迁移学习. 本节首先介绍基于文本规范化的自适应优化方法，其次讲解基于多任务的迁移自适应模型.

4.1　基于文本规范的自适应优化

为解决表 1 中口语化以及符号化等不规范化的语言现象，本文受 Butt 等人[31]启发，对目标领域数据进行字符级规范化处理. Butt 等人证明若对 TweetQA 中的表情、主题标签和“@”等符号进行删除，会对测试结果产生负面效果. 因此，本文在规范化目标领域数据样本时，保留了全部特殊字符. 此外，Huang 等人[8]经过消融实验验证了经过其分离（SPLIT）模块预处理的 TweetQA 样本，可以使模型性能获得显著的提升；van der Goot 等人[32]提出了一种词汇规范化模型 MoNoise，将规范任务分成候选生成和候选排序两个子任务，通过模块化开发，使每个模块负责不同的规范化操作.

本文结合 SPLIT 模块和 MoNoise 模型，将目标领域的文本转化成适应源领域正规化文本的格式. 具体步骤为：(1) 使用 SPLIT 模块将目标领域段落中的符号化混合词分离（如将“*@InStyle*”转化成“*@ In Style*”，将“*#RIPBrangelina*”分成“*# RIP Brangelina*”；(2) 使用 MoNoise 模型对口语化的目标领域段落进行规范化处理（如将简写“*u*”转化成“*you*”，将错写“*convo*”转化成“*conversation*”). 表 2 展示了部分规范化的效果.

表 2　进行规范化前后 TweetQA 文本格式的变化

未规范化： ***@InStyle***: *On KWs Cover: Beautiful statement. Thank* ***u*** *4 opening this* ***convo****. Its an important 1 that needs to be had. kerry washington (@kerrywashington) February 5, 2015*
规范化： ***@In Style***: *On KWs Cover: Beautiful statement. Thank* ***you*** *4 opening this* ***conversation****. Its an important 1 that needs to be had .kerry washington (@kerrywashington) February 5, 2015*
译文： *@InStyle：在 KWs 的采访上：漂亮的声明. 谢谢开启此次对话. 这一点很重要. 克里华盛顿（@kerrywa-shington）2015 年 2 月 5 号*

4.2　基于多任务学习的迁移自训练

多任务迁移自训练的目的是将源领域的知识迁移到无标注的目标领域上. 本文前期工作曾继承 Wang 等人[26]的思想，通过对抗训练寻找不同领域的共性表示，从而避免灾难性遗忘[33]. 相较而言，本文主要面向社交媒体领域提升 MRC 模型的迁移能力，因而与 Chung 等人[10]的目标一致，即尽可能地提取目标领域的个性表示，提升模型在目标领域的泛化性能，忽略模型在源领域上的性能变动.

本文提出的迁移自训练-筛选器-多任务学习框架（Transfer Self-Training with Filter and Multi-Task Learning，简称 TST-F-MTL）如图 1 所示. 其中，x 和 x' 分别为源领域和目标领域样本的问题段落，y 和 y' 分别为源领域和目标领域样本的答案，$\hat{y}'$ 为目标领域生成的伪答案，x'_h 为筛选后的目标领域

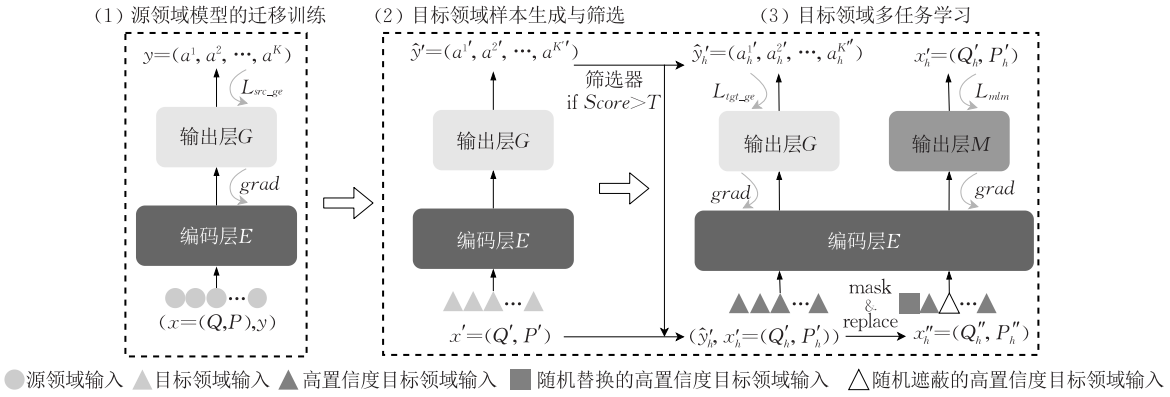


图 1 迁移自训练-筛选器-多任务学习(TST-F-MTL)模型框架

问题段落,$\hat{y}'_h$为筛选后的伪答案,而 x'_h经过随机遮蔽替换后得到具有迷惑性的问题段落 x''_h. K'是伪答案长度,K'_h是经过筛选后的伪答案长度.

本文使用的模型主要包含编码层 E 和两个输出层(生成式阅读理解任务输出层 G 和掩码预测任务输出层 M). 其中,编码层 E 包含分词预处理、词嵌入向量计算和 Transformer 编码三个模块,输出层 G 和输出层 M 则是两个不同的前馈层. 本文的训练方法分为三个阶段,即:

(1) 使用源领域数据(x,y),微调编码层 E 和生成式阅读理解输出层 G,得到源领域模型;

(2) 使用目标域样本 $x'=(Q',P')$,在源领域模型上进行测试,得到伪答案 $\hat{y}'$(其中,P'已经过文本规范化处理). 同时,利用筛选器精选置信度 $Score$ 高于设定阈值 T 的伪答案和对应的问题段落,并以此构成高置信度的目标领域样本$(x'_h,\hat{y}'_h)$;

(3) 结合迁移自训练和多任务机制,即通过$(x'_h,\hat{y}'_h)$微调源领域模型编码层 E 和生成式阅读理解输出层 G. 同时,使用(x''_h,x'_h)微调源领域模型编码层 E,并训练掩码预测输出层 M. 其中,输出层 G 与输出层 M 共享同一个编码层 E.

源领域模型的迁移训练. 首先使用有标注答案的源领域数据集对预训练语言模型进行微调. 本文实验对模型进行迁移,并将问题与段落拼接作为源序列 s_1,答案作为目标序列 s_2,构建端到端答案生成网络.

经过 12 层 Transformer 网络编码后,UniLMv2 将词嵌入向量换为最终的隐向量表示. 本文截取其最后的目标序列 s_2 对应的隐向量 $\boldsymbol{SH}^{12}_{s_2}$(上角标"12"代表第 12 层的 Transformer 网络编码),并将其输入前馈输出层 G:

$$\boldsymbol{SY}_{logit}=\mathrm{softmax}(G(\boldsymbol{SH}^{12}_{s_2})) \tag{1}$$

归一化(softmax)前馈输出层 G 用于得到生成答案的概率向量 $\boldsymbol{SY}_{logit}$,其对应的维度为 $n_{S_2}\times v$,n_{s_2} 表示源领域目标序列 s_2 的长度,v 为词表的大小. 在训练过程中,上述 UniLMv2 的风险损失计算如下(其中,f_{CE}表示交叉熵函数):

$$L_{src\_ge}=f_{CE}(\boldsymbol{SY}_{logit},y) \tag{2}$$

目标领域样本生成与筛选. 本文认为即使不同领域的数据分布不同,即 $D\neq D'$,它们之间也存在一部分共性特征. 具体表现为,在获得源领域模型后,用目标领域样本预测得到的伪答案中,有部分与原始标注的真实答案类似. 但是,如果不排除不相似的答案,由于其与真实答案的数据分布差异较大,势必会影响模型在目标领域的泛化能力. 对此,受 Cao 等人[11]工作的启发,本文使用了一种简单的筛选方法作为图 1 的筛选器. 通过设定阈值,筛选高置信度的样本,以此排除不适应目标领域伪答案,具体步骤如下:

首先,本文向已经在源领域数据上获得完备训练的 UniLMv2 模型,输入目标领域样本 x',测试得到其伪答案 $\hat{y}'$. 同时本文采用贪心解码的方式,并借鉴式(1),可得生成伪答案中的每个词的分数,即为这个词在整张词表中概率分布中的最高值:

$$P(w^k)=\max(\widehat{\boldsymbol{TY}}^k_{\mathrm{logit}}),\ k\in[1,K'] \tag{3}$$

其中,$\widehat{\boldsymbol{TY}}^k_{\mathrm{logit}}$表示伪答案第 k 个生成词在整张词表的概率分布. 最终的置信度分数计算方式如下:

$$Score=\prod_{k=1}^{K'}P(w^k) \tag{4}$$

通过设定阈值 T,保留 $Score>T$ 的样本$(x'_h,\hat{y}'_h)$,以此实现对目标领域样本降噪.

目标领域多任务学习. 为弥补筛选后的样本过少,导致的模型过拟合以及泛化能力不足的缺陷,同时保持迁移自训练能够降低模型困惑度的优势,本文使用了一种结合掩码位置预测任务和生成式阅读

理解任务的多任务学习框架. 该框架通过预测掩码遮蔽词，加强模型对深层语义信息的理解，同时充分提取目标领域的个性表示，借以克服上述缺陷.

生成式阅读理解任务的输入为筛选后的目标领域样本$(x'_h, \hat{y}'_h)$. 本文使用源领域模型的参数对目标领域模型(包括编码层 E 和输出层 G)进行初始化. 其他训练细节与源领域模型训练一致.

掩码预测任务与 BERT[5] 的预训练任务一致，即以一定概率随机遮蔽或替换输入 x'_h 中的词，并通过模型网络，还原被遮蔽或者被替换的词. 除输入不同外，其编码层 E 与源领域模型共享. 值得注意的是，掩码预测阶段仅使用源序列 x''_h 作为模型的输入，其对应的隐向量为 $\boldsymbol{TH}_{t_1}^{12}$. 其中，$n'$ 表示目标领域输入的长度，问题与段落拼接作为源序列 t_1. 归一化(softmax)的前馈输出层 M 用于直接预测(还原)被遮蔽或被替换的词项，其计算方法为

$$\boldsymbol{TX}_{logit} = \mathrm{softmax}(M(\boldsymbol{TH}_{t_1}^{12})) \tag{5}$$

其中，$\boldsymbol{TX}_{logit}$ 的维度为 $n'_{t_1} \times v$，n'_{t_1} 表示目标领域源序列的长度. 掩码预测损失的计算方式为

$$L_{mlm} = f_{CE}(\boldsymbol{TX}_{logit}, x'_h) \tag{6}$$

多任务训练时，本文先输入$(x'_h, \hat{y}'_h)$，得到目标领域生成式阅读理解的损失 L_{tgt_ge}，后输入 x''_h，得到掩码预测损失 L_{mlm}. 最终多任务学习损失的计算方式为(其中，λ 为调和系数，控制掩码预测任务的作用程度，且 $\lambda \in (0,1]$)：

$$L_{MTL} = L_{tgt\_ge} + \lambda L_{mlm} \tag{7}$$

目标领域样本测试. 本文使用贪心解码的方式，对目标领域开发和测试样本进行解码测试. 值得注意的是，在解码阶段并未进行掩码预测任务. 在测试之前，本文同样使用了 4.1 节提出的方法对开发或测试样本的段落进行文本规范化.

5　实　验

本节首先介绍实验环境，包括数据集、超参设置和评价方法，以及列入文本规范化、消融和对比实验的 MRC 模型. 然后，本节围绕消融实验进行分析，并与前人研究进行对比分析. 特别地，本节设置了十一组模型细节的验证和解释，包括(1)性能超越 SOTA 的解释；(2)部分领域自适应方法在生成式阅读理解方向的适应性验证；(3)部分消融实验性能反常下降的原因；(4)非归一化置信度设置的原理；(5)以 NarrativeQA 为源领域数据时，迁移模型的性能增幅极大的因由；(6)掩码预测任务的有效性及其解释；(7)源领域模型使用掩码预测的必要性；(8)相应模型在目标领域的普适性；(9)通用筛选方法的思考；(10)解释性样例及分析；(11)错误定性分析.

5.1　数据集

本文分别以 SQuAD、CoQA 和 NarrativeQA 作为源领域数据集，将 TweetQA 作为目标领域数据集进行实验分析.

(1) TweetQA①. Xiong 等人[15] 收集了新闻工作者撰写新闻文章的推文，并以众包的方式对推文标注问题和答案，构建了第一个面向社交媒体的数据集. 该数据集不同于传统的抽取式问答数据集，其答案并不是文章或语段中的一个连续文字片段，而是真实用户重新编辑的自由文本. 此外，TweetQA 的语段含有大量不规范的语言现象.

(2) SQuAD②. Rajpurkar 等人[2] 首次建设了基于跨度的阅读理解数据集 SQuAD，其文本段落从维基百科收集而来，且答案是文本中的连续片段.

(3) CoQA③. CoQA 是一个多轮对话数据集，其具有人类对话属性，并含有大量指代省略的现象. 特别地，某些对话的答案需要依据上文若干轮对话的内容才能推断，答案类型包含连续片段以及不在给定段落中的自由文本.

(4) NarrativeQA④. NarrativeQA 是一个生成式阅读理解数据集，其段落内容来自书籍以及电影脚本. 其答案均为人工撰写，从而不局限于段落中的片段. 训练机器阅读理解模型时，可用其全部故事文本或只使用其故事摘要. 本文基于摘要任务进行源领域生成式阅读理解任务.

由表 3 中数据集统计结果可知，相比于 TweetQA，源领域数据集(SQuAD、CoQA 和 NarrativeQA)中的训练样本更多，且其段落更长，因而具有的语义信息更丰富.

表 3　各数据集样本数以及平均长度统计

统计项	TweetQA	SQuAD	CoQA	NarrativeQA
训练集	10692	87599	10864	32747
验证集	1086	10570	7983	3461
测试集	1979	—	—	10557
段落平均	24.59	117.24	271	659
问题平均	6.95	10.08	5.5	9.83
答案平均	2.45	3.09	2.7	4.73

① https://tweetqa.github.io/
② https://github.com/rajpurkar/SQuAD-explorer
③ https://stanfordnlp.github.io/coqa/
④ https://github.com/deepmind/narrativeqa

5.2 超参设置和评价

本文初始模型采用 UniLMv2[①]. 由于上述数据集段落长短不一,因而设置了不同的超参. 表 4 为超参设置的不同部分. 除表 4 的参数项,其余超参设置相同. 其中,学习率设置为 2e-5,最大目标序列(max_tgt_len)长度为 24,训练回合数设置为 10,多任务调和系数 λ 设置为 1,学习率预热步数(warm-up_steps)设置为 500. 源序列或目标序列中超过规定的最大长度时,执行截断操作.

表 4 各数据集超参设置中的不同部分

参数项	TweetQA	SQuAD	CoQA	NarrativeQA
max_src_len	128	384	464	464
batch_size	12	12	6	6

对于 CoQA,由于其某轮对话的答案需要上文对话内容进行推理,本文保留了前两轮对话的内容作为对话历史,并与当前对话的问题拼接作为新的问题,以此再与给定段落拼接作为预训练语言模型的输入.

在设置筛选器阈值 T 的取值范围时,文本以 0.1 为步长,并在数值区间{0,0.9}进行取值,即 $T\in\{0,0.1,0.2,\cdots,0.9\}$. 对于掩码预测任务中随机掩码概率的设置,本文采用了 BERT 中掩码语言模型的设置方法.

本文实验采用三种评价方法对 MRC 性能进行检验,包括 BLEU-1[34]、METEOR[35] 和 ROUGE-L[36] 测度.

5.3 实验方法及对比对象

在数据层面,为验证文本规范化方法的有效性,本文使用 unilm1.2-base-uncased[②] 模型的参数初始化 UniLMv2 模型(简写为 UniLMv2-Base),再以不同的源领域数据集对其进行微调,最后使用 TweetQA 未规范化的数据以及经过规范化处理的数据,进行开发测试.

在模型层面,为验证本文结合多任务学习机制的迁移自训练方法是否能够解决领域适应性问题,本文在针对 TweetQA 数据完成本规范化的情况下,复现了部分经典领域自适应模型以及抽取式前沿领域自适应模型,并利用 4 种不同迁移方法建立消融实验,尝试对比了不同源领域数据对消融实验的影响. 以下是本文消融实验所使用的 4 种方法及其简称:

(1) ZERO 表示基线模型,即使用 TweetQA 数据在源领域数据集训练的模型上直接测试.

(2) TST 表示迁移自训练,即使用全部的伪答案以及对应的段落和问题对源领域模型进行微调.

(3) TST-F 表示在迁移自训练的基础上,附加筛选器. 通过设定阈值,筛选置信度分数大于阈值的目标领域样本对源领域模型进行微调.

(4) TST-F-MTL 表示在 TST-F 的基础上,融入多任务机制进行微调.

5.4 实验结果及分析

表 5 显示了附加规范化处理前后的 MRC 性能. 实验结果显示,经规范化处理后(标记为"√"),基线系统 ZERO 在 TweetQA 开发集上取得的稳定性能(性能波动极小时的状态),相比于未经使用规范化处理时的稳定性能(标记为"×"),在 BLEU-1、METEOR 和 ROUGE-L 测度上均有提升. 该实验结果验证了本文规范化方法的有效性,其一定程度上能将目标领域文本转化成适应源领域的规范格式.

表 5 未规范化与规范化的开发集性能对比(单位:%)

数据集	规范	BLEU-1	METEOR	ROUGE-L
SQuAD	×	72.76	69.26	74.75
	√	**73.82**	**70.03**	**75.78**
CoQA	×	70.89	67.64	72.59
	√	**73.06**	**69.66**	**74.89**
NarrativeQA	×	53.62	51.16	59.09
	√	**57.41**	**54.94**	**62.81**

表 6 展示了 TweetQA 数据集在 SOTA 模型(Huang 等人[8])上的 MRC 性能、基线模型 ZERO*(TweetQA 包含答案标注的训练集训练所得的模型)的 MRC 性能、经典无监督领域自适应方法(Gretton 等人[30])与抽取式机器阅读理解的前沿无监督领域适应方法(Wang 等人[26]、Yue 等人[29])应用到生成式 MRC 的性能以及本文方法(TST-F-MTL)及其简化版 TST 和 TST-F,从三种不同源领域数据集(SQuAD、CoQA 和 NarrativeQA)向目标领域数据(TweetQA)迁移后的 MRC 性能(注意,表 6 针对每种测度列举了两项性能指标,包括开发性能和测试性能,标记"|"左侧的数值为开发性能,右侧的数值为测试性能). 为增加性能对比的指向性,表 6 未显示阈值设定的细节,因而将其补充说明:使用 SQuAD、CoQA 和 NarrativeQA 作为源领域数据集时,TST-F 和 TST-F-MTL 设定的阈值 T 分别为 0.1、0.6、0.5.

① https://github.com/microsoft/unilm/tree/master/s2s-ft
② https://unilm.blob.core.windows.net/ckpt/unilm1.2-base-uncased.bin

表 6　TweetQA 在基于 UniLMv2_Base 的初始模型上的消融实验以及性能对比　（单位：%）

源领域数据集	方法	目标领域 MRC 性能（开发性能|测试性能）		
		BLEU-1	METEOR	ROUGE-L
TweetQA	Huang 等人[8]	78.2|76.1	73.3|72.1	79.6|77.9
	ZERO*	77.1|77.7	73.1|73.9	79.1|79.4
SQuAD	Gretton 等人[30]	73.1|73.1	68.9|69.8	74.9|75.6
	Wang 等人[26]	73.8|72.8	69.0|69.6	74.9|74.4
	Yue 等人[29]	72.2|73.5	68.5|69.7	74.3|75.5
	ZERO	73.8|74.6	70.0|70.8	75.7|76.5
	TST	75.1|75.3	71.5|71.6	77.2|77.1
	TST-F	75.5|76.0	71.5|72.2	77.5|77.7
	TST-F-MTL	**76.3**|**76.5**	**72.3**|**72.7**	**78.2**|**78.1**
CoQA	ZERO	73.0|74.1	69.6|70.3	72.5|75.7
	TST	75.2|75.2	71.6|71.4	76.9|76.8
	TST-F	75.0|74.8	71.5|71.3	76.4|76.2
	TST-F-MTL	**76.9**|**76.8**	**73.0**|**73.3**	**78.1**|**77.9**
NarrativeQA	ZERO	57.4|56.6	54.9|53.8	62.8|61.8
	TST	60.1|59.7	57.9|56.9	65.6|65.0
	TST-F	74.5|**74.5**	70.6|**70.4**	75.9|**75.8**
	TST-F-MTL	**75.5**|73.8	**71.2**|70.2	**76.7**|75.6

实验结果显示，本文方法及其简化版在上述三个迁移过程中，均取得优于基线模型 ZERO 的性能. 其中，TST-F-MTL 使用 NarrativeQA 作为源领域，并使用 TweetQA 作为目标领域进行迁移时，测试性能在 BLEU-1、METEOR 和 ROUGE-L 三种测度上分别优于基线系统 ZERO 30.3%、30.5%和 22.3%；使用 CoQA 为源领域时，分别提高 3.6%、4.2%和 2.9%；而使用 SQuAD 时，性能分别提升 2.5%、2.7%和 2.0%. 该实验结果证明了本文基于多任务学习的无监督自训练方法具有更好的迁移能力. 同时，TweetQA 测试集在以 SQuAD 和 CoQA 为源领域的 TST-F-MTL 模型上的性能已经超过 Huang 等人[8]提出的 SOTA，且已经登上榜单第一（见本文 P7 脚注①）. 此外，本文发现部分领域自适应方法无法直接适用于生成式阅读理解，复现模型的性能相较于 ZERO 均有降低. 表 6 的实验结果存在诸多细节值得探究，本文将在以下篇幅逐一进行介绍.

(1) SOTA 的特点及性能超过 SOTA 的原因

表 6 顶端的 SOTA 模型是 Huang 等人[8]在 2020 年开发完成的 NUT_RC 模型，该模型首先使用了四种规范化方法解决文本不规范的问题，然后通过判别器对候选答案进行打分（其中，候选答案为开发测试样本分别在抽取式和生成式模型上的预测结果），最后选取抽取和生成两类答案中分数较高者进行性能评估，从而结合抽取和生成的优势.

本文 TST-F-MTL 模型相较于有监督条件下的 SOTA 模型有以下两种优势：① 本文模型参数量较 SOTA 模型小. SOTA 包含生成式模型、抽取式模型以及二分类判别器，其分别使用 UniLMv1-BASE (110 M)、BERT-LARGE-WWM (336 M) 和 BERT-BASE(110 M)的参数进行初始化，较之本文模型仅使用的 UniLMv2-BASE(110 M)，参数量高出 405%；② 本文模型的训练较 SOTA 稳定. 本文在复现 SOTA 的实验中，发现二分类判别器的训练很不稳定. 在设置不同的随机种子训练判别器后，联合模型的性能便不能超过单个抽取或生成模型. 因此，SOTA 开发测试结果的优劣严重依赖于随机种子，这无疑加大了模型的训练难度，同时不稳定的训练也导致预测结果变得不稳定（SOTA 的开发和测试性能有最高两个百分点的落差）. 而本文模型仅通过一次训练即可达到表 6 的性能.

(2) 部分领域自适应方法在生成式阅读理解方向的适应性验证

本文又基于 UniLMv2-Base 模型，并以 SQuAD 为源领域数据集，TweetQA 为目标领域数据集，复现了 Gretton 等人[30]、Wang 等人[26]以及 Yue 等人[29]的无监督领域自适应模型. 因上述工作的研究任务与本文研究任务有所不同，本文在上述复现的模型的基础上做出一定调整. 下面分别对原模型以及其调整方法进行介绍.

对于 Gretton 等人[30]提出的最大均值差异[30] (MMD)的领域自适应方法，本文将其应用到源领域和目标领域样本经过 UniLMv2 所得的源序列表示上（其中，目标领域的训练样本为无答案标注的目标领域训练集在 ZERO 模型上的全部测试结果.

获得全部训练样本后，本文基于 UniLMv2 重新训练 MMD 模型). MMD 损失计算方式如下：

$$L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_s)=\frac{1}{|N|^2}\sum_{i=1}^{|N|}\sum_{j=1}^{|N|}f(\bar{\boldsymbol{h}}_s^i,\bar{\boldsymbol{h}}_s^j) \tag{8}$$

其中，N 表示源领域和目标领域样本个数之和，f 是高斯核函数[37]，$\bar{\boldsymbol{h}}_s$ 为截取的源序列隐向量经过平均池化得到的表示（源领域和目标领域不做区分，统一以 $\bar{\boldsymbol{h}}_s$ 标记）. 复现模型的生成式阅读理解的损失计算方式与 4.2 节源领域模型的迁移训练中介绍的一致. 最后模型的训练损失计算公式为

$$L_{\text{MMD\_MTL}}=L_{mix\_ge}^{mmd}+L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_s) \tag{9}$$

其中，$L_{mix\_ge}^{mmd}$ 表示混合样本在 MMD 模型上生成式阅读理解模块的损失，$L_{\text{MMD\_MTL}}$ 表示 MMD 模型最终的训练损失.

Wang 等人[26]于 2019 年提出了基于抽取式阅

读理解的无监督领域适应模型 AdaMRC(Adversal domain adaptation Machine Reading Comprehension).由于其解码方式不同于生成式,因而无法直接应用于生成式方向,所以本文保留了其梯度反转层[38](Gradient Reversal Layer,GRL),并将源领域和目标领域样本在 UniLMv2 上的源序列表示作为梯度反转层的输入.其中,目标领域训练样本不包含答案.

与原论文一致,本文只计算源领域训练样本在生成式阅读理解模块上所产生的生成损失,而目标领域的训练数据只参与梯度反转的损失计算.具体而言,本文使用两层感知机作为领域判别器,其输入为两种领域样本的源序列表示.因而,分类器训练损失的最终计算方式为

$$L_{Class}=\frac{1}{N}\sum_{i=1}^{N}\log P(d_i^m|Q_i^m,P_i^m) \tag{10}$$

其中,i 表示样本的序号,Q_i^m 和 P_i^m 分别表示混合样本的问题和段落,N 表示源领域和目标领域的样本总和,d_i^m 表示单个样本的领域标签(源领域样本为 0,目标领域为 1).复现的 AdaMRC 最后损失的计算方式为

$$L_{AdaMRC}=L_{src\_ge}^{adamrc}-\alpha L_{Class} \tag{11}$$

其中,$L_{src\_ge}^{adamrc}$ 表示源领域样本在 AdaMRC 上的生成损失,L_{AdaMRC} 表示复现的 AdaMRC 模型的最终损失,α 作为平衡因子,用于平衡两种任务的作用程度,且 α 的取值范围为(0,1].值得一提的是,复现的模型并未与 GAN[39]一致,交替更新生成器和鉴别器的参数,而是与原论文一致,采用多任务机制,通过 GRL 对模型进行优化.此外,因为本文使用的目标领域训练集包含标注问题的数据分布,所以删除了原模型的问题生成模块.

Yue 等人[29]通过生成模型生成目标领域伪答案以及伪问题,从而构造目标领域伪样本;同时结合 MMD 和对比学习,开发构建了基于抽取式阅读理解的无监督对比领域自适应模型 CAQA(Contrastive Domain Adaptation for Question Answering).由于本文工作包含问题的数据分布,因而删除了其问题生成模块.复现模型的训练样本由目标领域无答案标注的开发测试集在 ZERO 模型上的全部测试结果混合源领域全部训练集而来.

同时,为了让源领域的数据表示分布尽可能接近目标领域,与原论文一致,本文将 MMD 作用于两种不同领域的训练样本经过 UniLMv2 所得的源序列表示上.对于对比学习模块,原工作基于抽取式,其为了使模型更置信的区分段落中潜在的答案区间,依据数据集标注的开始和结束的位置下标,提取段落中的答案区间表示,同时结合 MMD,使答案的表示分布与对应的段落和问题的表示分布尽可能不相似.但是较于抽取式,生成的答案的取值范围为整个词表,不只局限于段落,抽取答案跨度表示的方法并不适用,所以本文使用经过平均池化的目标序列表示作为潜在的答案表示.

本文基于上述复现的 Gretton 等人[30]的工作,在最后损失上减去源序列和目标序列的对比损失,最终对比自适应损失的计算方法如下:

$$L_C=L_{mix\_ge}^{caqa}+L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_s)-L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_t) \tag{12}$$

其中,$L_{mix\_ge}^{caqa}$ 表示两种领域的样本在 CAQA 生成式阅读理解模块的生成损失,$L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_s)$ 的作用在于将两种领域样本的源序列表示分布尽可能的接近,而减去 $L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_t)$ 则是为了使样本的答案(目标序列)和问题段落(源序列)的表示分布尽可能的不一致.$L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_t)$ 具体的计算公式为

$$L_{mmd}(\bar{\boldsymbol{h}}_s,\bar{\boldsymbol{h}}_t)=\frac{1}{|N|^2}\sum_{i=1}^{|N|}\sum_{j=1}^{|N|}f(\bar{\boldsymbol{h}}_s^i,\bar{\boldsymbol{h}}_t^j) \tag{13}$$

其中,$\bar{\boldsymbol{h}}_t$ 是目标序列的平均池化表示.

Cao 等人[11]研究的是训练集样本中包含问题分布的领域自适应问题,与本文工作的研究任务近似.但由于其基于抽取式阅读理解,因而本文基于其论文给定的代码①,尝试改写为可应用到生成任务的领域自适应模型.

本文修改了其 CDAN 模块的输入.因抽取式阅读理解与生成式的解码方式不同,二者解码所得的隐向量的维度不同,所以区别于原论文以开始和结束向量的拼接表示与模型编码表示的点积向量作为 CDAN 模块的输入,本文分别对训练样本经过 UniLMv2 模型编码所得的源序列和目标序列的表示进行平均池化,再以二者的点积向量作为 CDAN 输入.但实验发现 CDAN 并不能在生成式方向起领域适应作用,随着训练进行,其动态生成的噪声样本进一步增加,最后仅在完全错误的伪样本上完成收敛.由于其在评估指标的数值均为 0,所以表 6 中未将其展示.

值得一提的是,上述复现的经典的以及抽取式阅读理解前沿领域自适应方法并不是完全适用于生成式.TweetQA 开发集和测试集在上述复现模型上的性能较本文提出的 ZERO 模型均有所降低.由此可知,部分领域自适应方法无法直接应用到生成式

① https://github.com/caoyu-noob/CASe

阅读理解上.

本文猜想上述方法在生成方向无法适用的原因有以下两点：① 最大均值差异或梯度反转的方法通常作用在 encoder 表示上，而生成式阅读理解是 encoder-decoder 结构，decoder 会一定程度上破坏已经提取的共性特征；② 因生成式解码的隐向量的维度与抽取式不同，本文在复现 Cao 等人[11]和 Yue 等人[29]的工作时，将目标序列的表示进行了平均池化. 但是，与经过 encoder 编码所得的上下文感知向量不同，生成式解码所得隐向量中各个词向量之间相互独立，且其解码的语义范围较抽取式更广（生成式从整张词表选词，而抽取式仅局限于段落），所以在上述模型中，对目标序列表示进行平均池化并不能达到理想预期. 未来工作将使用具有不同解码能力的 decoder 继续进行实验.

(3) 消融实验性能“反常”的原因

上述消融实验结果存在如下两项细节问题. 其一，TST-F 在源自 CoQA 的迁移过程中，相较于简化版本 TST，在性能上有所降低. 相对地，其在起始于 SQuAD 和 NarrativeQA 的迁移过程中，性能均优于 TST. 其二，TST-F-MTL 在源自 NarrativeQA 的迁移过程中，相较于其简化版本 TST-F，测试性能略微降低. 但在源自 SQuAD 和 NarrativeQA 的迁移过程中，其取得了优于 TST-F 的测试性能.

针对产生上述两种异常情况的因由，本文给出如下假设. 其一，使用多任务方法的前提是目标领域尽量没有噪声样本，相反过多的噪声对基于语义理解的掩码预测，将产生较大的负面影响，从而进一步削弱多任务框架下的 MRC 性能. 在 CoQA 的案例中，由于训练样本数量减少，TST-F 的模型泛化能力被减弱了，但是却基本保留了正确的伪答案. 所以，虽然 TST-F 的性能有所退化，但其导致的正确伪答案总量高于噪声数量，从而对基于多任务学习的 TST-F-MTL 提供了更多正面的影响，使其取得了性能优化. 其二，在 NarrativeQA（源领域）的案例中，由于 NarrativeQA 与 TweetQA（目标领域）的数据分布差异较大，当 TST-F 采用高阈值进行样本筛选时，得到保留的噪声样本总量较大，从而负面影响高于正面影响，对多任务框架下的 MRC 性能伤害较强. 且 NarrativeQA 的测试集的数据分布较开发集有一定的差异，所以尽管开发性能呈上升趋势，TST-F-MTL 的测试性能却较 TST-F 略微降低.

通过观察 CoQA 的 TST-F 和 TST-F-MTL 的性能变化趋势，可以进一步验证上述假设. 如图 2 所示，TST-F（灰色曲线）性能波动总体平缓，$T=0.4$ 后，其性能开始下降，但是比 $T=0$ 和表 6 中设定的 $T=0.5$ 的性能高，其证实 TST-F 筛选器的有效性. 相对地，TST-F-MTL（黑色曲线）在此阶段却从下降状态转为上升，且在 $T\geq 0.6$ 后，由于样本数量减少，两种方法性能虽都有下降，但使用多任务机制的性能较高，因此通过设定合适阈值排除噪声样本，会进一步发挥多任务学习的优势. 特别地，TST-F-MTL 在 $0.2\leq T\leq 0.4$ 之间性能下降，而 TST-F 却上升，其一定程度可验证噪声样本的负面效果会通过多任务学习进一步放大；尤其，TST-F-MTL 在 $0\leq T\leq 0.2$ 之间较 TST-F 的性能高，因为使用多任务学习机制后，正、负影响都会扩大，而本文为尽量保证高阈值下样本的正确性，使用的筛选方法较为严格. 因此，在低阈值下虽然噪声较多，但保留的正确样本居多，正面影响也相较明显. 未来工作将会研究不同比例的正负样例，在多任务学习机制下，对模型泛化能力的影响.

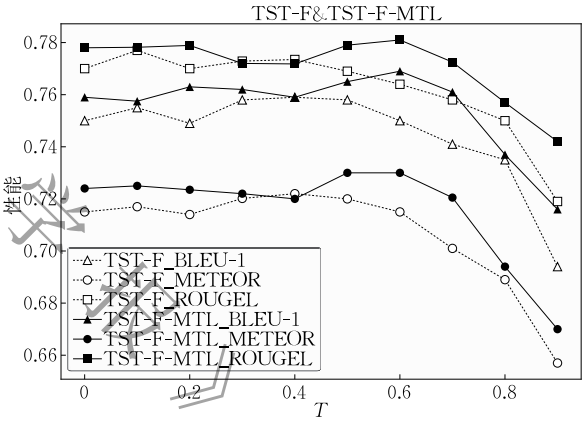


图 2　以 CoQA 为源领域数据集，不同阈值下，TST-F 和 TST-F-MTL 在 TweetQA 开发集的性能对比

(4) 置信度非归一化处理的原理

本文在前期研究中尝试采用归一化的置信度计算模式构建迁移自学习模型 TST-F 和 TST-F-MTL（如式(14)所示），但在本文汇报的工作中，却替代地采用了非归一化的计算模型（如式(4)）. 其原因是后者更为严格，对筛选目标领域伪样本的过滤力度更为明显.

$$Score=\sqrt[K']{\prod_{k=1}^{K'}P(w^k)} \tag{14}$$

图 3 显示了置信度归一化与否对样本数量（即筛选后目标领域伪样本的数量）的影响，该统计数据的获取，建立在自源领域 SQuAD 至目标领域 TweetQA 的模型迁移实验上. 如图 3 所示，归一化后，样

本数量变化稳定(白色柱体标记),其超过 1/3 的样本集中于 $T\geq 0.9$. 而未作归一化时,样本数量(灰色柱体标记)随阈值变化明显,且 $T\geq 0.9$ 的个数不及归一化的 50%.

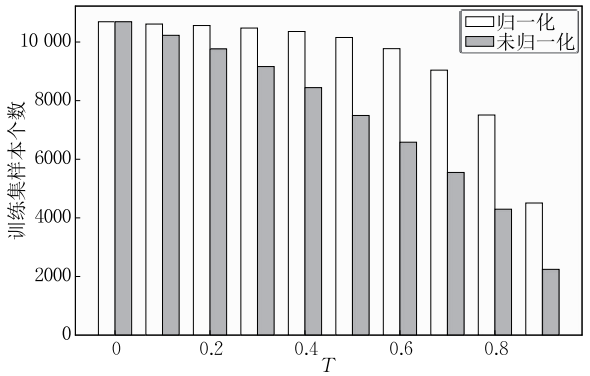


图 3 以 SQuAD 为源领域数据集,不同阈值下,归一化与未归一化筛选的 TweetQA 训练集样本个数

本文比较了以 SQuAD 为源领域数据集,目标领域 TweetQA 数据集分别在使用未归一化和归一化后在 TST-F-MTL 的性能表现. 评价指标为 BLEU-1、METEOR 和 ROUGE-L. 如表 7,未使用归一化(标记为×)筛选方法时,TweetQA 开发集在不同源领域模型上的性能,较使用归一化(标记为√)均有提升,尤其在以 NarrativeQA 作为源领域数据集上最为明显. 其中,未归一化设定的阈值与表 6 一致,归一化设定的阈值在使用 SQuAD、CoQA 和 NarrativeQA 作为源领域数据的条件下,分别为 0.9、0.5 和 0.9.

表 7 使用不同源领域数据,TweetQA 使用归一化与未归一化筛选方法后在 TST-F-MTL 上的开发性能

数据集	是否归一	BLEU-1	METEOR	ROUGE-L
SQuAD	×	**76.3**	**72.3**	**78.2**
	√	74.9	70.8	76.7
CoQA	×	**76.9**	**73.0**	**78.1**
	√	76.5	72.8	78.0
NarrativeQA	×	**75.5**	**71.2**	**76.7**
	√	70.0	66.1	71.6

(5)性能在 NarrativeQA 上增幅大的因由

使用基于 NarrativeQA 的源领域基线模型 ZERO 时,其在目标领域数据上的 MRC 性能较低. 原因是 NarrativeQA 中答案的平均长度相较于其他数据集最长,且几乎是 TweetQA 的两倍,因此相比于其它两种源领域数据集,由其训练得到的源领域模型更倾向于生成长答案. 从而,TST 的迁移过程将面临更多的噪声样本,使得迁移后的 MRC 模型性能提升不明显. 通过筛选器,可以排除大部分长答案,从而保证样本正确性. 使用目标领域数据集对源领域模型进行微调时,源领域模型会提取目标领域的个性表示,从而生成短答案,所以在使用 TST-F 后其性能大幅度提升.

为了验证上述论文,本文检查了 TST 生成的目标领域伪样本(即采用生成模式输出的目标领域的答案)的可靠性. 具体地,本文采用源领域数据集训练基于 TST 的 MRC 模型,然后将目标领域的训练集视作测试集,并检验 MRC 模型生成伪答案的质量,同样利用 BLEU-1、METEOR 和 ROUGE-L 进行质量评估,并设定三个测度取值为 1 时,代表伪样本正确,即伪答案与实际答案(Groud Truth)一致. 在此基础上,本文统计了筛选器在不同阈值时正确伪样本相对于全部伪样本的占比. 如表 8 所示,TST 可根据不同阈值产生质量不同的伪样本集合,当阈值为 0.5 的时候,正确伪样本在总体伪样本中的占比约为 74.2%.

表 8 使用 TST 和 TST-F 方法筛选的 TweetQA 训练集正确样本数、总样本数以及正确率统计

方法	正确样本数	总样本数	正确率/%
TST	3703	10069	34.6
TST-F	2296	3093	74.2

为继续验证提升幅度显著是否合理,本文使用筛选的 3093 个样本直接对 UniLMv2 进行微调,并用 TweetQA 开发集进行测试,得到的 BLEU-1、METEOR 和 ROUGE-L 性能指标分别为 69.6、66.2 和 71.4,相较于表 6 的 ZERO 方法提升约 10 个百分点,其验证了 TST-F 提升幅度巨大的合理性.

(6)掩码预测任务的有效性及其分析

由表 6 可知,尽管 TweetQA 在本文提出的结合生成式阅读理解和掩码预测任务的多任务模型(TST-F-MTL)上的开发测试性能较未使用掩码预测任务的模型(TST-F)在大部分情况下有提升,但也存在反常现象(TST-F-MTL 在源自 NarrativeQA 的迁移过程中,相较于 TST-F,测试性能略微降低),再者掩码预测任务的训练方式存在一定开销,所以其存在的必要性值得进一步验证. 为此,本文基于 UniLMv2 的 Large 模型继续进行实验.

如表 9,本文继使用 unilm2-large-uncased① 模型的参数初始化 UniLMv2 模型(简写为 UniLM-v2-Large),利用 5.3 节所提的 4 种迁移方法建立实验. 其中,使用 SQuAD、CoQA 以及 NarrativeQA 作为

① https://unilm.blob.core.windows.net/ckpt/unilm2-large-uncased.bin

源领域数据集时，TST-F 和 TST-F-MTL 设定的阈值分别为 0.5、0.6 和 0.5.结果所示，除了 NarrativeQA 到 TweetQA 的迁移过程中，开发集在 TST-F-MTL 上的性能较 TST-F 略低，其他部分均超过基线，本文方法在 large 模型上依然可以提升性能，其验证了掩码预测的有效性.

表 9　TweetQA 在基于 UniLMv2_Large 初始模型上的消融实验及性能对比　（单位：%）

源领域数据集	方法	目标领域 MRC 性能（开发性能|测试性能）		
		BLEU-1	METEOR	ROUGE-L
SQuAD	ZERO	75.1|75.2	71.0|71.5	77.2|77.3
	TST	76.4|76.5	72.4|72.6	78.3|78.4
	TST-F	76.7|77.1	72.4|73.2	78.5|78.6
	TST-F-MTL	**77.4**|**78.0**	**73.5**|**74.1**	**79.1**|**79.6**
CoQA	ZERO	74.6|74.3	70.7|70.2	76.5|76.1
	TST	76.9|76.8	73.0|72.8	78.4|78.5
	TST-F	76.9|76.4	73.1|72.6	78.2|77.8
	TST-F-MTL	**78.1**|**78.0**	**74.1**|**74.4**	**79.3**|**79.7**
NarrativeQA	ZERO	63.9|63.5	60.6|59.8	67.2|66.7
	TST	70.7|69.8	67.1|66.5	73.8|72.8
	TST-F	**76.7**|76.0	**72.7**|72.1	**78.3**|77.4
	TST-F-MTL	76.3|**76.7**	71.8|**72.4**	77.8|**78.2**
TweetQA*	ZERO	79.3|78.5	75.1|74.6	80.7|80.4

掩码预测作为一种有效的数据增强手段，在数据集样本极度稀缺的情况下，往往有着较好的性能提高. TweetQA 的训练样本数较少，所以数据增强后模型的性能有所提升（提升的显著程度较为明显且易于观测）.此外，本文采用的多任务机制其本身可以避免模型在单一任务上出现过拟合，从而模型能够形成更为鲁棒的预测能力，换言之，其在目标领域的泛化性得到加强.因此，本文提出的 TST-F-MTL 模型有着良好的领域迁移能力.

（7）源领域模型使用掩码预测的必要性

上述结果已经证实了结合掩码预测以及生成式阅读理解的多任务机制对目标领域模型训练的有效性，那是否在初始源领域模型训练过程中，也使用此多任务机制，会进一步提升性能？基于此，本文依然基于 UniLMv2-Base 模型，结合多任务机制，使用不同的源领域数据对其进行微调，同时基于 TweetQA 的规范化文本，在 ZERO 和 TST-F-MTL 的迁移方法上进行实验.

如表 10 所示，×表示在源领域模型训练过程中未进行掩码预测，√表示使用生成式阅读理解以及掩码预测的多任务机制训练源领域模型.其中，源领域模型训练未进行掩码预测时，TST-F-MTL 所设定的阈值与表 6 一致，而源领域模型训练过程中使用掩码预测时，以 SQuAD、CoQA 和 NarrativeQA 作为源领域数据，TST-F-MTL 设定的阈值分别为 0.7、0.3 和 0.7.结果表明，初期训练源领域模型时，使用多任务机制并不能达到预期中的效果，TweetQA 开发集在 ZERO 下的性能较不使用多任务的普遍降低，因而在生成目标领域伪样本时，会产生更多噪声样本，提高了过滤难度，从而不能有效筛选符合目标领域分布的伪样本，降低目标领域模型的泛化能力.

表 10　TweetQA 开发集在源领域模型使用以及未使用掩码预测任务情境下的性能对比

源领域数据集	源领域数据集是否掩码预测	方法	BLEU-1	METEOR	ROUGE-L
SQuAD	×	ZERO	73.8	70.0	75.7
	×	TST-F-MTL	76.3	72.3	78.2
	√	ZERO	73.6	69.8	75.9
	√	TST-F-MTL	75.4	71.5	76.9
CoQA	×	ZERO	73.0	69.6	72.5
	×	TST-F-MTL	**76.9**	**73.0**	**78.1**
	√	ZERO	72.1	68.7	73.8
	√	TST-F-MTL	75.0	71.3	76.5
NarrativeQA	×	ZERO	57.4	54.9	62.8
	×	TST-F-MTL	**75.5**	**71.2**	**76.7**
	√	ZERO	51.9	49.5	58.0
	√	TST-F-MTL	73.3	69.0	74.8

（8）目标领域的普适性

本文实验尝试分析如下问题，即相关方法在其它目标领域数据集上是否同样适用？为此，本文实验开展了通用性测试.

首先，本文使用 UniLMv2-Base 作为初始模型进行领域迁移实验.除上文所提到的 SQuAD 和 NarrativeQA 两种数据集，本文又添加了 MRQA-2019[40] 预处理过的 TriviaQA[41] 和 HotpotQA[42] 作为目标领域数据集.其中，SQuAD、TriviaQA 和 HotpotQA 的评价指标为 *F*1 和 *EM*，NarrativeQA 采用 BLEU-1、METEOR 和 ROUGE-L 进行评测.其次，实验依然使用未归一化的筛选器进行伪答案的筛选，即式(4)，并将其作用于目标领域 SQuAD、TriviaQA 和 HotpotQA 的数据上.对于 NarrativeQA，本文则使用归一化的筛选器（式(14)），因为 NarrativeQA 数据集的答案信息较为丰富，若筛选规则严格，则会排除此类长答案，导致筛选效果减弱.

表 11 和表 12 显示了通用性测试结果.通过比较不同数据集间的迁移性能，可以验证本文所提方法的通用性.其中的筛选器可以根据答案生成风格而替换成归一化或者未归一化.表 13 是表 11 和表 12 中各个目标领域数据集使用 TST-F-MTL 方法所设定的阈值.

表 11 目标领域数据集 SQuAD、TriviaQA 和 HotpotQA 的领域迁移性能对比 （单位：%）

迁移数据集	方法	评估指标	
		*F*1	*EM*
CoQA→SQuAD	ZERO	85.1	74.9
	TST-F-MTL	**87.7**	**79.1**
NarrativeQA→SQuAD	ZERO	73.0	54.2
	TST-F-MTL	**82.9**	**71.8**
SQuAD→SQuAD	ZERO	90.0	81.9
SQuAD→TriviaQA	ZERO	58.0	47.6
	TST-F-MTL	**63.0**	**53.5**
CoQA→TriviaQA	ZERO	57.5	47.9
	TST-F-MTL	**61.4**	**52.6**
NarrativeQA→TriviaQA	ZERO	55.5	45.6
	TST-F-MTL	**63.9**	**55.1**
TriviaQA→TriviaQA	ZERO	68.6	62.5
SQuAD→HotpotQA	ZERO	64.2	49.0
	TST-F-MTL	**67.8**	**52.7**
CoQA→HotpotQA	ZERO	59.9	45.3
	TST-F-MTL	**63.0**	**48.3**
NarrativeQA→HotpotQA	ZERO	57.9	41.2
	TST-F-MTL	**63.9**	**48.9**
HotpotQA→HotpotQA	ZERO	76.0	59.7

表 12 目标领域数据集 NarrativeQA 的领域迁移性能对比 （单位：%）

—	—	BLEU-1	METEOR	ROUGE-L
CoQA→NarrativeQA	ZERO	54.2	47.3	55.2
	TST-F-MTL	**55.6**	**48.5**	**56.5**
SQuAD→NarrativeQA	ZERO	51.6	45.6	53.0
	TST-F-MTL	**52.6**	**46.3**	**53.8**
NarrativeQA→NarrativeQA	ZERO	54.7	45.6	54.7

表 13 TST-F-MTL 设定的阈值 *T*

迁移数据集	阈值 *T*
CoQA→SQuAD	0.4
NarrativeQA→SQuAD	0.3
SQuAD→TriviaQA	0.6
CoQA→TriviaQA	0.5
NarrativeQA→TriviaQA	0.5
SQuAD→HotpotQA	0.6
CoQA→HotpotQA	0.7
NarrativeQA→HotpotQA	0.5
CoQA→NarrativeQA	0.7
SQuAD→NarrativeQA	0.8

(9)通用筛选方法的思考

值得额外思考的一个问题是，本文筛选器需要根据不同数据集设定不同的阈值，有没有更加通用的筛选方法？受 Wang 等人[43]启发，本文尝试过在模型训练过程中，使用每个目标领域样本的置信度乘以对应的损失，间接通过模型自动完成筛选. 但是在训练过程的前几轮，损失就接近于 0，使得模型丧失了学习能力，最终导致模型遗忘了所有领域知识. 从而，设置一种动态调整阈值将是未来研究的一项重要任务.

(10)迁移自训练性能提升大的因由

本文发现基于不同源领域数据集，TweetQA 在 TST-F 上的性能较 ZERO 都有较大幅度的提升，尤其以 NarrativeQA 作为源领域数据集时，提升最为明显. 但是源领域模型生成的伪样本对于其本身来说并没有多余的梯度进行反向传播，为什么还能在此基础上继续提升性能，并取得优良效果？

基于此，本文猜想由于模型事先未曾形成目标领域样本的语言表示，并没有建立目标领域问题段落与其对应答案的潜在联系，所以尽管部分答案预测正确，但答案的困惑度较高，模型绝大多数情况下处于不稳定状态. 为此，基于表 6 的实验，本文统计了 NarrativeQA 到 TweetQA 的迁移实验中，TweetQA 开发集在 ZERO 和 TST-F 两种模型下都预测正确（使用 BLEU-1、METEOR 和 ROUGE-L 进行质量评估，三个测度取值都为 1 时，代表样本正确）的 418 条样本的困惑度. 依据困惑度高低比较两种模型的领域适应能力. 其中，困惑度越低代表模型越能够形成更精确的语言表示，其领域适应能力越强. 困惑度计算方法如下：

$$Perplexity=\frac{1}{\sqrt[K']{\prod_{k=1}^{K'}P(w^k)}} \tag{15}$$

图 4 表示 418 条公共正确样本在两种方法下的困惑度对比. 其中，黑色线条表示正确样本在 ZERO 上的困惑度，而灰色线条表示在 TST-F 上的困惑度（阈值 T 设定为 0.5）. 横坐标表示样本序号（取值范围为 0～417），纵坐标表示困惑度. 由图 4 可见，大多数使用通过 ZERO 测试的样本的困惑度高于 TST-F. 在此基础上，本文统计了 ZERO 困惑度比 TST-F 高的样本个数，一共 349 个，在 418 条正确样本中占比 83.4%. 由此可见，同样是预测正确的

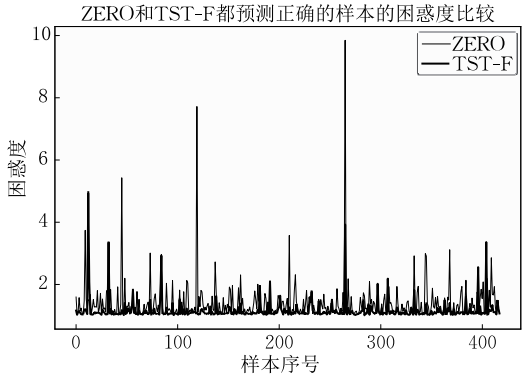


图 4 TweetQA 开发集在 ZERO 和 TST-F 上预测都正确的 418 条样本的困惑度对比

样本，其在 ZERO 上并不能形成鲁棒表示，从而导致困惑度较高，而通过迁移自训练的模型 可以一定程度上避免此种现象.

(11) 解释性样例及分析

为分析领域知识对模型迁移的贡献程度，本文提供了部分样例及其解释性文字. 其中，样例的来源是表 6 的 ZERO 模型和 TST-F-MTL 模型的开发结果，两者皆以 TweetQA 为目标领域，且以 SQuAD 为源领域. 值得说明的是，ZERO 模型的训练过程并未纳入跨领域迁移，相反，TSF-F-MTL 的训练过程纳入了本文所提的跨领域迁移学习.

样例 1

段落：*I won't be able to sleep until I know how this person figured out they were on the wrong flight. That's all I ask. 150 people have been majorly inconvenienced, please, just tell me — christine teigen (@chrissyteigen) December 27, 2017*
译文：*除非我知道这个人是怎么发现他们搭错了航班，否则我无法入睡. 这就是我所要求的. 150 人受到了极大的不便，请告诉我 — christine teigen (@chrissy-teigen) 2017 年 12 月 27 日*

问题：*who won't able to sleep?*
译文：*谁会睡不着?*

标准答案：*christine teigen*
译文：*christine teigen*

ZERO 预测答案：*i will not be able to sleep until i know how this person figured out they were on the wrong flight*
译文：*除非我知道这个人是如何发现他们搭错了航班，否则我无法入睡*

TST-F-MTL 预测答案：*christine teigen*
译文：*christine teigen*

样例 2

段落：*Forget nail polish, Cookie's FUR matches #theshoe the best! #Empire — Empire (@EmpireFOX) July 10, 2015*
译文：*忘了指甲油吧，Cookie's 毛皮衣服和这只鞋子是最好的搭配! #Empire — Empire (@emirefox) 2015 年 7 月 10 日*

问题：*what does cookie's fur match the best?*
译文：*Cookie's 毛皮衣服 和什么搭配最好*

标准答案：*theshoe*
译文：*这只鞋子*

ZERO 预测答案：*nail polish*
译文：*指甲油*

TST-F-MTL 预测答案：*theshoe*
译文：*这只鞋子*

样例 1 和样例 2 均体现了一种“命名实体背景知识缺失”导致的 MRC 误判问题. 样例 1 中 ZERO 由于缺乏“*christine teigen*”的人物实体知识，从而导致误判；样例 2 中 ZERO 缺少“*theshoe*”的名词实体知识，从而倾向于预测在源领域模型中已被认知为名词实体的“*nail polish*”，而使用 TST-F-MTL 引入领域知识后，模型均能预测正确，验证了领域知识对命名实体判别的贡献.

(12) 错误定性分析

本文基于表 9 中 CoQA 至 TweetQA 的迁移实验，在开发集测试结果中，筛选了全部 148 条错误样本进行分析(使用 BLEU-1、METEOR 和 ROUGE-L 进行质量评估，三个测度取值都为 0 时，代表样本错误). 如下文所示，本文将其为 6 大类.

① 外部知识(占比 36%)

段落：*BB, anyone could play a thousand notes and never say what you said in one. #RIP #BBKing Lenny Kravitz (@LennyKravitz) May 15, 2015*
译文：*B.B.KING，任何人都能演奏吉他，却不能将你想说的全说出来. #悼念 #BBKing 莱尼·克拉维茨(@莱尼·克拉维茨)*

问题：*what does bb king play?*
译文：*bb king 演奏的是什么*

标准答案：*guitar*
译文：*吉他*

预测答案：*notes*
译文：*音符*

此类问题需要一定的外部知识进行辅助. 如上例所示，段落中的“*play a thousand notes*”中文语义指的是“*演奏音乐*”. 如果仅从给定段落推断，那么似乎“*notes*”或者“*music*”才是标准答案. 但是，由外部常识知识可知，“*BB King*”是一名吉他手，所以答案应该是更为精确的“*吉他*”.

② 指代消歧(占比 18%)

段落：*YOU ARE IN LOVE *slams face into bathroom counter while getting ready* — Cattie ➳(@cattiehallway) May 5, 2015*
译文：*你恋爱了 *在准备的时候把脸撞到浴室柜台上* — 凯蒂 ➳(@凯蒂走廊) 2015 年 5 月 5 日*

问题：*who is the one in love?*
译文：*谁恋爱了?*

标准答案：*YOU*
译文：*你*

预测答案：*cattie*
译文：*凯蒂*

消歧类问题需要模型建立精准的指代关系，现有模型缺乏部分指代消歧的能力. 此例中，模型误将“*YOU*”和“*Cattie*”建立关联，导致答案预测错误. 段落中“—”符号有解释推文作者的特殊作用，其之后出现的人名即为其之前推文内容的作者名，所以推文既然是作者“*Cattie*”所写，那么内容中的“*YOU*”就无法指代作者. 事实上，“*YOU*”指代为看这条推文的人，在例子的语境下就是指“*你*”.

③ 误拆分或者未拆分(占比 16%)

段落：*Introducing the re-imagined "Do Ya Think I'm Sexy" ft @DNCE & a newly minted partnership with @republicrecords. Sir Rod Stewart (@rodstewart) August 23, 2017*
译文：*《Do Ya Think I'm Sexy》是由 ft @DNCE 重新设计的，并且是与 @共和国记录 的新合作. 罗德·斯图尔特爵士(@罗德·斯图尔特) 2017 年 8 月 23 日*

问题：*who was the partnership?*
译文：*谁是合伙人?*

标准答案：*republic records*
译文：*共和国记录*

预测答案：*republicrecords*
译文：*共和国记录*

部分特殊单词不应该被拆分的，却被拆分. 如4.1节表2的例子，从标准答案上看，“*InStyle*”不应该拆分却被拆分成“*In*”和“*Style*”，但文本规范化的方法将段落中的“*InStyle*”进行拆分，导致预测答案也进行了拆分. 另外一些应该拆分却未进行拆分，比如上例所示，段落中“*republicrecords*”并未拆分成“*republic*”和“*records*”，导致模型预测的答案也未进行拆分，从而与标准答案不一致，因此现有的基于规则方法的SPLIT模块需要进一步改进.

④ 答案正确却误判(占比14%)

段落：*The 3 goals the Predators scored in a 2：19 span in 3rd period are the fastest 3 goals the team has scored in their postseason history. — ESPN Stats & Info (@ESPNStatsInfo) April 18, 2015*
译文：*掠夺者队在第三节的2分19秒内进了3个球，这是他们季后赛历史上最快的3个进球. — ESPN统计信息(@ESPN统计信息) 2015年4月18日*

问题：*how many goals did the predators score?*
译文：*掠食者队进了多少球?*

标准答案：*3 goals*
译文：*3分*

预测答案：*three*
译文：*3*

比如预测答案是“*two*”，但是标准答案是“*2*”，或者如上例所示，标准答案“*3 goals*”，但是预测答案是“*three*”. 预测答案与标准答案语义上一致，按实际应用场景二者应该都判作正确，但现有的评估指标(BLEU-1、METEOR和ROUGE-L)不能对预测答案与标准答案是否具有同义关系作出有效评估. 为此，现阶段生成式阅读理解需要更灵活有效的评估指标.

⑤ 主、被动误判(占比5%)

段落：*I asked @ElectricMayhem for a song to release on Christmas. So in typical fashion, it's here a day late. #TheMuppets — Kermit the Frog (@KermitTheFrog) December 26, 2015*
译文：*我请求@布偶演播室 在圣诞节发行一首歌. 所以按照惯例，它晚到了一天. #布偶 — 科米蛙(@科米蛙) 2015年12月26日*

问题：*who is asked something?*
译文：*谁被问了问题?*

标准答案：*@electricmayhem*
译文：*布偶演播室*

预测答案：*kermit the frog*
译文：*科米蛙*

模型对主、被动关系不敏感. 此例中，问题所问“*谁被问了问题?*”为被动语态形式，模型误将问题提问者“*Kermit the Frog*”作为最终答案. 实际上，由本文引言部分可知，“@”符号后通常出现人物或机构等的名字的连写形式，所以“*@electricmayhem*”也附带“人”或“机构”的命名实体背景知识，可以合理的作为上例的标准答案. 但是，模型分不清“问”和“被问”的区别，并且依然缺乏一定的命名实体知识，导致误测为主动提问者.

⑥ 无法回答(占比11%)

段落：*2014*
译文：*2014*

问题：*what did akers just see in the sky?*
译文：*埃克斯刚才在天上看到了什么?*

标准答案：*a shooting star*
译文：*一个流星*

预测答案：*a sky*
译文：*一片天空*

无法回答类包括给定问题不可回答(比如样本的标准答案是“*I don't know*”，其问题无解)，或者给定的样本无意义. 无意义样本如上例所示，段落内容只有“*2014*”，并不包含其他有效信息，因而无法根据给定段落回答其给出的问题，同时因为缺乏诸如推文作者等外部知识的获取媒介，也无法通过添加外部知识丰富其语义信息.

6 结 论

本文专注于解决社交媒体机器阅读理解领域无监督的领域自适应的问题，将文本规范化作为数据预处理步骤，并提出了一种结合迁移自训练与多任务学习机制的模型框架. 通过在迁移自训练的基础上加入掩码预测任务与生成式阅读理解任务的多任务学习机制，借以加强模型对目标领域深层语义的理解以及特征建模，并进一步提高模型对目标领域的泛化能力. 本文在多个数据集上证实有效. 此外，迁移自训练后的模型的跨领域性能，与利用本领域训练数据进行训练所得的性能，具有较高的可比性.

本文发现噪声样本其负面影响会在多任务学习中进一步放大，为此，未来工作将着重研究构建动态阈值筛选器，通过在训练阶段设定动态变化的阈值，以达到同时保证样本质量和数量的效果，减轻多任务机制对模型的负面影响.

致 谢 感谢编辑部审稿人提供的宝贵意见!

参 考 文 献

[1] Liu S, Zhang X, Zhang S, et al. Neural machine reading comprehension: Methods and trends. Applied Sciences, 2019, 9(18): 3698

[2] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ questions for machine comprehension of text//Proceedings

of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 2383-2392

[3] Reddy S, Chen D, Manning C D, et al. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7: 249-266

[4] Kočiský T, Schwarz J, Blunsom P, et al. The NarrativeQA reading comprehension challenge. Transactions of the Association for Computational Linguistics, 2018, 6: 317-328

[5] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2019: 4171-4186

[6] Dong L, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation. arXiv preprint arXiv:1905.03197, 2019

[7] Bao H, Dong L, Wei F, et al. UniLMv2: Pseudo-masked language models for unified language model pre-training//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 642-652

[8] Huang R, Zou B, Hong Y, et al. NUT-RC: Noisy user-generated text-oriented reading comprehension//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain, 2020: 2687-2698

[9] Pan S J, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[10] Chung Y A, Lee H Y, Glass J. Supervised and unsupervised transfer learning for question answering. arXiv preprint arXiv:1711.05345, 2017

[11] Cao Y, Fang M, Yu B, et al. Unsupervised domain adaptation on reading comprehension//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(5): 7480-7487

[12] Long M, Cao Z, Wang J, et al. Conditional adversarial domain adaptation. arXiv preprint arXiv:1705.10667, 2017

[13] Liu X, He P, Chen W, et al. Multi-task deep neural networks for natural language understanding//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy, 2019: 4487-4496

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017

[15] Xiong W, Wu J, Wang H, et al. TweetQA: A social media focused question answering dataset. arXiv preprint arXiv: 1907.06292, 2019

[16] Nguyen T, Rosenberg M, Song X, et al. MS MARCO: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016

[17] Choi E, He H, Iyyer M, et al. QuAC: Question answering in context. arXiv preprint arXiv:1808.07036, 2018

[18] Dhingra B, Liu H, Yang Z, et al. Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549, 2016

[19] Kadlec R, Schmid M, Bajgar O, et al. Text understanding with the attention sum reader network. arXiv preprint arXiv: 1603.01547, 2016.

[20] Zhang S, Zhao H, Wu Y, et al. DCMN+: Dual co-matching network for multi-choice reading comprehension//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(5): 9563-9570

[21] Parikh S, Sai A B, Nema P, et al. ElimiNet: A model for eliminating options for reading comprehension with multiple choice questions. arXiv preprint arXiv:1904.02651, 2019

[22] Zhang Z, Wu Y, Zhou J, et al. SG-Net: Syntax-guided machine reading comprehension//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 9636-9643

[23] Yu A W, Dohan D, Luong M T, et al. QANet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541, 2018

[24] Qian Jin, Huang Rong-Tao, Zou Bo-Wei, et al. Generative reading comprehension via multi-task learning//Proceedings of the 19th Chinese National Conference on Computational Linguistics. Haikou, China, 2020: 301-312(in Chinese)
(钱锦，黄荣涛，邹博伟等. 基于多任务学习的生成式阅读理解//第 19 届中国计算语言学会议论文集. 海口，中国，2020: 301-312)

[25] Nishida K, Saito I, Nishida K, et al. Multi-style generative reading comprehension. arXiv preprint arXiv:1901.02262, 2019

[26] Wang H, Gan Z, Liu X, et al. Adversarial domain adaptation for machine reading comprehension. arXiv preprint arXiv: 1908.09209, 2019

[27] Shakeri S, Santos C N, Zhu H, et al. End-to-end synthetic data generation for domain adaptation of question answering systems. arXiv preprint arXiv:2010.06028, 2020

[28] Lee H G, Jang Y, Kim H. Machine reading comprehension framework based on self-training for domain adaptation. IEEE Access, 2021, 9: 21279-21285

[29] Yue Z, Kratzwald B, Feuerriegel S. Contrastive domain adaptation for question answering using limited text corpora. arXiv preprint arXiv:2108.13854, 2021

[30] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test. The Journal of Machine Learning Research, 2012, 13(1): 723-773

[31] Butt S, Ashraf N, Siddiqui M H F, et al. Transformer-based extractive social media question answering on TweetQA. Computación y Sistemas, 2021, 25(1): 23-32

[32] van der Goot R. MoNoise: A multi-lingual and easy-to-use lexical normalization tool//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy, 2019: 201-206

[33] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521-3526

[34] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[35] Denkowski M, Lavie A. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems//Proceedings of the Sixth Workshop on Statistical Machine Translation. Edinburgh, UK, 2011: 85-91

[36] Lin C Y. ROUGE: A package for automatic evaluation of summaries//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics Workshop. Stroudsburg. Barcelona, Spain, 2004: 74-81

[37] Hofmann T, Schölkopf B, Smola A J. Kernel methods in machine learning. The Annals of Statistics, 2008, 36(3): 1171-1220

[38] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 1180-1189

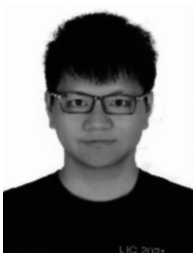
[39] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139-144

[40] Fisch A, Talmor A, Jia R, et al. MRQA 2019 shared task: Evaluating generalization in reading comprehension. arXiv preprint arXiv:1910.09753, 2019

[41] Joshi M, Choi E, Weld D S, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551, 2017

[42] Yang Z, Qi P, Zhang S, et al. HotpotQA: A dataset for diverse, explain-able multi-hop question answering. arXiv preprint arXiv:1809.09600, 2018

[43] Wang S, Zhang L. Self-adaptive re-weighted adversarial domain adaptation. arXiv preprint arXiv:2006.00223, 2020



LIU Hao, M. S. candidate. His main research interest is machine reading comprehension.

HONG Yu, Ph. D., professor. His research interests include machine reading, event extraction and so on.

ZHU Qiao-Ming, Ph. D., professor. His research interests includes machine reading comprehension, Chinese language processing and so on.

Background

The existing machine reading comprehension models still have the problem of domain adaptability, that is, in the absence of domain knowledge, it is difficult to form accurate feature representations of language phenomena in a specific domain (such as colloquial and symbolic expression languages in social media). In the modeling process, the adaptability problem is the under-fitting of distributed representation across domains. Obviously, building a substantial sample of annotations for the target domain can help solve this problem. However, the timeliness of manual annotation is often difficult to be guaranteed in the practical application scenarios of various fields.

The previous unsupervised domain adaptation work is mostly based on self-training. Data expansion is realized by constructing pseudo samples in the target domain, and optimization is carried out on this basis, which has achieved good results. However, there is not much work that incorporates based on the normative text domain, but the text domain multi-tasking mechanisms to further improve the generalization ability of the model. In addition, most of the previous work is based on specific language phenomena such as Twitter is scarce. Furthermore, the domain adaptation research based on generative reading comprehension is almost blank. To this end, this paper is based on generative machine reading comprehension, eliminates noise samples through filters, and incorporates the multi-task mechanism of the mask language model on the basis of transfer self-training, so as to further improve the adaptability of the model in the target domain.

This paper is supported by the National Key R&D Program of China (2020YFB1313601) and the National Natural Science Foundation of China (62076174, 61836007). These projects are mainly for natural language processing in different fields, including social fields and cross-language fields.