面向土木建筑信息领域的自然语言问题生成方法

朱磊，焦瑞，黑新宏，赵钦，姚燕妮，方潇颖，杨明松，盘隆

西安理工大学 Xi'an University of Technology

朱磊 leizhu@xaut.edu.cn
焦瑞 xdrshjr@qq.com

Abstract.

本文将Transformer和UniLM结合构建了序列学习模型，获取土木建筑信息领域的句子级语义信息，自动解码生成对应的自然语言问题。该模型是一个序列到序列模型，将大量开放域和土木专业领域中语料的语法、句法规则迁移到土木建筑信息问答领域，结合该领域内的少量人工标注数据集获取语义信息。通过对Transformer中不同模块进行随机采样分层训练，优化后生成良好的领域目标问句。实验结果表明，我们提出的模型不需要人为指定规则和设置复杂的自然语言处理管道，在机器评价和人工评价指标中都展现出更好的语义理解能力，最终生成高质量的土木建筑信息领域问题。

模型结构.

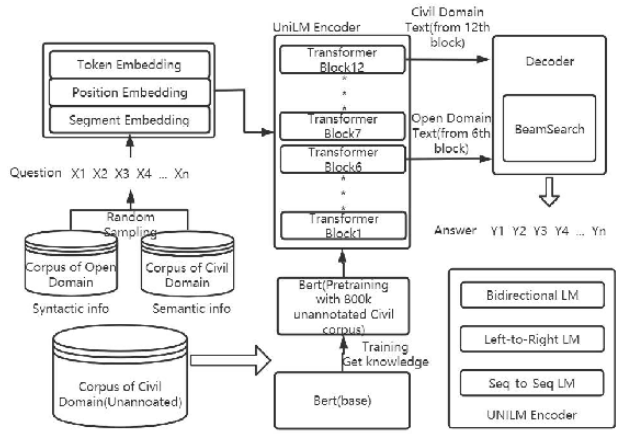

图1. 问题生成模型的框架

模型描述.

本文模型首先对基础的Bert(base)进行加载，获取Transformer最后一层的参数信息。并且将未标注的土木建筑信息领域的文本进行输入，进行无监督训练。这个训练过程是为了获取土木建筑信息领域的隐含知识信息。

然后，我们对开放域问答和土木建筑问答数据进行随机采样。不同的采样结果将通过Embedding方式输入到Transformer的不同层次模块中进行分层训练。嵌入方式包括了Token词嵌入、段嵌入和位置嵌入。在嵌入编码时采用了UniLM的遮盖词策略，目的是提升模型的自然语言生成能力。其中，引入开放域问答数据是因为基于Wiki百科语料的低参数Bert预训练模型可以获取常用百科知识信息，从而在语法和语义上增强真实文本的生成能力。

最后，在训练优化阶段，对于不同类型的采样数据，模型将取出不同层的Transformer模块进行梯度计算并反传优化。优化后的模型即可用于生成自然语言问题的推断，推断思路主要是采用了beam search技术。

类似Bert中多层堆叠的Transformer结构，不同层次的模块分别捕捉不同的文本嵌入信息：i）表层信息特征在底层网络中进行编码；ii）句法信息特征在中间层网络中进行编码；iii)语义特征则主要集中在高层网络中编码。因此，本文模型对开放域文本在低层次Transformer模块中进行训练，对土木建筑信息领域文本在高层Transformer模块中进行下游生成任务的微调。这种策略可以有效地增强模型的问题生成能力。

表7. 人工指定问题的结果

模型	每30个生成问题的平均有意义问题个数	每30个生成问题的平均有意义问题比例
LSTM-based Seq2seq	3	10%
Origin Bert + Our Model	17	56%
Low Paras Bert + Our Model	13	43%
PreTrain 800k data + Our Model	19	63%
PreTrain 800k data + Different data different train + Our Model	21	70%

问题定义.

问题生成可以形式化描述为：在给定输入序列X的基础上，生成和原始序列“高”相关的目标句子序列Y。通常X和Y分别定义为X=[$x_1, x_2, x_3, \ldots, x_{|X|}$]，Y=[$y_1, y_2, y_3, \ldots, y_{|Y|}$]。其中，|X|和|Y|分别表示输入和输出序列的长度。对于给定的输入序列S，通常需要训练神经网络模型来得到序列S的嵌入表示X。自然语言的问题生成可以被定义为:

$$\bar{Y} = \arg\max_{Y} P(Y \mid X)$$

(公式-1)

公式-1中，在给定序列S及其嵌入表示X的前提下，P是预测生成对应目标序列Y的最大化对数似然函数，从而输出相关性较高的输出序列。

数据集.

表 1. 土木建筑信息领域通过句子文本生成问题的实例

Sentence:特种结构是指具有特种用途的工程结构。

Question:什么是特种结构?

Sentence:深埋隧道多采取暗挖法施工，用圆形盾构开挖和钢筋混凝土管片支护。

Question:深埋区间隧道结构有什么特点?

Sentence:桥台是桥梁两端桥头的支承结构，是道路与桥梁的连接点。

Question:桥台的定义是什么?

表2. 训练数据集

数据集	问答对个数	示例
土木建筑信息领域的问答数据集	4000	**问句**：建筑物设计建筑标定人数的确定有固定座位等标明使用人数的建筑，应如何计算配套设施疏散通道和楼梯及安全出口的宽度？**答句**：建筑物设计建筑标定人数的确定有固定座位等标明使用人数的建筑计算配套设施疏散通道和楼梯及安全出口的宽度按照标定人数为基数计算。
开放域句子对数据集	400,000	**问句**：眼球突出，复视，视力减退到底是什么病在捣鬼？**答句**：眼球突出可能是近视，复视则是散光，建议到眼科或眼镜店检查 视力，及时进行矫正，防止视力问题更严重。
土木建筑规范数据集	800,000	**上半句**：{'text1': '抗震支吊架与建筑结构体牢固连接', **下半句**：'text2': '以地震力为主要荷载的抗震支撑设施。'}

表4. BLEU评估结果

Model	BLEU-4 80average	BLEU-4 80max	BLEU-3 80average	BLEU-3 80max	BLEU-2 80average	BLEU-2 80max	BLEU-1 80average	BLEU-1 80max
LSTM-based Seq2seq	13.41	73.61	13.23	74.8	14.36	75.95	18.09	76.91
Origin Bert + Our Model	20.58	87.15	31.64	94.1	25.67	93.54	31.64	94.10
Low Paras Bert + Our Model	21.42	81.93	32.13	90.01	26.67	87.02	32.13	91.02
PreTrain 800k data + Our Model	25.79	**89.42**	35.26	**96.10**	30.05	**94.05**	35.26	**96.10**
PreTrain 800k data + Different data with different training + Our Model	**26.19**	83.85	**37.88**	92.85	**30.53**	88.6	**37.88**	92.85

表5. ROUGE评价结果

Model	ROUGE1	ROUGE2	ROUGE3	ROUGE4	ROUGEL
LSTM-based Seq2seq	10.75	5.17	3.67	2.5	10.62
Origin Bert + Our Model	38.34	24.41	17.15	14.43	37.96
Low Paras Bert + Our Model	36.13	24.70	18.81	15.01	37.09
PreTrain 800k data + Our Model	37.82	22.07	12.68	9.31	37.34
PreTrain 800k data + Different data with different training + Our Model	**41.86**	**27.76**	**20.56**	**16.89**	**41.46**

本文针对土木建筑信息领域少量的问答训练数据进行了少样本训练方案的研究，使用了Transformer基础的编码器-解码器结构，并且结合了UniLM思想。同时，将开放域问答数据进行了预训练处理，随机采样并在Transformer堆叠模块的不同层分别计算损失函数并反传梯度进行模型训练。实验结果证明了模型在土木建筑信息领域生成问题的可行性和有效性，并且达到了较高的自然语言问题生成水准。