基于动态词遮掩的句子匹配预训练模型

宋挺[1,2], 郭展成[1,2], 何世柱[1,2], 刘康[1,2], 赵军[1,2], 刘升平[3]

[1] 中国科学院自动化研究所, [2] 中国科学院大学, [3] 云知声智能科技股份有限公司

ting.song@nlpr.ia.ac.cn

简介

背景

- 句子匹配任务
  - 句子匹配或复述识别，是NLP的重要任务之一，也是自动问答、聊天机器人、信息检索、机器翻译等应用的重要基础和关键模块。
  - 任务描述：一个典型的语义匹配任务，对于输入的两个句子，模型需要判断它们语义是否一致，是否表达了相同含义或相同意图。

图 1. 句子匹配任务示例。

- 预训练模型
  - BERT：通过MLM、NSP等自监督学习任务学习和建模通用语言规律，在序列标注、文本分类、阅读理解等自然语言理解任务中取得了良好效果。
  - **问题**：1、BERT的NSP任务不能建模句子的语义匹配关系，NSP任务得到的表示不适合应用于句子匹配任务；2、BERT的随机遮掩策略不能高效建模句子的关键信息，而关键信息是判断两个句子语义关系的重要特征。

本文贡献

- 针对BERT无法有效建模句子匹配关系的问题，提出了一种**面向句子匹配任务的预训练数据构造方法**。
- 针对BERT随机遮掩策略无法建模句子匹配任务的关键信息问题，提出了一种**高效建模关键信息的动态词遮掩策略**。
- 在4个句子匹配数据集上的实验结果表明：使用本文提出的预训练方法，RBT3和BERT base的效果都有一定提升，取得了当前最好效果。

方法

“句子对”预训练数据

- 构造与句子匹配任务数据分布接近的预训练数据：基于预训练模型获得大规模句子的向量表示，进而通过近似语义计算自动获取大规模“句子对”预训练数据。

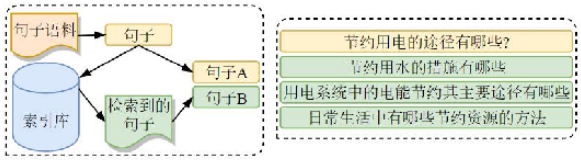

图 2. 构造预训练数据与“句子对”数据示例，左侧为构造预训练数据的流程，右侧为自动构造得到的“句子对”数据示例。

- 预训练数据构造步骤：
  1. 将所有句子用预训练BERT进行编码，得到所有句子的向量表示。
  2. 利用向量检索工具Hnswlib对所有句子的向量表示建立索引。
  3. 选取部分句子的向量表示检索相似句子，得到语义相似的句子对。

词遮掩策略

- 随机词遮掩（RAN）：BERT的随机遮掩策略。
- 关键词遮掩（KW）：利用TF-IDF算法抽取关键词，优先遮掩关键词。
- **动态词遮掩（MLM）**：优先遮掩句子中掩码恢复损失函数值高的关键词，增加MLM任务的难度，也能增强模型对关键信息的建模能力。

遮掩策略	遮掩后的句子
RAN	节约用电 [MASK] 途径 [MASK] 哪些?
KW	节约用电的途径有 [MASK] [MASK]?
MLM	节约 [MASK] [MASK] 的途径有哪些?

表 1. 以句子“节约用电的途径有哪些?”为例，3种遮掩策略的对比结果。

预训练实验设置

- 预训练任务：遮掩语言模型（Masked Language Model）。
- 超参数：batchSize=768, maxLen=48, LR=3e-5, epochs=5。
- 预训练模型：
  - RBT3（L=3，H=768，A=12，总参数量38M）。
  - BERT base（L=12，H=768，A=12，总参数量110M）。

实验

实验数据

- 预训练数据
  - 来源：百度知道的问句数据，包括多种领域。
  - 规模：近800万个问句数据，构造得到了约400万个“句子对”样本。
- 句子匹配数据
  - AFQMC、LCQMC、BQ、cMedQQ：蚂蚁金融问句匹配数据、大规模中文问句匹配数据、银行客服领域问句匹配数据、医疗领域问句匹配数据。

数据集	领域	训练集大小	开发集大小	测试集大小
AFQMC	互联网金融	34,334	4,316	-
LCQMC	开放域	238,766	8,802	12,500
BQ	银行客服	100,000	10,000	10,000
cMedQQ	医疗	16,071	1,793	1,935

表 2. 句子匹配数据的领域及规模。

实验结果

- 总实验结果：在4个句子匹配数据集上，基于本文提出的预训练方法，RBT3和BERT base的平均准确率分别提升了1%、0.6%，取得了当前最好的效果。

模型	AFQMC	LCQMC	BQ	cMedQQ	AVG
SiameseLSTM	65.11	73.50	73.51	72.11	71.06
DecAtt	66.11	80.04	77.78	72.96	74.22
ESIM	69.20	86.26	81.45	82.38	79.82
RBT3	70.57	85.97	81.70	84.39	80.66
RBT3 (Ours)	**71.52**	**86.70**	**82.30**	**86.25**	**81.69 (+1.0)**
BERT	73.47	85.94	84.29	86.61	82.58
RoBERTa	73.22	**87.38**	83.63	86.51	82.68
BERT (Ours)	**73.86**	87.00	**84.30**	**87.60**	**83.19 (+0.6)**

表 3. 基于动态词遮掩的句子匹配预训练方法，RBT3和BERT base在4个数据集上的准确率（%）。基线模型为Siamese-LSTM、DecAtt、ESIM等传统句子匹配模型以及RBT3、BERT base、RoBERTa等预训练模型。

- 对比实验：验证预训练数据构造方法与动态词遮掩策略。
  - +RAN（随机遮掩）：在预训练数据上使用BERT的随机遮掩策略。
  - +KW（关键词遮掩）：优先遮掩句子的关键词。
  - +MLM（动态词遮掩）：优先遮掩恢复损失函数值高的关键词。

模型	AFQMC	LCQMC	BQ	cMedQQ	AVG
RBT3	70.57	85.97	81.70	84.39	80.66
+RAN	70.62	86.14	82.13	84.91	80.95 (+0.3)
+KW	70.67	86.32	82.22	85.37	81.15 (+0.5)
+MLM (Ours)	**71.52**	**86.70**	**82.30**	**86.25**	**81.69 (+1.0)**
BERT	73.47	85.94	84.29	86.61	82.58
+RAN	72.94	86.94	84.10	86.22	82.55 (+0)
+KW	73.24	**87.14**	83.79	86.51	82.67 (+0.1)
+MLM (Ours)	**73.86**	87.00	**84.30**	**87.60**	**83.19 (+0.6)**

表 4. 相对于RBT3模型，使用动态词遮掩策略在4个数据集上均取得最好的效果，平均准确率提升了1%。使用随机遮掩策略平均准确率提升0.3%，这说明面向句子匹配的预训练数据构造方法是有效的。使用关键词遮掩策略平均准确率提升了0.5%，这说明在预训练阶段增强对关键词信息的建模能力有助于句子匹配任务。

总结与展望

总结：针对BERT在预训练阶段无法有效建模句子语义匹配信息及随机遮掩策略无法高效建模句子关键信息的问题，本文提出了基于动态词遮掩的句子匹配预训练方法。在4个句子匹配数据集上的实验表明，使用本文提出的预训练方法，基于RBT3和BERT base两个基础模型的效果都有一定提升，取得了当前最好效果。

展望：更好的预训练数据构造方式与高效建模句子的关键信息是未来的重点，如：以端到端的方式构造预训练数据、使用预训练模型直接提取句子的关键信息是以后的研究方向。

致谢

本论文受到国家重点研发计划（2017YFB1002101），国家自然科学基金项目（61533018, U1936207, 61976211, 61702512），模式识别国家重点实验室自主课题基金和中国科学院青年创新促进会资助。