Week8

本笔记对机器阅读理解(Machine Reading Comprehension，MRC)做简略总结梳理，便于以后复习。

一、MRC 简览

MRC 是一种利用算法使计算机理解文章语义并回答相关问题的技术。大部分机器阅读理解任务采用问答式测评：设计与文章内容相关的自然语言式问题，让模型理解问题并根据文章作答。

主要有多项选择式、区间答案式、自由问答式和完形填空式。一些数据集设计了“无答案”问题，即一个问题可能在文章中没有合适答案，需要模型输出“无法回答”（unanswerable）。在以上的答案形式中，多项选择和完形填空属于客观类答案，测评时可以将模型答案直接与正确答案比较，并以准确率作为评测标准，易于计算。

二、MRC 基础框架

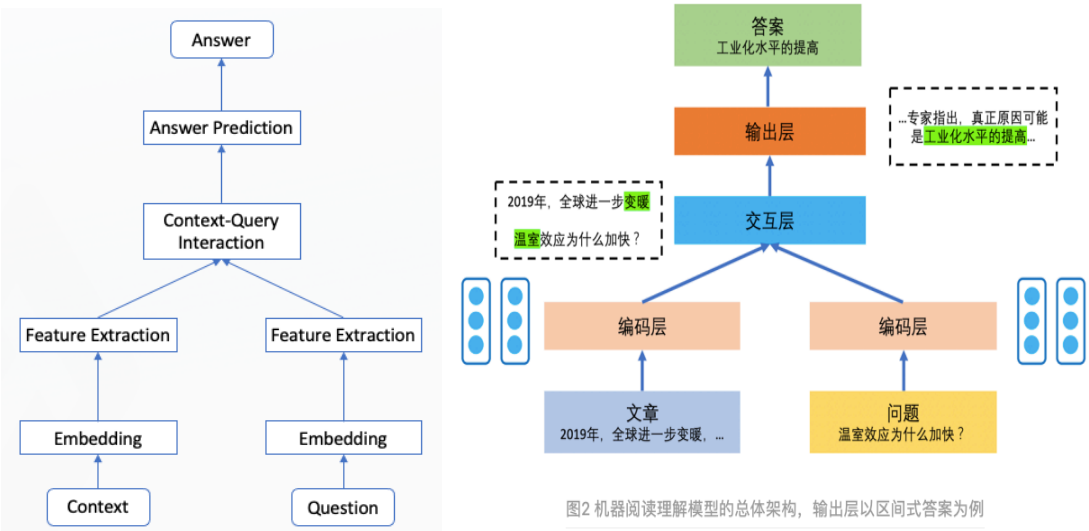

图2 机器阅读理解模型的总体架构，输出层以区间式答案为例

编码层(Embedding 和 Feature Extraction)：对 context 和 question 进行数字化编码，变成可以被计算机处理的信息单元。模型需要保留原有语句在文章中的语义。因此，每个单词、短语和句子的编码必须建立在理解上下文的基础上；

交互层(Context-Query Interaction)：由于文章和问题之间存在相关性，模型需要建立文章和问题之间的联系，主要用到自然语言处理中的注意力机制。阅读理解模型将文章和问题的语义结合在一起进行考量，进一步加深模型对于两者各自的理解；

输出层(Answer Prediction)：根据语义分析结果和答案的类型生成模型的答案输出，需要确定模型优化时的评估函数和损失函数。经过交互层，模型建立起文章和问题之间的语义联系，就可以预测问题的答案。完成预测功能的模块为输出层。

三、MRC 相关模型

1.Bert 之前算法

Attentive Reader& Impatient Reader: 应用了一维匹配和二维匹配模型；
BiDAF: 使用了双向注意力流；
R-Net: 使用了门机制，应用了 Pointer Network 以及注意力机制融入 RNN;
FusionNet: 应用单词历史和全关注注意力；
QANet: 应用 CNN+Self-Attention, 返译来实现数据扩增，开始接近 Bert；

2.Bert 之后算法

Bert: Pre-training+fine-tune，运用海量语料得到预训练模型，获得模型架构和权重，然后用任务语料调整权重，得到 fine-tune 结果。预训练任务有 Mask-LM 和 NSP，输入中加入 Positional-encoding;
ERNIE: 百度中文模型，应用了实体和短语 mask,构建了 Dialogue LM 任务；
Transformer-XL：应用了 RNN 循环机制以及相对位置编码；
XLNet: 应用了 Permutation Language Modeling(PLM)和双流自注意力；
StructureBert&T5: 前者是基于 Bert 预训练模型，从预训练任务的角度修改 bert。后者是一个通用的框架，做了大量前人想做却没做的实验，得到一些有意义的结论；
Roberta：应用了更大的模型参数量、batch size 和更多的训练数据，训练方法上去掉了 NSP 任务，应用了动态 Mask 和文本编码；
ALBERT: 应用了两种减少参数的方法，矩阵分解和参数共享。用 SOP 任务替换掉 NSP 任务，同时应用 n-gram Mask;
Retro-Reader: 基于 Bert+其他策略的组合阅读理解模型。

3.其他知识

模型集成：集成模型，提升效果；
模型蒸馏：新的小模型去学习大模型的预测结果以及泛化能力；

参考文献

1.课程讲义
2.朱晨光 NLP 专栏:一文读懂机器阅读理解,
https://www.jiqizhixin.com/articles/2020-04-30-3