**Reinforcement Learning from Human Feedback Summary**

沙之洲 2020012408

这次 talk 的主要工作是总结了当前 RLHF 用于 LLM finetune 中遇到的一些挑战和机遇。

首先，现在 RL 面临的最大问题就是 Truthfulness ，也即模型回答问题的真实性。这里更深层次的原因是 LLM next token prediction 的根本是最大化每个 token 的概率，换个角度看就是让整个句子尽可能接近人类真实说的话，但是关于输出正确性并没有被严格纳入考量。Talk 提出了两种可能解决的方法。其一，通过 SFT 加上 RL reward 改进的方式让模型学会说出自己的 Uncertainty；其二，对于单 token 预测任务，用模型对于 token 的预测概率，作为量化指标，只有 token 的预测概率大于某个给定阈值的时候，才会采纳模型的输出，否则认为模型对这个问题是 Uncertainty 的。

Talk 第二个关键的点在于模型中的 knowledge。Talk 认为模型在训练中已经在自己的权重中隐式维护了一个知识图谱，而模型只能够回答在这之中的问题。Talk 认为，RLHF 这个过程，只不过是让模型学会了如何利用内在的知识图谱回答某些规定范式的问题，但是对于知识图谱之外的问题，模型仍然会胡编乱造。进一步的，这种观点认为，reward model 的知识图谱不可能比原有的 LLM 更大，所以 RLHF 并不会对 LLM 的知识产生更新。事实上，LLM 缺少类似人类的反思性的过程，所以不能通过自我反思来对现有知识进行优化。所以，一种观点认为 LLM 的上限将是集成人类的所有知识，而不能 generate new knowledge

以下是对于上课观点讨论的一些总结。当我们做选择题的时候，我们是根据一个句子的整体来判断句子的正确性。现有的句子概率是将所有词的概率做简单的乘积，而这并不是一个合理的计算句子概率的方式。一种观点认为，人类做复杂逻辑推理的时候，只会关注于几个关键的 features，因此一种可能进行逻辑推理的方式是，利用 LLM 从句子中抽取出用来抽取的关键特征，再经过离散的推理网络得到复杂的推理结果。

另一点考量是关于模型权重中的知识图谱。实际上，我们对于内嵌在权重中的知识图谱所知甚少。所以我们很难对其作出合理的预测。同时，训练语料中可能会出现多方对于同一个问题冲突的观点。而对于模型来说，很肯能采取的是对所有观点采取 average的行为，我们无法预测这种情况出现的频率，也没法消除这种情况对于模型 truthfulness 的影响。因此，下一阶段的可能研究方向，是对于权重中知识图谱的提取和验证。