# 预训练模型学习情况周报 9

姚凯

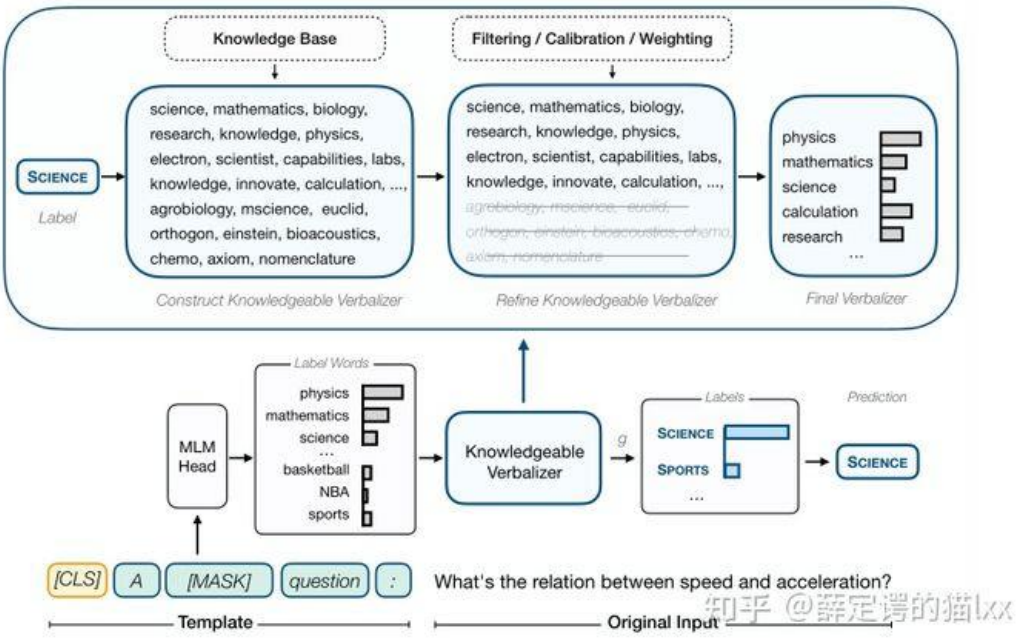
## 一、 本周学习:

### 论文阅读

### 阅读《Knowledgeable Prompt-tuning:Incorporating Knowledge into Prompt Verbalizer for Text Classification》

基于提示学习的文本分类

之前绝大部分的 Verbalizer 都是手工构造的，该论文重点在于如何融入外部的知识库信息以构建一个新的映射器（映射 MLM 预测的模型和文本数据集限制的分类类别)，整个流程:




KPT 主要包含三个步骤:

- 标签词的扩展
- 扩展标签词的精简
- Verbalizer 的使用

1) 标签词的扩展

知识库中，某单词的连接边权重分数大于阈值的点，作为标签词的扩展词

2) 扩展标签词的精简

a) 得到的扩展词不在 PLM 的单词空间中：

将词拆分成逐 token 的多部分，用 PLM 逐 token 预测的平均准确率，作为整个词的概率

b) PLM 对其空间中的稀有词，预测往往不准确

在扩展词表中删去概率低于阈值的低频词，D 为语料库

$$P_D(v) = \frac{1}{|\bar{C}|}\sum_{x\in\bar{C}} P_M([MASK] = v|x_p)$$

$$\mathbf{x}_\mathrm{p} = \texttt{[CLS]}\ \mathrm{A}\ \texttt{[MASK]}\ \mathrm{question}:\mathbf{x}$$

c) 标签词的先验分布有巨大的偏差

对于 b)中的 PLM 预测的概率，Zeroshot 用以下公式来校准预测的分布：

$$\tilde{P}_\mathcal{M}(\texttt{[MASK]}=v|\mathbf{x}_\mathrm{p}) \propto \frac{P_\mathcal{M}(\texttt{[MASK]}=v|\mathbf{x}_\mathrm{p})}{P_\mathcal{D}(v)} \quad (7)$$

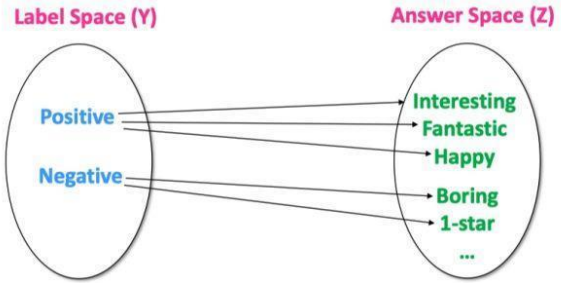
对于 fewshot，对每个标签词分配一个可学习的参数，参数在训练 PLM 模型参数时被同时训练

$$\alpha_v = \frac{exp(w_v)}{\sum_{u\in v_y} exp(w_u)}$$

Fewshot 不需要校准，训练过程中预测的分布会被训练到所需的范围。

$$\tilde{P}_\mathcal{M}(\texttt{[MASK]}=v|\mathbf{x}_\mathrm{p}) = P_\mathcal{M}(\texttt{[MASK]}=v|\mathbf{x}_\mathrm{p}).$$

3) Verbalizer 的使用




对 zeroshot，假定扩展词中每个词对于预测标签的贡献相同，简单平均，用预测分数的均值作为该标签的预测分数,最后取出预测分数最大的类别，作为最后的结果

$$\hat{y} = argmax_{y \in Y}(\frac{1}{|v_y|}\sum_{v \in v_y}\tilde{P}_M([MASK] = V|x_p))$$

Fewshot 情况

将 c）中设计的权重作为扩展词中每个词对于预测标签的贡献度，进行加权平均

$$\hat{y} = argmax_{y \in Y}\frac{exp(s(y|x_p))}{\sum_{y'} exp(s(y'|x_p))}$$

其中

$$s(y|x_p) = \sum_{v \in v_y}\alpha_v log P_M([MASK] = v|x_p)$$

参考：https://zhuanlan.zhihu.com/p/398009000

## 阅读《Bidirectional LSTM-CRF Models for Sequence Tagging》

命名实体识别的 baseline 文章，用 BILSTM+CRF 构建模型

BILSTM 和 CRF 独立均能进行语句单词标签预测，二者结合，BILSTM 使模型同时获取前后向的特征信息，CRF 层的引入可有效解决预测标签之间的强语法依赖问题，避免预测标签冲突情况

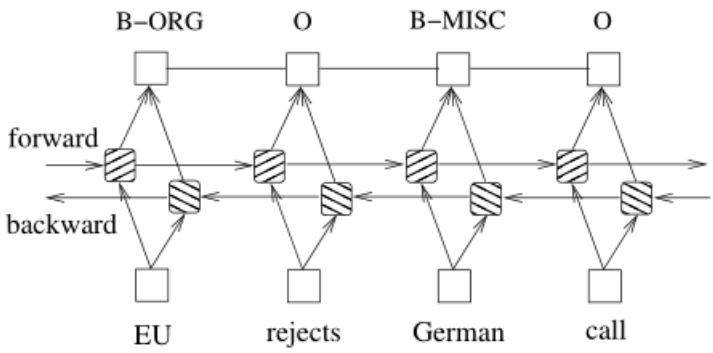



Figure 7: A BI-LSTM-CRF model.

设$[x]_1^T$和$[i]_1^T$分别表示有 T 个单词的句子以及对应的有 T 个的标签序列，$f_{\theta_{[i]_t,t}}$表示输入句子$[x]_1^T$第 t 个单词的第 i 个标注类型的得分，来自 BILSTM 的输出；$[A]_{[i]_{t-1},[i]_t}$表示标注序列状态从第 t-1 个转移到第 t 个的转移分数，A 是 CRF 层的参数，用动态规划计算。

$$s([x]_1^T, [i]_1^T, \tilde{\theta}) = \sum_{t=1}^{T}([A]_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t})$$

两者之和即为一个单词标签的分数，对 T 个单词求和。选取得分 s 最高的标注序列[i]即为所求。

模型的训练由经典反向传播算法更新参数。

参考：https://zhuanlan.zhihu.com/p/356329356

## 阅读《Exploring Pre-trained Language Models for Event Extraction and Generation》

论文主要内容为提出了一个事件抽取模型和一个事件生成方法

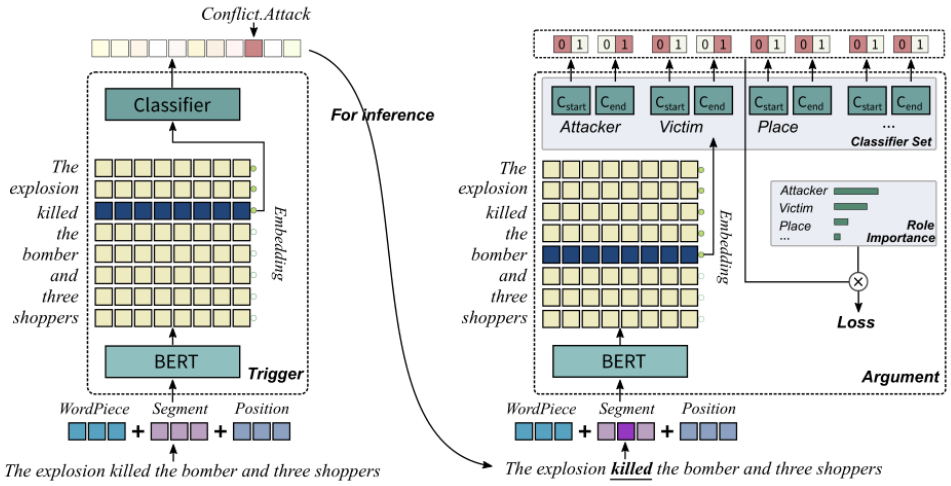



Figure 3: Illustration of the PLMEE architecture, including a trigger extractor and an argument extractor. The processing procedure of an event instance triggered by the word "killed" is also shown.

**事件抽取模型**：左边负责抽取触发词，右边负责抽取相关的论元，以及它们所扮演的全部角色。

因为论元大部分为长名词短语以及角色重叠问题，在 BERT 里添加多个二分类器，每一个分类器为一个角色提供服务，以确定扮演它的所有论元的跨度 span（start，end）。一个论元可以扮演多个角色，一个 token 可以属于不同的论元

**事件生成方法：**对 ACE 2005 数据集的事件，通过这两个步骤，生成新的事件，获得一个带标注的新句子。

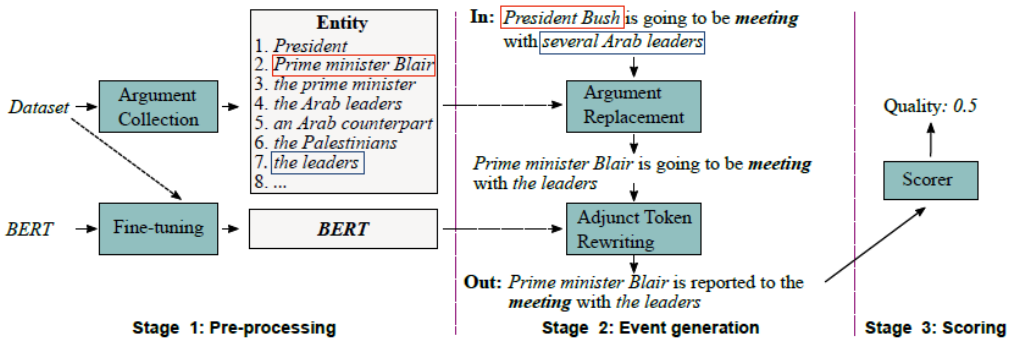



Figure 4: Flow chart of the generation approach.

1) 将原型数据的论元替换为相同角色的其他论元。

2) 使用微调的 BERT 重写附属 token（句子未被识别成触发词或角色的 token）

衡量生成样本质量指标：困惑度（重写的附属 token 的平均概率），以及与原始数据库的距离（句子与数据库的余弦相似度）

参考：https://blog.csdn.net/o11oo11o/article/details/120298139

https://www.jianshu.com/p/d4233193d3e0