# 周报

屈原斌

2020 年 6 月 15 日

## 1 上周任务

- 看论文；
- BLM 代码；

### 1.1 论文

- 论文题目：Contextual Embeddings: When Are They Worth It?
- 主要贡献：
  - 主要研究了训练数据规模和文本语言的特性对词向量的影响；
- 实验中进行对比的三种词向量：
  - Pretrained contextual embeddings: 上下文词向量，文中使用**Bert Embedding**；
  - Pretrained non-contextual embeddings：非上下文词向量，文中使用**GloVe**；
  - Random embeddings：随机词向量，使用循环随机矩阵；

#### 1.1.1 实验

- 训练数据规模
  - **实验结论**：在许多任务中，当提供充足的数据，GloVe这些词向量可匹配BERT。

- 文本语言的特性
  - 语言特性：
    - **Complexity of sentence structure**: 句子中不同单词的相互依赖程度；
    - **Ambiguity in word usage**: 训练过程中，单词的歧义性；
    - **Prevalence of unseen words**：训练过程中，未登录词的可能性。
  - **实验结论**：以 BERT 为代表的 Contextual embeddings 在解决一些文本结构复杂度高和单词歧义性方面有显著的效果，但是在未登录词方面 GloVe 代表的Non-Contextual embeddings 有不错的效果。

## 2 本周计划

- 继续写代码；
- 看论文；