阅读理解下一步实验计划

徐俊

2016 年 12 月 27 日

1 问题

CNN 数据集中，为了增加难度同时要求模型必须依赖于“上下文信息”获取答案，而将所有的数据中的实体（潜在答案）全部替换为 @entity，一篇文章中同一个实体使用同一个 @entity 编号但是不同文章中的同一 @entity 编号却可能表示不同的实体。

目前的 Reader 均将 @entity 编号作为一个个独立的“词”来处理的，它们拥有自己的 embedding，在 softmax 输出层（如果有的话）也是被当做单独的“词”来处理的。这样一来就有问题，在 NN 的模型中，embedding 和 softmax 输出层的权重向量均对应一个独立的词语，而 @entity 符号却并不具备“独立”这个属性，因为同一个 @entity 符号在不同文章中代表不同的实体，而实体才是具备“独立”属性的”词“。

数据集本意是要求 model 仅仅依赖于上下文信息来做判断，@entity 仅仅作为标示符出现而不含有任何语义。

但是实际操作过程中如 Attention Reader 中，最终的预测环节（softmax 输出层）却依赖于 @entity 之间的不同来做分类。这在逻辑上存在一定的悖论。而丹琦的实现中，将文章中的不同实体按照出现顺序依次标号（relabel），对于模型性能的提升有较大帮助，但这一点其实并不应该被利用。

在 Attensum Reader 中，虽然模型逻辑上没有悖论，但是实际操作中也是将不同的 @entity 使用不同的 embedding，间接的赋予 @entity 编号语义。为什么不遵从数据集原意，而将所有的 @entity 设置同一个 embedding，而用 mask 标识出同一个 @entity 标号在文中哪些位置出现?

2 动机

Reader 添加特征的实验一直没有进展，而多项实验从不同侧面暗示上述数据处理方式存在问题，为了验证数据处理方式的问题，遵从数据集合原意测试上述 Reader 的真实性能。即，测试 @entity 的编号以及在 NN 中的表示对于 Reader 的影响，毕竟这两者不应成为模型性能的来源。

3 需要做的实验

3.1 第一优先级的实验

1. **乐高代码迁移**：将模型转移到乐高上去，期待从根本上提速；

2. **数据构造**：取出 @entity 编号的频率影响，使得各个 @entity 编号出现次数相近，这样减少不同 @entity 编号之间的差异；

3. **Attention Reader 去除 relabel 操作下的性能实验**：验证 relabel 对于模型性能的影响；

4. **Attention Reader 使用新构造的数据**：验证 Attention Reader 对于 @entity 编号的依赖程度；

5. **Attensum Reader 中所有实体使用同一个 embeding，使用 mask 标识出同一个实体出现的位置**：完全排除掉 @entity 之间的语义差异，最符合数据集原意；

3.2 第二优先级的实验以及准备

1. **基于 CNN 数据集正在进行的实验**：等待在乐高上的模型运行成功；

2. **SVM**：SVM 的实验继续运行，及时很慢也需要有个结果出来；

3. **切换数据集**：由于之前的核心贡献点并不同数据集强依赖，如果证实 CNN 数据集有问题，迅速切换是比较合适的选择；