

**论文标题: SGM: Sequence Generation Model for Multi-Label Classification**  
**会议: COLING 2018**

3220180857 汪安平 计算机学院

**小错误:**

- 1 在论文的第一部分: 介绍中, 作者列出的关于本论文所做出的贡献方面, 有重复内容, 如第一点讲到论文解决预测标签的之间存在关联的问题, 第二点中, 再次说到不仅解决了第一点的问题, 还解决了某问题。第一点可以不要, 或者第二点可以不在重述第一点的会更加合适;
- 2 在论文 3.5 的结果部分, 其中关于 “Our proposed SGM model using global embedding achieves a reduction of 12.79% hamming-loss and an improvement of 2.33% micro-F1 score over the most commonly used baseline BR”, 这是两种实验结果的对比, 原表格中的计算结果与写在论文中的结果不符;
- 3 论文中采用了 2 部分的数据集, 对这两个部分的数据集解释十分清楚, 但是在对数据集进行训练集、验证集和测试集的划分上, 没有做任何说明, 甚至没有大概的比例说明, 这让想要做论文复现的学习者来说不友好。

**大错误:**

- 1 论文的摘要部分说 在对预测不同标签时, 文本不同部分对预测的贡献不同的这一想法, 在之前的工作结果中没有被考虑过, 这是不正确的, 在采用 RNN 的网络中, 因为 RNN 的特征, 不管文本各个部分的贡献问题有没有被提出, 实际上模型的原因, 都已经或多或少的被实践了的, 知识没有过多考虑到这一点。
- 2 论文的错误分析部分是存在较大问题的。该部分实际应该讲的是在实验过程中遇到的错误及对其解决分析过程, 而论文中该部分的内容讲解的是以往模型的不足和当前模型效果较好的对比, 这可以作为实验结果的一个对比。
- 3 论文的参考文献未使用尾注形式。从而导致论文中各处提及到的论文所对应的解决问题的思路和参考文献论文之间没有明确的对应关系。

## 阅读感想

本论文在前人在多分类任务的研究基础上，为了解决之前研究中的不足，主要是：

问题 1：各个类别之间是有关联关系的；

问题 2：文本的不同部分对预测分类的贡献不同，这是当前存在的模型没有考虑的问题。

提出了将多分类问题看作是序列生成模型，解码器结构的序列生成模型。其中值得注意的点是，该序列生成模型由编码器和含注意力机制的解码器组成。解码器使用 LSTM 来有序生成标签，在已知前面所有的预测的标签来预测当前标签，通过注意力机制的引入来解决问题 2。

相比较拥有注意力机制的 seq2seq 模型，本论文中的添加了全局嵌入的序列生成模型，在解码器结构中新加入了转化门，当前预测的  $y_t$  依赖于之前的预测所有输出  $y_1$  到  $y_{t-1}$ ，所以作者创新性的添加了转化门来衡量前面错误的预测对当前预测所造成的影响，被称作曝光误差。当前预测的状态获取在上一个隐藏状态和通过转化门的之前所有的输出  $y$  来决定，这也是解码模型的创新之处。

在验证在多分类问题在 encode-decode 模型中使用改进的使用了全局嵌入的解码器是否对模型的结果真正有改进方面，论文中作者通过改变转化门的参数，形成有无全局嵌入以及使用不同参数的全局嵌入的实验对比，通过结果的对比图表，再次证明退出这种创新模型的可行性以及进步性。给我们在今后的科学创新中，提供了验证自己创新结果的思路。

论文中不断验证模型在不同问题参数下的效果，例如为了验证所使用句子包含标签个数不同时，对模型预测结果的影响，作者使用将数据分为 标签序列长度不同的各个子集（在每个子集中，句子标签序列长度相等），然后在不同的测试集上面，验证模型效果，形成实验结果的对比。

该论文采用很大的数据集，通过充足的实验结果的论证，验证了采用序列生成模型并配合全局嵌入后，模型的效果相比较于之前多分类任务已有模型能取得更优的结果。为我们在解决多分类任务时，能采用该改进的 encode-decode 序列生成模型来获取更好的实验结果。