Large-Scale QA-SRL Parsing

QA-SRL：给定一个句子，对于句子中每个动词，提出若干个问题，每个问题的答案对应了一个semantic role。

数据标注分为两步：生成和验证。验证阶段标注者回答问题，若回答不了则标为无效

预处理：用CoreNLP标POS；用POS识别动词，用启发法过滤掉辅助动词，保留实义动词；

模型：

- Span detection：给定动词，从句子中选出一些span作为动词的参数；
- Question generation：对每个span预测出一个问题。
- 两部分都基于LSTM对句子编码: H_v；
- LSTM的输入是词向量+二元特征（表示这个词是不是当前要考虑的动词）；
- 两部分的LSTM参数互相独立。

Span Detection

- BIO
- Span-based：对句子中所有n^2个可能出现的span都预测它是不是动词的参数：对span(i,j)，将两端点位置的LSTM的输出向量连接$s_{vij} = [h_{vi}, h_{vj}]$得到的向量过MLP+全连接层+激活函数得到这个span是否为参数的概率
- Span-based效果更好

Question Generation

- 将问题划分为若干slot：Wh, Aux, Subj, Verb, Obj, Prep, Misc
- Local model：将span对应的s_{vij}向量过MLP+全连接层+softmax输出位置k的slot的概率分布（不同的k对应的权重参数不同，不同的slot之间互相独立）
- Sequence model：以slot为单位的LSTM，每个cell的输入是s_{vij}和前一个cell输出向量相连，输出过向量MLP+全连接层+激活函数+softmax得到slot的概率分布
- 结果：Sequence的exact match和partial match更高，local的Slot-level accuracy更高
- joint：span-based+seq效果好

数据扩展：

- 用模型对已有标注的句子生成问题，过滤掉重复的（答案和已标注答案重叠的，或问题和已标注问题一样的），剩下的问题就是标注时可能漏掉得的问题
- 在训练集上，把span detection的阈值调低再生成问题（为了得到更多潜在的问题）
- 用标数据的流水线来人工评测这些的问题，46017个（50%）被标注为有效
- 数据总量扩充20%
- 过滤掉扩展数据集中 答案与一个原始问题的答案有两处重叠的问题，过滤后的数据总量比原始扩充11.5%