DrQA

胡刚、梁达昌

答案生成

- 根据用户的提问生成答案，类似自动问答QA。
- 检索模块 Retriever: Tf-idf + hash 来对维基百科的数据进行检索。
- 答案生成模块 Reader：BiLSTM + attention 。

检索模块

- 维基百科爬取数据 整合成json
- 通过json文件存进数据库再导出来 得到一个db数据
- 通过db数据构建词-文档的tfidf矩阵，每一行代表一个词在各个文档的tfidf值
- 查询时只需把query分词，hash之后寻找对应的tfidf对应词的向量，并把他们相乘，遍历得到最大的即为相关度最高的文档。

生成模块

- Paragraph encoding
- Question encoding
- Prediction

Paragraph encoding

- bidirectional LSTM
- tokens ' features:
- Word embeddings
- Exact match
- Token features (POS,NER,TF)
- Aligned question embedding

$$a_{i,j} = \frac{\exp\left(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j))\right)}{\sum_{j'} \exp\left(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_{j'}))\right)}, \qquad f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j),$$

Question encoding

$$b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})},$$

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

建立在word embeding层之上，其中，w是待学习的参数，bj表示的是第j个词语的权重，qj表示的是question的第j个词。

Prediction

$$P_{start}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_s \mathbf{q})$$
$$P_{end}(i) \propto \exp(\mathbf{p}_i \mathbf{W}_e \mathbf{q})$$

单独训练2个分类器，分别预测答案开始的位置和结束的位置。Ws和We即是需要训练的矩阵。之后会选出长度小于15的p（start）*p(end) 的最大的起始和结束位置。

谢谢