阶段总结

张东旭

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

协助晓曦训练语言模型

- 目的：协助完成一些语言模型的训练，帮助我基本了解训练流程。

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

新词爬取与解析

- 目的：对词表进行定期更新
- 方法：
  - 1、爬取搜狗细胞词库的“网络新词”
  - 2、细胞词库解析成文本。
  - 3、将新词输入百度得到返回值，作为新词的权重
- 尚未解决的问题：
  - 1、这些新词只是作为候选词，并不能直接加入词表。需要在语料中计算词频、并结合百度返回结果，根据综合打分进行筛选。
  - 2、新词的实时性较强，如何删除词表中的新词也是一个问题。

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

v2.0词典

- 目的：改善1.0版本词典，使其更符合当前应用
- 方法：
  - 1、选用百度知道、百度hi、移动、四川移动、新浪微博语料进行预处理、分别统计词频。
  - 2、综合五个语料的大小和五个语言模型的插值系数得到权重，对词频进行加权求和。按照词频由高到低排序，取前20万作为候选词集合。
  - 3、爬取候选词集的百度返回值，按照返回值选取前15万个词作为最终的词。
  - 4、将返回值与词频信息加权求得最终每个词的权重。加权系数利用SGD拟合到v1.0的系数。
  - 5、英文词典根据词频简单排序得到，并最终人工过滤得到5000个词。

v2.0词典

- 不足：
  1、腾讯分词工具的粒度存在一定的问题。导致最终词典中地名人名不足。
  2、权重的计算方法仍然存在不合理性。

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

LSTM-RNN

- 目的：通过LSTM-RNN提高语言模型性能
- 方法：
  - 1、词典按照词频分类，去低频词。
  - 2、使用rwthlm工具训练语言模型
  - 3、使用n-best方法进行测试。
- 效果与尚未解决的问题：
  - 最终的WER较RNN有一定差距，并没有得到提高。
  - 1、工具不够成熟。2、仍需详细比较RNNlm和rwthlm的不同。

- 协助晓曦训练语言模型
- 新词爬取与解析
- v2.0词典
- LSTM-RNN
- 知识向量（Knowledge Vector）研究

知识向量（Knowledge Vector）研究

- 目的：找到一种将知识向量化的方法
- 方法：
  - 1、使用wikipedia的词条作为知识的载体
  - 2、利用wiki词条的结构层次，构建图结构
  - 3、利用图结构、词条所在页面的文本信息、连接信息，训练出词条的向量表示。
- 测试方法：
  - 对一些实体对的相关程度进行多人打分得到测试集。
  - 通过词条的向量计算距离，最终根据与打分序列的相关度评价系统性能。
- 改进方向：
  - Wiki的层次结构错综复杂，随着树规模的增大，会导致性能的下降。
  - 目标函数可能过于依赖树结构，存在一定的不合理性。

希望大家批评指正！

谢谢