CHIP2018

基于多视图学习的文本匹配

deadline

HITSZ

团队成员：

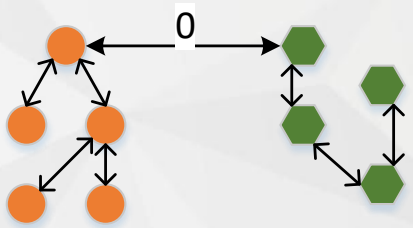
李心雨， 沈叶丹， 张伟林，陈红燕， 俞奕斐

CHIP2018

目录
contents

数据预处理

- 根据图特征进行数据扩展
  - 构造样本之间的连通图
  - 根据连通图构造出新的样本
    1）每个联通子图当中的样本之间两两构成正例
    2）若一个负例的两个句子分别出现在两个联通子图当中，则这两个联通子图当中的句子之间可以构成负例



- 平衡的数据扩展
  - 若一个句子只在正例（负例）中出现过，那么利用图特征为其生成负例（正例）
  - 对于无法生成正例的以自身作为相似问法生成正例

- 生成含有UNK的数据
  - 对于每条样本，用不存在于训练集中的字符随机替换其中的几个字/词，构造一倍的数据量，增强模型鲁棒性

模型简介


01
模型微调
最后一层拼接简单特征
学习率衰减
Embedding层优化
ESIM、BPool
BIMPM、DISAN
02
模型融合
字、词分别训练模型，二者投票
十折交叉
根据模型的线下效果求加权平均结果
字词信息的融合
多视图的匹配模型
03

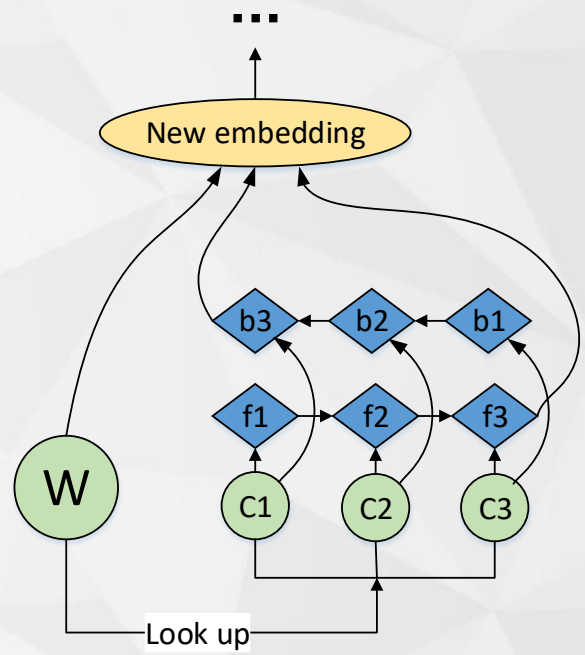
Embedding层的优化

CHIP2018

- 词向量的增量训练
  - 使用elmo训练语言模型，其中以给定的词向量作为初始输入
  - 最终保存训练后的词向量作为增量训练词向量

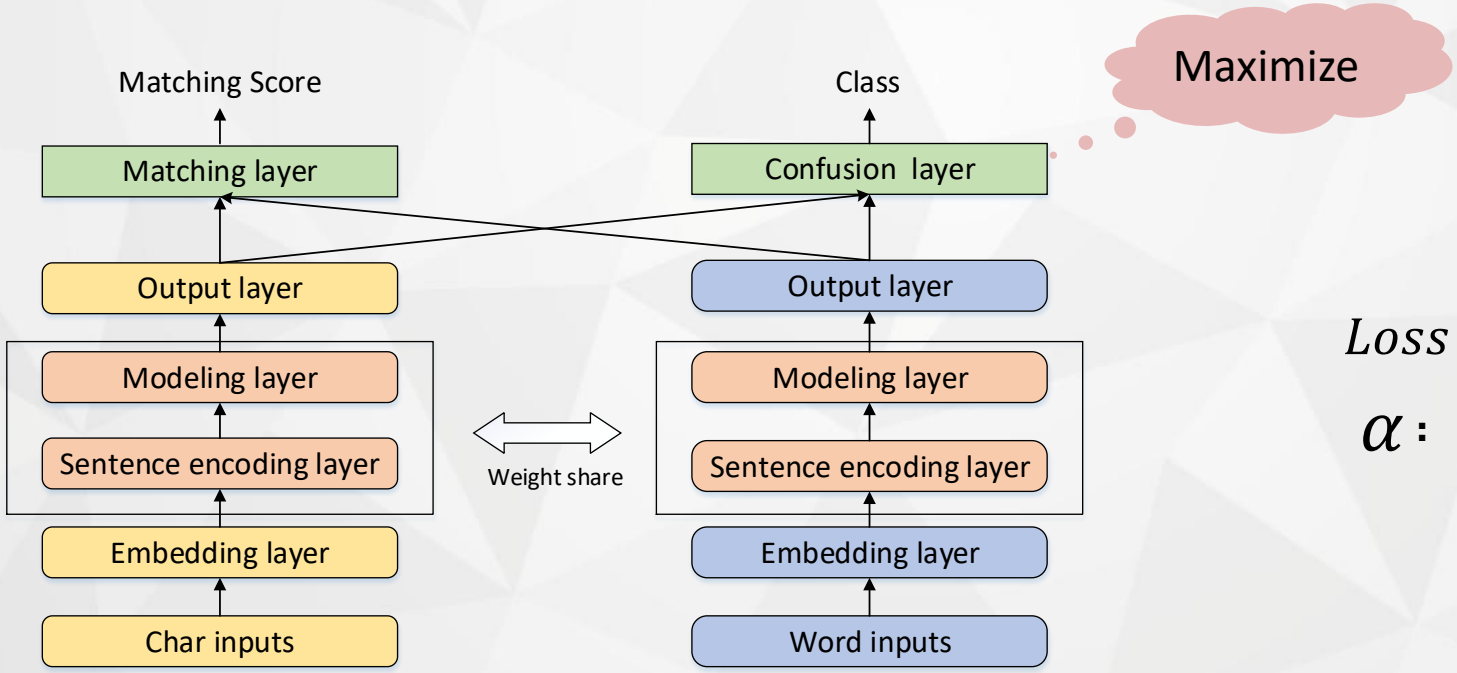
- Word-char融合表示
  - 根据数据的字表示与词表示，找出每个词所对应的字的组合
  - 将word对应的chars用BiLSTM表示并拼接在word的表示后面作为新的embedding

- 拼接可训练词向量
  - 对于每个字/词，在固定的词向量之后拼接一个较小维度的可训练词向量



多视图学习的匹配模型

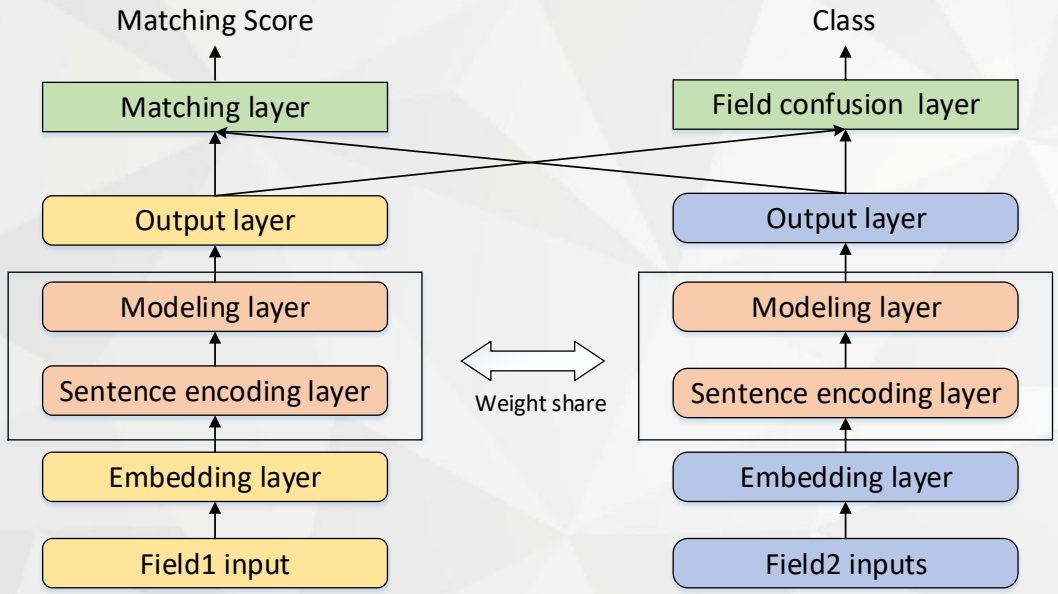
- 训练数据量较小，需要同时运用字、词信息
- 两个句子是否匹配只与两个句子之间的匹配关系相关，与视图本身无关
- 同时使用字、词两个视图当中的数据输入网络进行匹配任务参数学习



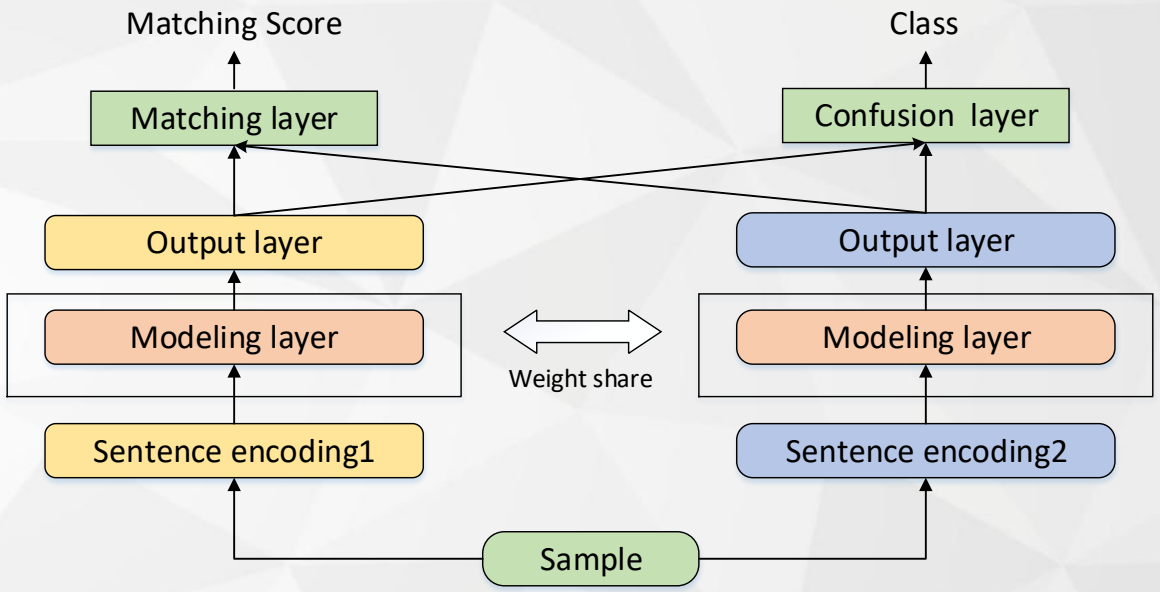
$$Loss = loss_M - \alpha loss_C$$

α：平衡因子

多视图学习的匹配模型



不同领域的文本输入



给相同的输入编码出不同的句子表示

谢谢您的观看！

汇报人：李心雨