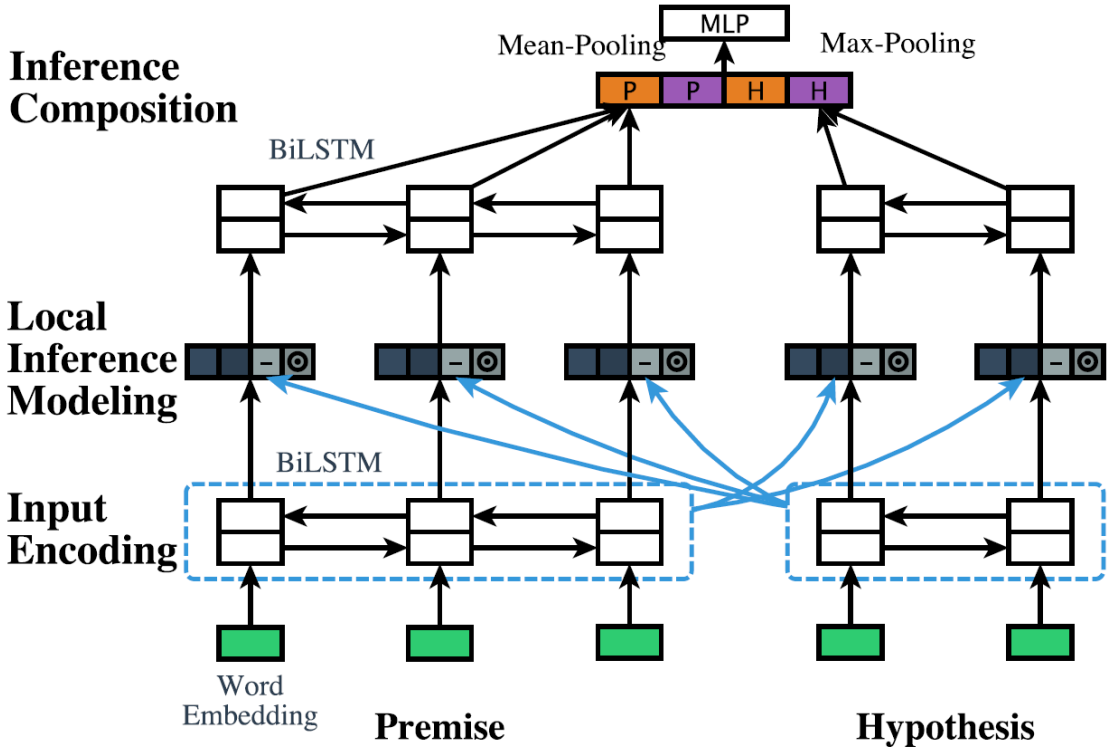
DOCUMENT-BASED CHINESE QA

ZHANG CHEN

方法：ESIM模型



- BiLSTM分别对问题和句子进行contextual encoding
- 用attention矩阵进行问题和句子的交互
- 将交互后的表示再输入到一层BiLSTM中
- 对问题和句子的表示进行average pooling和max pooling
- 将pooling结果输入到全连接中进行二分类

Xu et al. *Enhanced LSTM for natural language inference*. ACL2017

方法：输入

- 使用综合语料上预训练的中文词向量
  - https://github.com/Embedding/Chinese-Word-Vectors

- 尝试了**词向量**和**字向量**两种细粒度的表示
  - 字：以字符为单位分词，得到~8000个字
  - 词：使用jieba分词，得到>80000个词

方法：解决标签不平衡

- 数据集标签不平衡
  - 训练集94.8%标签为0，验证集95.68%标签为0
  - 使用Cross Entropy Loss训练不动，Accuracy稳定在“全预测为0”的水平

- 从Loss的角度平衡标签：使用Focal Loss
  - 对容易分类的负例进行down weighting，聚焦于数量少而难分类的正例
  - *alpha*为权重因子，我们取0.5； *gamma*为调节参数，我们取2

$$L_{fl} = -\alpha(1-\hat{y})^{\gamma} y \log \hat{y} - (1-\alpha)\hat{y}^{\gamma}(1-y)\log(1-\hat{y})$$

方法：尝试数据增强

- 使用CMRC2018阅读理解数据集尝试data augmentation
  - 文本来源和问题形式与我们的任务相近，猜测使用更多的数据能提升模型效果
  - 拆成answer sentence selection的形式加入到训练集，正例~10k，负例~100k
  - 尝试“只加正例”和“同时加正负例”两种形式

实验

	Precision	Recall	F1	MAP	MRR
All-0	0.00	0.00	0.00	25.30	25.81
BERT-base	77.99	89.33	83.27	**93.73**	**93.83**
ESIM (Word)	70.02	38.32	49.53	81.72	81.96
ESIM (Char)	67.64	77.57	72.27	**90.33**	**90.48**
ESIM (Char + DA Pos)	47.90	77.62	59.28	87.78	87.95
ESIM (Char + DA All)	67.25	75.59	71.18	89.18	89.34

- “全预测为0”baseline：MAP和MRR能达到25
- BERT baseline：使用谷歌预训练的中文bert-base和官方run_classifier.py脚本跑的，MAP和MRR达到93

实验

	Precision	Recall	F1	MAP	MRR
All-0	0.00	0.00	0.00	25.30	25.81
BERT-base	77.99	89.33	83.27	**93.73**	**93.83**
ESIM (Word)	70.02	38.32	49.53	81.72	81.96
ESIM (Char)	67.64	77.57	72.27	**90.33**	**90.48**
ESIM (Char + DA Pos)	47.90	77.62	59.28	87.78	87.95
ESIM (Char + DA All)	67.25	75.59	71.18	89.18	89.34

- 字向量 + ESIM + Focal Loss达到MAP 90.33 MRR 90.48，仅比BERT低3个点
  - 传统的序列模型也能接近预训练语言模型的表现
- 字向量表现远高于词向量
  - 原因：分词准确性的制约；字向量适合中文的word-level matching

实验

	Precision	Recall	F1	MAP	MRR
All-0	0.00	0.00	0.00	25.30	25.81
BERT-base	77.99	89.33	83.27	**93.73**	**93.83**
ESIM (Word)	70.02	38.32	49.53	81.72	81.96
ESIM (Char)	67.64	77.57	72.27	**90.33**	**90.48**
ESIM (Char + DA Pos)	47.90	77.62	59.28	87.78	87.95
ESIM (Char + DA All)	67.25	75.59	71.18	89.18	89.34

- 暴力Data Augmentation不可取
  - 两种方式表现均降
  - 存在数据分布的gap，需要更精细的数据增强方法

谢谢!