CCL2018
中文文本蕴含识别评测
总结报告

北京语言大学
于东 金天华 刘春花
2018.10

提纲

- 评测问题
- 数据准备
- 基线系统
- 评测结果
- 展望

评测问题描述

- 文本蕴含关系是自然语言中广泛存在的语义关系。

- 评测将中文文本蕴含识别看作一个分类问题：每个输入样本为2个句子，分别是“前提句Premise”和“假设句Hypothesis”，要求参评系统判断两者之间的蕴含类别，包括：
  - 蕴含(E, entailment)：P能合理推理得到H,反之不一定；
  - 矛盾(C，contradiction)：P和H不能同时成立；
  - 无关(N, neutral)： P和H不存在必然的联系。

- 采用准确率作为评价标准

评测数据集

- 来自SNLI、MultiNLI两个数据集，
  - 保持原有数据集中各个类别的比例，
  - SNLI：78259句，MultiNLI：31741句
- 人工转译和标注：
  - 长句和过短的句子，直接由人工翻译
  - 词数为7~10词的句子，先经过机器翻译，再人工整理
  - 转译过程会产生语义误差，约有1.24%的句子，转译后蕴含标签改变

	train	dev	test
蕴含	29738	3486	3476
矛盾	28937	3416	3343
中立	31325	3098	3183
总计	90000	10000	10000

特殊情况

- 特殊的情况的例子和标签：
  - 重述认为是蕴含
  - P：登山者正在攀登一座山崖。 H：一名登山者正在攀登山峰
- 有关代词的界定，认为P和H中的指代是一致的。
  - P：一个穿着大衣的小女孩。
  - H：她有一件外套。
- 删除提到图片的数据
  - P：一个年轻的女人在沙滩上带着许多彩色围巾。
  - H：这张照片里有一个女人，她在外面。

基线系统

ESIM 模型流程图

Output Layer	Avg&Max Pooling -> Softmax
Aggregation Layer	BiLSTM
Matching Layer	Heuristic matching
Attention Layer	Dot product attention
Encoding Layer	BiLSTM
Embedding Layer	SGNS embedding

拼接
逐元素点乘
逐元素减

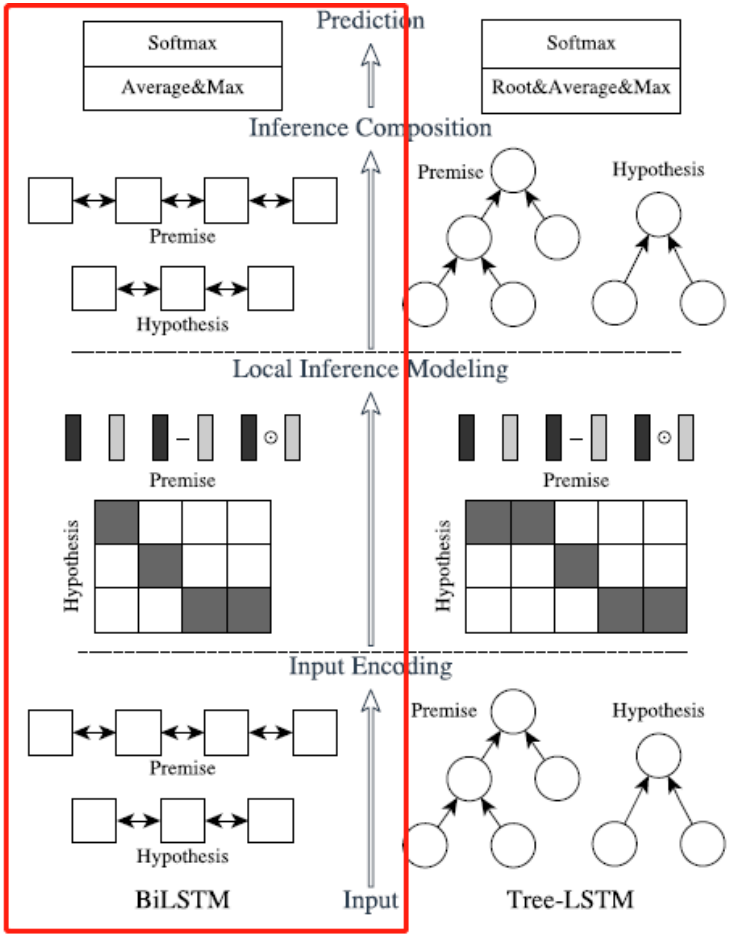


Figure 1: A high-level view of our hybrid neural inference networks.

Chen, Qian et al. "Enhanced LSTM for Natural Language Inference." *ACL* (2017). https://arxiv.org/pdf/1609.06038.pdf

基线系统

Decomposable Attention 模型流程图

Output Layer	Sum Pooling -> Softmax
Aggregation Layer	2层FNN
Matching Layer	Matching → 拼接
Attention Layer	Dot product attention
Encoding Layer	2层FNN
Embedding Layer	SGNS embedding

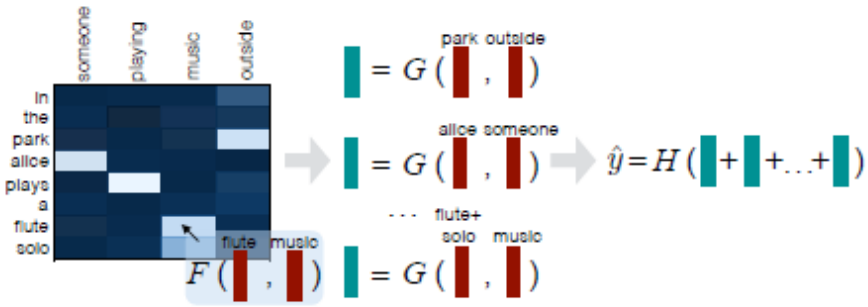


Figure 1: Pictoral overview of the approach, showing the *Attend* (left), *Compare* (center) and *Aggregate* (right) steps.

Parikh, Ankur P. et al. "A Decomposable Attention Model for Natural Language Inference." *EMNLP* (2016). https://arxiv.org/pdf/1606.01933.pdf

评测过程

- 2018.4~2018.6 数据集建设
- 2018.6~2018.8 发布baseline和train、dev数据
- 2018.9.10~9.17 发布test数据，评测

- 共53支队伍报名
- 11支队伍提交了有效结果

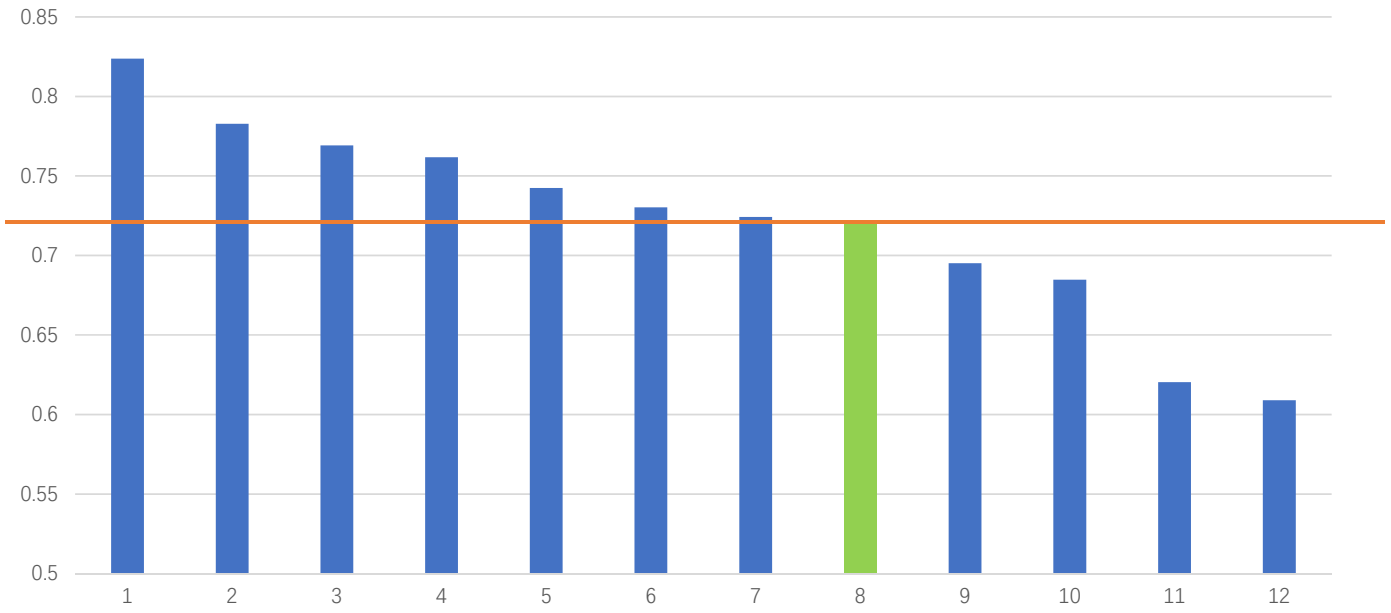
评测结果分析

- 总成绩排序

Results				
#	User	Entries	Date of Last Entry	accuracy ▲
1	water123	3	09/17/18	0.8238 (1)
2	nlpc	11	09/17/18	0.7828 (2)
3	ShawnNg	46	09/16/18	0.7692 (3)
4	Kunxun_Qi	57	09/16/18	0.7618 (4)
5	ray_li	1	09/12/18	0.7425 (5)
6	eedanny	6	09/10/18	0.7303 (6)
7	Parkhaeju	20	09/17/18	0.7242 (7)
8	BLCU-nlp	10	09/16/18	0.7222 (8)
9	firend2	8	09/17/18	0.6952 (9)
10	__503	1	09/17/18	0.6848 (10)
11	lyb3b	6	09/13/18	0.6203 (11)
12	oliver_arrow	5	09/17/18	0.6090 (12)

评测结果分析

- 超过baseline的比例：7/11， 63.6%
- 提交系统报告： 6/11， 54.5%
- 柱状图分析：



评测方法分析

team	embed	encode	attention	match	prediction
1	word word2vec	CNN	no	融合	FNN
2	word word2vec	FNN	plain attention	注意力向量拼接	FNN
3	word+char word2vec OOV embed	BiLSTM	co-attention self-attention	多个特征融合后拼接	FNN
4	ELMo	BiLSTM	co-attention self-attention	多个特征融合后拼接	pooling 后FNN
5	char word2vec	Siamese BiLSTM	decomposable attention	注意力向量拼接	FNN
6	word+char word2vec	BiLSTM+CNN	plain attention	注意力向量拼接	FNN

获奖情况

奖项	队伍名称	单位
一等奖	water	北京拓尔思信息技术股份有限公司，北京信息科技大学
二等奖	zzunlp2018	郑州大学
	zhizhu	吴晓晖，尹存祥，骆金昌，钟辉强
	GDUFSER	广东外语外贸大学
三等奖	狂奔	朱洪银
	严明，张真练	武汉科技大学
	李永彬	遵义医学院医学信息工程学院

展望

- 文本蕴含、语言推断，将成为NLP的基础问题
- 发展方向一：文本蕴含知识是什么？在哪？
- 发展方向二：语言推断有哪些类型？哪些种类？
- 发展方向三：可解释的语言推断？

相关资源

- Codalab在线评测网址：

https://competitions.codalab.org/competitions/19911

- Github代码和数据网址：

https://github.com/blcunlp/CNLI