

NLPCC2021: FewCLUE

预训练模型的中文小样本学习评测

队伍：姜汁可乐 北京邮电大学

王泽元 魏志宇

2021-07-04



目录

01

赛题介绍

02

技术方案

03

总结





赛题介绍





预训练模型的中文小样本学习

- 目标
结合预训练语言模型通用和强大的泛化能力基础上，探索小样本学习最佳模型和中文上的实践。评测包括单、双句分类和阅读理解等多个任务。
- 利用预训练模型的泛化能力
 - gpt 3 : 在zero-shot下取得不错的结果，在很少样本训练实现媲美sota的结果；（模型很大，很难部署）
 - pet / entailment : 通过将nlp任务转化为和预训练任务相似；
- 基于元学习的小样本学习
 - 测试集为少样本任务；训练集包含较多与测试集相似的数据，通过元学习获得更加通用的数据表示

Corpus	Train	Dev	Test Public	cat num	Task	Source
	Single Sentence Tasks					
EPRSTMT	32	32	610	2	SntmntAnalysis	E-CommrceReview
CSLDCP	536	536	1784	67	LongTextClassify	AcademicCNKI
TNEWS	240	240	2010	15	ShortTextClassify	NewsTitle
IFLYTEK	928	690	1749	119	LongTextClassify	AppDesc

	Sentence Pair Tasks					
OCNLI	32	32	2520	3	NLI	5Genres
BUSTM	32	32	1772	2	SemanticSmlarty	AlVirtualAssistant
	Reading Comprhnsn Task					
CHID	42	42	2002	7	MultipleChoice,idiom	Novel,EssayNews
CSL	32	32	2828	2	KeywordRecogntn	AcademicCNKI
CLUEWSC	32	32	976	2	CorefResolution	ChineseFictionBooks

赛道：模型限制赛道，不能ensemble，预训练模型layer<=24层....



技术方案





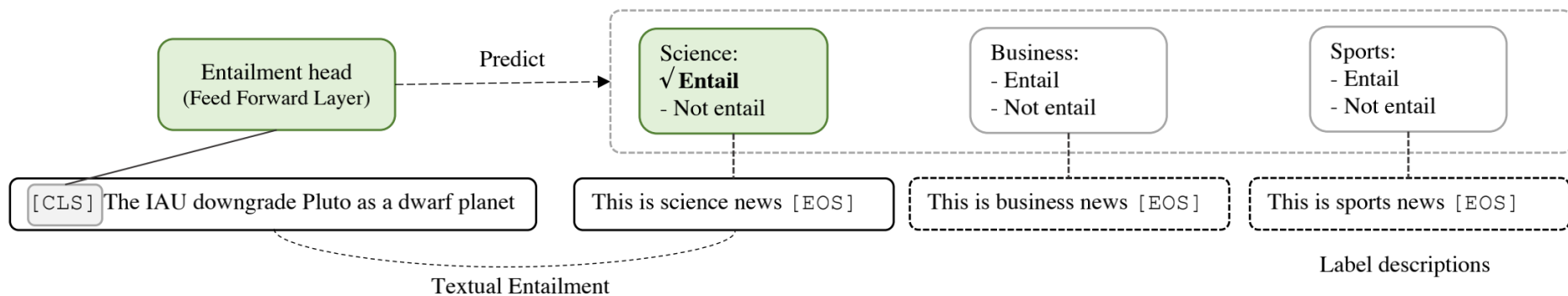
EFL: Entailment as Few-Shot Learner (Facebook)

➤ 主要思路

文本分类可以被修正为一个蕴含任务（判断两句话的关系）；

$x_{in} = [\text{CLS}] S_1 [\text{SEP}] S_2 [\text{EOS}]$, where $S_2 = \text{This indicates positive user sentiment}$,

情感分类



(c) Entailment-based method (our approach). Each class has a label description, and we choose the class with maximum probability of entailment between input sentence and label description.

➤ Benefits

利用公开的蕴含任务的数据集（如CMNLI）进行中间训练；

主要流程

1. Intermediate training with entailment task with [cmnli data](#)
2. Finetune with few-shot entailment samples.



任务模板

任务名	模板
EPRSTMT	这表达了xxxx的情感
CSLDCP	这篇论文阐述了xxxx
TNEWS	下面报道一条xxxx新闻
IFLYTEK	这段文本的应用描述主题是xxxx
OCNLI	None
BUSTM	None
CHID	#idiom#位置的成语是xxxx
CSL	xxxx为真实关键词
CLUEWSC	pronoun是指noun

	score	eprstmt	bustm	ocnli	csldcp	tnews	wsc	ifytek	csl	chid
EFL (1:8)	53.4	85.6	67.6	67.5	46.7	53.5	54.2	44.0	61.6	28.2
PET	57.36	87.2	64.0	43.9	56.9	53.7	59.2	35.1	55.0	61.3
EFL (1:1)	50.3	79.1	67.6	67.5	26.0	50.2	51.2	41.3	51.6	18.4

chid-task

➤ 问题

据俄罗斯卫星通讯社3月15日报道,新八国联军#idiom#逼近附近海域,但是军舰却遭岸舰导弹锁定,英承认今非昔比。最近一段时间,北约多个国家开始频繁进行军事演习,来对其他国家进行威慑。3月12日当天,英国出动了兰开斯特号、威斯敏斯特号..."

➤ original template

#idiom#位置的成语是气势汹汹

➤ new template

sentence1: "据俄罗斯卫星通讯社3月15日报道,新八国联军

sentence2: 气势汹汹逼近附近海域,但是军舰却遭岸舰导弹锁定,英承认今非昔比。最近一段时间,北约多个国家开始频繁进行军事演习,来对其他国家进行威慑。3月12日当天,英国出动了兰开斯特号、威斯敏斯特号...

28% ———> 48%

无效尝试: 使用成语解释, 使用特殊token标志成语;

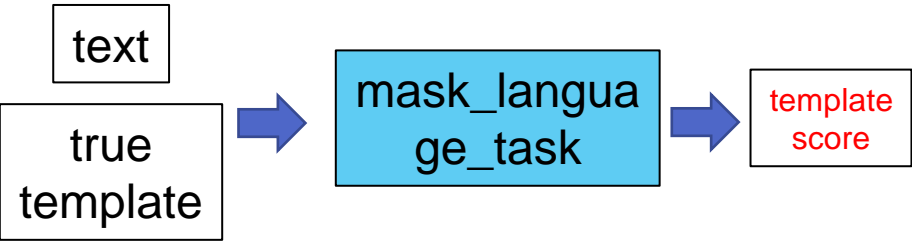
EFL：模版的探索

模版的顺序

模版在后：
模版在前：

EPRSTMT	CSLDCP	TNEWSF	IFLYTEKF
83.45(1.3,84.3)	48.86(1.17,48.2)	70.07(2.73,68.9)	41.75(1.19,49.1)
84.46(1.3,84.6)	47.41(0.78,37.8)	69.84(2.7,49.5)	41.92(1.02,49.3)

模版的选择



eprstmt	lm_score	acc
这表达了"+value+"的情感	2.65	85.05
评论表达了"+value+"的情感	2.3	0.8704
这是"+value+"的情感	2.91	84.6

任务名	模板（修改）
EPRSTMT	评论表达了xxxx的情感
CSLDCP	这篇论文是关于xxxx
TNEWS	这条新闻是关于xxxx
IFLYTEK	文本描述的应用是关于xxxx



未采用的方案

- Bootstrap (利用无标注数据)
 - 参数敏感, 调参代价大
在1000条无标注数据, 每次抽取置信度top5%, 抽取2次, 在test_public能带来3%的提升; (EPSRMT)
在10000条无标注数据, 每次抽取置信度top5%, 抽取2次, 在test_public基本没有提升; (EPSRMT)
- focal loss (针对类别不平衡的问题)
 - 无明显提升
 - 发现模型输出概率会平滑很多 (0.99, 0.01) => (0.9, 0.1)
- 复用分类器 (针对模型参数增加的问题)
 - 无明显提升, 甚至在chid任务上导致不拟合;
- 通过在少样本上多次dropout, 并用loss平均值优化模型;
 - 无明显提升;
- 通过融入masked language的任务
 - 无明显提升;

上下句关系



是否替换



[CLS] 这个手机太棒了[SEP]这[是]积极的情感[SEP]



尚未尝试的方案

- 多个模版融合
 - 利用动态路由、GNN融合句子和多个模板的表示;
- 利用对比学习优化模型表示, 然后再拟合mlp;



03

总结



- Entailment的特点
 - 将分类任务转换为蕴含任务，该任务为句子粒度；
 - 句子粒度的研究会更加成熟一些，算法落地的优化空间更大；
 - 在多类别分类任务上，因为负采样会导致信息的缺失，并且推理时间会随类别数增加；
 - 适应于类别少，短句的分类场景；
 - 最简单直接的收益为模板的修改或组合；
 - 少样本任务的参数较为敏感；

Score	EPRSTMT	CSLDCP	TNEWSF	IFLYTEKF	OCNLIF	BUSTM	CHIDF	CSLF	CLUEWSCF
60.024	84.2(2.17,83.1)	48.46(1.33,52.9)	68.97(2.19,74.7)	41.35(4.38,13.4)	66.23(1.41,69.4)	72.67(1.57,78)	50.43(4.2,64.8)	60.23(1.37,61.9)	56.55(9.19,65.9)

感谢聆听

北京邮电大学

魏志宇 王泽元