组会

曾双

2020.8.2

Content

- One ACL2020 best paper
- One ACL2020 honorable mention paper
- SciERC dataset (multi-task setup for NER, RE, CR)
- NYT SOTA paper
- ICML 2020 flooding loss

ACL 2020 Best Paper

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro[1] **Tongshuang Wu**[2] **Carlos Guestrin**[2] **Sameer Singh**[3]
[1]Microsoft Research [2]University of Washington [3]University of California, Irvine
marcotcr@gmail.com {wtshuang,guestrin}@cs.uw.edu sameer@uci.edu

Motivation

- 计算测试集(held-out)上的accuracy是现在评价泛化性能的主流方法，但通常都高估了模型的表现
  - 标准做法：Train-Validation-Test split / Leadboard
- 原因：
  - 测试集并不全面，无法覆盖现实生活中所有的情况
  - 测试集可能包含和训练集相同的bias
  - 将模型性能用一个数字来表示，很难去发现模型不会做什么，也很难想到要怎么去解决

- 本文使用类似于软件测试中的行为（黑盒）测试的评价方法，针对模型设计CheckList（类似于OJ的测试用例），无需知道模型的内部结构，就能知道模型会什么不会什么（类似于找bug，然后debug）

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** — Labels: negative, positive, neutral			
`Template: I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** — Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos → neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral → neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** — Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg → neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg → neutral	X
...			
Failure rate = 34.6%			

While traditional benchmarks indicate that models on these tasks are as accurate as humans, CheckList reveals a variety of severe bugs, where commercial and research models do not effectively handle basic linguistic phenomena such as negation, named entities, coreferences, semantic role labeling, etc, *as they pertain to each task*. Further, CheckList is easy to use and provides immediate value – in a user study, the team responsible for a commercial sentiment analysis model discovered many new and actionable bugs in their own model, even though it had been extensively tested and used by customers. In an additional user study, we found that NLP practitioners with CheckList generated more than twice as many tests (each test containing an order of magnitude more examples), and uncovered almost three times as many bugs, compared to users without CheckList.

ACL2020 Honorable Mention Paper

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan[†] **Ana Marasović**[†◇] **Swabha Swayamdipta**[†]
Kyle Lo[†] **Iz Beltagy**[†] **Doug Downey**[†] **Noah A. Smith**[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

Motivation

- Language models pre-trained on text from a wide variety of sources form the foundation of today's NLP.
- We investigate whether it is still helpful to tailor a pre-trained model to the domain of a target task.

⇨ Multiphase Adaptive Pre-Training
on 4 domains and 8 classification task

Conclusion

- Domain-Adaptive Pre-Training leads to performance gains, under both high- and low-resource settings.
- Adapting to the task's unlabeled data (Task-Adaptive Pre-Training) improves performance even after domain-adaptive pre-training.

Domain	Pretraining Corpus	# Tokens	Size	$\mathcal{L}_{\text{RoB.}}$	$\mathcal{L}_{\text{DAPT}}$
BioMed	2.68M full-text papers from S2ORC (Lo et al., 2020)	7.55B	47GB	1.32	0.99
CS	2.22M full-text papers from S2ORC (Lo et al., 2020)	8.10B	48GB	1.63	1.34
News	11.90M articles from RealNews (Zellers et al., 2019)	6.66B	39GB	1.08	1.16
Reviews	24.75M Amazon reviews (He and McAuley, 2016)	2.11B	11GB	2.10	1.93
RoBERTa (baseline)	see Appendix §A.1	N/A	160GB	‡1.19	-

Table 1: List of the domain-specific unlabeled datasets. In columns 5 and 6, we report RoBERTa's masked LM loss on 50K randomly sampled held-out documents from each domain before ($\mathcal{L}_{\text{RoB.}}$) and after ($\mathcal{L}_{\text{DAPT}}$) DAPT (lower implies a better fit on the sample). ‡ indicates that the masked LM loss is estimated on data sampled from sources *similar* to RoBERTa's pretraining corpus.

Figure 2: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to RoBERTa's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Domain	Task	Label Type	Train (Lab.)	Train (Unl.)	Dev.	Test	Classes
BioMed	ChemProt	relation classification	4169	-	2427	3469	13
	†RCT	abstract sent. roles	18040	-	30212	30135	5
CS	ACL-ARC	citation intent	1688	-	114	139	6
	SciERC	relation classification	3219	-	455	974	7
News	HyperPartisan	partisanship	515	5000	65	65	2
	†AGNews	topic	115000	-	5000	7600	4
Reviews	†Helpfulness	review helpfulness	115251	-	5000	25000	2
	†IMDB	review sentiment	20000	50000	5000	25000	2

Table 2: Specifications of the various target task datasets. † indicates high-resource settings. Sources: ChemProt (Kringelum et al., 2016), RCT (Dernoncourt and Lee, 2017), ACL-ARC (Jurgens et al., 2018), SciERC (Luan et al., 2018), HyperPartisan (Kiesel et al., 2019), AGNews (Zhang et al., 2015), Helpfulness (McAuley et al., 2015), IMDB (Maas et al., 2011).

Dom.	Task	RoBa.	DAPT	¬DAPT
BM	ChemProt	$81.9_{1.0}$	$\mathbf{84.2}_{0.2}$	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	$\mathbf{87.6}_{0.1}$	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$\mathbf{75.4}_{2.5}$	$66.4_{4.1}$
	SciERC	$77.3_{1.9}$	$\mathbf{80.8}_{1.5}$	$79.2_{0.9}$
News	HyP.	$86.6_{0.9}$	$\mathbf{88.2}_{5.9}$	$76.4_{4.9}$
	†AGNews	$\mathbf{93.9}_{0.2}$	$\mathbf{93.9}_{0.2}$	$93.5_{0.2}$
Rev.	†Helpful.	$65.1_{3.4}$	$\mathbf{66.5}_{1.4}$	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	$\mathbf{95.4}_{0.2}$	$94.1_{0.4}$

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Table 5: Results on different phases of adaptive pretraining compared to the baseline RoBERTa (col. 1). Our approaches are DAPT (col. 2, §3), TAPT (col. 3, §4), and a combination of both (col. 4). Reported results follow the same format as Table 3. State-of-the-art results we can compare to: ChemProt (84.6), RCT (92.9), ACL-ARC (71.0), SciERC (81.8), HyperPartisan (94.8), AGNews (95.5), IMDB (96.2); references in §A.2.

BioMed	RCT	ChemProt
TAPT	$87.7_{0.1}$	$82.6_{0.5}$
Transfer-TAPT	$87.1_{0.4}$ (↓0.6)	$80.4_{0.6}$ (↓2.2)

CS	ACL-ARC	SciERC
TAPT	$67.4_{1.8}$	$79.3_{1.5}$
Transfer-TAPT	$64.1_{2.7}$ (↓3.3)	$79.1_{2.5}$ (↓0.2)

News	HyperPartisan	AGNews
TAPT	$89.9_{9.5}$	$94.5_{0.1}$
Transfer-TAPT	$82.2_{7.7}$ (↓7.7)	$93.9_{0.2}$ (↓0.6)

Reviews	Helpfulness	IMDB
TAPT	$68.5_{1.9}$	$95.7_{0.1}$
Transfer-TAPT	$65.0_{2.6}$ (↓3.5)	$95.0_{0.1}$ (↓0.7)

Table 6: Though TAPT is effective (Table 5), it is harmful when applied *across* tasks. These findings illustrate differences in task distributions within a domain.

Honorable Mention Papers – Main Conference

Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey and Noah A. Smith

Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Nitika Mathur, Timothy Baldwin and Trevor Cohn

To summarise, our key recommendations are:

- When evaluating metrics, use the technique outlined in Section 4.2 to remove outliers before computing Pearson's r.
- When evaluating MT systems, stop using BLEU or TER for evaluation of MT, and instead use CHRF, YISI-1, or ESIM;
- Stop using small changes in evaluation metrics as the sole basis to draw important empirical conclusions, and make sure these are supported by manual evaluation.

EMNLP2018

Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction

Yi Luan   **Luheng He**   **Mari Ostendorf**   **Hannaneh Hajishirzi**
University of Washington
{luanyi, luheng, ostendor, hannaneh}@uw.edu

Motivation

- We introduce a multi-task setup of identifying and classifying entities, relations, and coreference clusters in scientific articles.
- We create SCIERC, a dataset that includes annotations for all three tasks and develop a unified framework called Scientific Information Extractor (SCIIE) for with shared span representations.
- The multi-task setup reduces cascading errors between tasks and leverages cross-sentence relations through coreference links.

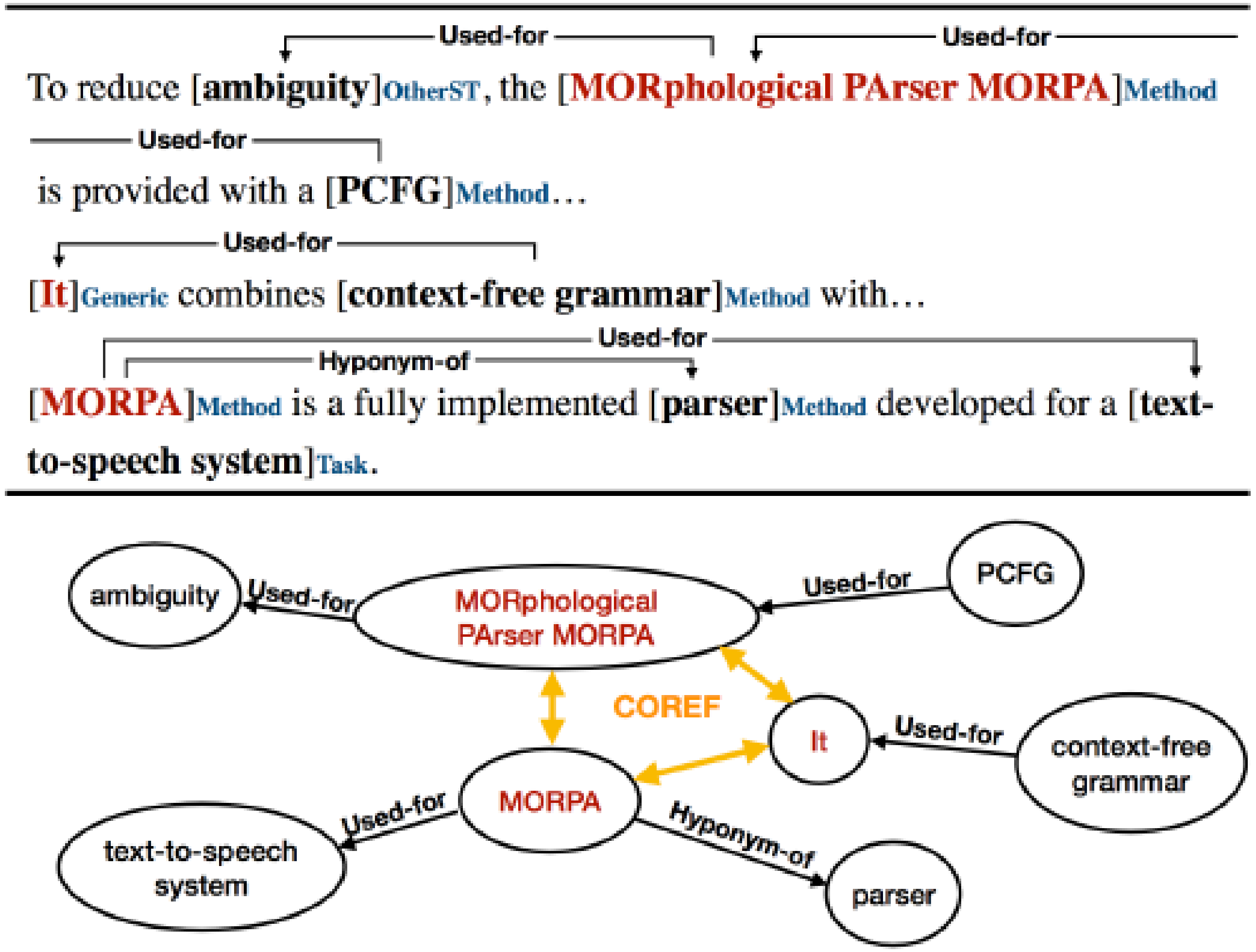


Figure 1: Example annotation: phrases that refer to the same scientific concept are annotated into the same coreference cluster, such as *MORphological PAser MORPA*, *it* and *MORPA* (marked as red).

Statistics	SciERC	SemEval 17	SemEval 18
#Entities	8089	9946	7483
#Relations	4716	672	1595
#Relations/Doc	9.4	1.3	3.2
#Coref links	2752	-	-
#Coref clusters	1023	-	-

Table 1: Dataset statistics for our dataset SciERC and two previous datasets on scientific information extraction. All datasets annotate 500 documents.

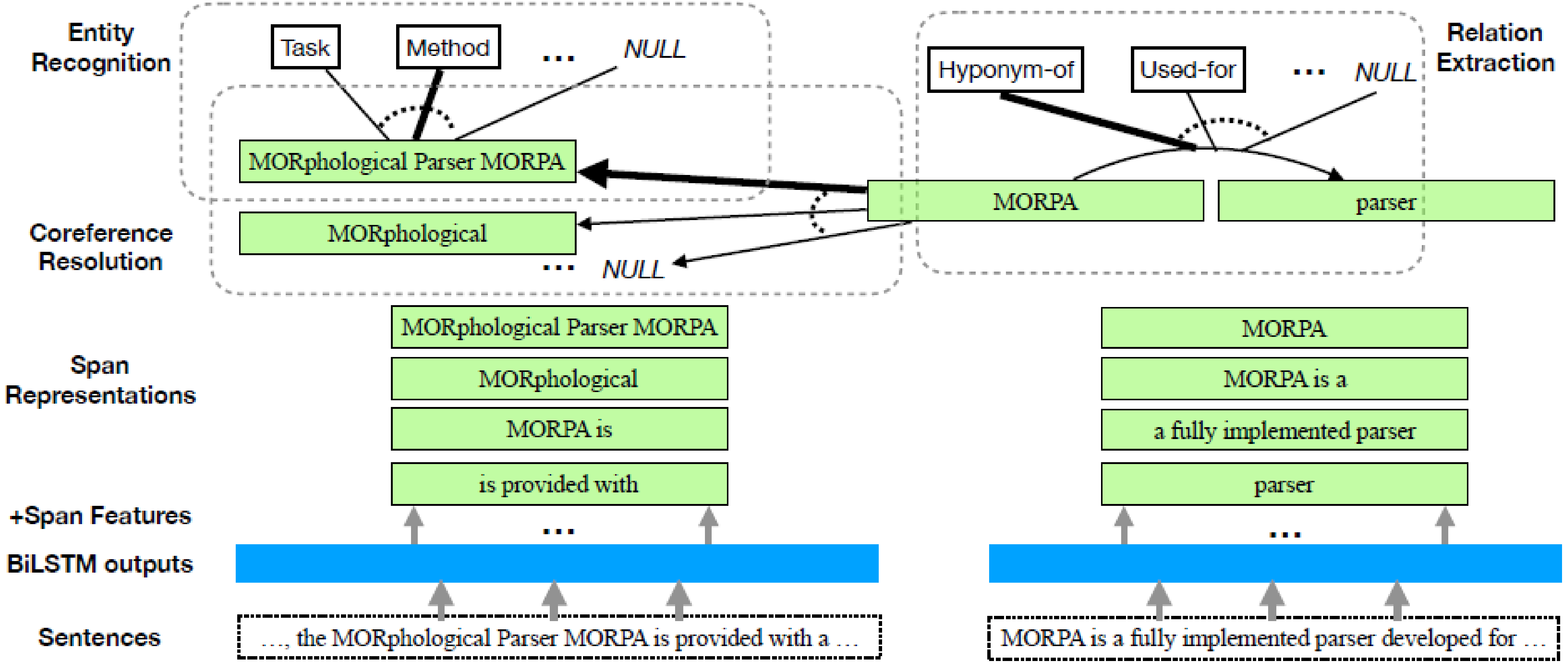


Figure 2: Overview of the multitask setup, where all three tasks are treated as classification problems on top of shared span representations. Dotted arcs indicate the normalization space for each task.

Model	Dev P	Dev R	Dev F1	Test P	Test R	Test F1
LSTM+CRF	67.2	65.8	66.5	62.9	61.1	62.0
LSTM+CRF+ELMo	68.1	66.3	67.2	63.8	63.2	63.5
E2E Rel(Pipeline)	66.7	65.9	66.3	60.8	61.2	61.0
E2E Rel	64.3	68.6	66.4	60.6	61.9	61.2
E2E Rel+ELMo	67.5	66.3	66.9	63.5	63.9	63.7
SciIE	70.0	66.3	**68.1**	67.2	61.5	**64.2**

(a) Entity recognition.

Model	Dev P	Dev R	Dev F1	Test P	Test R	Test F1
E2E Rel(Pipeline)	34.2	33.7	33.9	37.8	34.2	35.9
E2E Rel	37.3	33.5	35.3	37.1	32.2	34.1
E2E Rel+ELMo	38.5	36.4	37.4	38.4	34.9	36.6
SciIE	45.4	34.9	**39.5**	47.6	33.5	**39.3**

(b) Relation extraction.

Model	Dev P	Dev R	Dev F1	Test P	Test R	Test F1
E2E Coref	59.4	52.0	55.4	60.9	37.3	46.2
SciIE	61.5	54.8	**58.0**	52.0	44.9	**48.2**

(c) Coreference resolution.

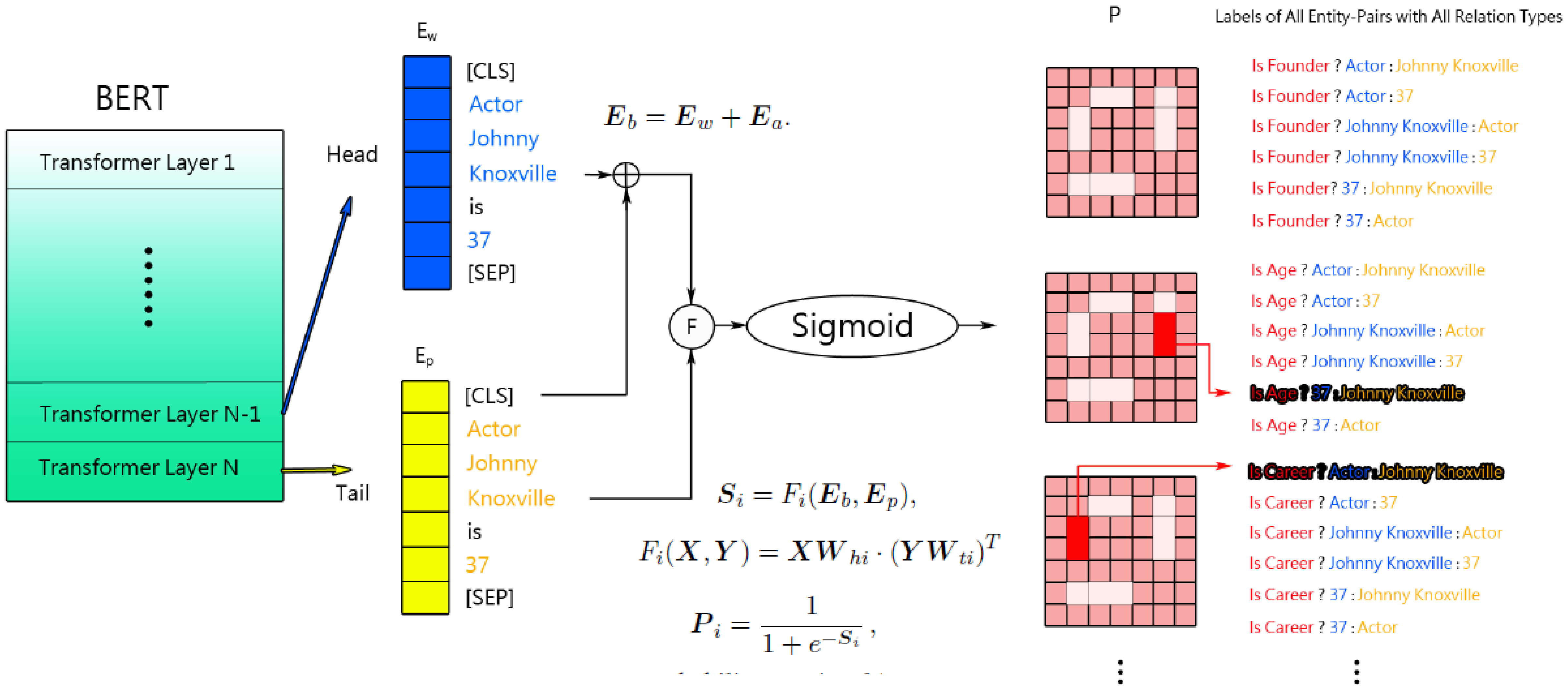
Task	Entity Rec.	Relation	Coref.
Multi Task (SciIE)	68.1	39.5	58.0
Single Task	65.7	37.9	55.3
+Entity Rec.	-	38.9	57.1
+Relation	66.8	-	57.6
+Coreference	67.5	39.5	-

Table 3: Ablation study for multitask learning on SciERC development set. Each column shows results for the target task.

Arxiv 202004

Downstream Model Design of Pre-trained Language Model for Relation Extraction Task

Cheng Li, Ye Tian
AI Application Research Center, Huawei Technologies, Shenzhen, China
licheng81@huawei.com


BERT
Transformer Layer 1
Transformer Layer N-1
Transformer Layer N
Head
Tail
E_w
[CLS]
Actor
Johnny
Knoxville
is
37
[SEP]
E_p
[CLS]
Actor
Johnny
Knoxville
is
37
[SEP]
$E_b = E_w + E_a.$
F
Sigmoid
$S_i = F_i(E_b, E_p),$
$F_i(X, Y) = XW_{hi} \cdot (YW_{ti})^T$
$P_i = \frac{1}{1 + e^{-S_i}},$
P
Labels of All Entity-Pairs with All Relation Types
Is Founder ? Actor : Johnny Knoxville
Is Founder ? Actor : 37
Is Founder ? Johnny Knoxville : Actor
Is Founder ? Johnny Knoxville : 37
Is Founder? 37 : Johnny Knoxville
Is Founder ? 37 : Actor
Is Age ? Actor : Johnny Knoxville
Is Age ? Actor : 37
Is Age ? Johnny Knoxville : Actor
Is Age ? Johnny Knoxville : 37
Is Age ? 37 Johnny Knoxville
Is Age ? 37 : Actor
Is Career ? Actor Johnny Knoxville
Is Career ? Actor : 37
Is Career ? Johnny Knoxville : Actor
Is Career ? Johnny Knoxville : 37
Is Career ? 37 : Johnny Knoxville
Is Career ? 37 : Actor

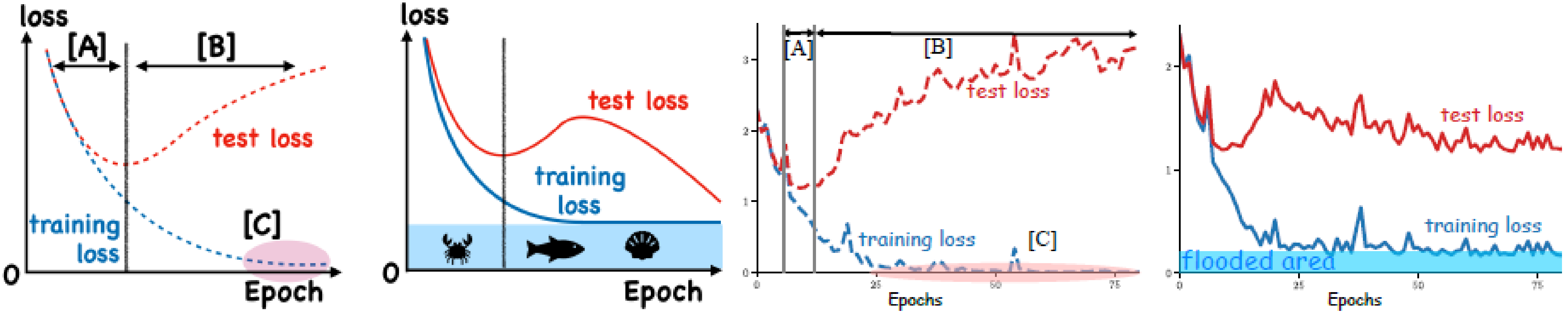
Methods	**SemEval**	**NYT**	**WebNLG**
C-AGGCN [4]	85.7	–	–
GraphRel2p [2]	–	61.9	42.9
BERT_{EM}-MTB [25]	89.5	–	–
HBT [27]	–	87.5	88.8
ours	**91.0**	**89.8**	**96.3**

ICML 2020

Do We Need Zero Training Loss After Achieving Zero Training Error?

Takashi Ishida[1,2] Ikko Yamane[1] Tomoya Sakai[3]
Gang Niu[2] Masashi Sugiyama[2,1]

[1]The University of Tokyo [2]RIKEN [3]NEC Corporation



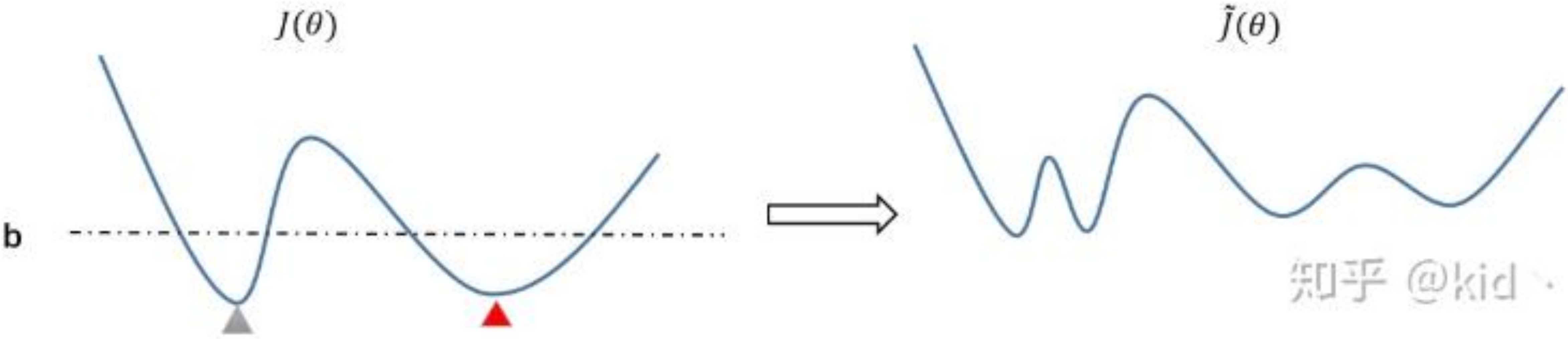
(a) w/o Flooding **(b)** w/ Flooding **(c)** CIFAR-10 w/o Flooding **(d)** CIFAR-10 w/ Flooding

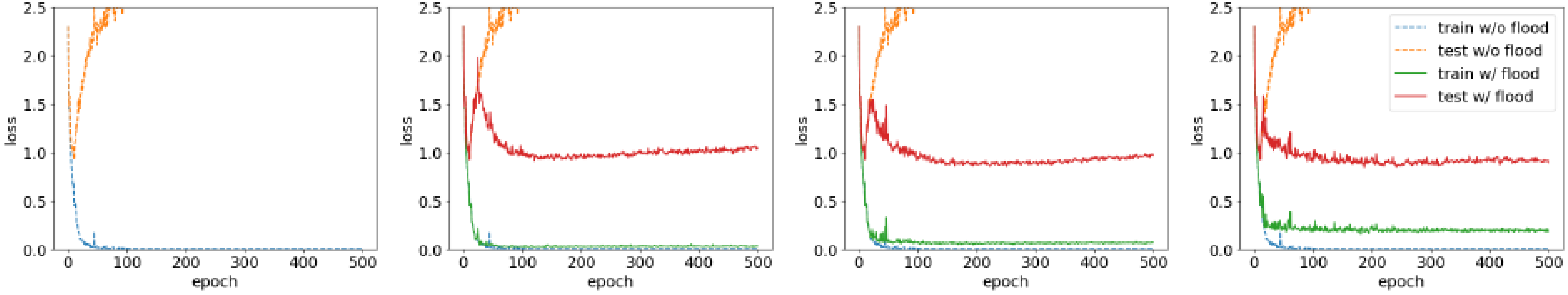
Figure 1: (a) shows 3 different concepts related to overfitting. [A] shows the generalization gap increases, while training & test losses decrease. [B] also shows the increasing gap, but the test loss starts to rise. [C] shows the training loss becoming (near-)zero. We avoid [C] by *flooding* the bottom area, visualized in (b), which forces the training loss to stay around a constant. This leads to a decreasing test loss once again. We confirm these claims in experiments with CIFAR-10 shown in (c)–(d).

$$\widetilde{J}(\theta) = |J(\theta) - b| + b,$$

$$\widetilde{J}(\boldsymbol{\theta}) = |J(\boldsymbol{\theta}) - b| + b,$$

Gravity V.S. buoyancy





(a) CIFAR-10 (0.00) **(b)** CIFAR-10 (0.03) **(c)** CIFAR-10 (0.07) **(d)** CIFAR-10 (0.20)

Figure 2: Learning curves of training and test loss for training/validation proportion of 0.8. (a) shows the learning curves without flooding. (b), (c), and (d) show the learning curves with different flooding levels.

Thanks!