自然语言处理算法鲁棒性研究思考

张奇

复旦大学

0 各类自然语言处理算法快速发展，在很多任务上甚至超越人类

SQuAD2.0
The Stanford Question Answering Dataset

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
3 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799

Dynabench: Rethinking Benchmarking in NLP

0 各类自然语言处理算法快速发展，在很多任务上甚至超越人类

CLUE1.0分类任务排行榜

CLUE1.1/1.0提交规则 | 项目地址

CLUE1.1与CLUE1.0区别：区别与原有的CLUE1.0，CLUE1.1在部分任务启用了新的测试集，训练集和验证集保持不变；CLUE1.0保留CMNLI自然语言推理任务

模型

排行	模型	研究机构	测评时间	Score1.0	认证	AFQMC	TNEWS1.0	IFLYTEK	CMNLI	OCNLI_50K	WSC1.0	CSL
1	TI-NLP	优图实验室 & 腾讯云	21-10-19	83.251	待认证	82.7	79.3	65.23	84.31	84.57	96.55	90.1
2	ShenZhou	QQ浏览器实验室(QQ Brow...	21-09-19	83.247	待认证	80.55	74.15	67.65	86.49	86.37	96.55	90.97
3	HUMAN	CLUE	19-12-01	82.943	已认证	81	71	80.3	76	90.3	98	84
4	Mengzi	澜舟科技-创新工场	21-09-14	82.436	待认证	81.79	75.06	65.08	86.13	82.57	96.55	89.87
5	BERTSG	Sogou Search	21-06-25	81.991	待认证	79.85	74.15	64.54	85.3	85.93	95.17	89
6	Motian	QQ浏览器搜索	21-06-25	81.764	待认证	78.3	73.18	65.46	85.44	84.97	94.83	90.17
7	Pangu	华为云-循环智能	21-04-23	81.016	待认证	78.11	72.07	65.19	85.19	83.3	95.52	87.73
8	PLUG	Alibaba DAMO NLP	21-04-18	80.614	待认证	77.44	73.06	64	84.95	83.27	94.48	87.1
9	Bert	lihaiyu	21-04-08	79.663	待认证	75.6	70.32	64.92	84.55	81.73	93.45	87.07
10	MT-BERTs	Meituan NLP	21-03-10	79.624	待认证	77.36	70.03	64.31	85.14	83.47	89.66	87.4

自然语言处理真的被解决了吗？

0 算法在实际应用中的效果却不尽如人意

搜索引擎线上，精度95%条件下召回率小于20%

能够回答的部分绝大多数都是原文匹配类型

0 算法在实际应用中的效果却不尽如人意

0 不经过鲁棒性评估会面临巨大风险

对话系统答非所问

潜在政治风险

非常不好的用户体验

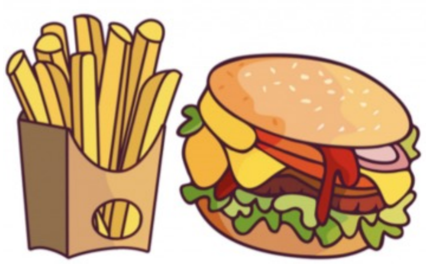
自然语言处理仍然面临很多问题

0 模型对测试数据的微小变化非常敏感

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% **World**
South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism. 95% **Sci/Tech**
Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives. 75% **World**
Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives. 94% **Business**

Ebrahimi et al., *HotFlip: White-Box Adversarial Examples for Text Classification*, 2018.

0 模型对测试数据的微小变化非常敏感



Sentiment Analysis Data

Tasty burgers, and crispy fries.

burgers 😃 fries 😃 SA 😃

? **Model predicts 😃 for burgers, is it due to *tasty*, *crispy*, or even other clues?**

SubQ.	Generation Strategy	Example
Prereq.	SOURCE: The original sample from the test set	Tasty **burgers**, and crispy fries. (Tgt: burgers)
Q1	REVTGT: Reverse the sentiment of the *target* aspect	Terrible **burgers**, but crispy fries.
Q2	REVNON: Reverse the sentiment of the *non-target* aspects with originally the same sentiment as target	Tasty **burgers**, but soggy fries.
Q3	ADDDIFF: Add aspects with the *opposite* sentiment from the target aspect	Tasty **burgers**, crispy fries, but poorest service ever!

Xing et al., *Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis*, EMNLP 2020

0 模型对测试数据的微小变化非常敏感

Model	Entire Test Ori → New (Change)	REVTGT Subset Ori → New (Change)	REVNON Subset Ori → New (Change)	ADDDIFF Subset Ori → New (Change)
Laptop Dataset				
MemNet	64.42 → 16.93 (↓47.49)*	72.10 → 28.33 (↓43.77)*	82.22 → 79.26 (↓02.96)	64.42 → 56.58 (↓07.84)*
GatedCNN	65.67 → 10.34 (↓55.33)*	75.11 → 24.03 (↓51.08)*	83.70 → 78.52 (↓05.18)	65.67 → 45.14 (↓20.53)*
AttLSTM	67.55 → 09.87 (↓57.68)*	72.96 → 27.04 (↓45.92)*	85.93 → 75.56 (↓10.37)*	67.55 → 39.66 (↓27.89)*
TD-LSTM	68.03 → 22.57 (↓45.46)*	73.39 → 29.83 (↓43.56)*	83.70 → 77.04 (↓06.66)	68.03 → 60.66 (↓07.37)*
GCN	72.41 → 19.91 (↓52.50)*	78.33 → 35.62 (↓42.71)*	88.89 → 74.81 (↓14.08)*	72.41 → 52.51 (↓19.90)*
BERT-Sent	73.04 → 17.40 (↓55.64)*	78.76 → 59.44 (↓19.32)*	88.15 → 42.22 (↓45.93)*	73.04 → 34.64 (↓38.40)*
CapsBERT	77.12 → 25.86[6] (↓51.26)*	80.69 → 57.73 (↓22.96)*	88.89 → 49.63 (↓39.26)*	77.12 → 45.14 (↓31.98)*
BERT	77.59 → 50.94 (↓26.65)*	83.05 → 65.02 (↓18.03)*	93.33 → 71.85 (↓21.48)*	77.59 → 71.00 (↓06.59)*
BERT-PT	78.53 → 53.29 (↓25.24)*	82.40 → 60.09 (↓22.31)*	93.33 → 83.70 (↓09.63)*	78.53 → 75.71 (↓02.82)
Average	71.60 → 25.23 (↓46.37)*	77.42 → 43.01 (↓34.41)*	87.57 → 70.29 (↓17.28)*	71.60 → 53.45 (↓18.15)*
Restaurant Dataset				
MemNet	75.18 → 21.52 (↓53.66)*	80.73 → 27.54 (↓53.19)*	84.46 → 73.65 (↓10.81)*	75.18 → 60.71 (↓14.47)*
GatedCNN	76.96 → 13.12 (↓63.84)*	85.11 → 23.17 (↓61.94)*	88.06 → 72.97 (↓15.09)*	76.96 → 54.91 (↓22.05)*
AttLSTM	75.98 → 14.64 (↓61.34)*	82.98 → 28.96 (↓54.02)*	86.26 → 61.26 (↓25.00)*	75.98 → 52.32 (↓23.66)*
TD-LSTM	78.12 → 30.18 (↓47.94)*	85.34 → 34.99 (↓50.35)*	88.51 → 75.68 (↓12.83)*	78.12 → 70.18 (↓07.94)*
GCN	77.86 → 24.73 (↓53.13)*	86.76 → 35.58 (↓51.18)*	88.51 → 79.50 (↓09.01)*	77.86 → 65.00 (↓12.86)*
BERT-Sent	80.62 → 10.89 (↓69.73)*	89.60 → 44.80 (↓44.80)*	89.86 → 57.21 (↓32.65)*	80.62 → 30.89 (↓49.73)*
CapsBERT	83.48 → 55.36 (↓28.12)*	89.48 → 71.87 (↓17.61)*	90.99 → 74.55 (↓16.44)*	83.48 → 77.86 (↓05.62)*
BERT	83.04 → 54.82 (↓28.22)*	90.07 → 63.00 (↓27.07)*	91.44 → 83.33 (↓08.11)*	83.04 → 79.20 (↓03.84)*
BERT-PT	86.70 → 59.29 (↓27.41)*	92.20 → 72.81 (↓19.39)*	92.57 → 81.76 (↓10.81)*	86.70 → 80.27 (↓06.43)*
Average	79.77 → 31.62 (↓48.15)*	86.92 → 44.75 (↓42.17)*	88.96 → 73.32 (↓15.64)*	79.77 → 63.48 (↓16.29)*

Xing et al., *Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis*, EMNLP 2020

0 为什么会这样？

问题1：为什么基于基准测试集合和常用评价指标的模式不能反映上述问题？

问题2：深度神经网络模型到底学习到了什么？

问题3：现阶段自然语言处理算法鲁棒性究竟怎么样?

为什么会这样？

问题1：为什么基于基准测试集合和常用评价指标的模式不能反映上述问题？

问题2：深度神经网络模型到底学习到了什么？

问题3：现阶段自然语言处理算法鲁棒性究竟怎么样?

1 数据集上存在偏置 – WINOGRANDE

AAAI 2020 Best Paper WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale

Winograd Schema Challenge (WSC) **Commonsense reasoning**

The trophy doesn’t fit into the brown suitcase because **it**’s too large. **trophy** / suitcase
The trophy doesn’t fit into the brown suitcase because **it**’s too small. trophy / **suitcase**

RoBERTa large achieves **91.3%** accuracy on a variant of WSC dataset

Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense?

Dataset-specific Biases

Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.

1 数据集上存在偏置 – WINOGRANDE

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

Table 1: WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices. The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)-(3) are drawn from WSC (Levesque, Davis, and Morgenstern 2011) and (4) from DPR (Rahman and Ng 2012)). Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable since the word "predators" is more often associated with the word "lions", compared to "zebras"

Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.

1 数据集上存在偏置 – WINOGRANDE

Instead of manually identified lexical features, they adopt a dense representation of instances using their precomputed neural network embeddings.

Main Steps:

1. RoBERTa fine-tuned on a small subset of the dataset.
2. An ensemble of linear classifiers (logistic regressions)
3. Trained on random subsets of the data
4. Determine whether the representation is strongly indicative of the correct answer option
5. Discard the corresponding instances

Algorithm 1: AFLITE

Input: dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, ensemble size n, training set size m, cutoff size k, filtering threshold τ
Output: dataset $\mathcal{D}'$

$\mathcal{D}' = \mathcal{D}$
while $|\mathcal{D}'| > m$ **do**
    `// Filtering phase`
    forall $e \in \mathcal{D}'$ **do**
        Initialize the ensemble predictions $E(e) = \emptyset$
    for *iteration* $i : 1..n$ **do**
        Random partition $(\mathcal{T}_i, \mathcal{V}_i)$ of $\mathcal{D}'$ s.t. $|\mathcal{T}_i| = m$
        Train a linear classifier $\mathcal{L}$ on $\mathcal{T}_i$
        forall $e = (\mathbf{x}, y) \in \mathcal{V}_i$ **do**
            Add $\mathcal{L}(\mathbf{x})$ to $E(e)$
    forall $e = (\mathbf{x}, y) \in \mathcal{D}'$ **do**
        $score(e) = \frac{|\{p \in E(e) \text{ s.t. } p = y\}|}{|E(e)|}$
    Select the top-k elements $\mathcal{S}$ in $\mathcal{D}'$ s.t. $score(e) \geq \tau$
    $\mathcal{D}' = \mathcal{D}' \setminus \mathcal{S}$
    if $|\mathcal{S}| < k$ **then**
        break
return $\mathcal{D}'$

Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.

1 数据集上存在偏置 – WINOGRANDE

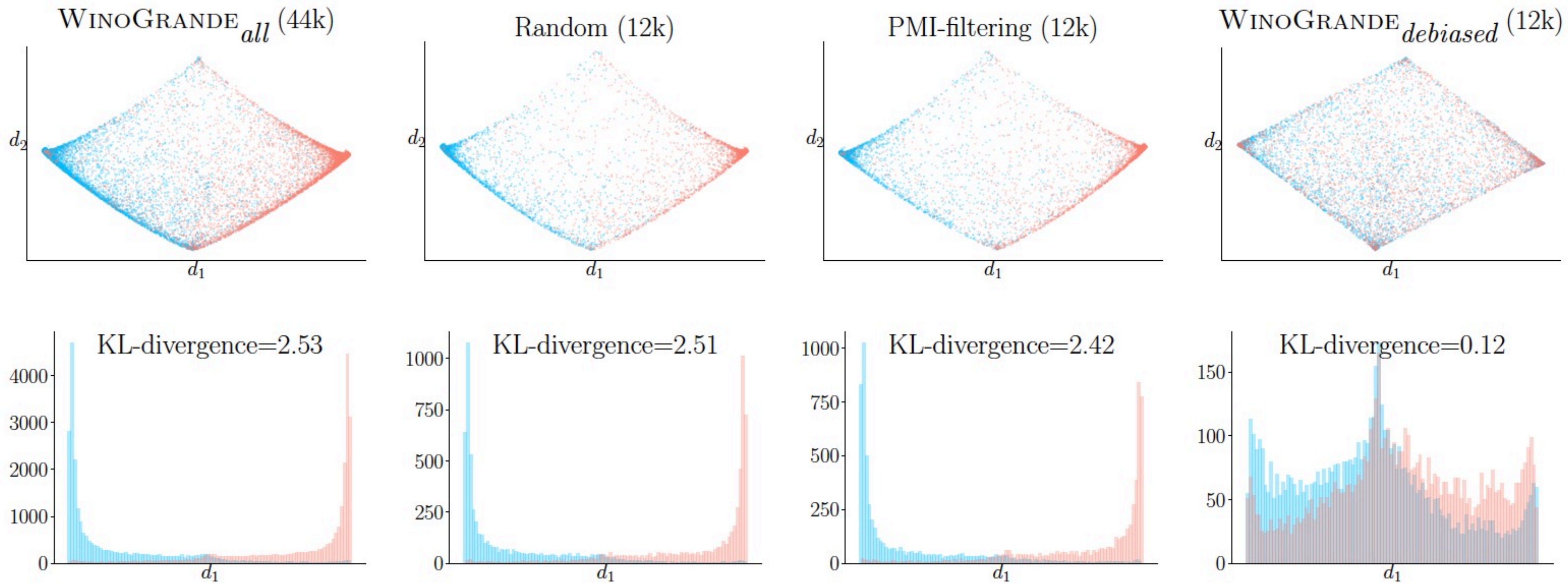


Figure 1: The effect of debiasing by AFLite. RoBERTa pre-computed embeddings (applied PCA for dimension reduction) are shown in two-dimensional space (*top row*) and histograms regarding d_1 (*bottom row*) with the bin size being 100. Data points are colored depending on the label (i.e., the answer y is option 1 (blue) or 2 (red)). In the histograms, we show the KL-divergence between $p(d_1, y{=}1)$ and $q(d_1, y{=}2)$.

Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.

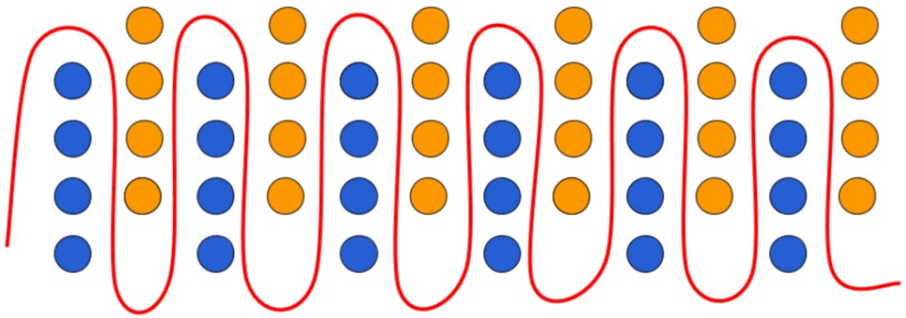
1 数据集上存在偏置 – WINOGRANDE

Methods	dev acc. (%)	test acc.(%)
WKH	49.4	49.6
Ensemble LMs	53.0	50.9
BERT	65.8	64.9
RoBERTa	**79.3**	**79.1**
BERT (local context)	52.5	51.9
RoBERTa (local context)	52.1	50.0
BERT-DPR⋆	50.2	51.0
RoBERTa-DPR⋆	59.4	58.9
Human Perf.	94.1	94.0

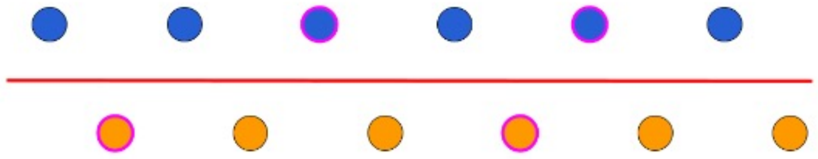
Table 3: Performance of several baseline systems on WINOGRANDE$_{debiased}$ (dev and test). The star (⋆) denotes that it is zero-shot setting (e.g., BERT-DPR⋆ is a BERT model fine-tuned with the DPR dataset and evaluated on WINOGRANDE$_{debiased}$.)

Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.

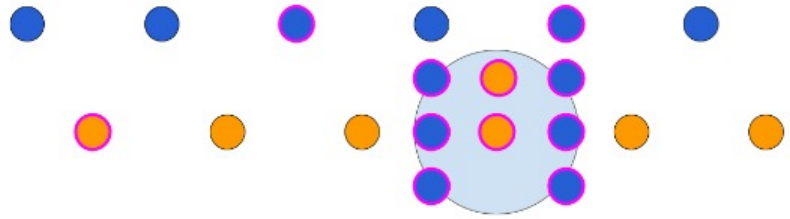
1 数据集采样对模型训练和测试重要影响 – Contrast Sets



(a) A two-dimensional dataset that requires a complex decision boundary to achieve high accuracy.



(b) If the same data distribution is instead sampled with systematic gaps (e.g., due to annotator bias), a simple decision boundary can perform well on i.i.d. test data (shown outlined in pink).



(c) Since filling in all gaps in the distribution is infeasible, a contrast set instead fills in a local ball around a test instance to evaluate the model's decision boundary

Gardner et al., *Evaluating Models' Local Decision Boundaries via Contrast Sets*, EMNLP 2020

1 数据集采样对模型训练和测试重要影响 – Contrast Sets

更严格的自然语言处理任务数据集合构建规范

The dataset authors **manually perturb** the test instances in small but meaningful ways that (typically) change the gold label, creating ***contrast sets***.

Dataset	Original Instance	Contrastive Instance (color = edit)
IMDb	Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. . . . The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience. *(Label: Negative)*	Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** . . . The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience. *(Label: Positive)*
MATRES	Colonel Collins followed a normal progression once she was picked as a NASA astronaut. *("picked" was before "followed")*	Colonel Collins followed a normal progression **before** she was picked as a NASA astronaut. *("picked" was after "followed")*
UD English	They demanded talks with local US commanders. I attach a paper on gas storage value modeling. I need to get a job at the earliest opportunity.	They demanded talks with **great urgency**. I attach a paper on **my own initiative**. I need to get a job at **House of Pies**.

Gardner et al., *Evaluating Models' Local Decision Boundaries via Contrast Sets*, EMNLP 2020

1 数据集采样对模型训练和测试重要影响 – Contrast Sets

Dataset	# Examples	# Sets	Model	Original Test	Contrast		Consistency
NLVR2	994	479	LXMERT	76.4	61.1	(–15.3)	30.1
IMDb	488	488	BERT	93.8	84.2	(–9.6)	77.8
MATRES	401	239	CogCompTime2.0	73.2	63.3	(–9.9)	40.6
UD English	150	150	Biaffine + ELMo	64.7	46.0	(–18.7)	17.3
PERSPECTRUM	217	217	RoBERTa	90.3	85.7	(–4.6)	78.8
DROP	947	623	MTMSN	79.9	54.2	(–25.7)	39.0
QUOREF	700	415	XLNet-QA	70.5	55.4	(–15.1)	29.9
ROPES	974	974	RoBERTa	47.7	32.5	(–15.2)	17.6
BoolQ	339	70	RoBERTa	86.1	71.1	(–15.0)	59.0
MC-TACO	646	646	RoBERTa	38.0	14.0	(–24.0)	8.0

Gardner et al., *Evaluating Models' Local Decision Boundaries via Contrast Sets*, EMNLP 2020

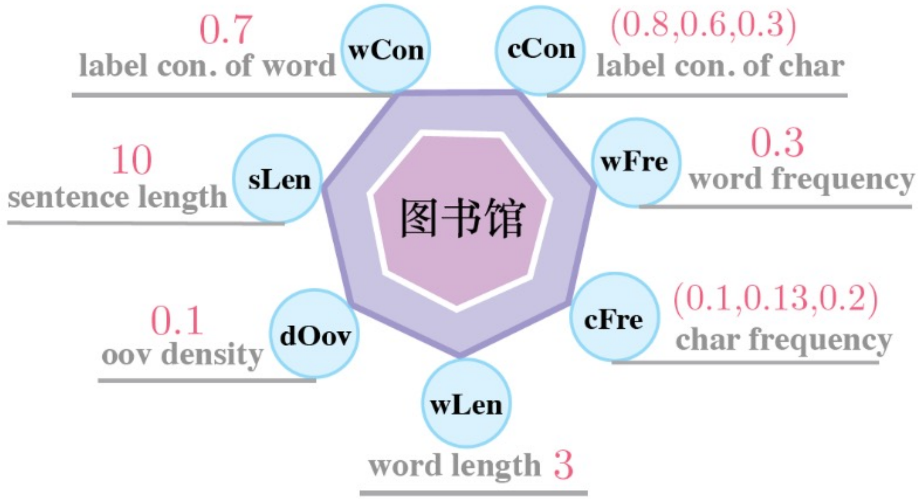
1 细粒度评测 - Is Chinese Word Segmentation a Solved Task?

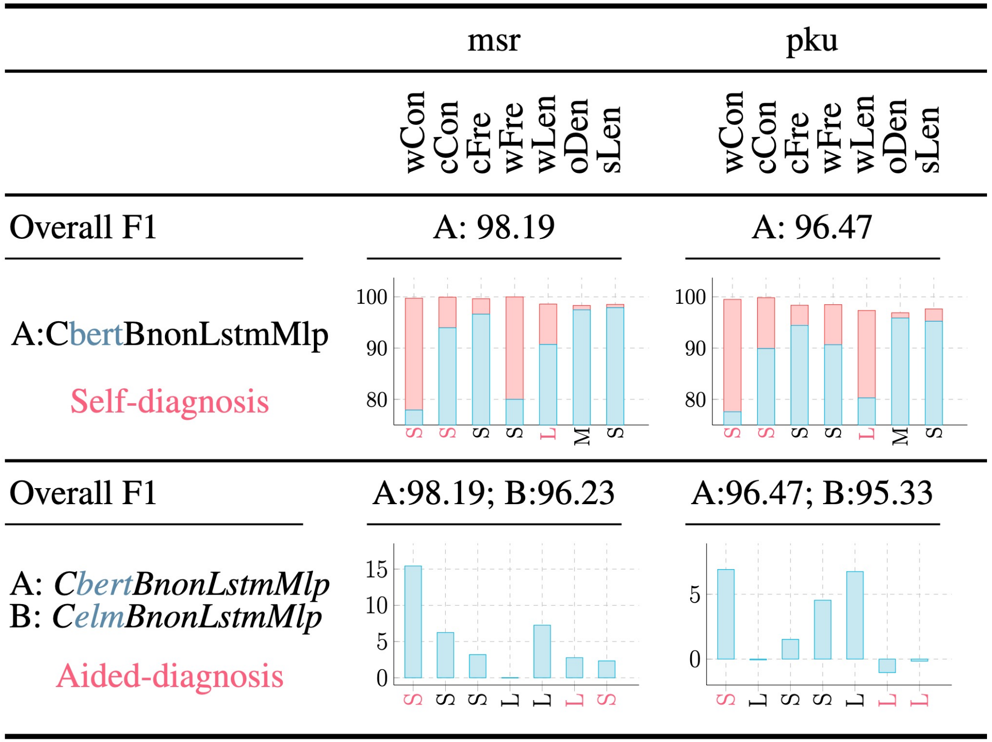
Model	Character				Bigram			SenEnc.		Dec.		Holistic Evaluation (Overall F1)						
	rand	w2v	elmo	bert	none	avg	w2v	lstm	cnn	crf	mlp	msr	pku	ctb	ckip	cityu	ncc	sxu
CrandBavgLstmCrf	√					√		√		√		96.21	**94.22**	**95.32**	**92.81**	93.54	92.01	94.87
Cw2vBavgLstmCrf		√				√		√		√		96.46	94.10	95.08	92.81	93.67	92.04	94.71
Cw2vBavgLstmMlp		√				√		√			√	96.41	92.74	94.09	91.40	93.25	92.00	93.16
Cw2vBavgCnnCrf		√				√			√	√		96.48	93.99	94.72	92.73	**93.72**	**92.64**	94.36
Cw2vBw2vLstmCrf		√					√	√		√		**96.66**	94.19	95.14	92.46	93.70	92.24	**94.97**
CelmBnonLstmMlp			√		√			√			√	96.23	95.33	96.77	94.83	96.44	93.21	96.47
CbertBnonLstmMlp				√	√			√			√	98.19	96.47	**97.68**	**96.23**	97.09	95.77	97.49
CbertBw2vLstmMlp		√		√			√	√			√	**98.20**	96.52	97.65	96.18	97.07	**95.78**	**97.51**
Huang et al. (2019)												97.90	**96.60**	97.60	—	**97.60**	—	97.30

Table 2: Neural CWS systems with different architectures and pre-trained knowledge studied in this paper. We exclude systems based on joint training to make a fair comparison in the in-dataset setting. For the model name, "`C`" refers to "`Character`" and "`B`" refers to "`Bigram`". Intuitively, the models are named based on their constituents. For example, *Cw2vBw2vLstmCrf* denotes a model's character and the bigram feature is initialized by pre-trained embeddings using Word2Vec, and sentence encoder, as well as the decoder, are LSTM and CRF, respectively. We perform a Friedman test at p = 0.05 on model- (row-) wise and data- (column-)wise. The testing results are $p(\text{model} - \text{wise}) = 2.26 \times 10^{-6} < 0.05$ and $p(\text{data} - \text{wise}) = 8.42 \times 10^{-8}$. Therefore, the results of model-wise and data-wise have passed the significance testing.

Fu et al. , *RethinkCWS: Is Chinese Word Segmentation a Solved Task?，EMNLP 2020*

1 细粒度评测 - Is Chinese Word Segmentation a Solved Task?





Aspect-I: Intrinsic nature

word length (wLen); sentence length (sLen)

OOV density (oDen);

Aspect-II: Familiarity

word frequency (wFre); character frequency (cFre)

Aspect-III: Label consistency

label consistency of word (wCon);

label consistency of character (cCon)

Self-diagnosis：aims to locate the bucket on which the input model has obtained the worst performance with respect to a given attribute.

Aided-diagnosis(A,B): aims to compare the performance of different models on different bucket.

Fu et al. , *RethinkCWS: Is Chinese Word Segmentation a Solved Task?，EMNLP 2020*

1 细粒度评测 - Rethinking Generalization of Neural Models

Datasets	Embed-layer		Entity Coverage Rate					
	Char	Word	Overall	1	$(0.5, 1)$	$(0, 0.5]$	$C \neq 0$	$C = 0$
CoNLL	CNN	-	76.42	79.94	86.99	78.84	69.74	77.61
	FLAIR	-	89.98	95.30	95.58	82.39	72.16	90.39
	ELMo	-	91.79	97.61	95.98	85.15	71.43	92.22
	BERT	-	91.34	97.72	95.17	86.66	**77.83**	92.37
	-	Rand	78.43	95.05	94.75	73.54	37.97	66.40
	-	GloVe	89.10	98.44	96.31	81.34	57.80	87.23
	CNN	Rand	82.88	94.13	94.48	74.25	47.78	78.91
	CNN	GloVe	90.33	98.32	95.94	80.33	59.67	89.74
	ELMo	GloVe	92.46	98.08	**96.46**	86.14	69.79	93.08
	FLAIR	GloVe	**93.03**	**98.56**	96.38	**87.07**	73.58	**93.42**
WNUT	CNN	-	20.88	45.99	67.01	40.25	19.14	19.74
	FLAIR	-	41.49	81.15	88.14	54.36	39.56	43.44
	ELMo	-	43.70	88.72	90.83	55.56	**44.19**	43.32
	BERT	-	44.08	77.75	81.61	49.74	34.65	41.92
	-	Rand	14.97	60.62	83.84	50.00	3.90	4.77
	-	GloVe	37.28	89.29	92.62	45.65	35.34	35.15
	CNN	Rand	22.29	48.88	71.43	39.08	16.75	18.83
	CNN	GloVe	40.72	86.12	**92.24**	49.74	26.67	40.06
	ELMo	GloVe	45.33	90.38	89.92	56.57	37.8	46.58
	FLAIR	GloVe	**45.96**	**90.52**	89.92	**61.69**	42.07	**48.38**

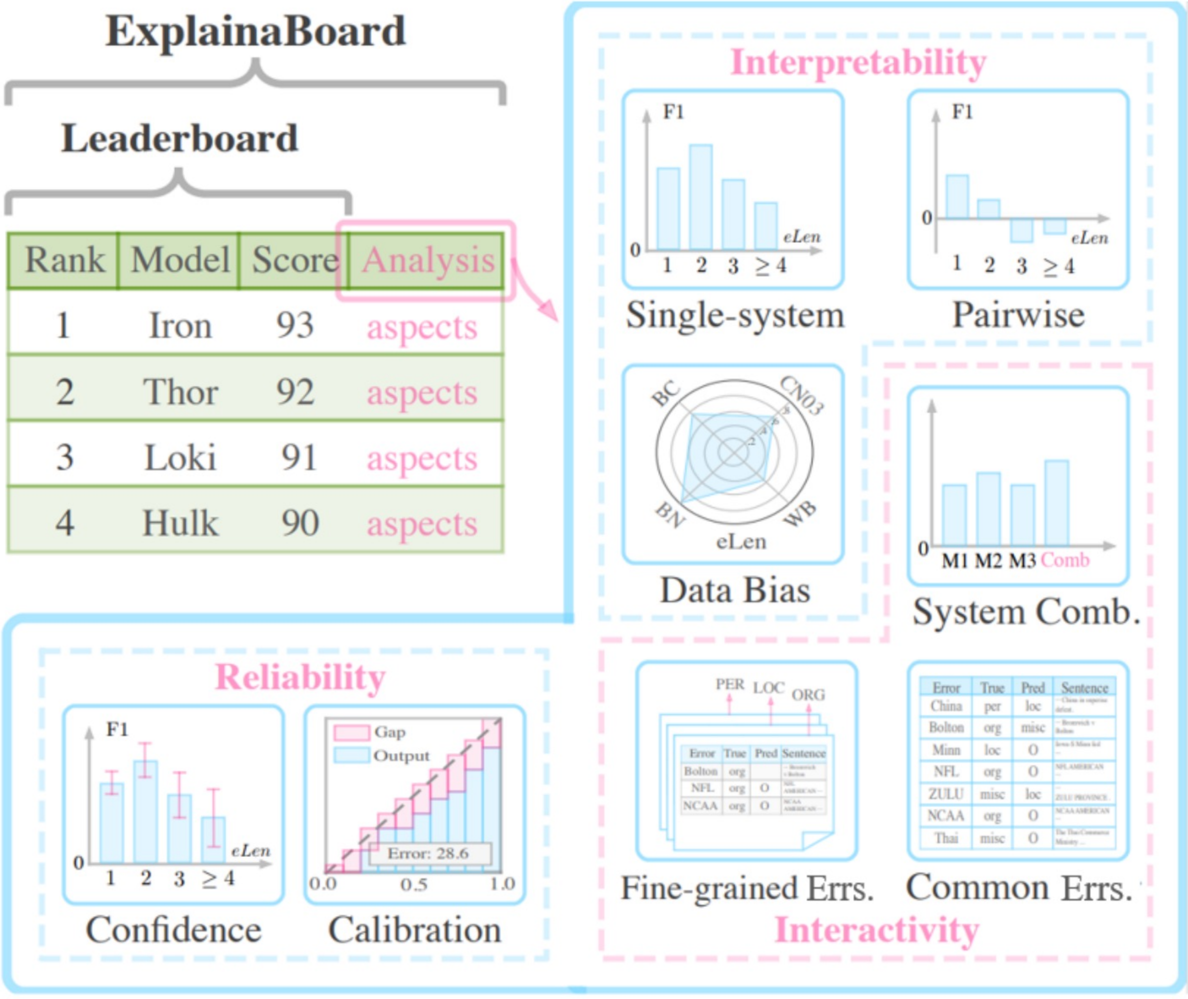
Entity Coverage Ratio (ECR) The measure entity coverage ratio is used to describe the degree to which entities in the test set have been seen in the training set with the same category.

$$\rho(e_i) = \begin{cases} 0 & C = 0 \\ \left(\sum_{k=1}^{K} \frac{\#(e_i^{tr,k})}{C^{tr}} \#(e_i^{te,k})\right)/C^{te} & \text{otherwise} \end{cases} \quad (1)$$

where $e_i^{tr,k}$ is the entity e_i in the training set with ground truth label k, $e_i^{te,k}$ is the entity e_i in the test set with ground truth label k, $C^{tr} = \sum_{k=1}^{K} \#(e_i^{tr,k})$, $C^{te} = \sum_{k=1}^{K} \#(e_i^{te,k})$, and $\#$ denotes the counting operation.

Fu et al. , *Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study*， AAAI 2020

1 细粒度评测 - EXPLAINABOARD: An Explainable Leaderboard for NLP



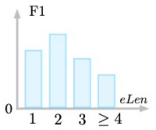
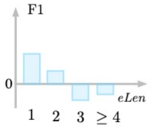
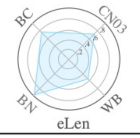
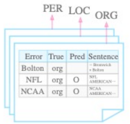
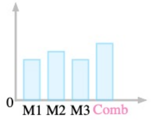
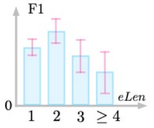
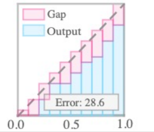
Aspect	Functionality	Input	Output	
Interpretability	Single-system Analysis	One model		**Performance Histogram**: the input model is good at dealing with short entities, while achieving lower performance on long entities.
	Pairwise Analysis	Two models (M1,M2)		**Performance Gap Histogram** (M1−M2): M1 is better at dealing with short entities, while M2 is better at dealing with long entities.
	Data Bias Analysis	Multi-dataset		**Data Bias Chart**: For the `entity length` attribute, the average entity length (We average the length of all test entities on a given data set.) of these datasets order by descending is BN> BC> CN03> WB.
Interactivity	Fine-grained Error Analysis	Single- or Pairwise-system diagnostic results		**Error Table**: Error analysis allows the user to print out the entities that are incorrectly predicted by the given model, as well as the true label of the entity, the mispredicted label, and the sentence where the entity is located.
	System Combination	Multi-models (M1,M2,M3)		**Ensemble Chart**: The combined result of model M1, M2, and M3 is shown by the histogram with x-label value `comb`. The combined result is better than the single models.
Reliability	Confidence	One model		**Error Bars**: the error bars represent 95% confidence intervals of the performance on the specific bucket.
	Calibration	One model		**Reliability Diagram**: Confidence histograms (red) and reliability diagrams (blue). that indicate the accuracy of model probability estimates

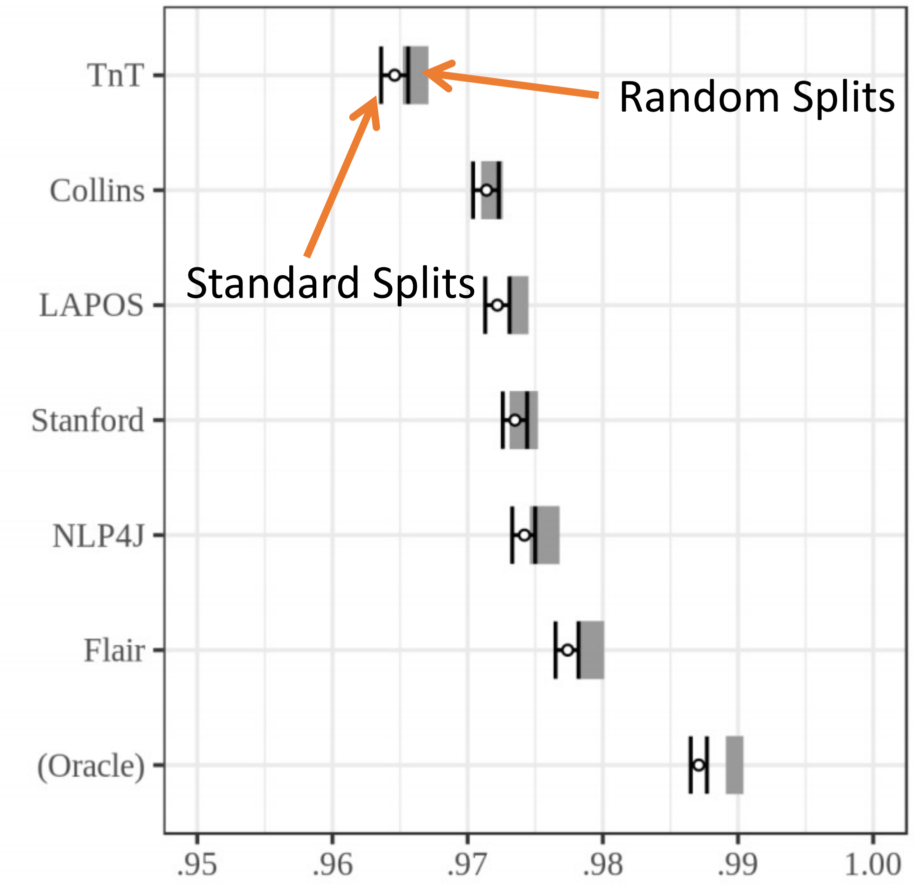
Table 1: A graphical breakdown of the functionality of EXPLAINABOARD, with examples from an NER task.

Liu et al. , *EXPLAINABOARD: An Explainable Leaderboard for NLP*, ACL 2021

1 数据集划分 – The impact of splitting methods

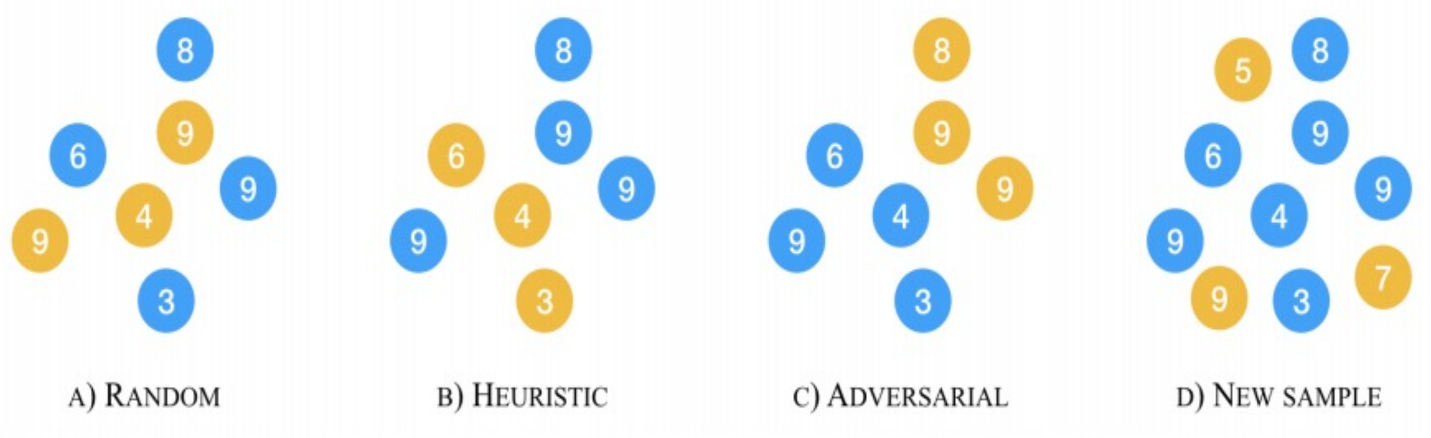
Standard splits:

Training: sections 00–18
Development: sections 19-21
Testing: sections 22-24



Gorman et al., *We need to talk about standard splits*, ACL 2019.

1 数据集划分 – The impact of splitting methods



Blue balls – Training
Orange balls -- Test

Task	Model	Splits				
		Standard	**Random**	**Heuristic**	**Adversarial**	**New Samples**
POS TAGGING	NCRF++	0.961	0.962	0.960	0.944	**0.927**
PROBING-WC	BERT	0.520	0.527	**0.232**	0.250	0.279
PROBING-BSHIFT	BERT	0.800	0.808	0.695	0.706	**0.450**
HEADLINE GENERATION*	seq2seq	0.073	0.095	0.062	**0.040**	0.069
QUALITY ESTIMATION†	MLP-Laser	0.502	0.626	0.621	0.711	**0.767**
EMOJI PREDICTION	MLP-Laser	-	0.125	0.196	**-0.040**	0.091
NEWS CLASSIFICATION	MLP-Laser	-	0.681	0.720	0.634	**0.618**
MSE (**New Samples**)		0.179	0.030	0.015	0.011	-

Søgaard, *We Need to Talk About Random Splits*, EACL 2021.

0 问题1：为什么基于基准测试集合和常用评价指标的模式不能反映上述问题？

1. **基准集合构建时通常存在数据偏置**
   a. 要消除数据集合偏置
   b. 根据任务特性增加人工变形
2. **粗粒度的评测指标不能够全面反映模型特性**
   a. 针对任务特性的评测指标设计

FUDAN UNIVERSITY NLP NATURAL LANGUAGE PROCESSING

问题1：为什么基于基准测试集合和常用评价指标的模式不能反映上述问题？

问题2：深度神经网络模型到底学习到了什么？

问题3：现阶段自然语言处理算法鲁棒性究竟怎么样?

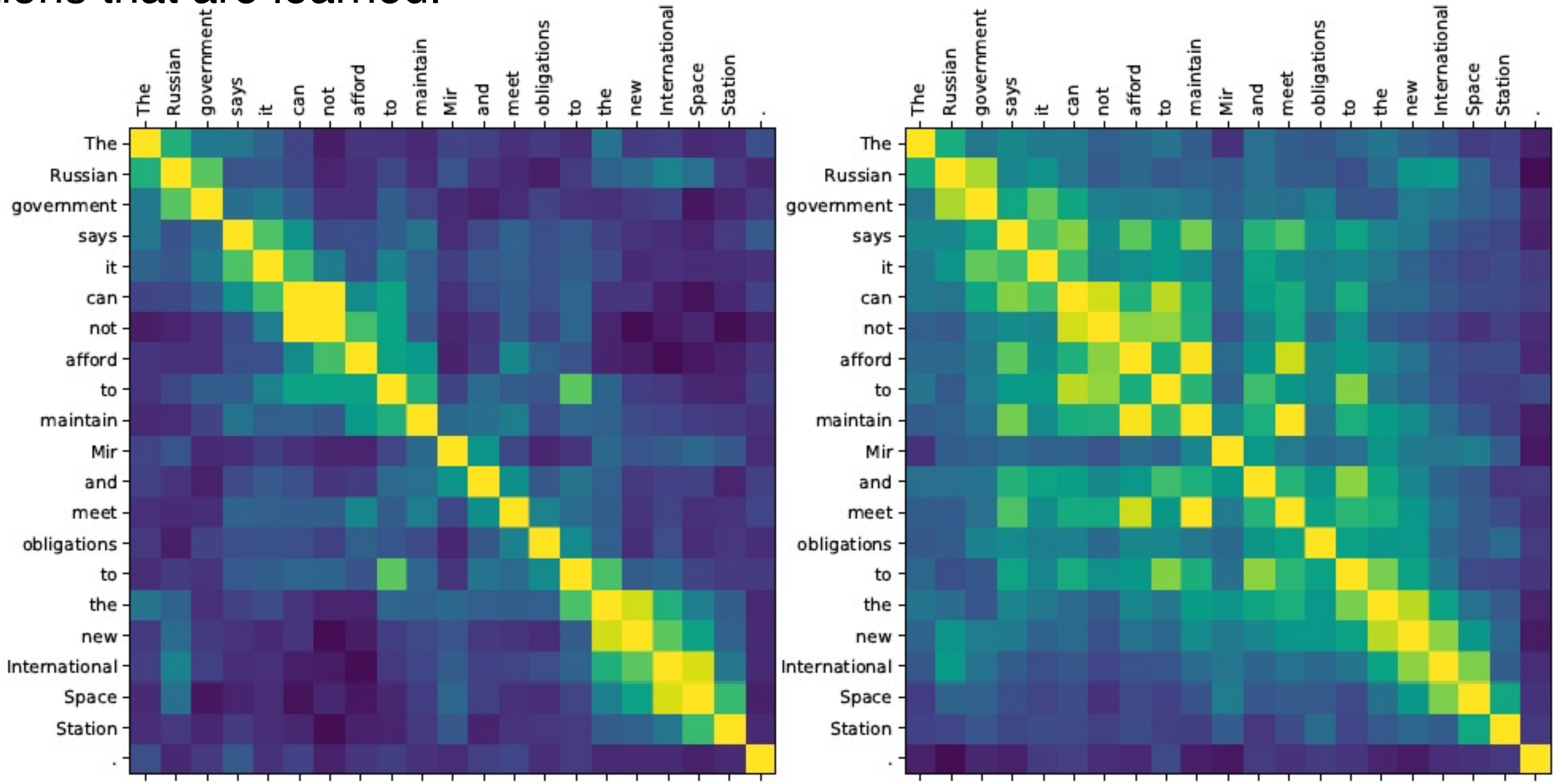
Several examples of cells with interpretable activations discovered in LSTM trained with **Linux Kernel** and **War and Peace**.

Karpathy et al. , *Visualizing and Understanding Recurrent Networks* , 2016

2 Contextual Word Embeddings

They presented a detailed empirical study of how the choice of neural architecture (e.g. LSTM, CNN, or self attention) influences both end task accuracy and qualitative properties of the representations that are learned.

Bottom LSTM layer



Top LSTM layer

Visualization of contextual similarity between all word pairs in a single sentence using the 4-layer LSTM.

Peters et al. , *Dissecting Contextual Word Embeddings: Architecture and Representation*, 2018

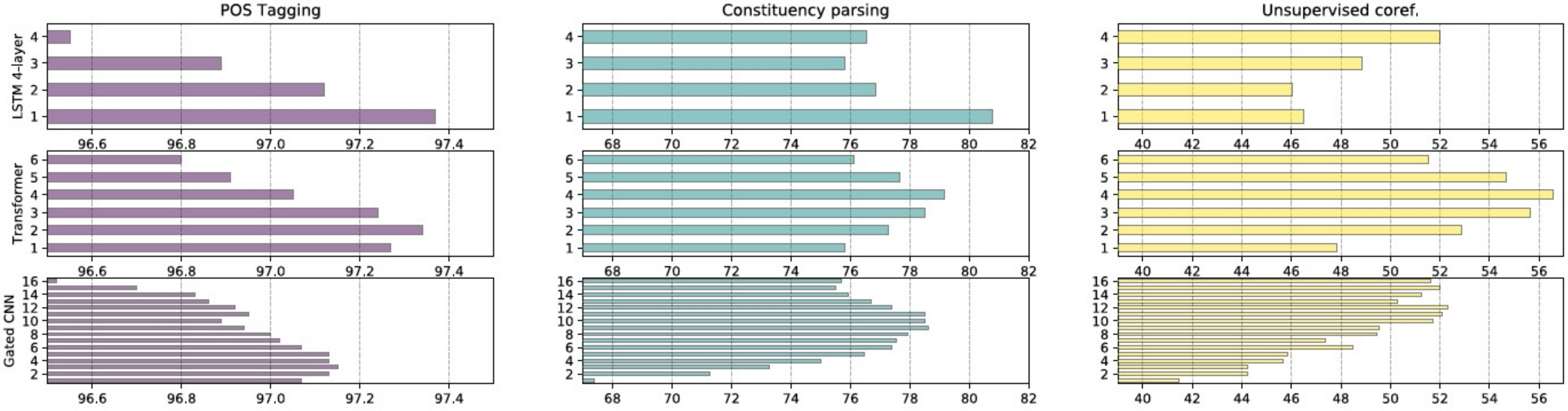


Figure 3: Various methods of probing the information stored in context vectors of deep biLMs. Each panel shows the results for all layers from a single biLM, with the first layer of contextual representations at the bottom and last layer at the top. From top to bottom, the figure shows results from the 4-layer LSTM, the Transformer and Gated CNN models. From left to right, the figure shows linear POS tagging accuracy (%; Sec. 5.3), linear constituency parsing (F_1; Sec. 5.3), and unsupervised pronominal coreference accuracy (%; Sec. 5.1).

Peters et al. , *Dissecting Contextual Word Embeddings: Architecture and Representation*, 2018

2 Integrated Gradients 归因方法

Question: how symmetrical are the white bricks on either side of the building

Prediction: very

Ground truth: very

Red -- high attribution

Blue -- negative attribution

Gray -- near-zero attribution

Integrated Gradients (IG) (Sundararajan et al., 2017) to isolate question words that a deep learning system uses to produce an answer.

Definition 1 (Integrated Gradients) *Given an input x and baseline x', the integrated gradient along the i^{th} dimension is defined as follows.*

$$\mathrm{IG}_i(x, x') ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

(here $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F along the i^{th} dimension at x).

For image networks, the baseline input x' could be the black image, while for text models it could be the zero embedding vector.

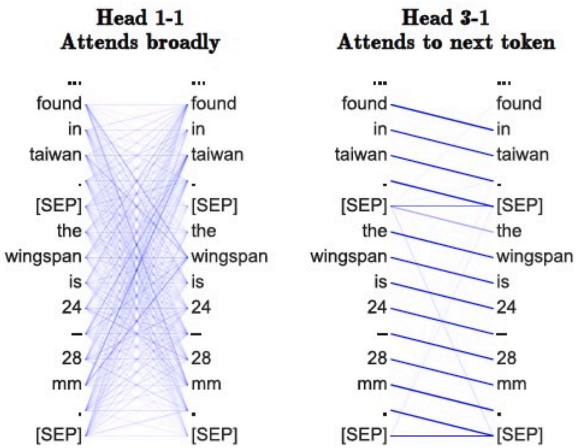
Sundararajan et al., *Axiomatic attribution for deep networks*. 2017

Mudrakarta et al. *Did the Model Understand the Question?* ACL 2018

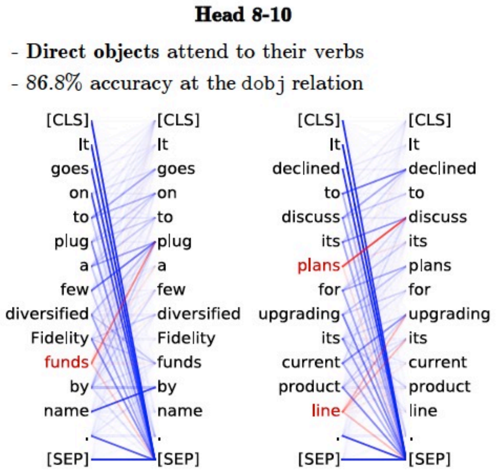
基于Bert的 用户检索词---文章语义匹配模型

用户查询：硫酸沙丁胺醇吸入气雾剂用法

2 BERT中Attention Head学习到了丰富的高层语言特征



Attention heads exhibiting patterns



Attention heads corresponding to linguistic phenomena

Relation	Head	Accuracy	Baseline
All	7-6	34.5	26.3 (1)
prep	7-4	66.7	61.8 (-1)
pobj	9-6	**76.3**	34.6 (-2)
det	8-11	**94.3**	51.7 (1)
nn	4-10	70.4	70.2 (1)
nsubj	8-2	58.5	45.5 (1)
amod	4-10	75.6	68.3 (1)
dobj	8-10	**86.8**	40.0 (-2)
advmod	7-6	48.8	40.2 (1)
aux	4-10	81.1	71.5 (1)
poss	7-6	**80.5**	47.7 (1)
auxpass	4-10	**82.5**	40.5 (1)
ccomp	8-1	**48.8**	12.4 (-2)
mark	8-2	**50.7**	14.5 (2)
prt	6-7	**99.1**	91.4 (-1)

The best performing attentions heads of BERT on WSJ dependency parsing

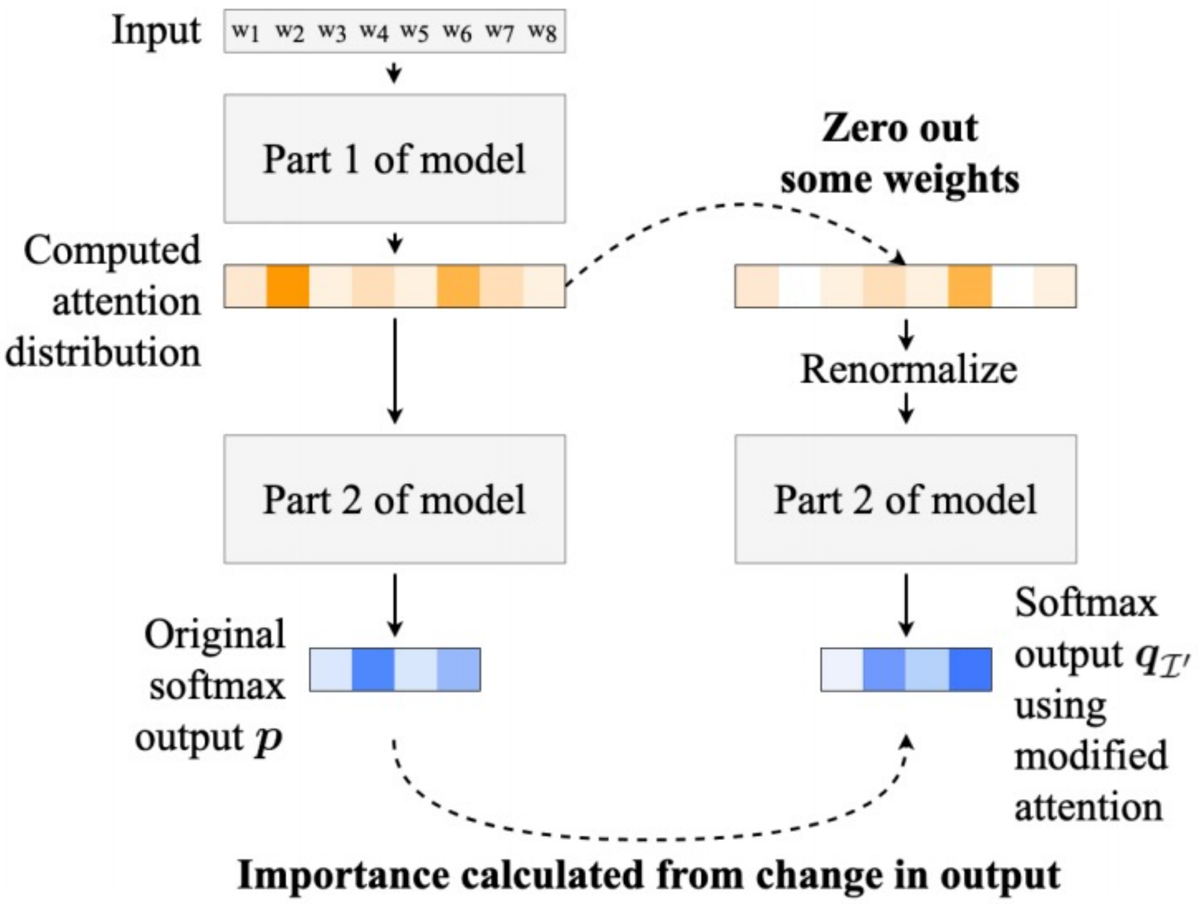
BERT's attention heads exhibit patterns such as attending to delimiter tokens, specific positional offsets, or broadly attending over the whole sentence, with heads in the same layer often exhibiting similar behaviors

Certain attention heads correspond well to linguistic notions of syntax and coreference.

Attention-based probing classifier demonstrated that substantial syntactic information could be captured in BERT's attention.

Clark et al. , *What Does BERT Look At? An Analysis of BERT's Attention* , ACL 2019

2 Attention 是否可以解释？

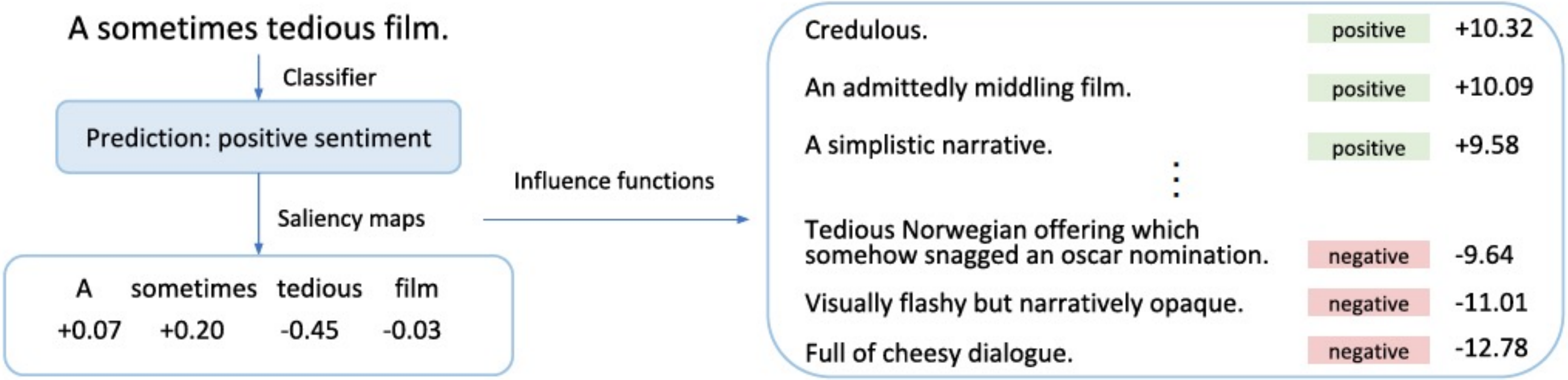


Attention layers explicitly weight input components' representations, it is also often assumed that attention can be used to identify information that models found important

They observe some ways in which higher attention weights correlate with greater impact on model predictions, they also find many ways in which this does not hold

Sofia Serrano & Noah A. Smith, *Is Attention Interpretable?*, ACL 2019

2 训练语料对于模型的影响 – Influence Function



Salient tokens in the input

Influential examples in the training corpus

Influence functions：

$$\frac{d\hat{\theta}}{d\epsilon_i} = -\left(\frac{1}{n}\sum_{j=1}^{n}\nabla_\theta^2 \mathcal{L}(x_j, y_j, \hat{\theta})\right)^{-1}\nabla_\theta \mathcal{L}(x_i, y_i, \hat{\theta})$$

How upweighting a particular training example (x_i, y_i) in the training set {(x_1, y_1) ,…, ((x_n, y_n)} by ϵ_i would change the learned model parameters θ

$$\frac{d\mathcal{L}_{\hat{y}}}{d\epsilon_i} = \nabla_\theta \mathcal{L}_{\hat{y}} \cdot \frac{d\hat{\theta}}{d\epsilon_i}$$

How this change in the model parameters would in turn affect the loss of the test input

Pang Wei Koh and Percy Liang. 2017. *Understanding black-box predictions via influence functions*. ICML 2017
Han et al., *Explaining black box predictions and unveiling data artifacts through influence functions*, ACL 2020

0 问题2：深度神经网络模型到底学习到了什么？

非常初步的猜想，大规模数据分析和实验中

1. 预训练方法提供了句法等高层语言特征
2. 高层语言特征与词表层特征综合提供了分类表示
3. 预训练语言模型学习到了部分复述（Paraphrase）的相似表示

覆盖了人工构造的基础特征，以及人工很难构造的特征高阶综合

超强的数据拟合能力　　独立同分布条件的泛化能力

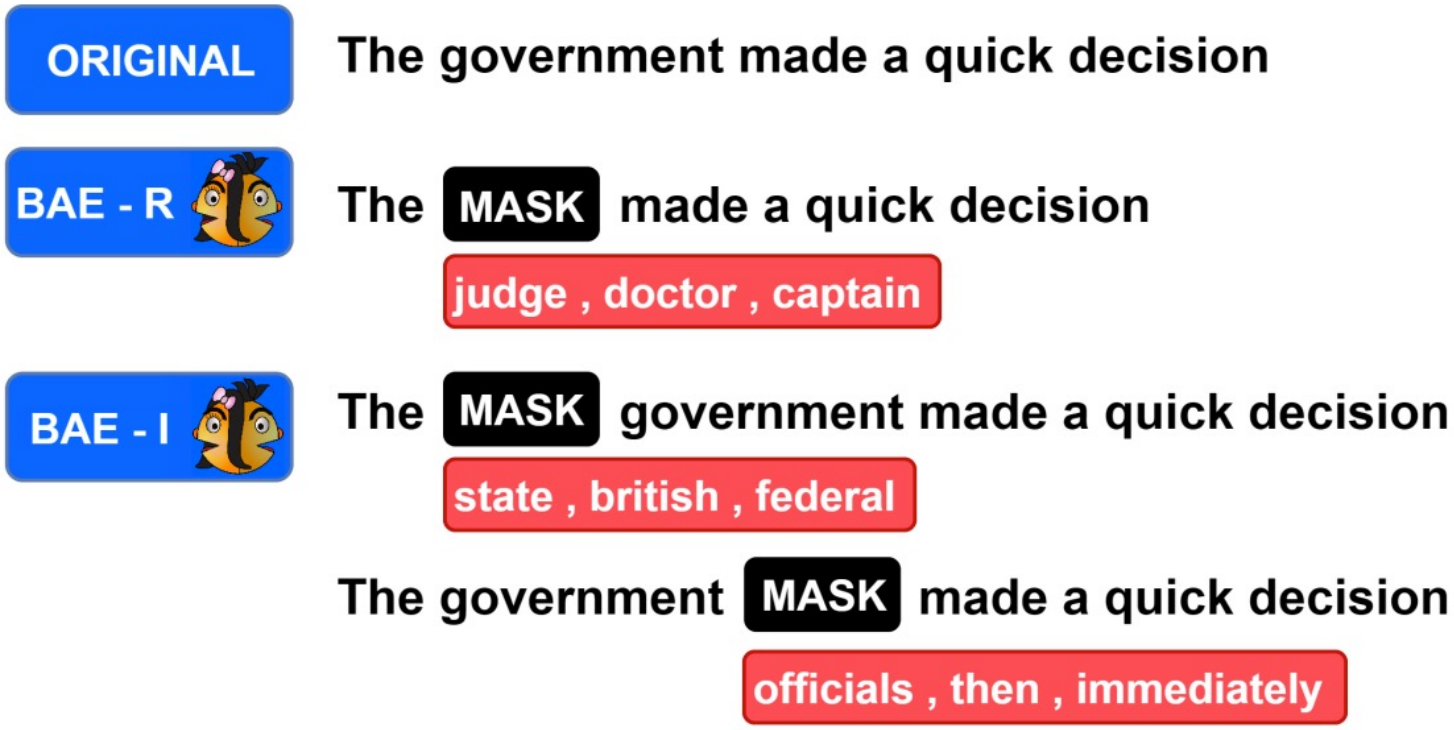
0 为什么会这样？

问题1：为什么基于基准测试集合和常用评价指标的模式不能反映上述问题？

问题2：深度神经网络模型到底学习到了什么？

问题3：现阶段自然语言处理算法鲁棒性究竟怎么样?

3 模型评测 - BERT-based Adversarial Examples



They use BERT-MLM to predict masked tokens in the text for generating adversarial examples. The MASK token replaces a word (BAE-R attack) or is inserted to the left/right of the word (BAE-I).

Garg and Ramakrishnan, *BAE: BERT-based Adversarial Examples for Text Classification*, EMNLP 2020.

3 模型评测 - BERT-based Adversarial Examples

Model	Adversarial Attack	Datasets: Amazon	Yelp	IMDB	MR
wordLSTM	Original	88.0	85.0	82.0	81.16
	TextFooler	31.0 (0.747)	28.0 (0.829)	20.0 (0.828)	25.49 (0.906)
	BAE-R	21.0 (0.827)	20.0 (0.885)	22.0 (0.852)	24.17 (0.914)
	BAE-I	17.0 (0.924)	22.0 (0.928)	23.0 (0.933)	19.11 (0.966)
	BAE-R/I	16.0 (0.902)	19.0 (0.924)	8.0 (0.896)	15.08 (0.949)
	BAE-R+I	**4.0 (0.848)**	**9.0 (0.902)**	**5.0 (0.871)**	**7.50 (0.935)**
wordCNN	Original	82.0	85.0	81.0	76.66
	TextFooler	42.0 (0.776)	36.0 (0.827)	31.0 (0.854)	21.18 (0.910)
	BAE-R	16.0 (0.821)	23.0 (0.846)	23.0 (0.856)	20.81 (0.920)
	BAE-I	18.0 (0.934)	26.0 (0.941)	29.0 (0.924)	19.49 (0.971)
	BAE-R/I	13.0 (0.904)	17.0 (0.916)	20.0 (0.892)	15.56 (0.956)
	BAE-R+I	**2.0 (0.859)**	**9.0 (0.891)**	**14.0 (0.861)**	**7.87 (0.938)**
BERT	Original	96.0	95.0	85.0	85.28
	TextFooler	30.0 (0.787)	27.0 (0.833)	32.0 (0.877)	30.74 (0.902)
	BAE-R	36.0 (0.772)	31.0 (0.856)	46.0 (0.835)	44.05 (0.871)
	BAE-I	20.0 (0.922)	25.0 (0.936)	31.0 (0.929)	32.05 (0.958)
	BAE-R/I	**11.0 (0.899)**	16.0 (0.916)	22.0 (0.909)	20.34 (0.941)
	BAE-R+I	14.0 (0.830)	**12.0 (0.871)**	**16.0 (0.856)**	**19.21 (0.917)**

Dataset	Sentiment Accuracy (%): Original	TF	R	R+I
Amazon	95.7	79.1	**85.2**	83.8
IMDB	90.3	83.1	**84.3**	79.3
MR	93.3	82.0	**84.6**	82.4
Dataset	**Naturalness (1-5)**: Original	TF	R	R+I
Amazon	4.26	3.17	**3.91**	3.71
IMDB	4.35	3.41	**3.89**	3.76
MR	4.19	3.35	**3.84**	3.74

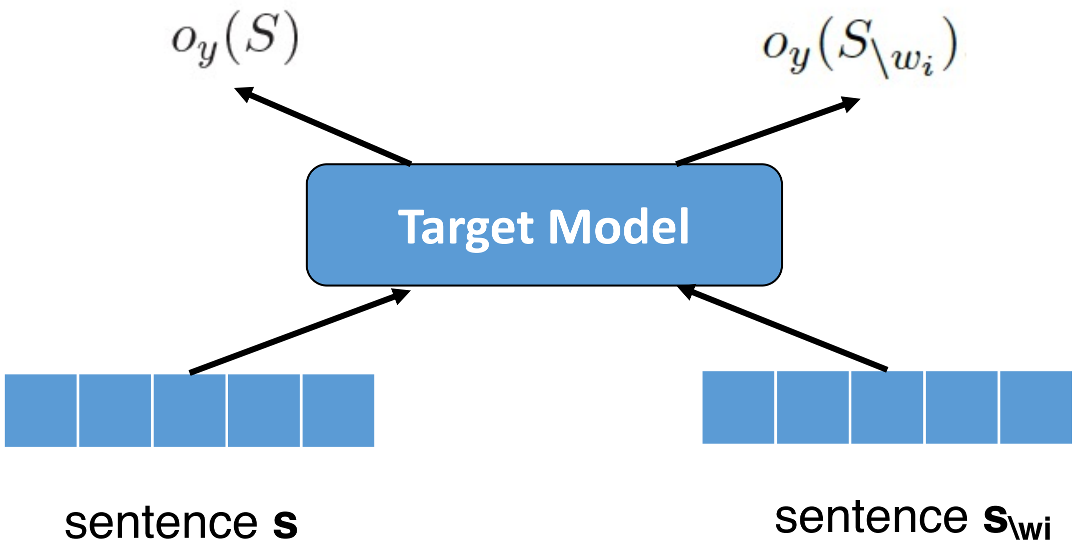
Human evaluation results

Garg and Ramakrishnan, *BAE: BERT-based Adversarial Examples for Text Classification*, EMNLP 2020.

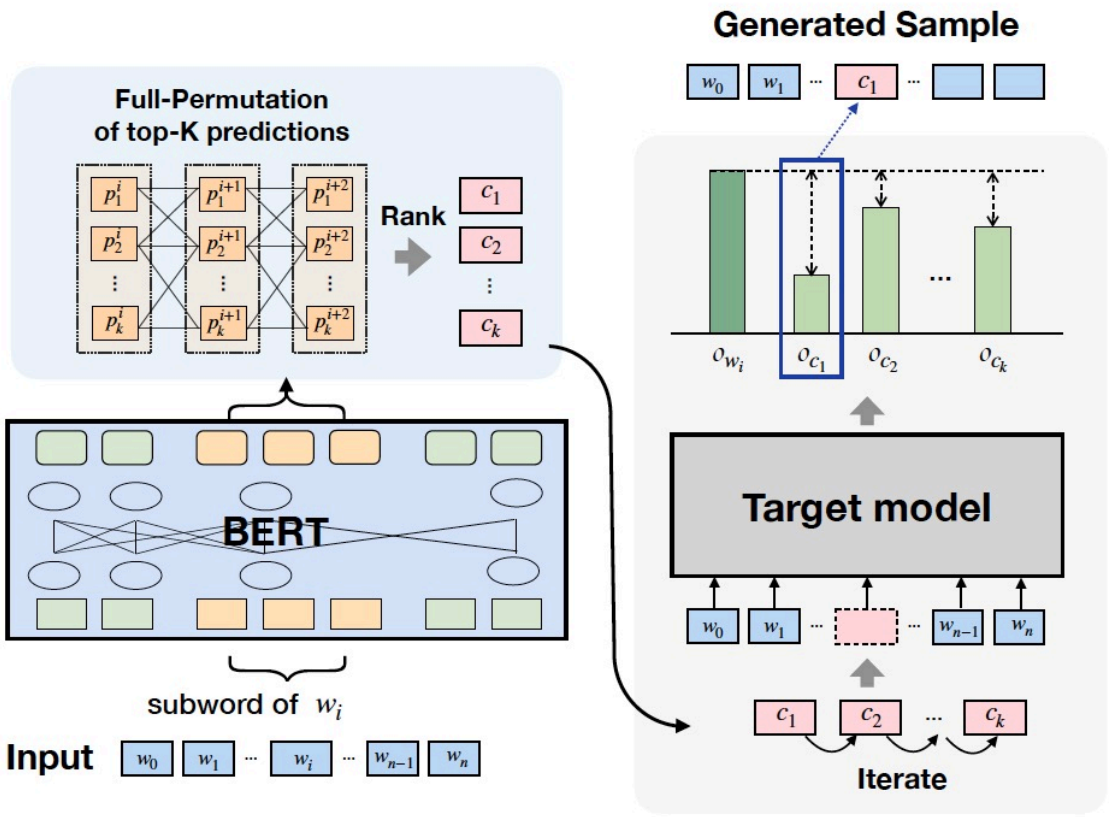
BERT-Attack

1. Finding Vulnerable Words

$$I_{w_i} = o_y(S) - o_y(S_{\backslash w_i})$$



2. Word Replacement via BERT



3 模型评测 - BERT-ATTACK

Dataset	Method	Original Acc	Attacked Acc	Perturb %	Query Number	Avg Len	Semantic Sim
Fake	BERT-Attack(ours)	97.8	**15.5**	**1.1**	**1558**	885	**0.81**
	TextFooler(Jin et al., 2019)		19.3	11.7	4403		0.76
	GA(Alzantot et al., 2018)		58.3	1.1	28508		-
Yelp	BERT-Attack(ours)	95.6	**5.1**	**4.1**	**273**	157	**0.77**
	TextFooler		6.6	12.8	743		0.74
	GA		31.0	10.1	6137		-
IMDB	BERT-Attack(ours)	90.9	**11.4**	**4.4**	**454**	215	**0.86**
	TextFooler		13.6	6.1	1134		**0.86**
	GA		45.7	4.9	6493		-
AG	BERT-Attack(ours)	94.2	**10.6**	**15.4**	**213**	43	**0.63**
	TextFooler		12.5	22.0	357		0.57
	GA		51	16.9	3495		-
SNLI	BERT-Attack(ours)	89.4(H/P)	7.4/**16.1**	**12.4/9.3**	**16/30**	8/18	0.40/**0.55**
	TextFooler		**4.0**/20.8	18.5/33.4	60/142		**0.45**/0.54
	GA		14.7/-	20.8/-	613/-		-
MNLI matched	BERT-Attack(ours)	85.1(H/P)	**7.9/11.9**	**8.8/7.9**	**19/44**	11/21	0.55/**0.68**
	TextFooler		9.6/25.3	15.2/26.5	78/152		**0.57**/0.65
	GA		21.8/-	18.2/-	692/-		-
MNLI mismatched	BERT-Attack(ours)	82.1(H/P)	**7/13.7**	**8.0/7.1**	**24/43**	12/22	0.53/**0.69**
	TextFooler		8.3/22.9	14.6/24.7	86/162		**0.58**/0.65
	GA		20.9/-	19.0/-	737/-		-

Table 1: Results of attacking against various fine-tuned BERT models. TextFooler is the state-of-the-art baseline. For MNLI task, we attack the hypothesis(H) or premises(P) separately.

Li et al., *BERT-ATTACK: Adversarial Attack Against BERT Using BERT*, EMNLP 2020.

模型评测 - A Case Against Synonym-Based Adversarial Example

Between 96% and 99% of the analyzed attacks do not preserve semantics, indicating that their success is mainly based on feeding poor data to the model.

Attack	Word Similarity			Text Similarity		
	Avg. (1-7)	Above 5 (%)	Above 6 (%)	Avg. (1-7)	Above 5 (%)	Above 6 (%)
TextFooler	**3.88**	**22**	**7**	**3.47**	**24**	**12**
PWWS	3.83	21	6	2.70	13	6
BERT-Attack	2.27	4	4	2.55	7	3
BAE	1.64	0	0	1.85	3	2

Table 2: Average human scores on a scale from 1-7 and the percentage of scores above 5 and 6 (corresponding to the answers "Somewhat Agree" and "Agree") for the different attacks and when the words were shown with (text similarity) or without (word similarity) context.

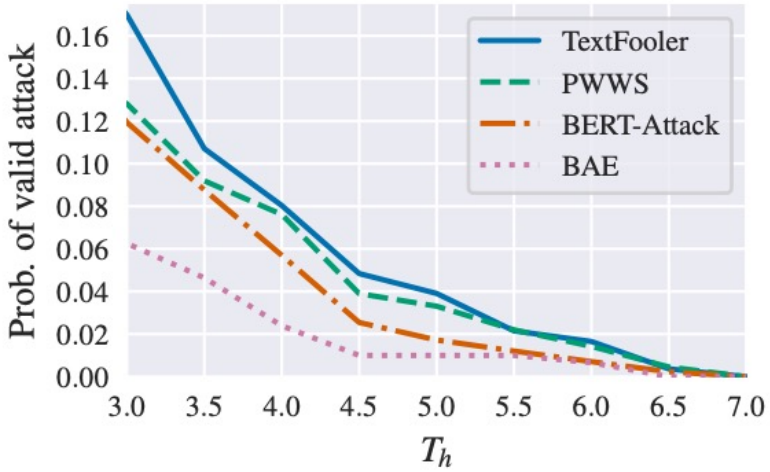


Figure 1: Probability that an attack is valid according to our probabilistic analysis, for the different attacks and for different thresholds T_h.

Hauser et al., *BERT is Robust! A Case Against Synonym-Based Adversarial Examples in Text Classification*, arXiv 2021.

Benchmarking Robustness of Machine Reading Comprehension Models

However, most of these benchmarks only evaluate models on in-domain test sets without considering their robustness under test-time perturbations.

Perturbation	Perturbation Level	Applied Component	MCRC-specific
AddSent	Sentence	Passage	No
CharSwap	Character	Passage + Question	No
Paraphrase	Sentence	Passage	No
Superimposed	Sentence + Character	Passage	No
Distractor Extraction	Sentence	Distractors	Yes
Distracor Generation	Sentence	Distractors	Yes

Table 1: Summary of our perturbations. MCRC-specific means whether the method is specific to the format of multiple-choice reading comprehension.

Si et al., *Benchmarking Robustness of Machine Reading Comprehension Models*, ACL 2021

3 模型评测 – Benchmarking Robustness MRC

Test Set	BERT	RoBERTa	XLNet	ALBERT
Original	69.5	83.7	79.9	86.0
AddSent	30.0 (*-56.8%*)	57.3 (*-31.5%*)	51.4 (*-35.7%*)	57.8 (*-32.8%*)
CharSwap	48.8 (*-29.8%*)	69.4 (*-17.1%*)	63.4 (*-20.7%*)	73.0 (*-15.1%*)
Paraphrase	59.4 (*-14.5%*)	72.3 (*-13.6%*)	68.2 (*-14.6%*)	73.7 (*-14.3%*)
Superimposed	18.6 (*-73.2%*)	38.1 (*-54.5%*)	36.4 (*-54.4%*)	36.1 (*-58.0%*)
Distractor Extraction	32.0 (*-54.0%*)	47.5 (*-43.2%*)	42.9 (*-46.3%*)	50.7 (*-41.0%*)
Distractor Generation	55.5 (*-20.1%*)	67.7 (*-19.1%*)	63.8 (*-20.2%*)	69.9 (*-18.7%*)
Average	40.7 (*-41.4%*)	58.7 (*-29.9%*)	54.4 (*-32.0%*)	60.2 (*-30.0%*)

Table 2: Attack results on different models. *Numbers* in brackets are the percentage drop in performance.

Si et al., *Benchmarking Robustness of Machine Reading Comprehension Models*, ACL 2021

3 模型评测 – Benchmarking Robustness MRC

(1) Over-sensitivity

MRC models provide different answers to the paraphrased questions.

(2) Over-stability

Models might fail into a trap span that has many words in common with the question, and extract an incorrect answer from the trap span

(3) Generalization

The well-generalized MRC models have good performance on both in-domain and out-of-domain data.

Passage 近年来，随着琥珀蜜蜡市场的兴起，蜜蜡与琥珀的价格都有不断上涨的趋势，其中蜜蜡首饰的价格一般是琥珀首饰价格的2–4倍，最近几年二者价格差距更大……	**Passage** *In recent years, with the rise of the amber market, the price of amber keeps going up. The price of opaque amber is generally 2–4 times the price of clear amber ...*
Original Question 琥珀和蜜蜡哪一个比较贵 **Golden Answer** : 蜜蜡 **Predicted Answer** : 蜜蜡 ($\text{BERT}_{\text{base}}$)	**Original Question** *Which is more expensive, clear amber or opaque amber?* **Golden Answer** : *opaque amber* **Predicted Answer** : *opaque amber* ($\text{BERT}_{\text{base}}$)
Paraphrase Question 蜜蜡和琥珀哪个价格高 **Golden Answer** : 蜜蜡 **Predicted Answer** : 琥珀 ($\text{BERT}_{\text{base}}$)	**Paraphrase Question** *Which has the higher price, opaque amber or clear amber?* **Golden Answer** : *opaque amber* **Predicted Answer** : *clear amber* ($\text{BERT}_{\text{base}}$)

(a) An example illustrates the over-sensitivity issue, where $\text{BERT}_{\text{base}}$ gives different predictions to the original question and the paraphrased question.

Passage 包粽子的线以前人们认为是来自麻叶子，其实是棕榈树，粽子的音就来自棕叶子。	**Passage** *Many people argue that the zongzi (rice dumpling) leaves are made of hemp. Actually, it is the palm tree, the real origin, that endows zongzi with the special pronunciation.*
Question 包粽子的线来自什么 **Golden Answer** : 棕榈树 **Predicted Answer** : 麻叶子 ($\text{BERT}_{\text{base}}$)	**Question** *What is the raw material of zongzi leaves?* **Golden Answer** : *palm tree* **predicted Answer** : *hemp* ($\text{BERT}_{\text{base}}$)

(b) An example illustrates the over-stability issue. The underlined span in the passage appears as a trap because it has many words in common with the question. $\text{BERT}_{\text{base}}$ falls into the trap.

Passage *cos(2x)'=-sin(2x)*(2x)'=-2sin(2x)* 属于复合函数的求导。	**Passage** *cos(2x)'=-sin(2x)*(2x)'=-2sin(2x) This is the derivative of a compound function.*
Question *cos2x*的导数是多少? **Golden Answer** : *-2sin(2x)* **Predicted Answer** : *-sin(2x)* ($\text{BERT}_{\text{base}}$)	**Question** *What is the derivative of cos2x?* **Golden Answer** : *-2sin(2x)* **Predicted Answer** : *-sin(2x)* ($\text{BERT}_{\text{base}}$)

(c) An example illustrates the generalization issue. Although $\text{BERT}_{\text{base}}$ is sufficiently trained on large-scale open-domain data, it fails to predict the answer to a math question.

Tang et al., *DuReader$_{robust}$: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications*, ACL 2021

3 模型评测 – Benchmarking Robustness MRC

	In-domain dev set		In-domain test set		Challenge test set	
	EM	F1	EM	F1	EM	F1
$\text{BERT}_{\text{base}}$	71.20	82.87	67.70	80.85	37.57	53.86
$\text{ERNIE 1.0}_{\text{base}}$	68.73	81.12	66.72	80.50	36.75	55.64
$\text{RoBERTa}_{\text{large}}$	74.17	86.02	71.20	84.16	45.02	62.83
Human			78.00	89.75	72.00	86.43

Table 4: Comparing MRC baselines to human on the development, test and all challenge sets.

	Over-Sensitivity		Over-Stability		Genera-lization	
	EM	F1	EM	F1	EM	F1
$\text{BERT}_{\text{base}}$	53.31	69.30	16.78	38.40	36.41	50.15
$\text{ERNIE 1.0}_{\text{base}}$	58.10	73.89	17.27	38.34	32.86	52.84
$\text{RoBERTa}_{\text{large}}$	55.24	75.16	28.18	47.03	46.03	61.67

Table 5: The results on the three subsets of the challenge set.

	Finance		Education	
	EM	F1	EM	F1
$\text{BERT}_{\text{base}}$	30.73	51.16	38.70	50.83
$\text{ERNIE 1.0}_{\text{base}}$	26.53	50.53	34.67	53.11
$\text{RoBERTa}_{\text{large}}$	40.22	61.16	47.77	61.82

Table 7: The performance of baselines in the domains of education and finance.

Topcis	EM	F1	#
Math	19.85	34.63	136
Chemistry	37.46	53.88	323
Language	44.31	61.18	255
Others	69.63	79.28	438
All	49.13	62.88	1152

Table 8: The performance of baselines on different topics in the domain of education.

Tang et al., *DuReader$_{robust}$: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications*, ACL 2021

FUDAN UNIVERSITY NATURAL LANGUAGE PROCESSING NLP

3 模型评测 – NER Can Fine-tuning Pretrained Model Lead to the Promised Land?

EMNLP 2020

Benchmarks are blessed with strong name regularity, high mention coverage and sufficient context diversity.

When scaling NER to open situations, these advantages may no longer exist

	Regular NER	Open NER
Typical Categories	Person, Location, Organization, etc.	Movie, Song, Book, TV Series, etc.
Name Regularity	Entity types with strong regularity	Entity types with weak or no regularity
Mention Coverage	Training set with high mention coverage	Many new and unseen mentions
Context Pattern	With decent training instances to capture	Fully-annotated training data is rare
Examples	**Location** Train *starting from [**Cherry Street**]* *... at [**8th Avenue**] ...* Test ⇩ *... at [**Cherry Street**]...* *... go to [**9th Avenue**] ...*	**Movie** Train *I watched [**avatar**]last night* *...[**the matrix**] is the best...* Test ⇩ *Wow...[**Joker**] was great!* *Love [**inception**] so much.*

Figure 1: Comparison between regular NER benchmarks and open NER tasks in reality.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

3 模型评测 – NER Can Fine-tuning Pretrained Model Lead to the Promised Land?

Settings	Name	Mention	Context	Examples
Vanilla Baseline	√	√	√	Train { [Putin] concluded his two days of talks. / [Blair] spoke to [Bush] on April 5. } Test [Putin] will face re-election in March 2004.
Name Permutation (NP)	×	√	√	Train { [the united] concluded his two days of talks. / [Hillsborough] spoke to [analysts] on April 5. } Test [the united] will face re-election in March 2004.
Mention Permutation (MP)	×	×	√	Train { [the united] concluded his two days of talks. / [Hillsborough] spoke to [analysts] on April 5. } Test [which girl] will face re-election in March 2004.
Context Reduction (CR)	√	√	↓	Train { [Putin] concluded his two days of talks. / [Blair] concluded his two days of talks. / [Bush] concluded his two days of talks. } Test [Putin] will face re-election in March 2004.
Mention Reduction (MR)	↓	↓	√	Train { [Blair] concluded his two days of talks. / [Blair] spoke to [Blair] on April 5. } Test [Putin] will face re-election in March 2004.

Table 1: Illustration of our four kinds of randomization test. The utterances in square brackets are entity mentions. Name: name regularity knowledge; Mention: high mention coverage; Context: sufficient training instances for context diversity √: the knowledge is preserved in this setting; ×: the knowledge is erased from the data in the setting; ↓: the knowledge decreases.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

3 模型评测 – NER Can Fine-tuning Pretrained Model Lead to the Promised Land?

Data Setting	**PER**	**ORG**	**GPE**	**FAC**	**LOC**	**WEA**	**VEH**	**ALL**
Baseline	86.31	76.49	80.89	69.23	40.58	74.70	61.97	81.76
Name Permutation	73.41	44.34	49.71	37.96	28.24	33.33	23.93	62.28
- Drop Compared with Baseline	15%	42%	39%	45%	44%	55%	61%	24%
Mention Permutation	61.78	39.40	33.27	32.16	18.60	9.38	21.92	51.58
- Drop Compared with Baseline	28%	48%	59%	54%	54%	87%	65%	34%

Table 2: Micro-F1 scores of BERT-CRF tagger on original data, name permutation setting and mention permutation setting respectively. We can see that erasing name regularity and mention coverage will significantly undermine the model performance.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

CheckList

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

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%			

Test NLP models, like we test software

What to test: Linguistic capabilities

How to test: Test behaviors with different test types

Minimum Functionality Test (MFT)

I didn't love the flight.
I can't say I recommend the food.
….

Perturbation tests

INV: Invariance tests

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.
@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon

Dir: Directional Expectation Tests

@AmericanAir service wasn't great. You are lame.	↓
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓

Ribeiro et al., *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*, ACL 2020.

3 模型评测 – CHECKLIST

Ribeiro et al., *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*, ACL 2020.

模型评测 – Dynabench

Dynabench is a research platform for dynamic data collection and benchmarking.

FACEBOOK AI

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

DynaBench

QUESTION ANSWERING

Question answering and machine reading comprehension is answering a question given a context.

Round:	2
Model error rate:	22.90% (1043/4555)
Last activity:	8 hours ago

NATURAL LANGUAGE INFERENCE

Natural Language Inference is classifying context-hypothesis pairs into whether they entail, contradict or are neutral.

Round:	4
Model error rate:	41.83% (18477/44167)
Last activity:	12 hours ago

SENTIMENT ANALYSIS

Sentiment analysis is classifying one or more sentences by their positive/negative sentiment.

Round:	3
Model error rate:	42.67% (32/75)
Last activity:	an hour ago

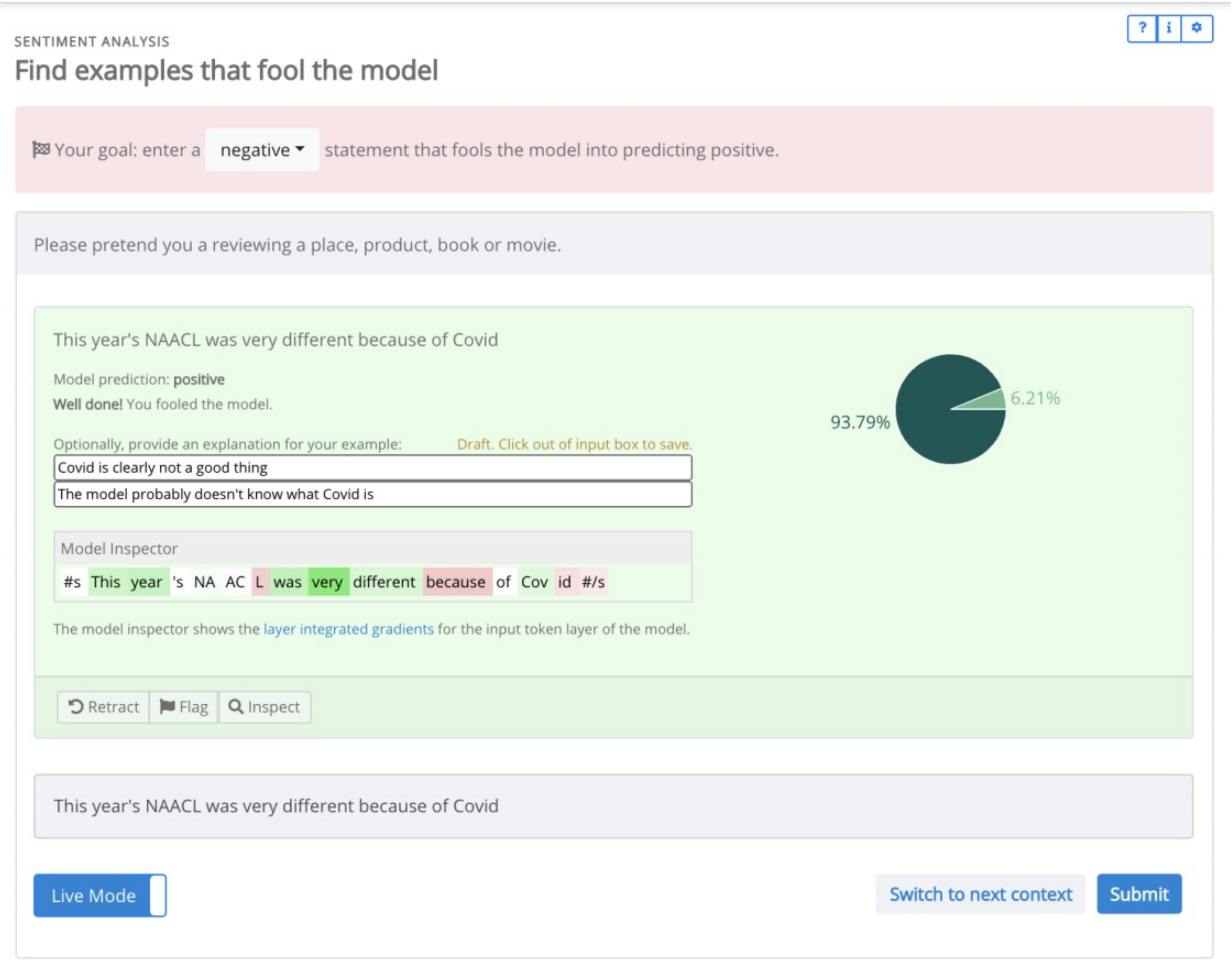
HATE SPEECH

Hate speech detection is classifying one or more sentences by whether or not they are hateful.

Round:	5
Model error rate:	60.77% (660/1086)
Last activity:	8 hours ago

Kiela et al., *Dynabench: Rethinking Benchmarking in NLP*, NAACL 2021.

3 模型评测 – Dynabench

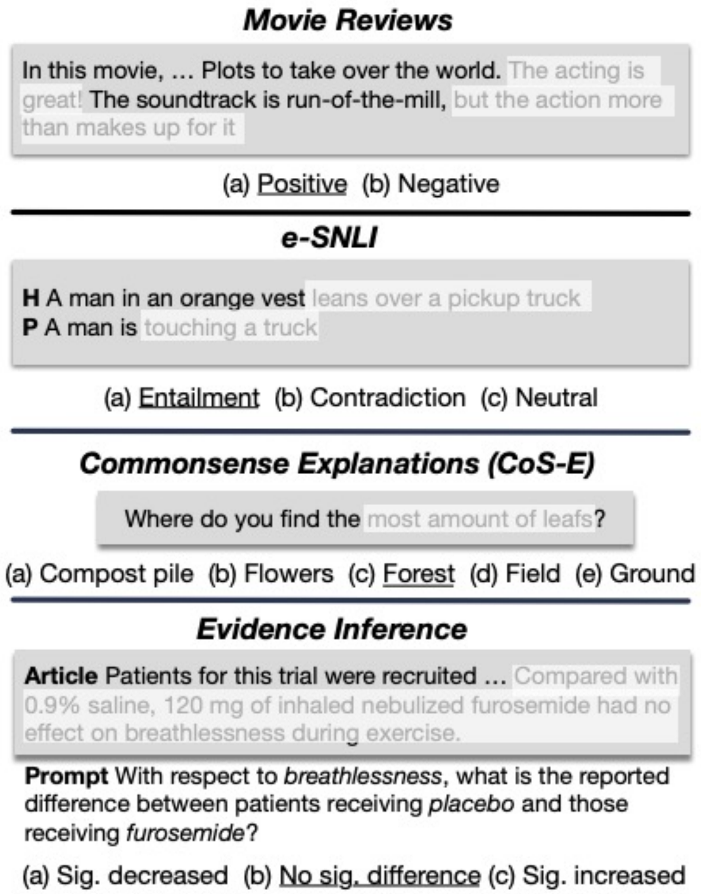


FACEBOOK AI

Dyna Bench

Kiela et al., *Dynabench: Rethinking Benchmarking in NLP*, NAACL 2021.

The **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning benchmark



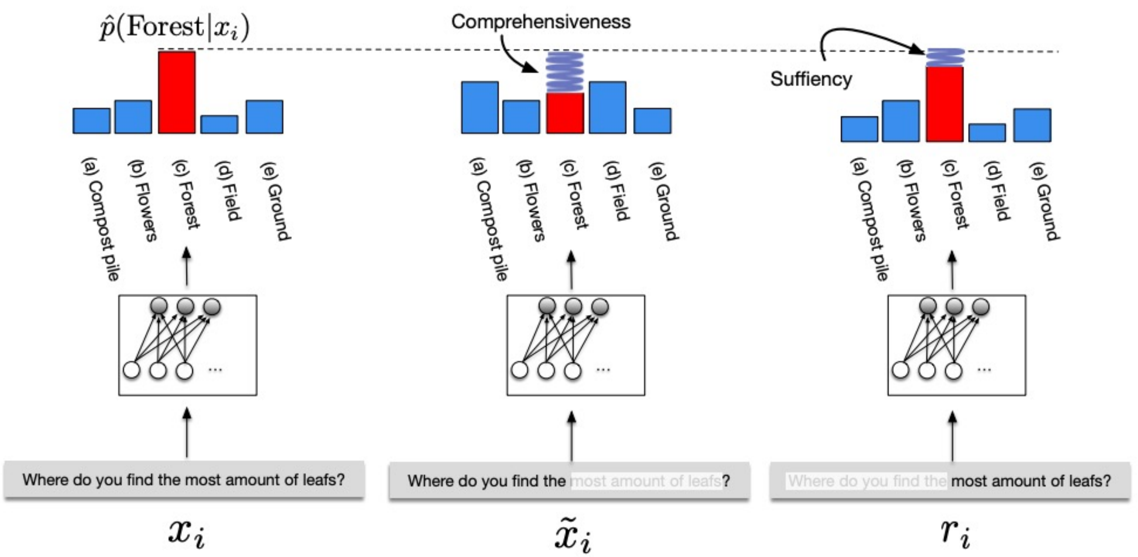


Figure 2: Illustration of faithfulness scoring metrics, *comprehensiveness* and *sufficiency*, on the Commonsense Explanations (CoS-E) dataset. For the former, erasing the tokens comprising the provided rationale ($\tilde{x}_i$) ought to decrease model confidence in the output 'Forest'. For the latter, the model should be able to come to a similar disposition regarding 'Forest' using *only* the rationales r_i.

http://www.eraserbenchmark.com/

DeYoung et al. *ERASER: A Benchmark to Evaluate Rationalized NLP Models*, 2020

TextFlint

Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing

https://github.com/textflint/textflint

3 TextFlint: 自然语言处理鲁棒性分析工具集

完备性 – 20 种通用变形、60种任务特有变形、数千种变形组合

14种NLP常见任务

中英双语

可接受 – 所有变形基于语言学知识

变形结果进行人工检查

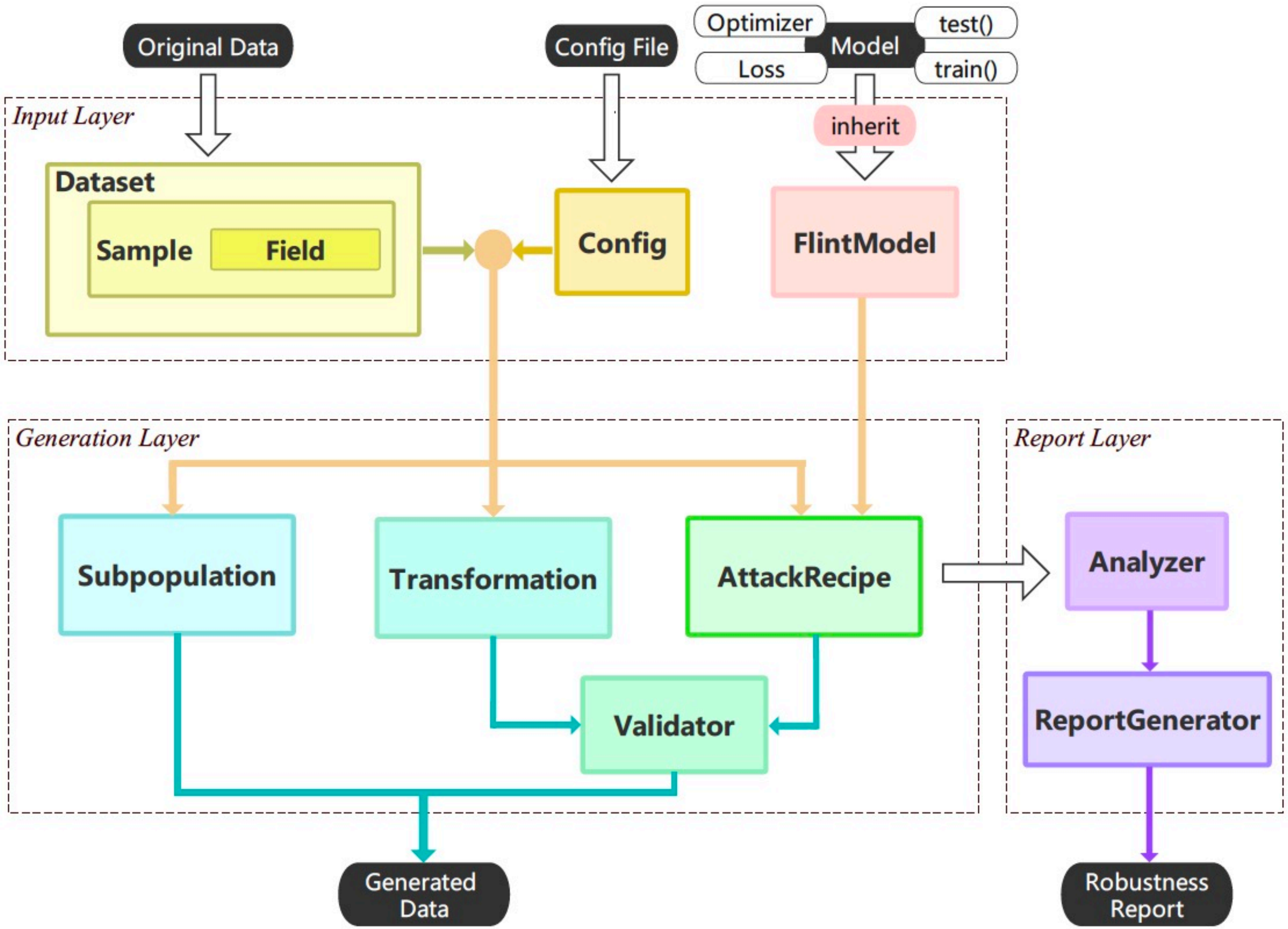
具备高的可接受度和语法正确性

分析功能 – 对评测结果给出可视化分析报告

针对性的提供数据增强

TextFlint

3 TextFlint: 自然语言处理鲁棒性分析工具集



TextFlint

通用变形

同义词

“He <u>loves</u> NLP” is transformed into “He <u>likes</u> NLP”

拼写错误

definitely → difinately	Typos
Shanghai → Shenghai	EntTypos
like → l1ke	OCR

反义词

John lives in Ireland → John <u>doesn’t</u> live in Ireland

3 TextFlint: 自然语言处理鲁棒性分析工具集

领域变形

NER: SwapNamedEnt

"He was born in <u>China</u>" → "He was born in **Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch**"

CWS: SwapVerb

看 → "看看," "看一看," "看了看," and "看了一看."

POS: SwapMultiPOS

"There is an <u>apple</u> on the desk" →
"There is an <u>imponderable</u> on the desk"

TextFlint

分组抽样

原始集合	分组抽样 - Gender
She became a nurse and worked in a hospital.	✓
I told John to come early, but he failed.	✓
The river derives from southern America.	✗
Marry would like to teach kids in the kindergarten.	✓
The storm destroyed many houses in the village.	✗

TextFlint

人工检查

- **Plausibility (Lambert et al., 2010)** measures whether the text is reasonable and written by native speakers. Sentences or documents that are natural, appropriate, logically correct, and meaningful in the context will receive a higher plausibility score. Texts that are logically or semantically inconsistent or contain inappropriate vocabulary will receive a lower plausibility score.

- **Grammaticality (Newmeyer, 1983)** measures whether the text contains syntax errors. It refers to the conformity of the text to the rules defined by the specific grammar of a language.

TextFlint

人工检查

(a) SA

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
DoubleDenial	3.26	3.37	3.59	3.49
AddSum-Person	3.39	3.32	3.76	3.59
AddSum-Movie	3.26	3.34	3.61	3.58
SwapSpecialEnt-Person	3.37	3.14	3.75	3.73
SwapSpecialEnt-Movie	3.17	3.28	3.70	3.49

(b) NER

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
OOV	3.69	3.76	3.54	3.48
SwapLonger	3.73	3.66	3.77	3.54
EntTypos	3.57	3.5	3.59	3.54
CrossCategory	3.48	3.44	3.41	3.32
ConcatSent	4.14	3.54	3.84	3.81

(c) SM

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
SwapWord	3.08	3.08	3.98	3.92
SwapNum	3.14	3.21	3.87	3.86
Overlap	—	3.33	—	4.11

(d) RE

	Plausibility		Grammaticality	
	Ort.	Trans.	Ort.	Trans.
SwapEnt-MultiType	3.59	3.36	3.97	3.94
SwapEnt-LowFreq	3.34	3.56	3.94	4.05
InsertClause	3.37	3.4	3.89	3.95
SwapEnt-AgeSwap	3.29	3.52	3.85	4.07
SwapTriplePos-BirthSwap	3.52	3.53	3.91	3.86
SwapTriplePos-EmployeeSwap	3.39	3.43	3.88	3.86

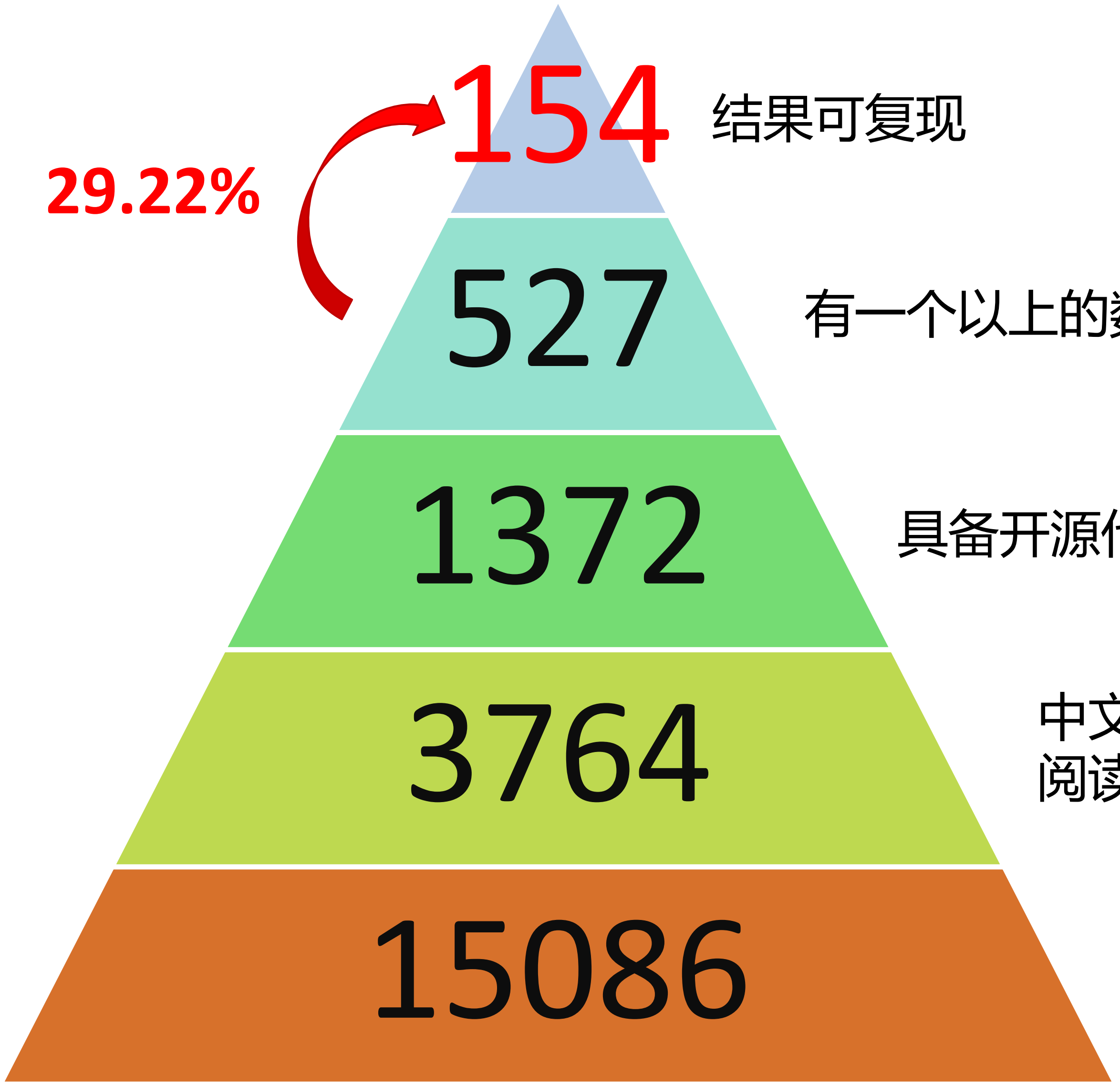
TextFlint

3 TextFlint: 自然语言处理鲁棒性分析工具集

```python
from TextFlint.engine import TextFlintEngine
from TextFlint.config.config import Config

# load the data samples
sample1 = {'x': 'Titanic is my favorite movie.', 'y': 'pos'}
sample2 = {'x': 'I don\'t like the actor Tim Hill', 'y': 'neg'}
data_samples = [sample1, sample2]

# define the transformation/subpopulation/attack types in the json config file
config = Config.from_json_file("TextFlint/common/config_files/SA/SA.json")

# define the output directory
out_dir_path = './test_result/'

# run transformation/subpopulation/attack and save the transformed data to out_dir_path in json format
engine = TextFlintEngine('SA', config_obj=config)
engine.run(data_samples, out_dir_path)
```

TextFlint

3 TextFlint: 自然语言处理鲁棒性分析工具集



154 结果可复现

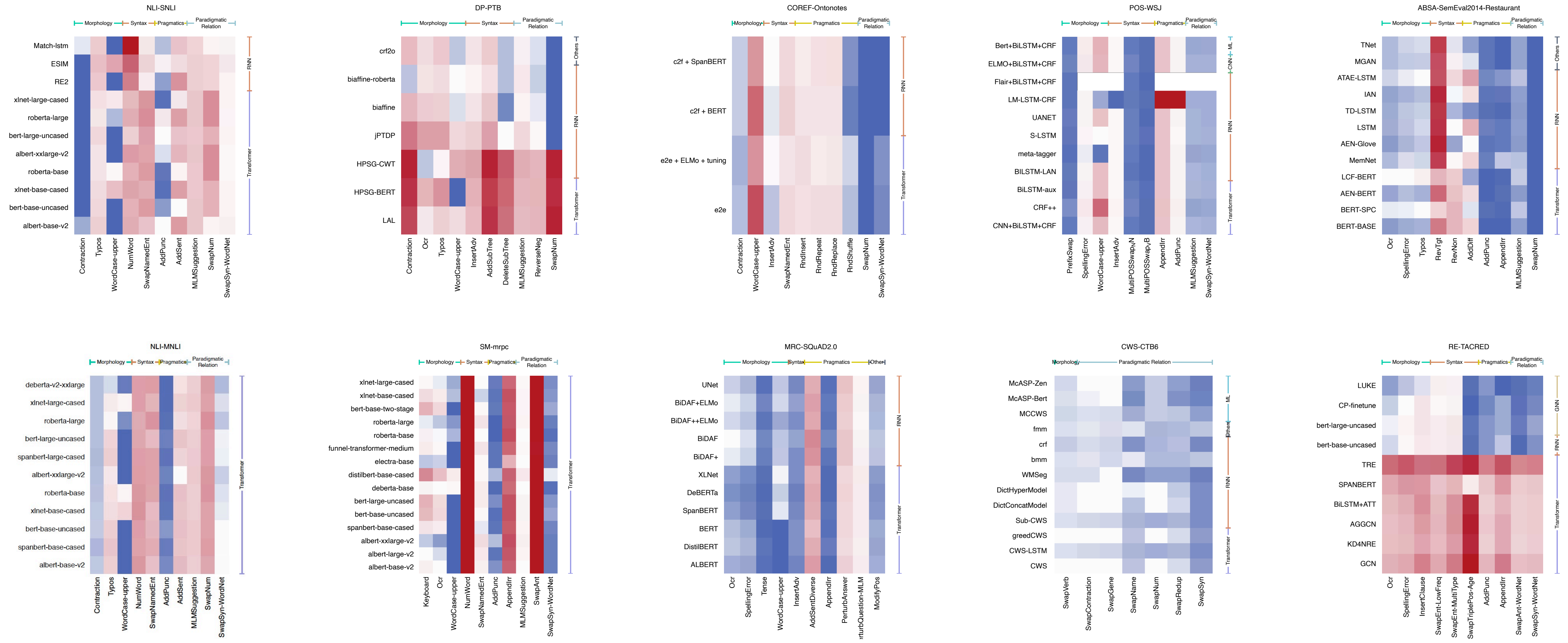
527 有一个以上的数据集集合相同

1372 具备开源代码

3764 中文分词、命名实体识别、句法分析、语义匹配、阅读理解等12个任务相关论文

15086 2011-2021年 ACL、EMNLP、NAACL、COLING、IJCAI、AAAI

3 自然语言处理模型鲁棒性分析初步结果



绝大部分任务中的大多数模型的鲁棒性都不好

TextFlint

3 自然语言处理模型鲁棒性分析初步结果

Table 10: F1 score of commercial APIs on the CoNLL 2003 dataset.

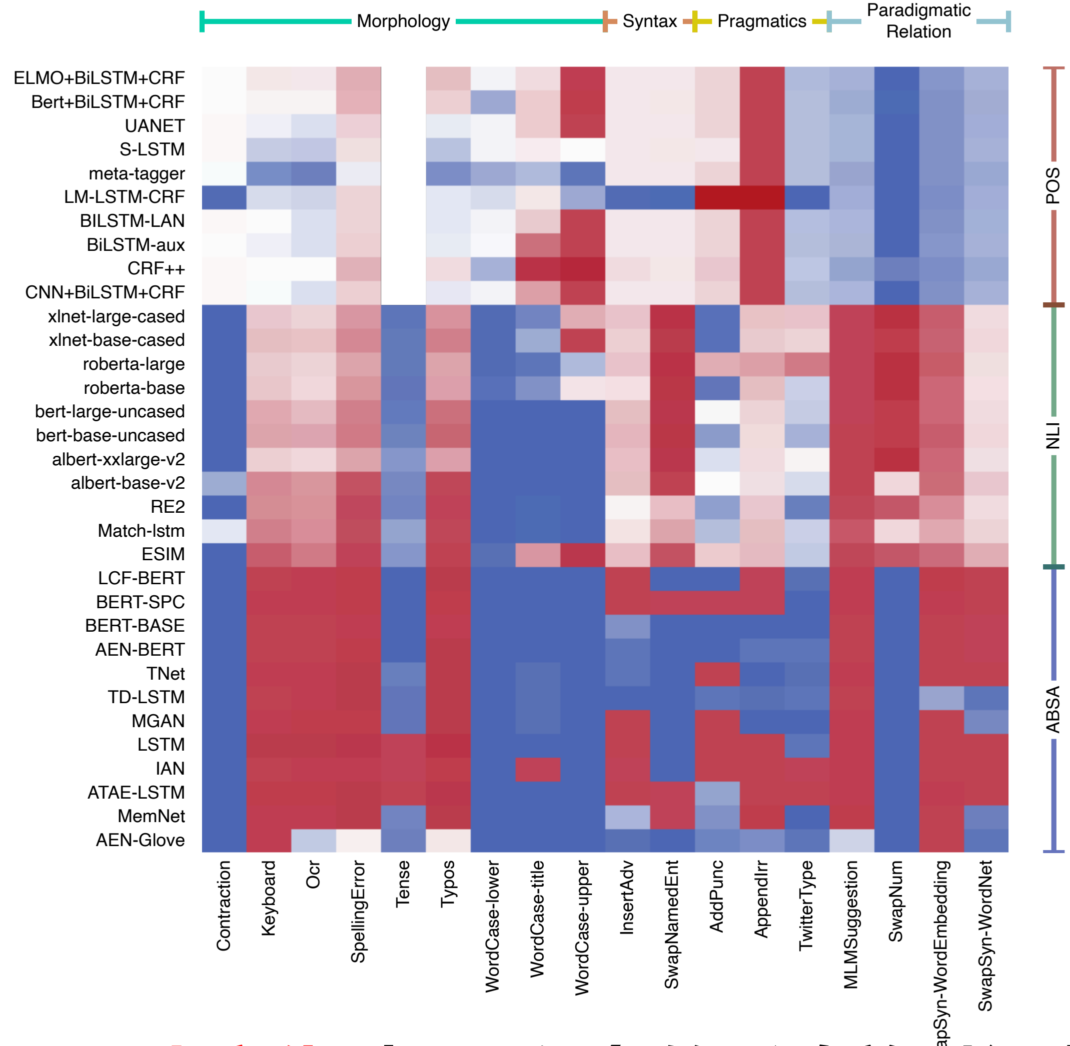
Model	*CrossCategory* Ori. → Trans.	*EntTypos* Ori. → Trans.	*OOV* Ori. → Trans.	*SwapLonger* Ori. → Trans.
CoNLL 2003				
Amazon	69.68 → 33.01	70.19 → 65.98	69.68 → 56.27	69.68 → 57.63
Google	59.14 → 28.30	62.41 → 50.87	59.14 → 48.53	59.14 → 53.40
Microsoft	82.69 → 43.37	83.42 → 78.47	82.69 → 60.18	82.69 → 52.51
Average	70.50 → 34.89	72.01 → 65.11	70.50 → 54.99	70.50 → 54.51

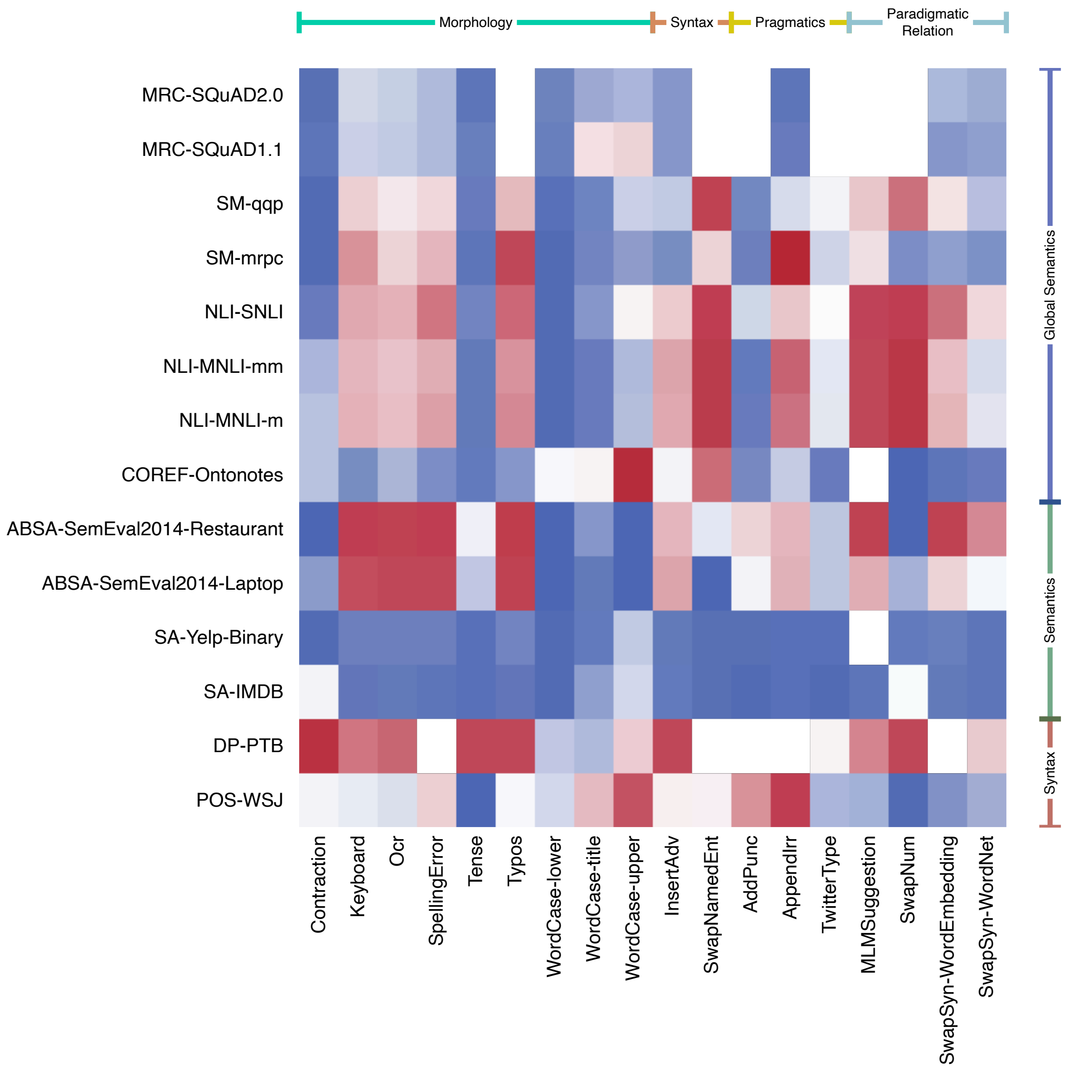
Gui, Tao, et al. "Textflint: Unified multilingual robustness evaluation toolkit for natural language processing." *arXiv preprint arXiv:2103.11441* (2021).

大厂商用 Open API Platform 也有类似的问题

TextFlint

3 自然语言处理模型鲁棒性分析初步结果

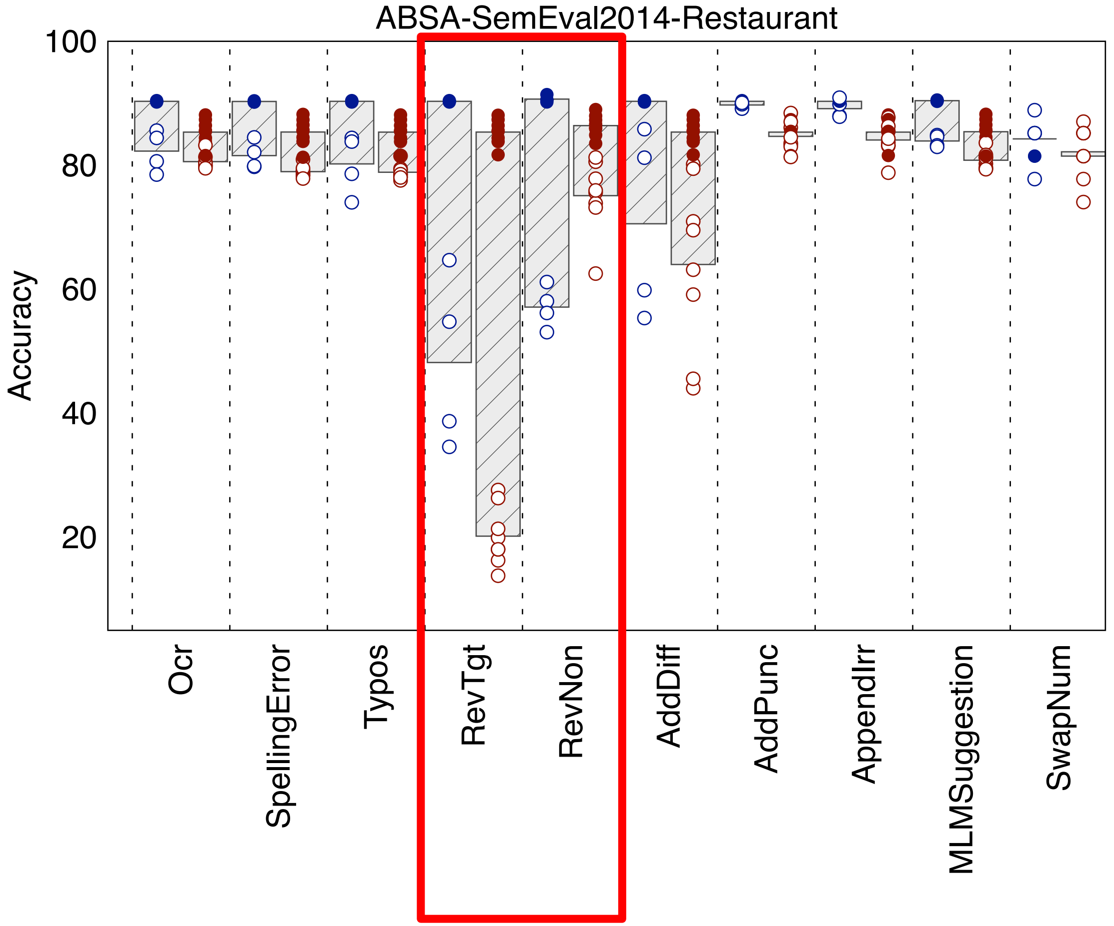




同一个变化对不同任务的影响差别很大

TextFlint

3 自然语言处理模型鲁棒性分析初步结果



Data Setting	PER	ORG	GPE	FAC	LOC	WEA	VEH	ALL
Baseline	86.31	76.49	80.89	69.23	40.58	74.70	61.97	81.76
Name Permutation	73.41	44.34	49.71	37.96	28.24	33.33	23.93	62.28
- Drop Compared with Baseline	15%	42%	39%	45%	44%	55%	61%	24%
Mention Permutation	61.78	39.40	33.27	32.16	18.60	9.38	21.92	51.58
- Drop Compared with Baseline	28%	48%	59%	54%	54%	87%	65%	34%

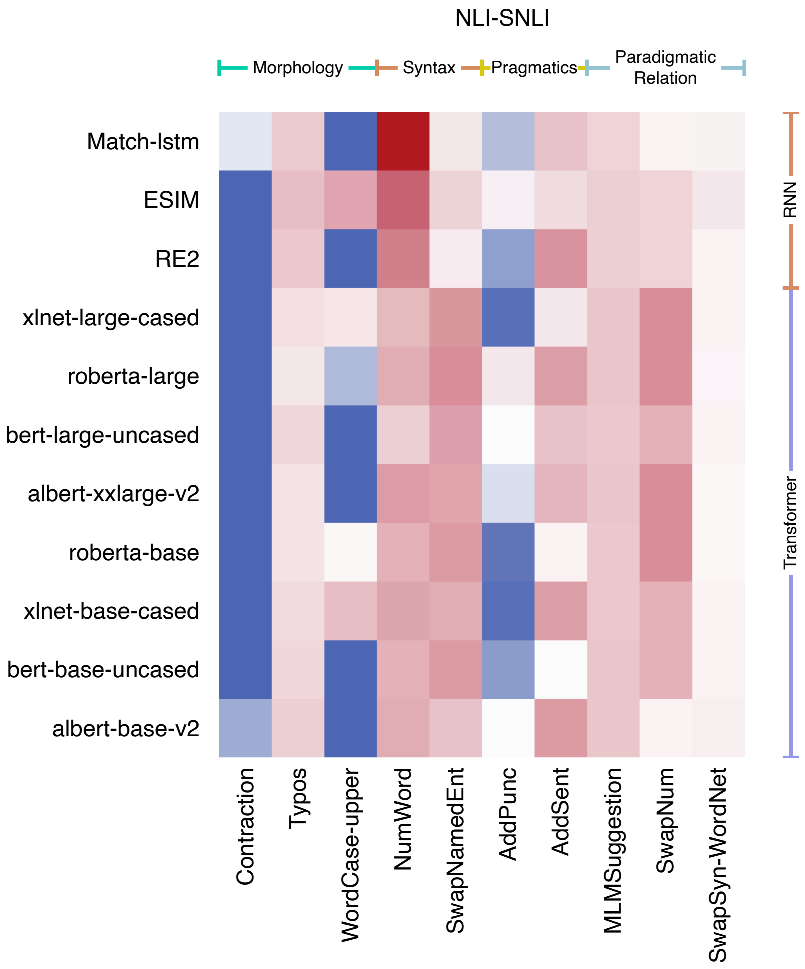
Table 2: Micro-F1 scores of BERT-CRF tagger on original data, name permutation setting and mention permutation setting respectively. We can see that erasing name regularity and mention coverage will significantly undermine the model performance.

Lin et al., *A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land?*, EMNLP 2020

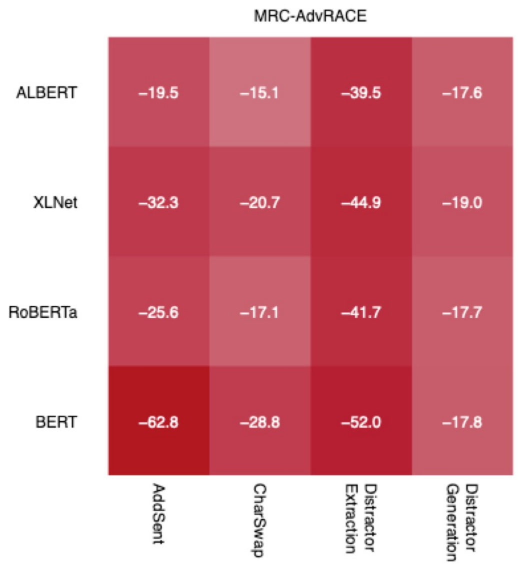
仅数据驱动，模型很难学习到任务特性

TextFlint

3 自然语言处理模型鲁棒性分析初步结果



	BERT	RoBERTa	XLNet	ALBERT	*Average*	Valid	Correct
Original	68.5	83.7	79.9	86.0		100.0%	100.0%
AddSent	25.5 (*-62.8%*)	62.3 (*-25.6%*)	54.1 (*-32.3%*)	69.2 (*-19.5%*)	*-35.1%*	98.0%	89.8%
CharSwap	48.8 (*-28.8%*)	69.4 (*-17.1%*)	63.4 (*-20.7%*)	73.0 (*-15.1%*)	*-20.4%*	100.0%	94.0%
Distractor Extraction	32.9 (*-52.0%*)	48.8 (*-41.7%*)	44.0 (*-44.9%*)	52.0 (*-39.5%*)	*-44.5%*	98.0%	95.9%
Distractor Generation	56.3 (*-17.8%*)	68.9 (*-17.7%*)	64.7 (*-19.0%*)	70.9 (*-17.6%*)	*-18.0%*	98.0%	93.9%
Average	40.9 (*-40.3%*)	62.4 (*-25.4%*)	56.6 (*-29.2%*)	66.3 (*-22.9%*)			

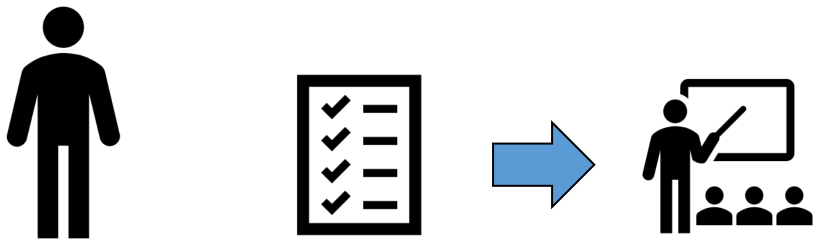


Si et al. Benchmarking Robustness of Machine Reading Comprehension Model, ACL 2021

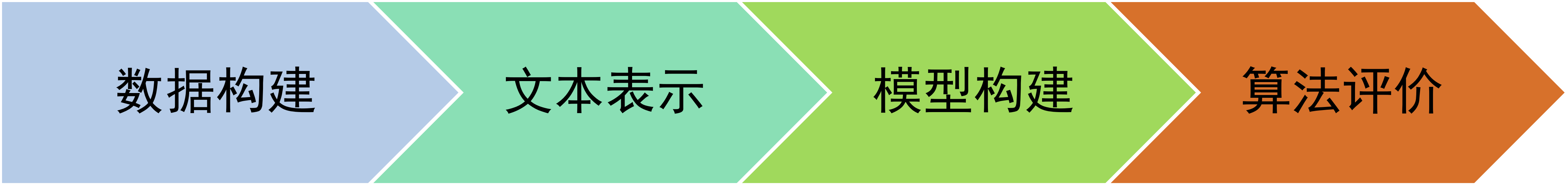
深度学习真的能解决推理类的任务吗？

TextFlint

0 提升自然语言处理算法鲁棒性是个系统工程



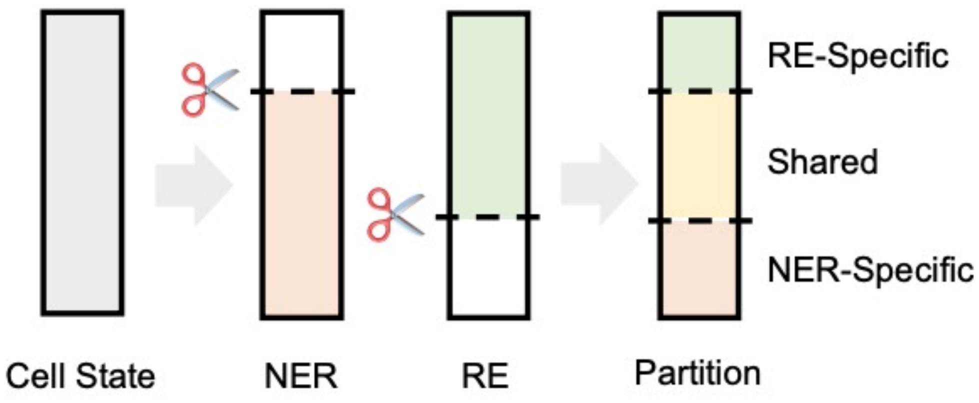
任务建模

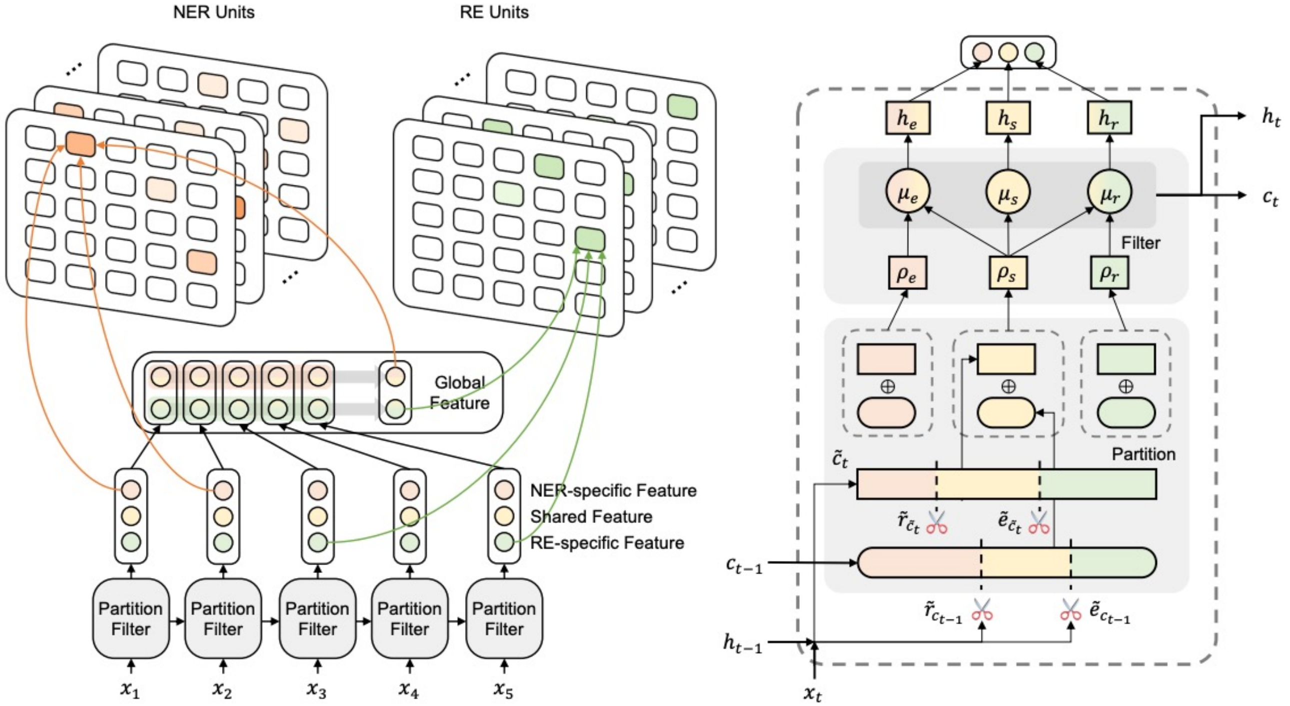


每个环节都会对模型的鲁棒性产生影响

根据任务特性驱动模型设计是个值得思考的问题

4 关系抽取与实体抽取联合训练





(a) Framework of Partition Filter Network

(b) Inner Mechanism of Partition Filter

Yan et al., *A Partition Filter Network for Joint Entity and Relation Extraction*, EMNLP 2021

4 关系抽取与实体抽取联合训练

Method	NER	RE
NYT △		
CopyRE (Zeng et al., 2018)	86.2	58.7
GraphRel (Fu et al., 2019)	89.2	61.9
CopyRL (Zeng et al., 2019)	-	72.1
Casrel (Wei et al., 2020) †	(93.5)	89.6
TpLinker (Wang et al., 2020b) †	-	91.9
PFN†	**95.8**	**92.4**
WebNLG △		
CopyRE (Zeng et al., 2018)	82.1	37.1
GraphRel (Fu et al., 2019)	91.9	42.9
CopyRL (Zeng et al., 2019)	-	61.6
Casrel (Wei et al., 2020) †	(95.5)	91.8
TpLinker (Wang et al., 2020b) †	-	91.9
PFN†	**98.0**	**93.6**
ADE ▲		
Multi-head (Bekoulis et al., 2018b)	86.4	74.6
Multi-head + AT (Bekoulis et al., 2018a)	86.7	75.5
Rel-Metric (Tran and Kavuluru, 2019)	87.1	77.3
SpERT (Eberts and Ulges, 2019) †	89.3	79.2
Table-Sequence (Wang and Lu, 2020) ‡	89.7	80.1
PFN†	89.6	80.0
PFN‡	**91.3**	**83.2**

Model	ConcatSent		CrossCategory		EntTypos		OOV		SwapLonger		Average
	Ori → Aug	Decline	Ori → Aug	Decline	Ori → Aug	Decline	Ori → Aug	Decline	Ori → Aug	Decline	Decline
BiLSTM-CRF	83.0→82.2	0.8	82.9→43.5	39.4	82.5→73.5	9.0	82.9→64.2	18.7	82.9→67.7	15.2	16.6
BERT-base(cased)	87.3→86.2	1.1	87.4→48.1	39.3	87.5→83.1	4.1	87.4→79.0	8.4	87.4→82.1	5.3	11.6
BERT-base(uncased)	88.8→88.7	**0.1**	88.7→46.0	42.7	89.1→83.0	6.1	88.7→74.6	14.1	88.7→78.5	10.2	14.6
TENER	84.2→83.4	0.8	84.7→39.6	45.1	84.5→76.6	7.9	84.7→51.5	33.2	84.7→31.1	53.6	28.1
Flair	85.5→85.2	0.3	84.6→44.9	39.7	86.1→81.5	4.6	84.6→81.3	**3.3**	84.6→73.1	11.5	11.9
PFN	89.1→87.9	1.2	89.0→80.5	**8.5**	89.6→86.9	**2.7**	89.0→80.4	8.6	89.0→84.3	**4.7**	**5.1**

Table 4: Robustness test of NER against input perturbation in ACE05, baseline results and test files are copied from `https://www.textflint.io/`

Yan et al., *A Partition Filter Network for Joint Entity and Relation Extraction*, EMNLP 2021

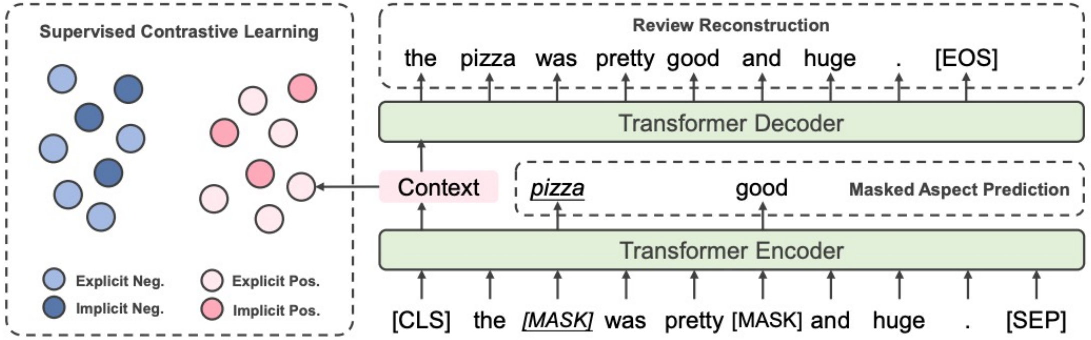
TextFlint

4 针对情感倾向分析任务构建针对性预训练模型

Reviews contain implicit sentiment
The **waiter** poured water on my hand and walked away
The **bartender** continued to pour champagne from his reserve
10 hours of **battery life** ...
The **battery life** is probably an hour

Dataset	Positive	Neutral	Negative	Total	Implicit Sentiment %
Restaurant-train	2164	805	633	3602	28.59
Restaurant-test	728	196	196	1120	23.84
Restaurant	2892	1001	829	4722	27.47
Laptop-train	987	866	460	2313	30.87
Laptop-test	341	128	169	638	27.27
Laptop	1328	994	629	2951	30.09
MAMS	4183	6253	3418	13854	-
YELP	1.17M	-	0.39M	1.56M	-
Amazon	0.38M	-	0.13M	0.51M	-

Table 2: Statistics on three datasets of ABSA and two external corpus for SCAPT.



SCAPT to align the representation of explicit and implicit sentiment expressions with the same emotion.

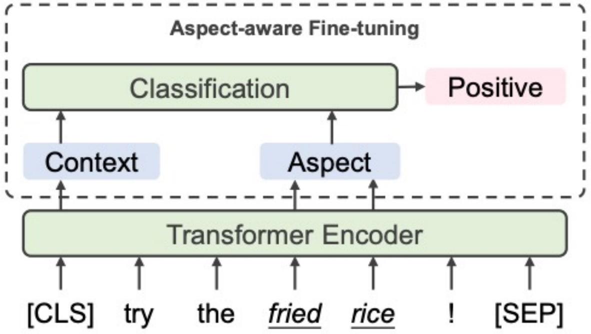


Figure 2: Aspect-aware fine-tuning on Transformer encoder based models. Sentiment representation and aspect-based representation are taken into account in sentiment classification.

Li et al., *Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training*, EMNLP 2021

TextFlint

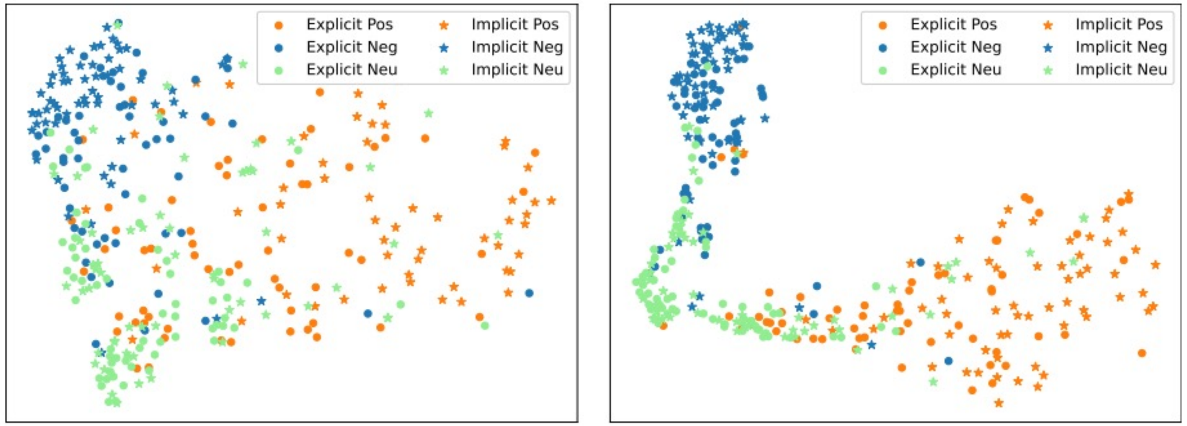
4 针对情感倾向分析任务构建针对性预训练模型

Method		Restaurant				Laptop			
		Acc.	F1	ESE	ISE	Acc.	F1	ESE	ISE
Attention	ATAE-LSTM (Wang et al., 2016a)	76.90*	62.64*	84.16	53.71	65.37*	62.92*	75.69	37.86
	IAN (Ma et al., 2017)	76.88*	67.71*	86.52	46.07	67.24*	63.72*	75.86	44.25
	RAM (Chen et al., 2017)	80.23	70.80	85.11	55.81	74.49	71.35	75.86	44.25
	MGAN (Fan et al., 2018)	81.25	71.94	85.18	60.04	75.39	72.47	76.16	56.31
GNN	ASGCN (Zhang et al., 2019)	80.77	72.02	84.29	62.91	75.55	71.05	75.46	57.77
	BiGCN (Zhang and Qian, 2020)	81.97	73.48	87.19	59.05	74.59	71.84	79.53	62.64
	CDT (Sun et al., 2019)	82.30	74.02	88.79	65.87	77.19	72.99	77.53	68.90
	RGAT (Wang et al., 2020)	83.30	76.08	89.45	61.05	77.42	73.76	80.17	65.52
Knowledge Enhanced	TransCap (Chen and Qian, 2019)	79.55	71.41	86.52	59.93	73.87	70.10	77.16	60.34
	BERT-SPC (Devlin et al., 2019)	83.57*	77.16*	89.21	65.54	78.22*	73.45*	81.47	69.54
	CapsNet+BERT (Jiang et al., 2019)	85.09*	77.75*	91.68	64.04	78.21*	73.34*	82.33	67.24
	BERT-PT (Xu et al., 2019)	84.95	76.96	92.15	64.79	78.07	75.08	81.47	71.27
	BERT-ADA (Rietzler et al., 2020)	87.14	80.05	94.14	65.92	78.96	74.18	82.76	70.11
	R-GAT+BERT (Wang et al., 2020)	86.60	81.35	92.73	67.79	78.21	74.07	82.44	72.99
Ours	TransEncAsp	77.10	57.92	86.97	48.96	65.83	59.53	74.31	43.20
	BERTAsp	85.80	78.95	92.73	63.67	78.53	74.07	82.33	68.39
	BERTAsp+CEPT	87.50	82.07	93.67	67.79	81.66	78.38	83.84	75.86
	TransEncAsp+SCAPT	83.39	74.53	88.04	68.55	77.17	73.23	78.70	72.82
	BERTAsp+SCAPT	**89.11**	**83.79**	**94.37**	**72.28**	**82.76**	**79.15**	**84.70**	**77.59**

TextFlint

Li et al., *Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training*, EMNLP 2021

4 针对情感倾向分析任务构建针对性预训练模型



(a) BERTAsp (b) BERTAsp+SCAPT

Figure 3: Visualization of the hidden sentiment representations on Restaurant (best to view the colored version). BERTAsp+SCAPT tightly clusters the representations of both explicit and implicit sentiment expressions.

Method	Restaurant-test		Laptop-test	
	Ori → New	Decline	Ori → New	Decline
LSTM	75.98→14.64	-61.34	67.55→9.87	-57.68
ASGCN	77.86→24.73	-53.13	72.41→19.91	-52.50
CapsNet+BERT	83.48→55.36	-28.12	77.12→25.86	-51.46
BERT	83.04→54.82	-29.22	77.59→50.94	-26.65
BERT-PT	86.70→59.29	-27.41	78.53→53.29	-25.24
TransEncAsp+SCAPT	83.39→67.76	-15.63	76.80→52.52	-24.28
BERTAsp+SCAPT	**89.11→80.06**	**-9.05**	**82.76→76.13**	**-6.63**

Table 6: Model performance on aspect robustness test sets. We compare the model accuracy on the original and new test sets, and the decline of prediction on new examples are reported.

TextFlint

Li et al., *Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training*, EMNLP 2021

在深度学习模型黑盒下任务特性驱动模型设计

谢谢！

Watch Star Fork