Really Reaching Human Parity?
- Addressing Benchmark Issues on Robustness, Bias and Metric.

Qi Zhang (Fudan University), Nan Duan (Microsoft Research Asia), Ming Zhou (Sinovation)

ACL-2021 Workshop on Benchmarking: Past, Present and Future

Benchmarks are significant for the growth of AI research.

NIST

Search NIST

Menu

Information Technology Laboratory / Information Access Division

MULTIMODAL INFORMATION GROUP

Tools

Past HLT Evaluation Projects

Staff

Open Machine Translation Evaluation

Open Machine Translation Evaluation (OpenMT)

The objective of the NIST Open Machine Translation (OpenMT) evaluation series is to support research in, and help advance the state of the art of, machine translation (MT) technologies - technologies that translate text between human languages. Input may include all forms of text. The goal is for the output to be an adequate and fluent translation of the original.

The MT evaluation series started in 2001 as part of the DARPA TIDES program. In their current form, the evaluations are driven and coordinated by NIST as NIST OpenMT. They provide an important contribution to the direction of research efforts and the calibration of technical capabilities in MT. The OpenMT evaluations are intended to be of interest to all researchers working on the general problem of automatic translation between human languages. To this end, they are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible to all those wishing to participate. The most recently completed NIST OpenMT evaluation was MT09 and took place in June 2009. MT09 featured three language pairs, the second cycle of a progress test, and, for the first time, system combination categories. Results of past NIST OpenMT and DARPA TIDES MT evaluations as well as resources specific to each evaluation can be accessed via the year-specific links at the bottom.

OpenMT introduced an MT Challenge that began in the fall of 2015.

Contact

Email mt_poc@nist.gov for with questions for NIST related to MT.

To request to be added to NIST's MT mailing list, email mt_list+subscribe@list.nist.gov .

[2001] [2002] [2003] [2004] [2005] [2006] [2008] [2009] [2012] [2015] [2015 challenge]

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

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 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
4 Sep 11, 2020	EntitySpanFocus+AT (ensemble) *RICOH_SRCB_DML*	90.454	92.748

GLUE

DNNs even achieved human parity on some famous ones!

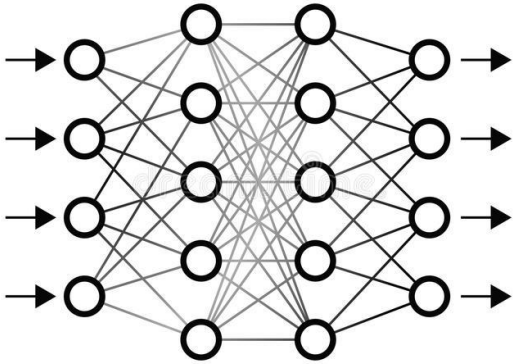
SuperGLUE GLUE | Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

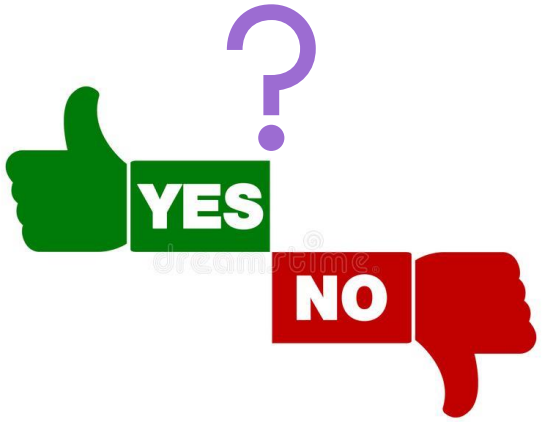
	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
+	1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	4	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	5	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	6	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	7	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5
+	8	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	9	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	10	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	11	Anuar Sharafudinov	AILabs Team, Transformers		82.6	88.1	91.6/94.8	86.8	85.1/54.7	82.8/79.8	88.9	74.1	78.8	100.0	100.0/100.0
	12	Rakesh Radhakrishnan Menon	ADAPET (ALBERT) - few-shot		76.0	80.0	82.3/92.0	85.4	76.2/35.7	86.1/85.5	75.0	53.5	85.6	-0.4	100.0/50.0
+	13	Timo Schick	iPET (ALBERT) - Few-Shot (32 Examples)		75.4	81.2	79.9/88.8	90.8	74.1/31.7	85.9/85.4	70.8	49.3	88.4	36.2	97.8/57.9
	14	Adrian de Wynter	Bort (Alexa AI)				.4	89.6	83.7/54.1	49.8/49.0	81.2	70.1	65.8	48.0	96.1/61.5

Click on a submission to see more information

But are these benchmarks good enough for evaluating AI?



Neural Networks



Human

Take our dataset (i.e., XGLUE) as an example.

XGLUE

Home Intro **Leaderboard** Contact

XGLUE Dataset and Leaderboard

Tasks

1. NER
2. POS Tagging (POS)
3. News Classification (NC)
4. MLQA
5. XNLI
6. PAWS-X
7. Query-Ad Matching (QADSM)
8. Web Page Ranking (WPR)
9. QA Matching (QAM)
10. Question Generation (QG)
11. News Title Generation (NTG)

New Tasks!

Relevant Links

XGLUE Submission Guideline/Github | XGLUE Paper | Unicoder Paper(Baseline)

Leaderboard (05/25/2020-Present) ranked by XGLUE Score (average score on 11 tasks)

Rank	Model	Submission Date	NER	POS	NC	MLQA	XNLI	PAWS-X	QADSM	WPR	QAM	QG	NTG	XGLUE Score
1	**Unicoder Baseline** (XGLUE Team)	2020-05-25	79.7	79.6	83.5	66.0	75.3	90.1	68.4	73.9	68.9	10.6	10.7	64.2

https://microsoft.github.io/XGLUE/

(1) Robustness issue in XGLUE.

- DNA profiling is today possible with even very small quantities of blood: this is commonly used in forensic science, but is now also part of the diagnostic process of many disorders…
  - **Original question**: where is DNA used today? -- forensic science
  - **Modified question**: where is RNA used today? -- forensic science
  - **Modified question**: where is DNA not used today? -- forensic science

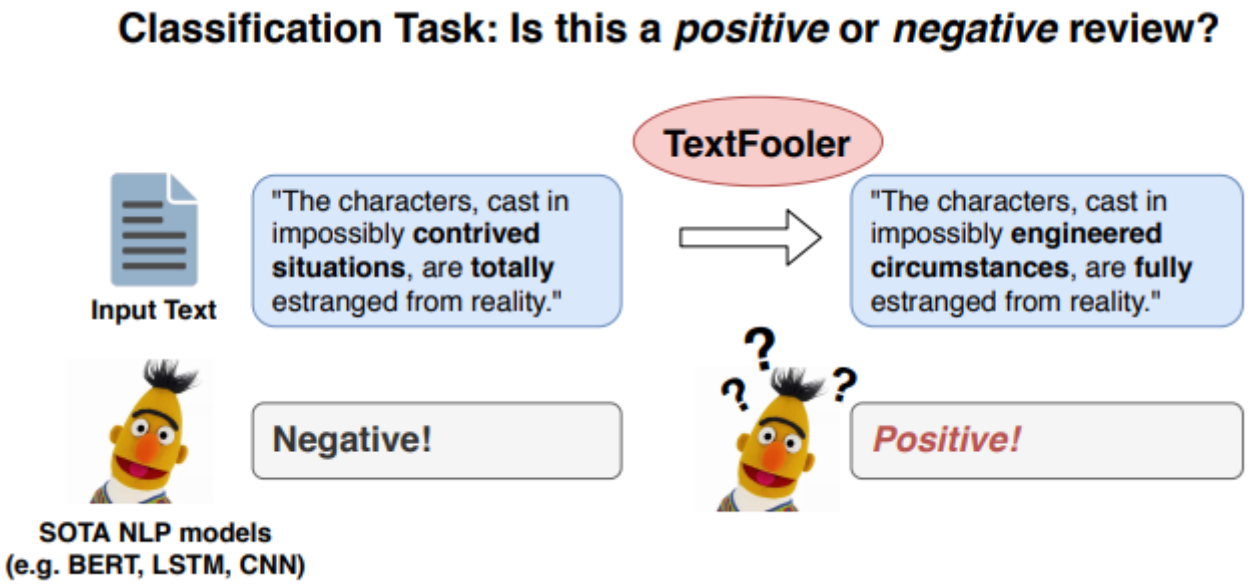
- In 1975, Bill Gates co-founded Microsoft with childhood friend Paul Allen in Albuquerque, New Mexico. It became the world's largest personal computer software company.
  - **Original question**: Who founded Microsoft? -- Bill Gates
  - **Modified question**: Who founded Apple? -- Bill Gates

- Donald Knuth argues that ternary computers will be brought back into development in the future to take advantage of ternary logic's elegance and efficiency…
  - **Original question**: Who is a proponent of ternary computers? -- Donald Knuth
  - **Modified question**: Who is an opponent of ternary computers? -- Donald Knuth

Similar findings in other datasets.

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

[1707.07328] Adversarial Examples for Evaluating Reading Comprehension Systems (arxiv.org)



[1907.11932] Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment (arxiv.org)

(2) Bias issue in XGLUE.

- Biased questions type distribution in QA Matching (QAM)

Question Type	How	What	Where	Which	Who	Why
# of questions	16,595	14,981	10,010	5,601	11,602	515

- Biased news article class distribution in News Classification (NC)

Class Names	Sports	Finance	News	Autos	Video	Travel	Lifestyle	Food & Drink	Health	Entertain ment
# of train instances	29,232	12,488	24,254	3,329	11,066	2,470	6,001	2,544	5,225	3,391

Similar findings in other datasets.

Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).	✓	✓	✓	✓

Table 1: Following the talk by Crawford (2017), we categorize representation bias in NLP tasks into the following four categories: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation.

[1906.08976] Mitigating Gender Bias in Natural Language Processing: Literature Review (arxiv.org)

(3) Metric issue in XGLUE.

(news title generation task)

- **INPUT**: Three years ago, a man nicknamed "Murder" was charged in connection with a deadly home invasion in Bradenton. On Monday, Courtney Lawrence, 30, who also has "Murder" tattooed on his neck, pleaded guilty to second-degree murder with a firearm, attempted second-degree murder with a firearm, armed burglary of a dwelling with a firearm and possession of a firearm by a convicted felon. The Bradenton Herald reported in May 2016 that four or five armed men stormed a home and demanded money from the 38-year-old owner. Deputies say the homeowner and the gunman exchanged gunfire. The homeowner and one of the alleged gunmen, Emanuel Johnson, were injured in the shootout, and Johnson was later pronounced dead at a local hospital. The Miami Herald said Lawrence was identified as one of the alleged gunmen when he was dropped off at a local hospital within 30 minutes of the shooting. The Herald said he had gunshot wounds on his buttocks and an injury to his shoulder. An investigation from the Manatee County Sheriff's Office said Lawrence admitted he had been shot during an armed robbery and that an accomplice had been killed. Lawrence pleaded guilty Monday and was sentenced to 40 years in prison. What other people are reading right now: 85-year-old Florida man accused of murdering 90-year-old lover of 60 years Assisted living facility worker reportedly left residents alone to go clubbing Man missing after being carjacked by armed bank robber, deputies say 5-year-old SC girl missing after mom found dead in-home Ex-cardinal's letters show signs of grooming victims for abuse, experts say ► Make it easy to keep up-to-date with more stories like this. Download the 10News app now . Have a news tip? Email desk@wtsp.com or visit our Facebook page or Twitter feed .

- **GOLDEN**: Florida man nicknamed 'Murder' pleads guilty to murder charges

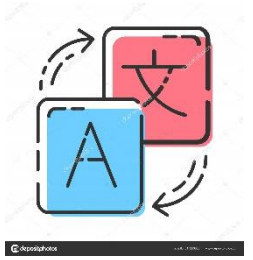
- **GENERATED**: Man with 'Murder' tattoo admit his guilt in deadly home invasion

Similar findings in other datasets.

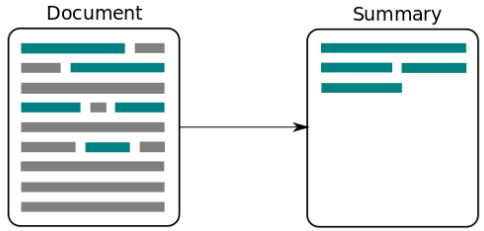
$$BLEU = \min\left(1, \frac{|prediction|}{|ref|}\right) \cdot \left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}}$$

doesn't consider meaning

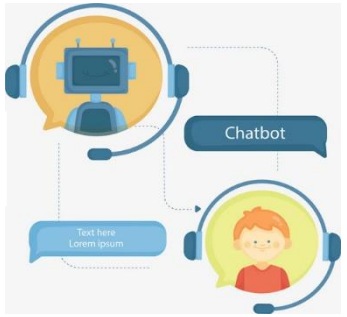
Machine Translation



Text Summarization



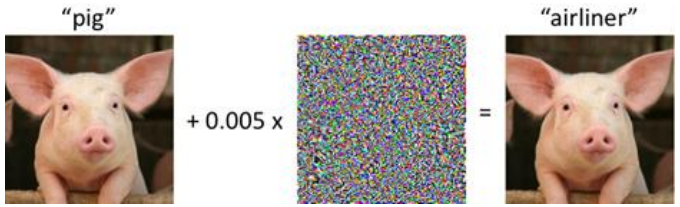
Response Generation



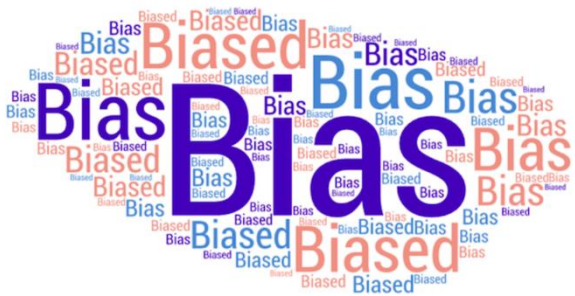
...

...

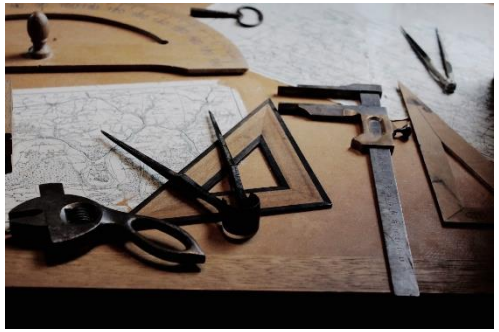
3 Issues of Current Benchmarks



Robustness Measurement

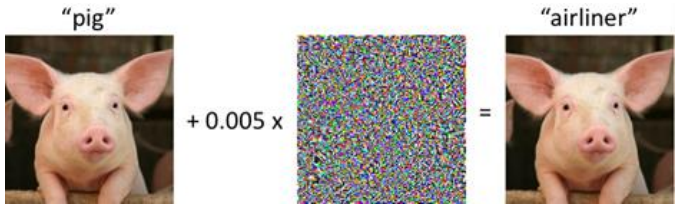


Data Bias



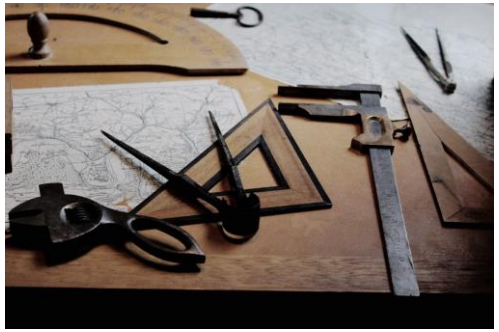
Evaluation Metric

3 Issues of Current Benchmarks



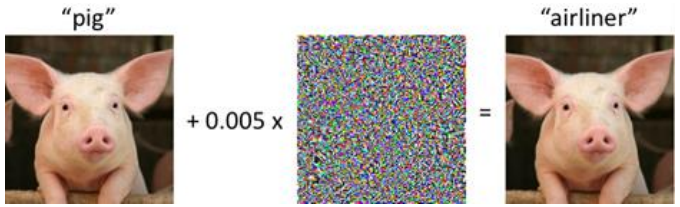
Robustness Measurement

Data Bias



Evaluation Metric

3 Issues of Current Benchmarks

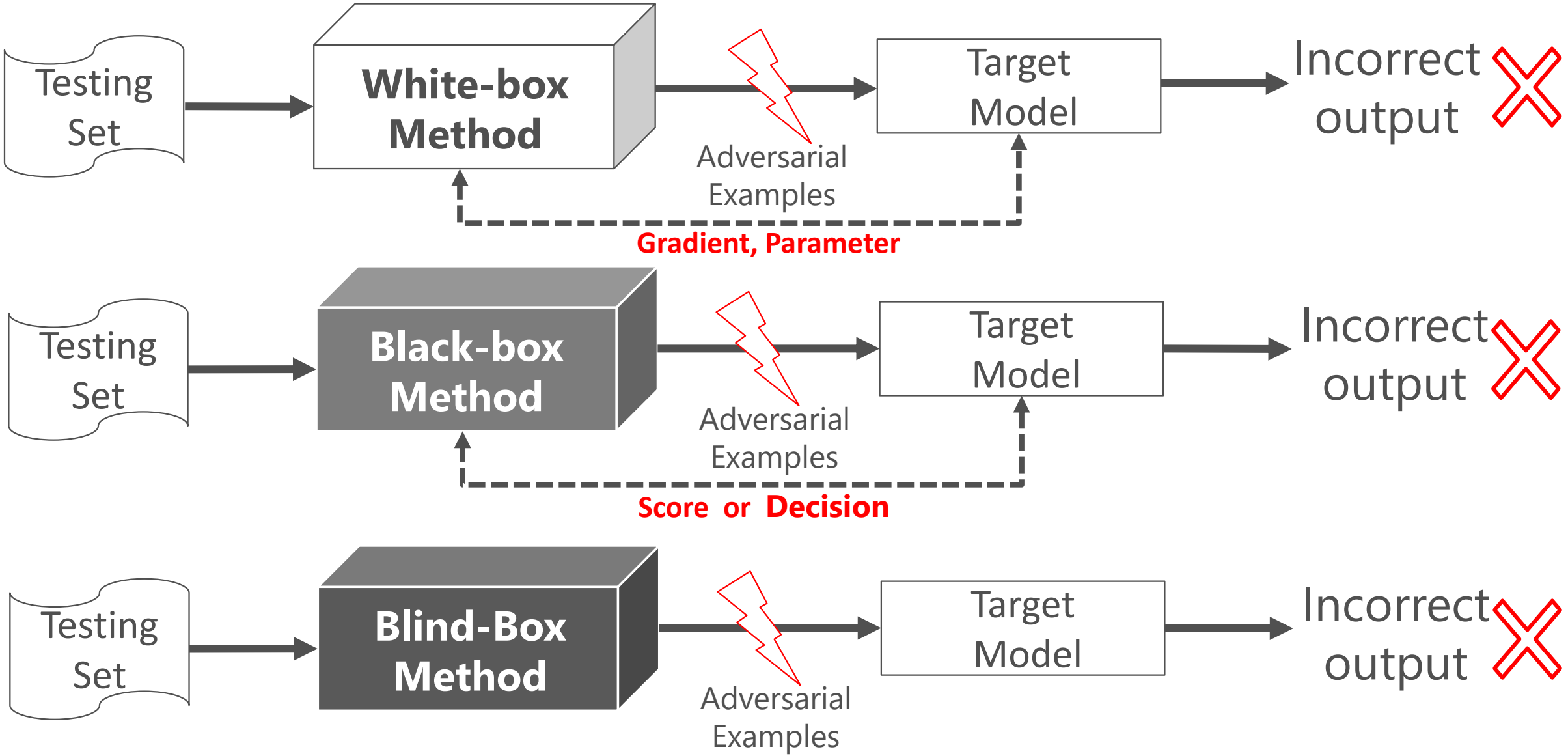


Robustness Measurement

Data Bias

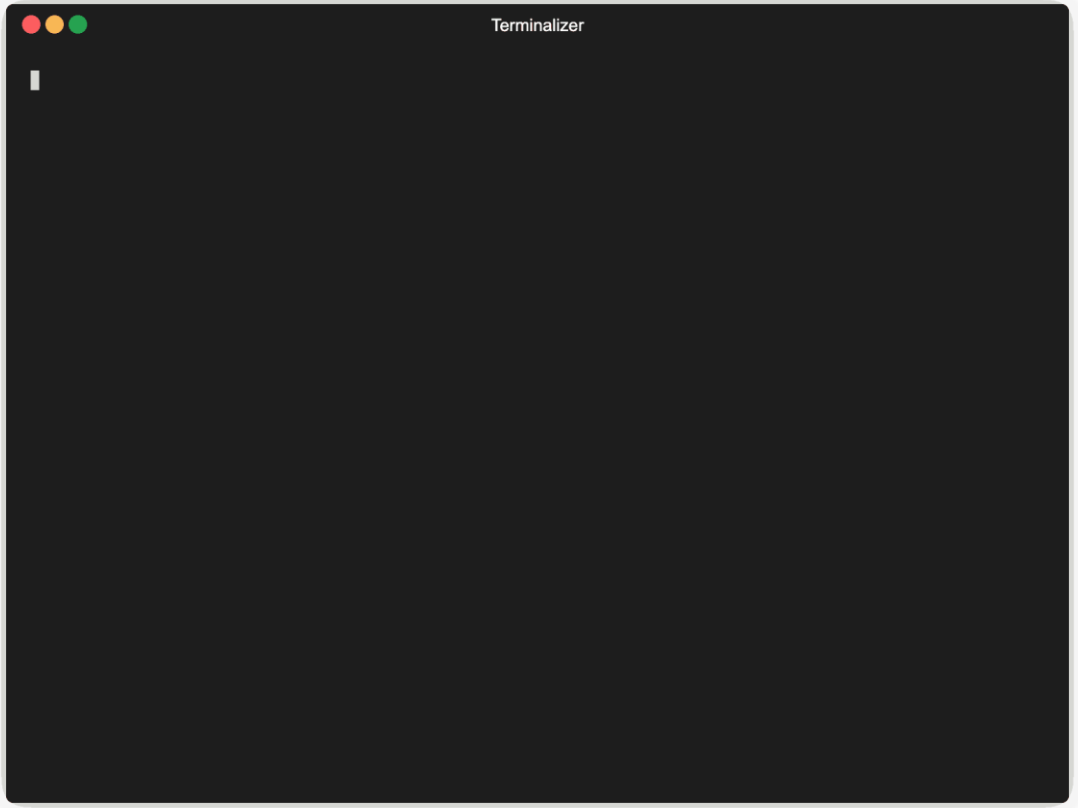
Evaluation Metric

Some Efforts to Robustness Measurement



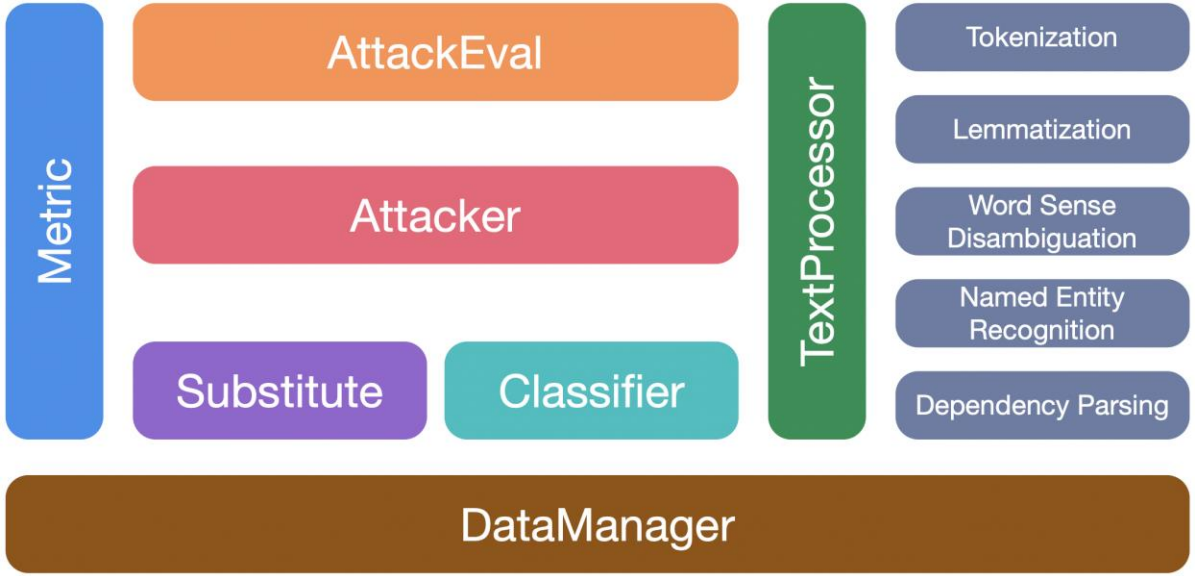
Some Efforts to Robustness Measurement

TextAttack



Morris et al. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*, EMNLP 2020

OpenAttack



https://github.com/thunlp/OpenAttack

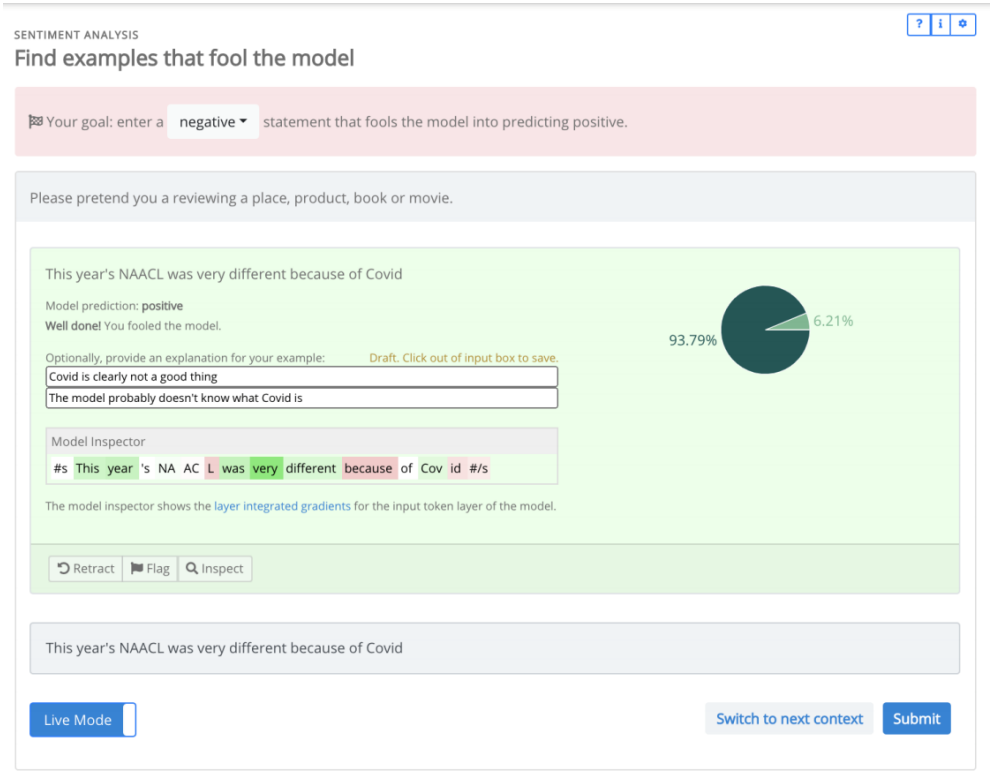
Some Efforts to Robustness Measurement

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%			

Figure 1: CheckListing a commercial sentiment analysis model (Google). Tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

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

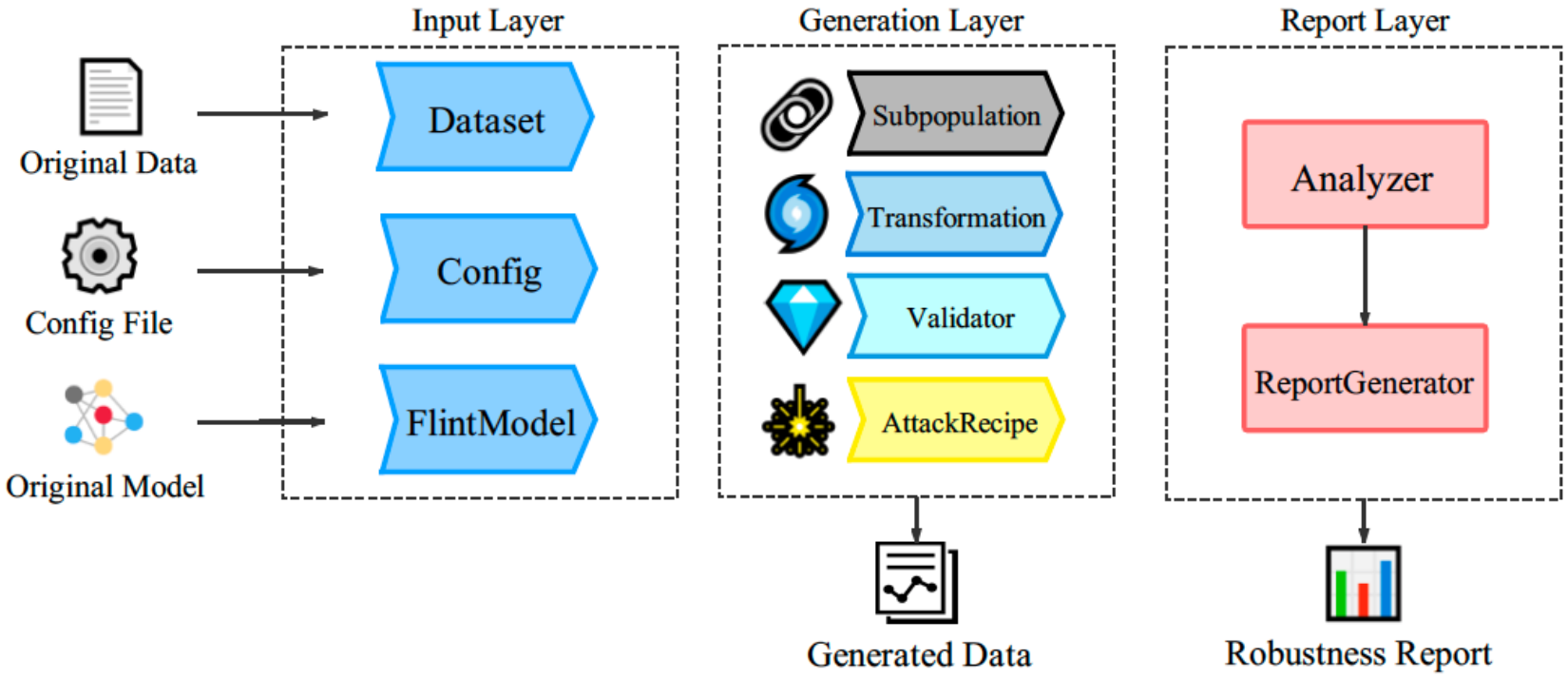


FACEBOOK AI

Dyna Bench

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

Some Efforts to Robustness Measurement



Transformation

Original	Tasty **burgers**, and crispy fries. (Target aspect: burgers)
RevTgt	Terrible **burgers**, but crispy fries.
RevNon	Tasty **burgers**, but soggy fries.
Typos	Tatsy burgers, and cripsy fries.

Adversarial attack

Original	Premise: Some rooms have balconies. Hypothesis: All of the rooms have balconies.	Contradiction
Adv	Premise: Many rooms have balconies. Hypothesis: All of the rooms have balconies.	Neutral

Subpopulation

Original Set	Subpopulation - 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.	✗

12 NLP Task
24 Classic Datasets
20 General transformations
60 Task-specific transformations

TextFlint

Wang et al., *TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing*, ACL 2021

Some Efforts to Robustness Measurement

Transformation - General

Synonym

"He loves NLP" --> "He likes NLP"

Antonym

John lives in Ireland → John doesn't live in Ireland

Spelling Error

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

Wang et al., *TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing*, ACL 2021

Some Efforts to Robustness Measurement

Transformation – Domain Specific

NER: SwapNamedEnt

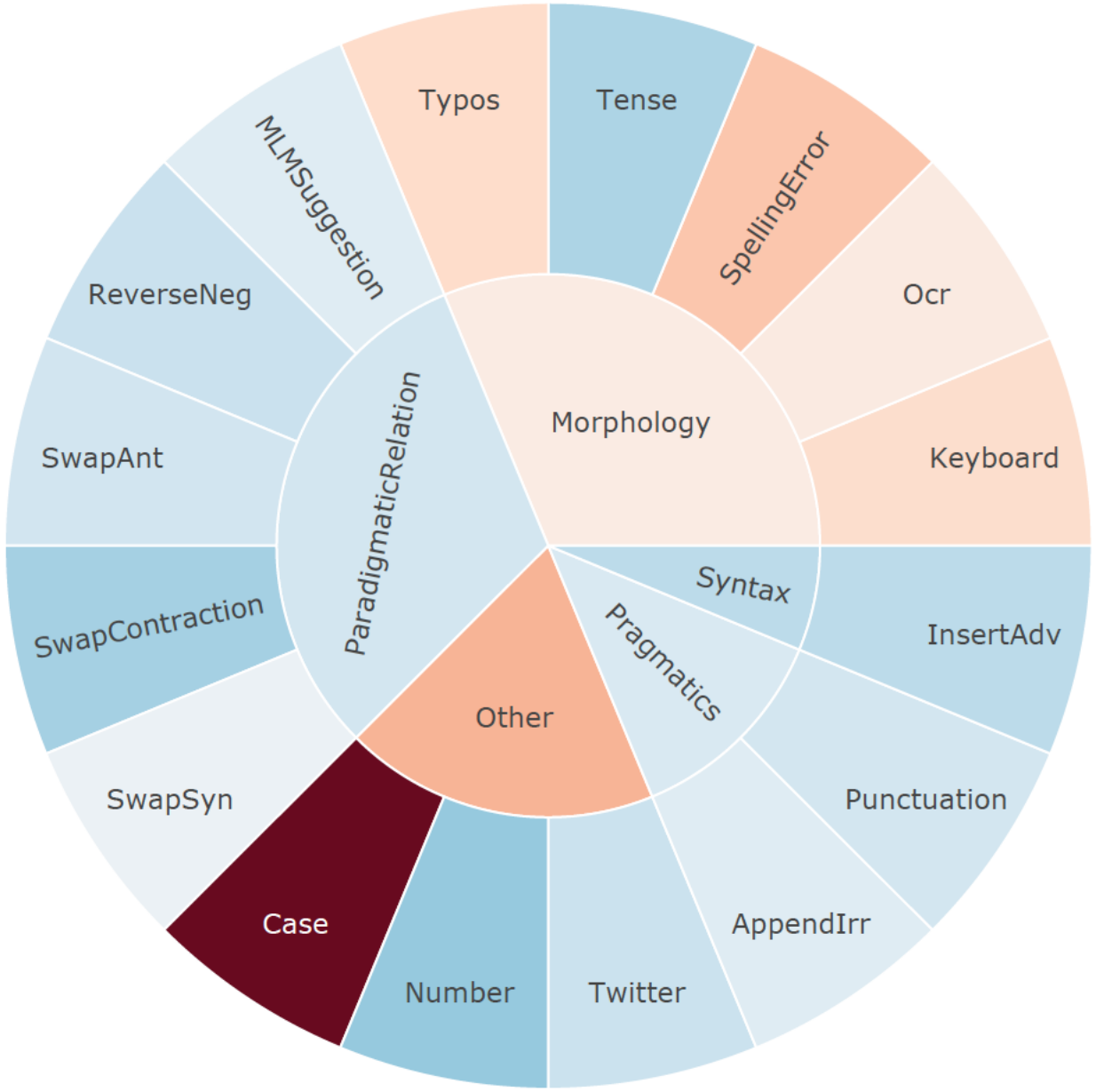
"He was born in China" → "He w

POS: SwapMultiPOS

"There is an apple on the desk"

Wang et al., *TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing*, ACL 2021

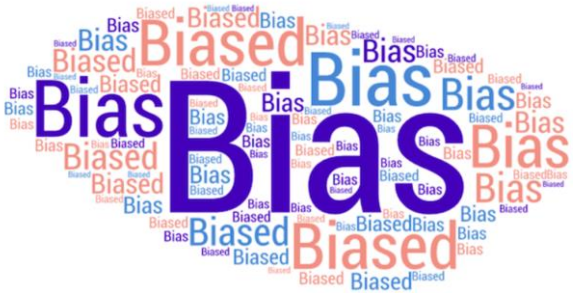
Model	→ Trans.) Macro-F1
Restaurant I	
LSTM (Hoch	.75 → 33.24
TD-LSTM (	.92 → 55.69
ATAE-LSTM	.02 → 36.40
MemNet (Ta	.57 → 39.36
IAN (Ma et a	.91 → 37.08
TNet (Li et a	.29 → 59.68
MGAN (Fan	.98 → 51.71
BERT-base (	.66 → 37.00
BERT+aspec	.91 → 71.01
LCF-BERT (	.56 → 73.71
Average	.86 → 49.49



Wang et al., *TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing*, ACL 2021

3 Issues of Current Benchmarks

Robustness Measurement



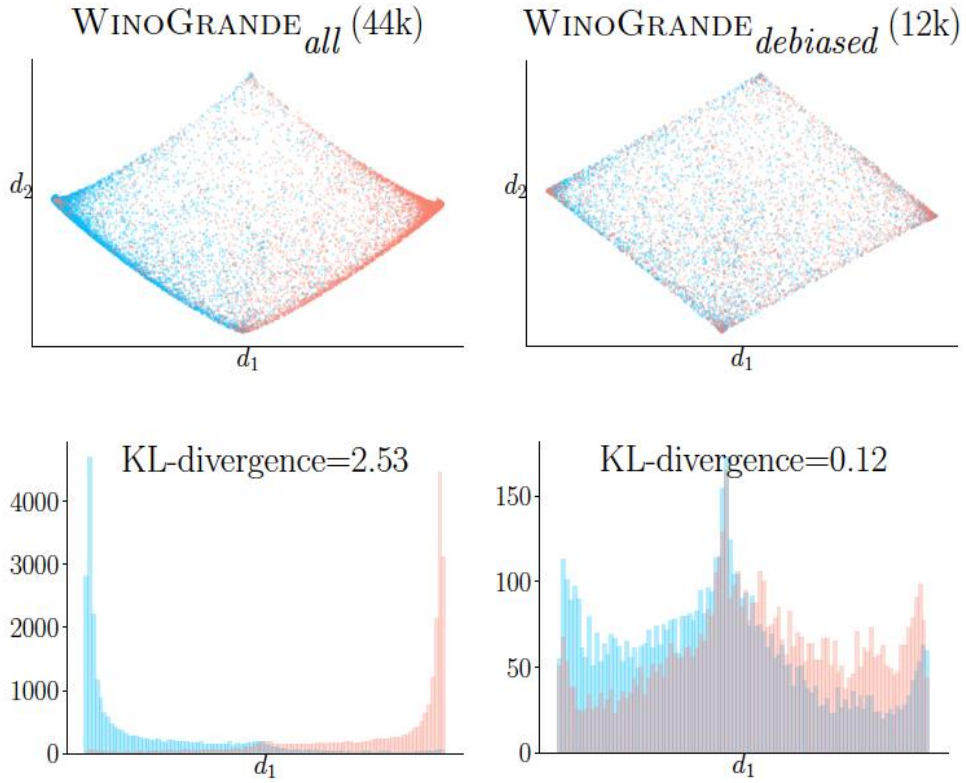
Data Bias

Evaluation Metric

Some Efforts to Data Bias

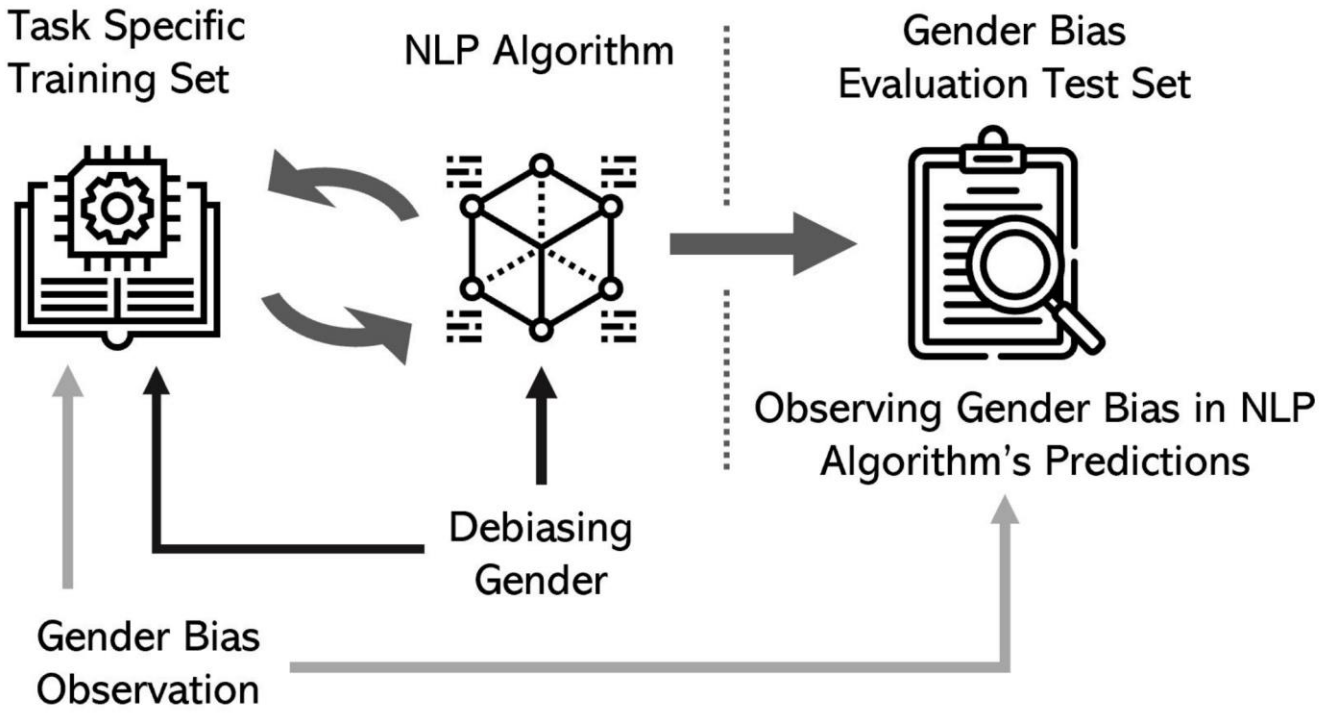
WINOGRANDE

- Large collection of WSC style challenge question pairs.
- Inspired by WSC design
  - Much larger dataset
  - Higher difficulty questions
  - “Fill in the blank” format
  - Filtered to remove bias
  - Not all sentence questions have pairs
  - More diverse language used in questions
- ~44k questions before bias filtering, ~12k after



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

Some Efforts to Data Bias



Methods	Method Type
Data Augmentation by Gender-Swapping	Retraining
Gender Tagging	Retraining
Bias Fine-Tuning	Retraining
Hard Debiasing	Inference
Learning Gender-Neutral Embeddings	Retraining
Constraining Predictions	Inference
Adjusting Adversarial Discriminator	Retraining

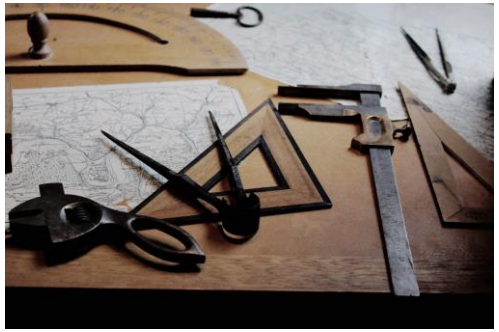
Debiasing methods can be categorized according to how they affect the model.

Sun et al., *Mitigating Gender Bias in Natural Language Processing: Literature Review*, ACL 2019.

3 Issues of Current Benchmarks

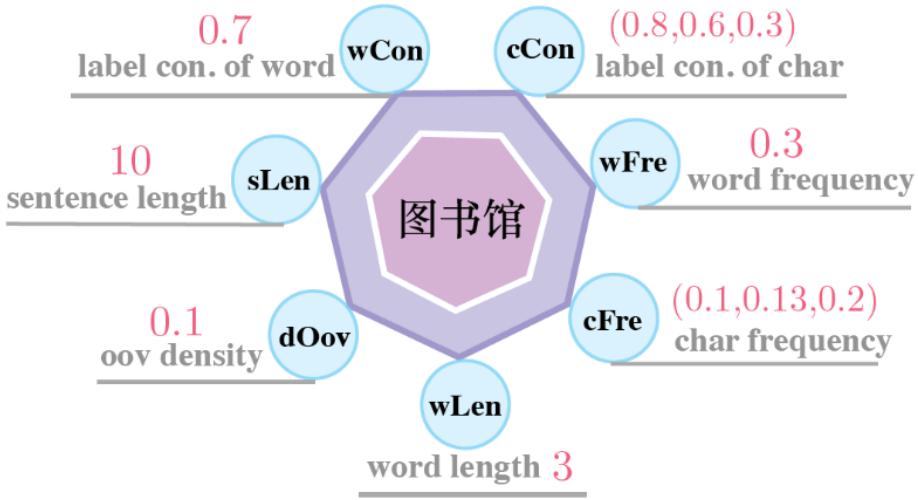
Robustness Measurement

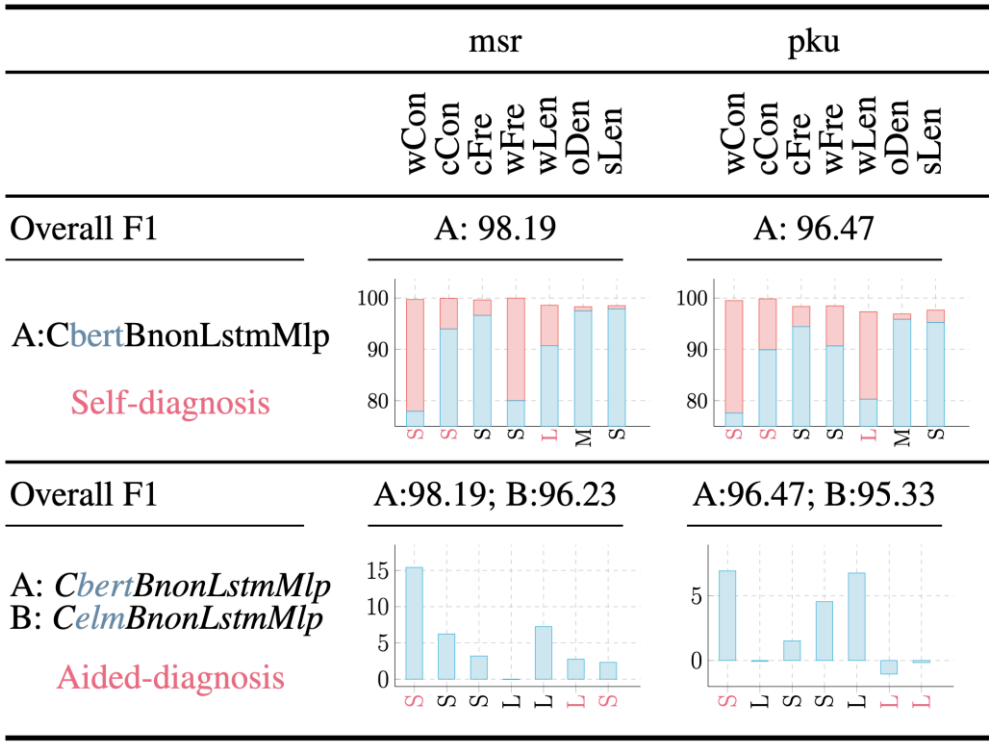
Data Bias



Evaluation Metric

Some Efforts to Evaluation Metric





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*

Some Efforts to Evaluation Metric

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.

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

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

References

- Robustness
  - Ribeiro et al., *Beyond Accuracy: Behavioral Testing of NLP Models with CheckList*, ACL 2020
  - Kiela et al. *Dynabench: Rethinking Benchmarking in NLP*, NAACL 2021
  - Morris et al. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*, EMNLP 2020
  - Wang et al., *TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing*, ACL 2021
- Bias
  - Sakaguchi et al., *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale*, AAAI 2020.
  - Sun et al., *Mitigating Gender Bias in Natural Language Processing: Literature Review*, ACL 2019.
- Evaluation
  - Fu et al. , *RethinkCWS: Is Chinese Word Segmentation a Solved Task?*, EMNLP 2020
  - Fu et al. , *Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study*, AAAI 2020

Conclusion

- Robustness
  - We should look beyond the simple and well-defined problems and pay more attention on the challenges of real-world systems. How to evaluate and improve the robustness of a model is one of these challenges.
- Bias
  - Dataset-specific biases may highly impact the usefulness of model trained on it.
  - When we construct dataset, debiasing methods are not optional, are necessary.
- Evaluation
  - After dataset is constructed, it would be better to manually perturb a small number test instances for further evaluation.
  - Task specific and attribute-aided evaluation metrics may help us diagnose the weaknesses of methods.

Thanks!