TextFlint

Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing

Tao Gui, Qi Zhang

Fudan University

FNLP

0 NLP Progress

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

Penn Treebank

Model	POS	UAS	LAS
Label Attention Layer + HPSG + XLNet (Mrini et al., 2019)	97.3	97.42	96.26
HPSG Parser (Joint) + XLNet (Zhou and Zhao, 2019)	97.3	97.20	95.72
HPSG Parser (Joint) + BERT (Zhou and Zhao, 2019)	97.3	97.00	95.43
CVT + Multi-Task (Clark et al., 2018)	97.74	96.61	95.02
CRF Parser (Zhang et al., 2020)	-	96.14	94.49
Left-to-Right Pointer Network (Fernández-González and Gómez-Rodríguez, 2019)	97.3	96.04	94.43
Graph-based parser with GNNs (Ji et al., 2019)	97.3	95.97	94.31
Deep Biaffine (Dozat and Manning, 2017)	97.3	95.74	94.08
jPTDP (Nguyen and Verspoor, 2018)	97.97	94.51	92.87
Andor et al. (2016)	97.44	94.61	92.79
Distilled neural FOG (Kuncoro et al., 2016)	97.3	94.26	92.06
Distilled transition-based parser (Liu et al., 2018)	97.3	94.05	92.14
Weiss et al. (2015)	97.44	93.99	92.05

IMDb

Model	Accuracy
XLNet (Yang et al., 2019)	96.21
BERT_large+ITPT (Sun et al., 2019)	95.79
BERT_base+ITPT (Sun et al., 2019)	95.63
ULMFiT (Howard and Ruder, 2018)	95.4
Block-sparse LSTM (Gray et al., 2017)	94.99
oh-LSTM (Johnson and Zhang, 2016)	94.1
Virtual adversarial training (Miyato et al., 2016)	94.1
BCN+Char+CoVe (McCann et al., 2017)	91.8

https://github.com/sebastianruder/NLP-progress

FNLP

0 NLP Progress

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*	86.831	89.452
1 Apr 06, 2020			
2 May 05, 2020			
2 Apr 05, 2020	*Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694		
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

Penn Treebank

Model	POS	UAS	LAS
Label Attention Layer + HPSG + XLNet (Mrini et al., 2019)	97.3	97.42	96.26
HPSG Parser (Joint) + XLNet (Zhou and Zhao, 2019)	97.3	97.20	95.72
HPSG Parser (Joint) + BERT (Zhou and Zhao, 2019)	97.3	97.00	95.43
Deep Biaffine (Dozat and Manning, 2017)	97.3	95.74	94.08
jPTDP (Nguyen and Verspoor, 2018)	97.97	94.51	92.87
Andor et al. (2016)	97.44	94.61	92.79
Distilled neural FOG (Kuncoro et al., 2016)	97.3	94.26	92.06
Distilled transition-based parser (Liu et al., 2018)	97.3	94.05	92.14
Weiss et al. (2015)	97.44	93.99	92.05

IMDb

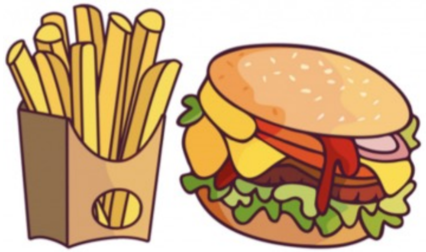
Model	Accuracy
XLNet (Yang et al., 2019)	96.21
BERT_large+ITPT (Sun et al., 2019)	95.79
BERT_base+ITPT (Sun et al., 2019)	95.63
ULMFiT (Howard and Ruder, 2018)	95.4
BCN+Char+CoVe (McCann et al., 2017)	91.8

Whether these models can maintain good performance **in real applications**?

https://github.com/sebastianruder/NLP-progress

FNLP

0 Accuracy is NOT the Sole Metric



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

FNLP

0 Accuracy is NOT the Sole Metric

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

FNLP

1. Is there a simple TOOLKIT that can comprehensively evaluate the robustness of existing models?
2. Is this robustness evaluation REASONABLE?
3. Is this phenomenon COMMON in real experiments?
4. How can we BENEFIT from the use of this toolkit?

1 TextFlint: Robustness Evaluation Toolkit

TextFlint

Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing

FNLP

1 TextFlint: Robustness Evaluation Toolkit

Integrity

TextFlint offers 20 general transformations, 60 task-specific transformations and thousands of their combinations, and provides over 67,000 evaluation results generated by the transformation on 24 classic datasets from 12 tasks, basically covers all aspects of text transformations to comprehensively evaluate the robustness of a model.

Acceptability

Only when the new generated texts conforms to human language, can the robustness result obtained by the verification be credible. Transformation methods provided by TextFlint are scored in plausibility and grammaticality by human evaluation. The results of human and model evaluation can be found on this website.

Analyzability

TextFlint can give a standard analysis report from the lexics, syntax, semantic levels. All evaluation results can be displayed with visualization and tabulation, so that users can accurately grasp the shortcomings of the model. More evaluation results and related analysis are in the paper.

TextFlint

1 TextFlint: Robustness Evaluation Toolkit

Transformation - General

Synonym

"He <u>loves</u> NLP" is transformed into "He <u>likes</u> NLP"

Spelling Error

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

Antonym

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

TextFlint

1 TextFlint: Robustness Evaluation Toolkit

Transformation – Domain Specific

NER: SwapNamedEnt

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

CWS: SwapVerb

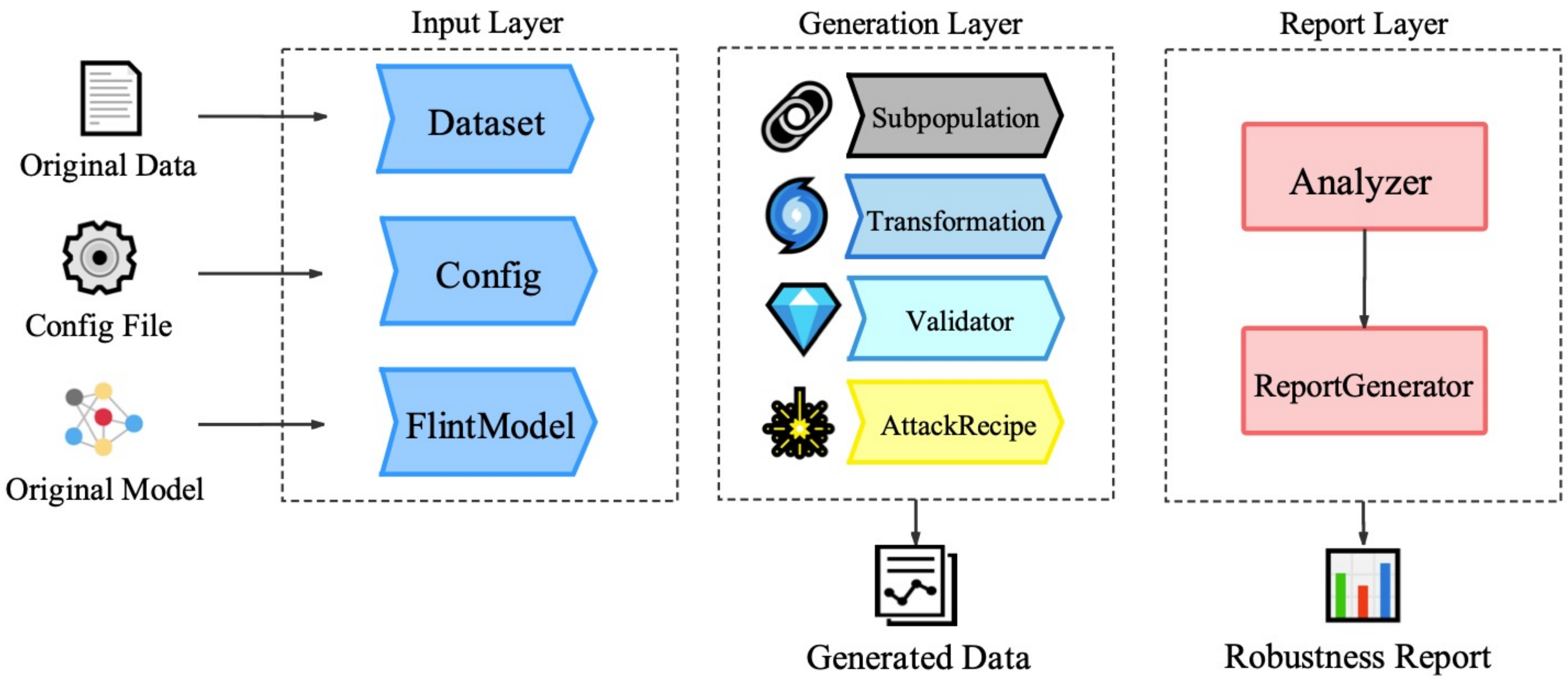
看 → “看看,” “看一看,” “看了看,” and “看了一看.”

POS: SwapMultiPOS

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

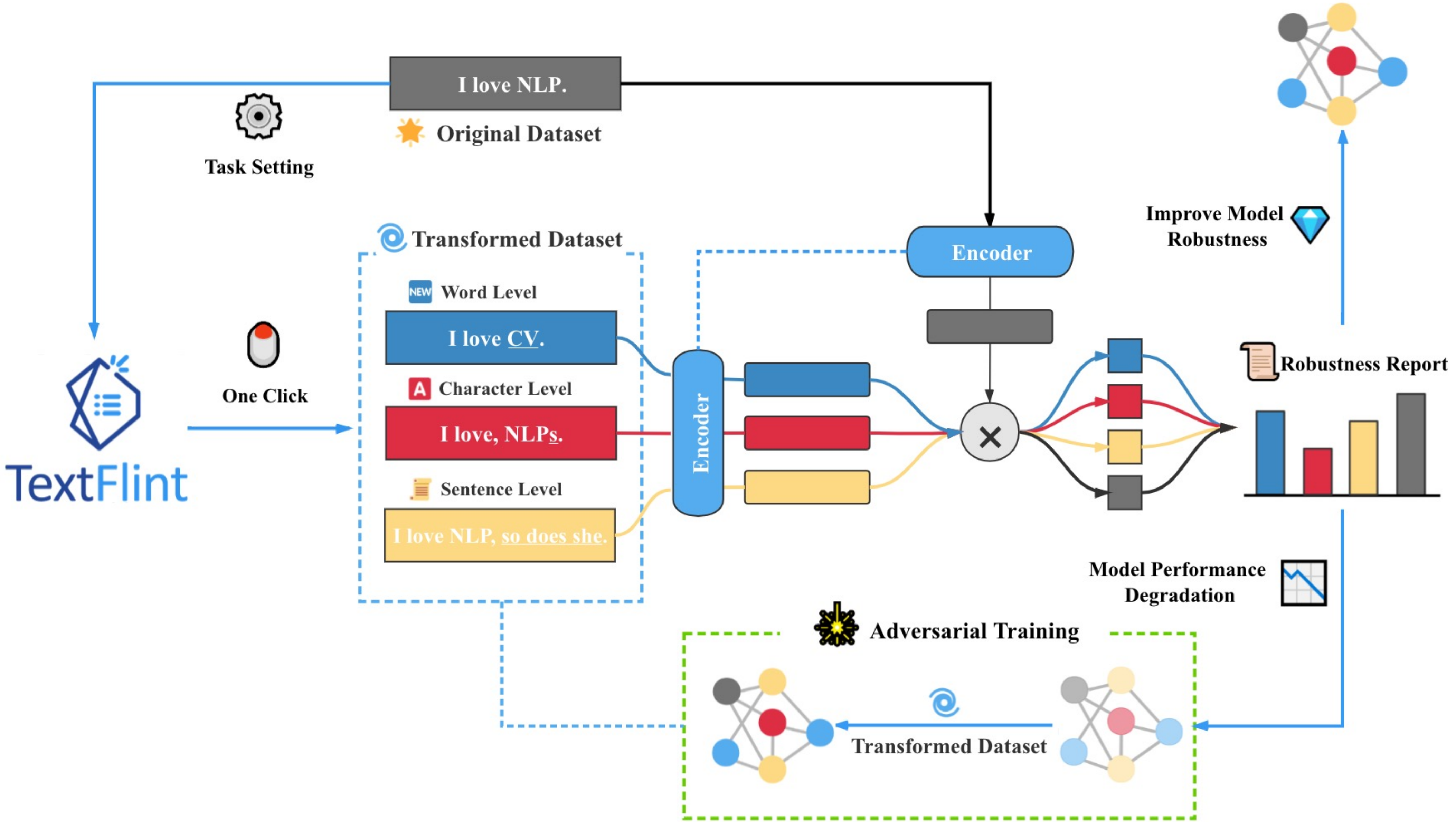
TextFlint

1 TextFlint: Robustness Evaluation Toolkit



TextFlint

1 TextFlint: Robustness Evaluation Toolkit



1 TextFlint: Robustness Evaluation Toolkit

```
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

4 TextFlint: Robustness Evaluation Toolkit

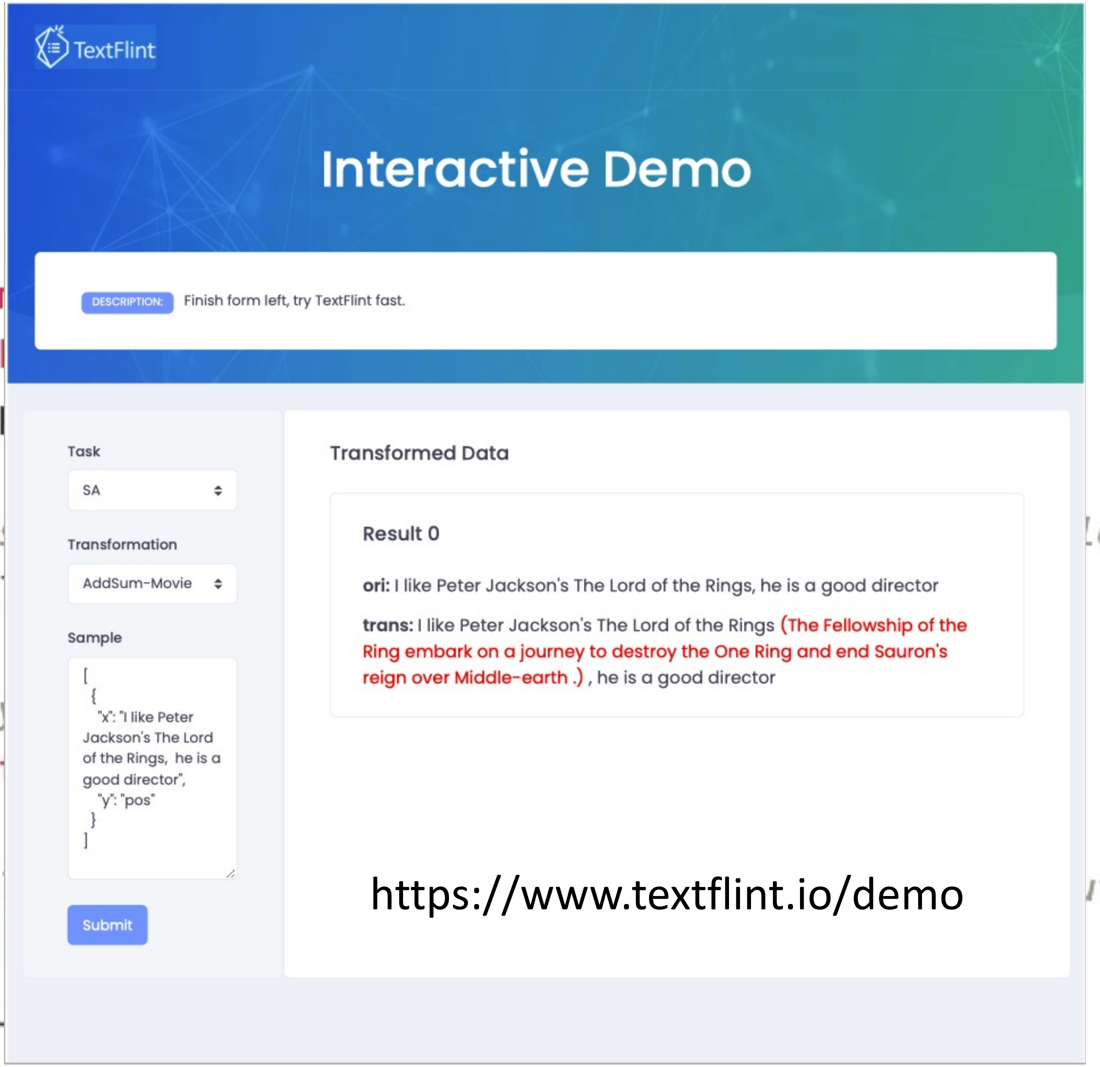
```
from TextFlint.engine import
from TextFlint.config.config

# load the data samples
sample1 = {'x': 'Titanic is r
sample2 = {'x': 'I don\'t lil
data_samples = [sample1, samp

# define the transformation/s                                    le
config = Config.from_json_fi

# define the output directory
out_dir_path = './test_result

# run transformation/subpopu                                      ut_dir_path in json format
engine = TextFlintEngine('SA
engine.run(data_samples, out_
```



https://www.textflint.io/demo

TextFlint

1. Is there a simple TOOLKIT that can comprehensively evaluate the robustness of existing models?

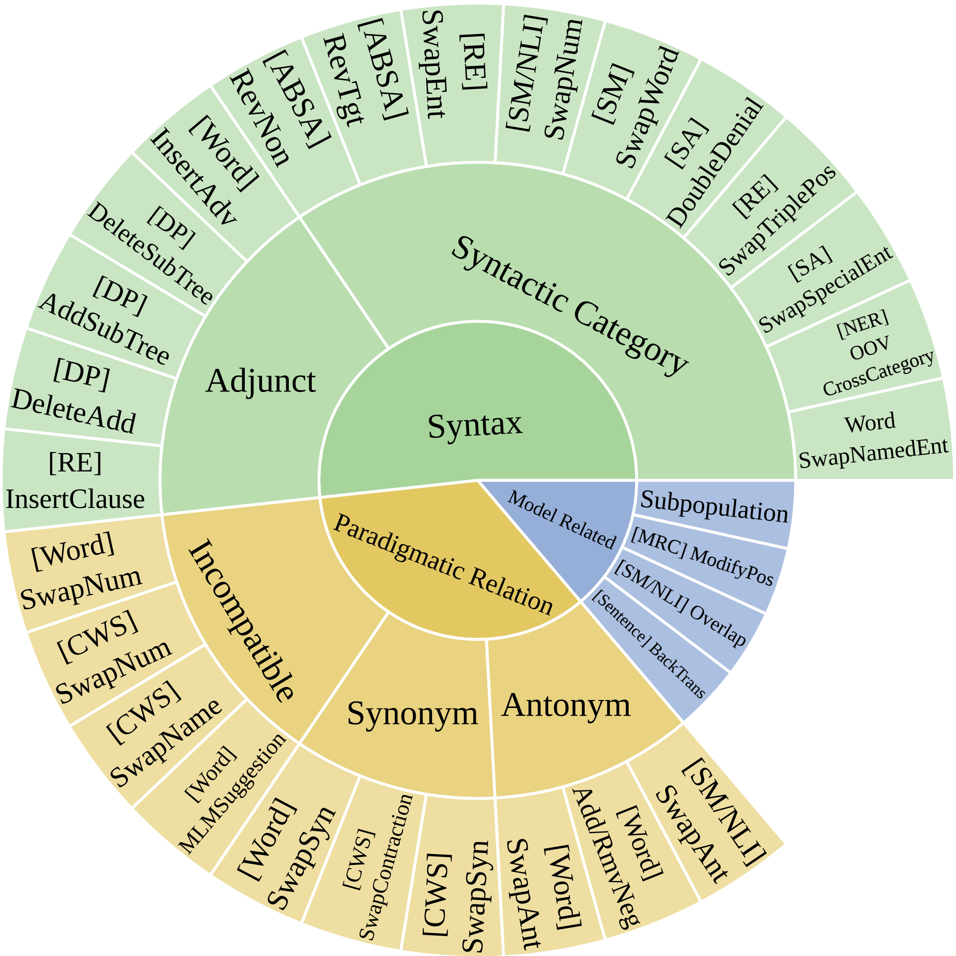
2. Is this robustness evaluation REASONABLE?

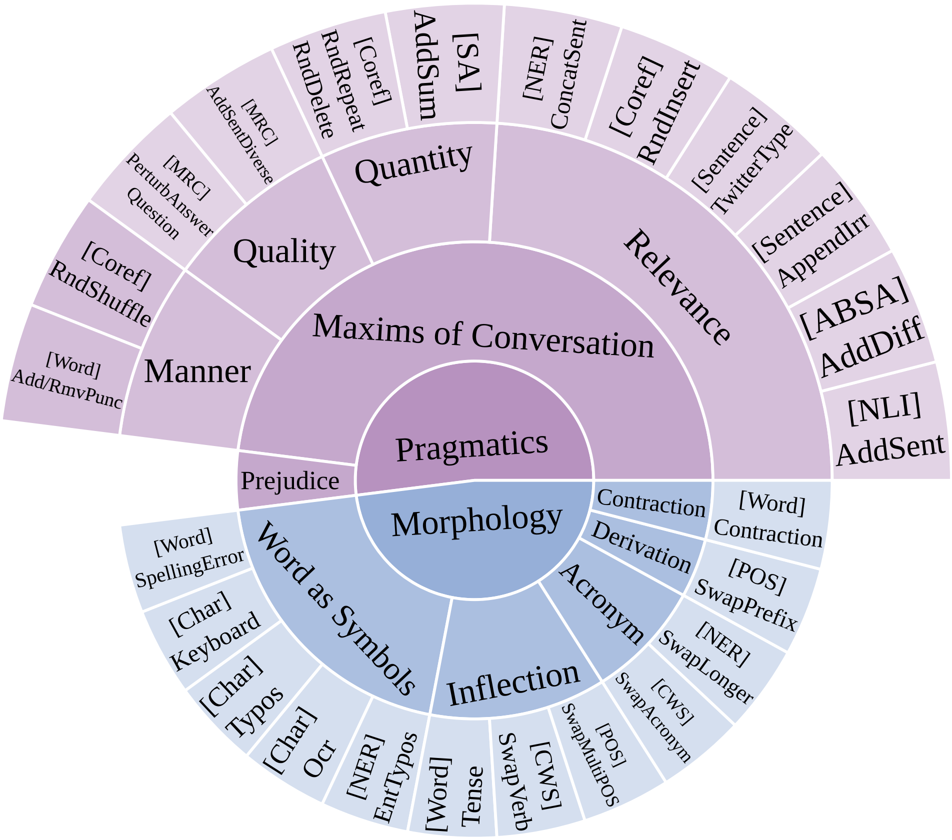
3. Is this phenomenon COMMON in real experiments?

4. How can we BENEFIT from the use of this toolkit?

PHOTO BY ZHOUYANG

2 TextFlint: Robustness Evaluation Toolkit





TextFlint

2 TextFlint: Robustness Evaluation Toolkit

Human Evaluation

- **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

Human Evaluation

Table 2: Human evaluation results for task-specific transformation. Ort. and Trans. represent the original text and the transformed text, respectively. These metrics are rated on a 1-5 scale (5 for the best).

(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

1. Is there a simple TOOLKIT that can comprehensively evaluate the robustness of existing models?

2. Is this robustness evaluation REASONABLE?

3. Is this phenomenon COMMON in real experiments?

4. How can we BENEFIT from the use of this toolkit?

PHOTO BY ZHOUYANG

3 TextFlint: Robustness Evaluation Toolkit

Table 4: F1 score on the CoNLL 2003 dataset.

Model	*ConcatSent* Ori. → Trans.	*CrossCategory* Ori. → Trans.	*EntTypos* Ori. → Trans.	*OOV* Ori. → Trans.	*SwapLonger* Ori. → Trans.
CoNLL 2003					
CNN-LSTM-CRF (Ma and Hovy, 2016)	90.61 → 87.99	90.59 → 44.18	91.25 → 79.10	90.59 → 58.99	90.59 → 61.15
LSTM-CRF (Lample et al., 2016)	88.49 → 86.88	88.48 → 41.33	89.31 → 74.32	88.48 → 43.55	88.48 → 54.50
LM-LSTM-CRF (Liu et al., 2018)	90.89 → 88.21	90.88 → 44.28	91.54 → 82.90	90.88 → 70.40	90.88 → 65.43
Elmo (Peters et al., 2018)	91.80 → 90.67	91.79 → 44.13	92.48 → 86.19	91.79 → 68.10	91.79 → 61.82
Flair (Akbik et al., 2018)	92.25 → 90.73	92.24 → 45.30	93.05 → 86.78	92.24 → 73.45	92.24 → 66.13
Pooled-Flair (Akbik et al., 2019)	91.90 → 90.45	91.88 → 43.64	92.72 → 86.38	91.88 → 71.70	91.88 → 67.92
TENER (Yan et al., 2019)	91.36 → 90.27	91.35 → 45.43	92.01 → 82.26	91.35 → 55.67	91.35 → 51.10
GRN (Chen et al., 2019)	91.57 → 89.30	91.56 → 42.90	92.29 → 82.72	91.56 → 68.20	91.56 → 65.38
BERT-base (cased) (Devlin et al., 2019)	91.43 → 89.91	91.42 → 44.42	92.20 → 85.02	91.42 → 68.71	91.42 → 79.28
BERT-base (uncased) (Devlin et al., 2019)	90.41 → 90.05	90.40 → 47.19	91.25 → 81.25	90.40 → 64.46	90.40 → 78.26
Average	91.07 → 89.45	91.06 → 44.28	91.81 → 82.69	91.06 → 64.32	91.06 → 65.10

TextFlint

Table 5: Exact Match (EM) and F1 score on the SQuAD 1.0 dataset.

Model	*ModifyPos* (Ori.→Trans.)		*AddSentDiverse* (Ori.→Trans.)		*PerturbAnswer* (Ori.→Trans.)	
	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match	F1 Score
SQuAD 1.0						
BiDAF (Seo et al., 2016)	68.93 → 68.64	78.09 → 77.52	68.10 → 22.68	77.45 → 26.07	68.27 → 51.24	77.50 → 63.76
BiDAF^{+} (Seo et al., 2016)	69.60 → 67.58	78.91 → 76.72	68.88 → 22.71	78.21 → 26.60	68.91 → 52.19	78.24 → 64.55
DrQA (Chen et al., 2017)	70.99 → 69.99	80.20 → 78.67	70.34 → 35.34	79.62 → 40.56	70.19 → 52.32	79.52 → 64.85
R-Net (Wang et al., 2017)	72.06 → 70.79	80.56 → 78.96	71.31 → 26.55	79.83 → 30.63	71.35 → 54.15	79.87 → 66.13
FusionNet (Huang et al., 2018)	73.00 → 71.60	82.01 → 80.38	72.21 → 34.40	81.28 → 39.33	72.47 → 54.90	81.44 → 67.49
QANet (Yu et al., 2018)	71.52 → 71.27	79.98 → 79.79	70.67 → 19.34	79.32 → 22.09	70.86 → 55.13	79.45 → 67.36
BERT (Devlin et al., 2019)	79.95 → 79.81	87.68 → 87.25	79.25 → 27.93	87.09 → 32.47	79.30 → 62.48	87.13 → 75.40
ALBERT-V2 (Lan et al., 2019)	85.31 → 84.24	91.76 → 90.82	84.70 → 35.87	91.27 → 40.45	84.63 → 68.80	91.26 → 80.52
XLNet (Yang et al., 2019b)	81.79 → 81.13	89.81 → 88.94	81.37 → 32.12	89.50 → 37.48	81.30 → 67.15	89.45 → 80.15
DistillBERT (Sanh et al., 2019)	79.96 → 79.10	87.56 → 86.69	79.43 → 25.53	87.10 → 29.60	79.35 → 62.21	87.04 → 74.92
Average	75.31 → 74.42	83.65 → 82.57	74.63 → 28.25	83.07 → 32.53	74.66 → 58.06	83.09 → 70.51

TextFlint

3 TextFlint: Robustness Evaluation Toolkit

Table 6: Model accuracy on the MultiNLI dataset.

Model	*SwapAnt* Ori. → Trans.	*AddSent* Ori. → Trans.	*NumWord* Ori. → Trans.	*Overlap* Ori. → Trans.
MultiNLI				
BERT-base (Devlin et al., 2019)	85.10 → 55.69	84.43 → 55.27	82.97 → 49.16	None → 62.67
BERT-large (Devlin et al., 2019)	87.84 → 61.18	86.36 → 58.19	85.42 → 54.19	None → 70.65
XLNet-base(Yang et al., 2019b)	87.45 → 70.98	86.33 → 57.65	85.55 → 48.77	None → 70.35
XLNet-large(Yang et al., 2019b)	89.41 → 75.69	88.63 → 63.37	86.84 → 51.35	None → 78.09
RoBERTa-base(Delobelle et al., 2020)	87.45 → 63.53	87.13 → 57.25	86.58 → 50.32	None → 75.49
RoBERTa-large(Delobelle et al., 2020)	92.16 → 74.90	90.12 → 67.73	88.65 → 54.71	None → 73.14
ALBERT-base-v2 (Lan et al., 2019)	87.45 → 50.20	84.09 → 53.59	82.97 → 49.42	None → 67.15
ALBERT-xxlarge-v2 (Lan et al., 2019)	91.76 → 69.80	89.89 → 79.11	89.03 → 46.84	None → 74.92
Average	88.58 → 65.25	87.12 → 61.52	86.00 → 50.60	None → 71.56

TextFlint

Table 7: F1 score on the CTB6 dataset.

Model	*SwapName* Ori. → Trans.	*SwapNum* Ori. → Trans.	*SwapVerb* Ori. → Trans.	*SwapContraction* Ori. → Trans.	*SwapSyn* Ori. → Trans.
CTB6					
FMM[1]	82.13 → 78.39	83.62 → 79.88	82.03 → 78.14	84.25 → 79.11	83.97 → 79.26
BMM[1]	83.21 → 79.28	83.91 → 80.11	82.45 → 78.61	84.82 → 79.51	84.41 → 79.75
CRF[2]	93.80 → 91.70	93.30 → 89.33	91.13 → 87.32	94.20 → 87.83	93.50 → 92.00
CWS-LSTM (Chen et al., 2015)	94.87 → 91.56	95.25 → 91.32	93.16 → 88.91	95.47 → 88.88	94.84 → 93.01
CWS (Cai and Zhao, 2016)	94.96 → 91.31	94.12 → 86.42	92.42 → 87.92	94.91 → 91.02	94.02 → 92.85
GreedyCWS (Cai et al., 2017)	95.18 → 91.74	94.04 → 86.75	93.27 → 88.54	94.83 → 88.58	94.61 → 93.07
Sub-CWS (Yang et al., 2019a)	95.72 → 92.92	96.92 → 92.26	94.01 → 89.26	96.51 → 89.49	96.15 → 94.75
MCCWS (Qiu et al., 2020)	92.30 → 89.97	92.85 → 88.94	89.60 → 85.76	93.12 → 87.03	92.36 → 89.77
Average	91.52 → 88.36	91.75 → 86.88	89.76 → 85.56	92.26 → 86.43	91.73 → 89.31

TextFlint

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

TextFlint

3 TextFlint: Robustness Evaluation Toolkit

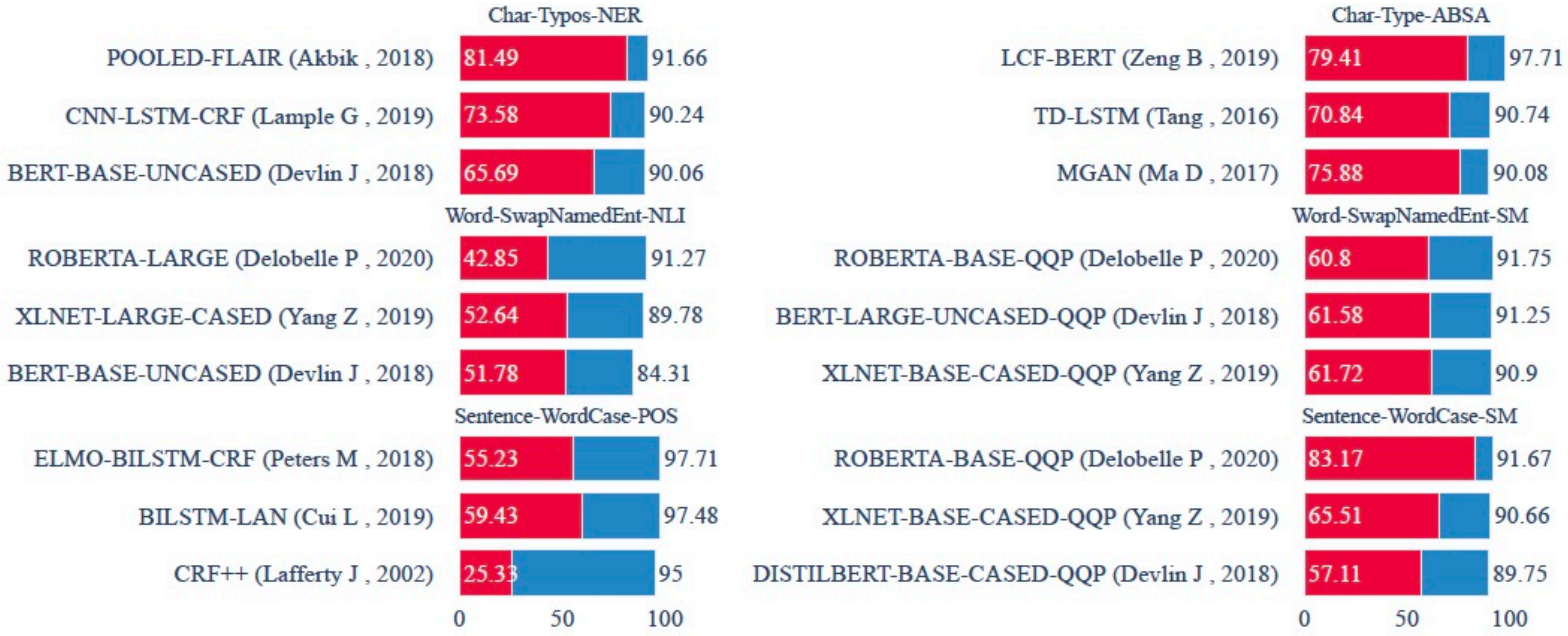


Figure 4: Accuracy results of multi-granularity universal transformations (UT). We choose **Typos**, **SwapNamedEnt**, and **WordCase** for character-level, word-level, and sentence-level UT, respectively.

3 TextFlint: Robustness Evaluation Toolkit

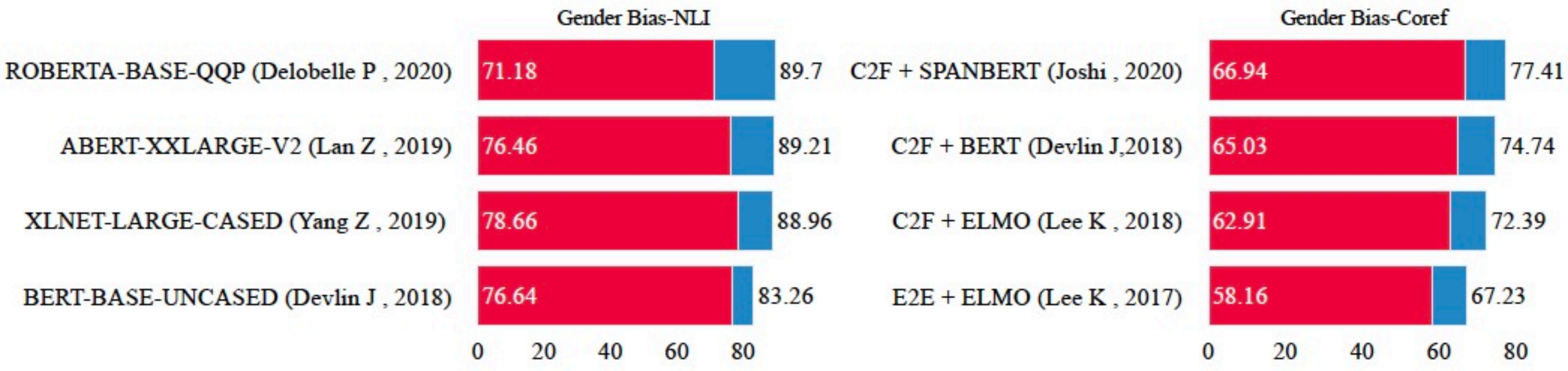


Figure 5: Results of gender bias transformations. We replace human names by female names and perform robustness evaluation in NLI and Coref tasks.

TextFlint

1. Is there a simple TOOLKIT that can comprehensively evaluate the robustness of existing models?

2. Is this robustness evaluation REASONABLE?

3. Is this phenomenon COMMON in real experiments?

4. How can we BENEFIT from the use of this toolkit?

PHOTO BY ZHOUYANG

4 TextFlint: Robustness Evaluation Toolkit

TextFlint

IMDB

The IMDb dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or negative. The dataset contains an even number of positive and negative reviews. Only highly polarizing reviews are considered. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. No more than 30 reviews are included per movie. Models are evaluated based on accuracy.

MODEL	PAPER	GITHUB	ACCURACY	F1
XLNET XLNet: Generalized Autoregressive Pretraining for Language Understanding			96.17	88
BERT How to Fine-Tune BERT for Text Classification?			95.3	
ULMFIT Universal Language Model Fine-tuning for Text Classification			94.6	
OH-LSTM Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings			93.86	
ADV Adversarial Training Methods for Semi-Supervised Text Classification			93.26	

xtFlint

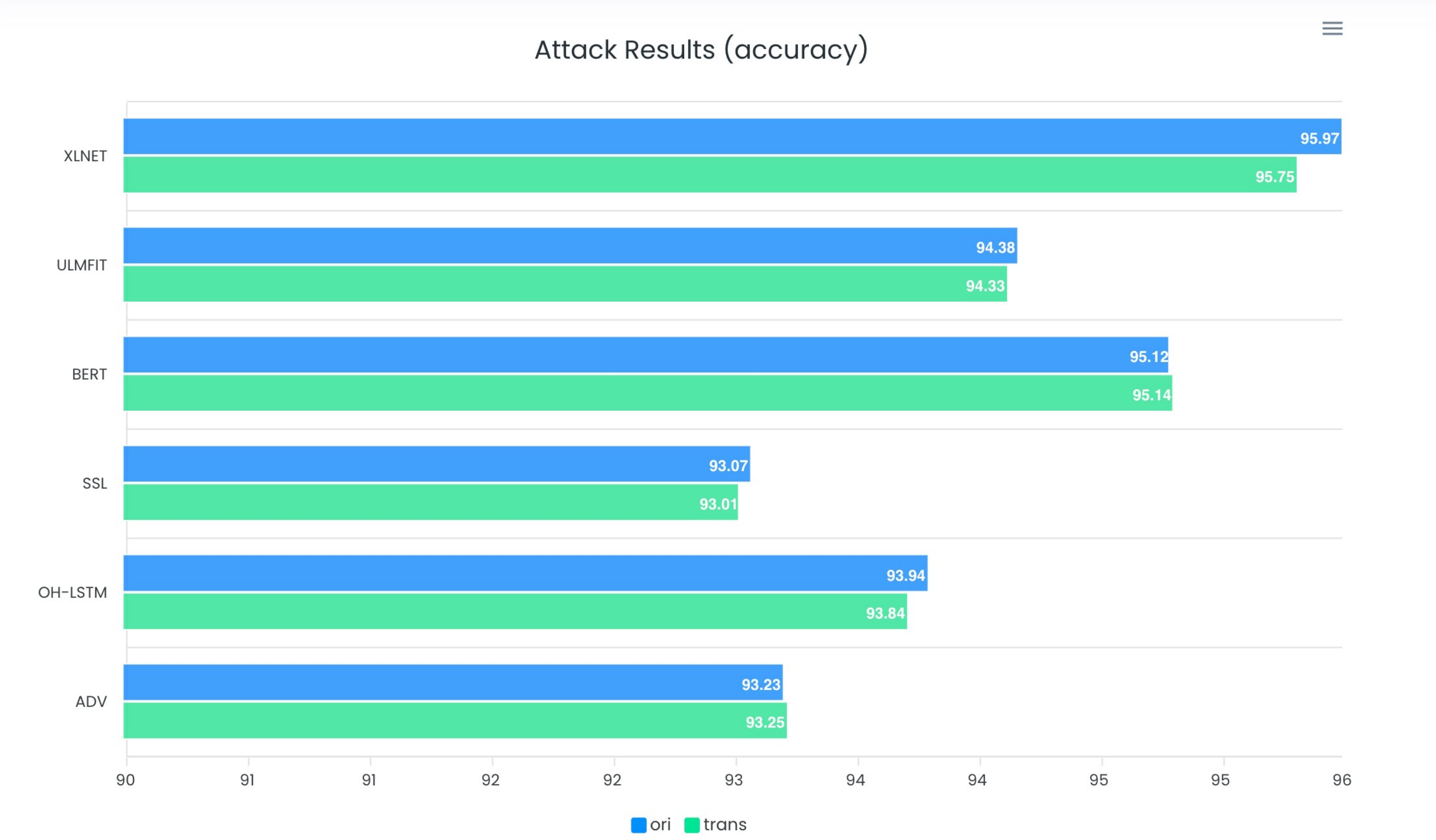
Domain Specific

5 domain specific transformations are available

IMDB | Yelp-Binary

Model/accuracy	SpecialEntityReplace(Movie)		SpecialEntityReplace(Person)		AddSummary(Movie)	
	ori	trans	ori	trans	ori	trans
XLNET	95.97	95.75	95.92	95.85	95.97	95.38
ULMFIT	94.38	94.33	94.76	94.7	94.38	94.03
BERT	95.12	95.14	95.27	95.23	95.12	94.97
SSL	93.07	93.01	93.31	93.28	93.07	92.81
OH-LSTM	93.94	93.84	94.07	94.13	93.94	93.49
ADV	93.23	93.25	93.22	93.28	93.23	92.56

4 TextFlint: Robustness Evaluation Toolkit



TextFlint

4 TextFlint: Robustness Evaluation Toolkit

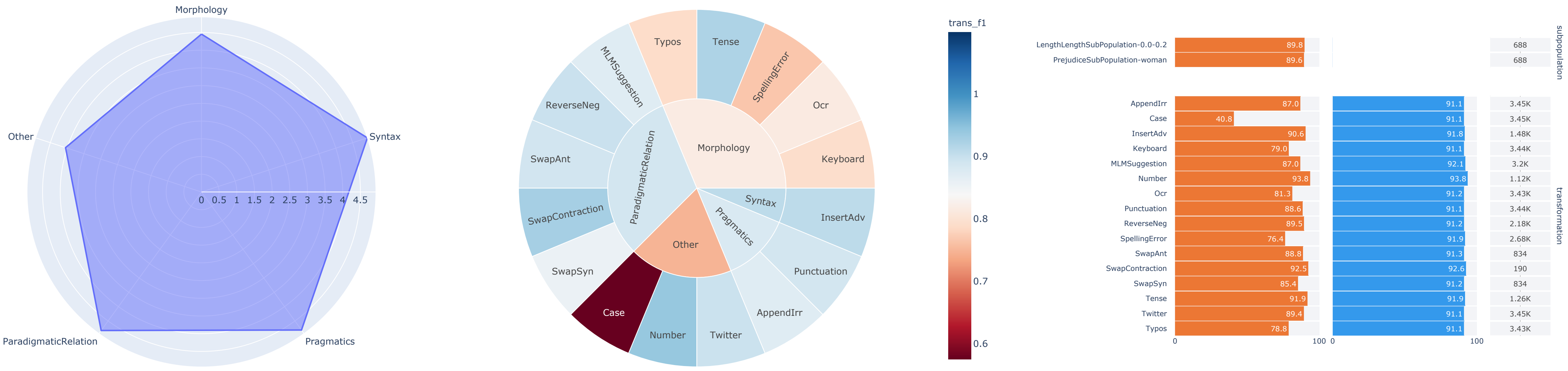
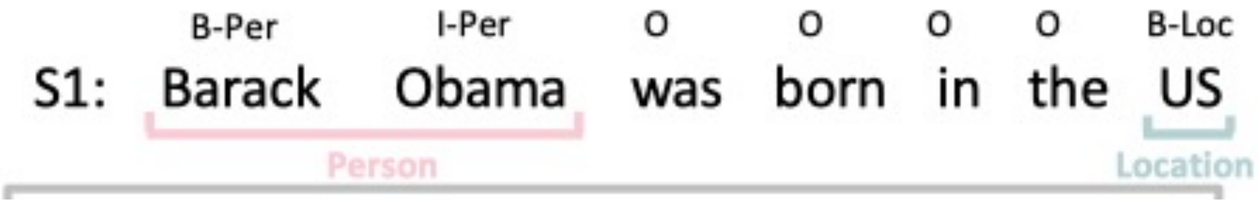


Figure 2: Robustness reports of BERT base model on CONLL 2003 dataset. The first one, namely the radar report, provides an overview of the linguistic ability of the target model. The middle chart gives an intuitive result on each transformation categorized by linguistics. The last bar-chart reveals the details of model performance towards every single generation method.

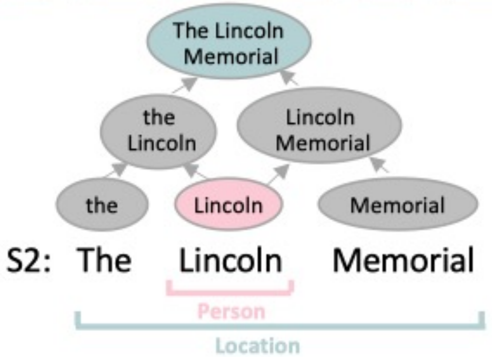
Text Modelling – Named Entity Recognition

Flat NER

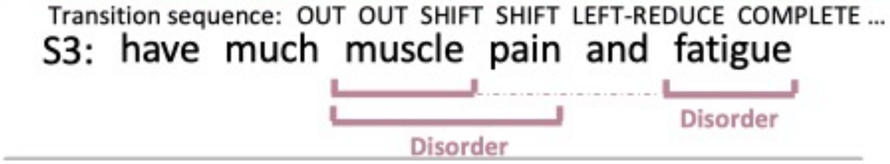


Real Application

Nested NER



Discontinuous NER

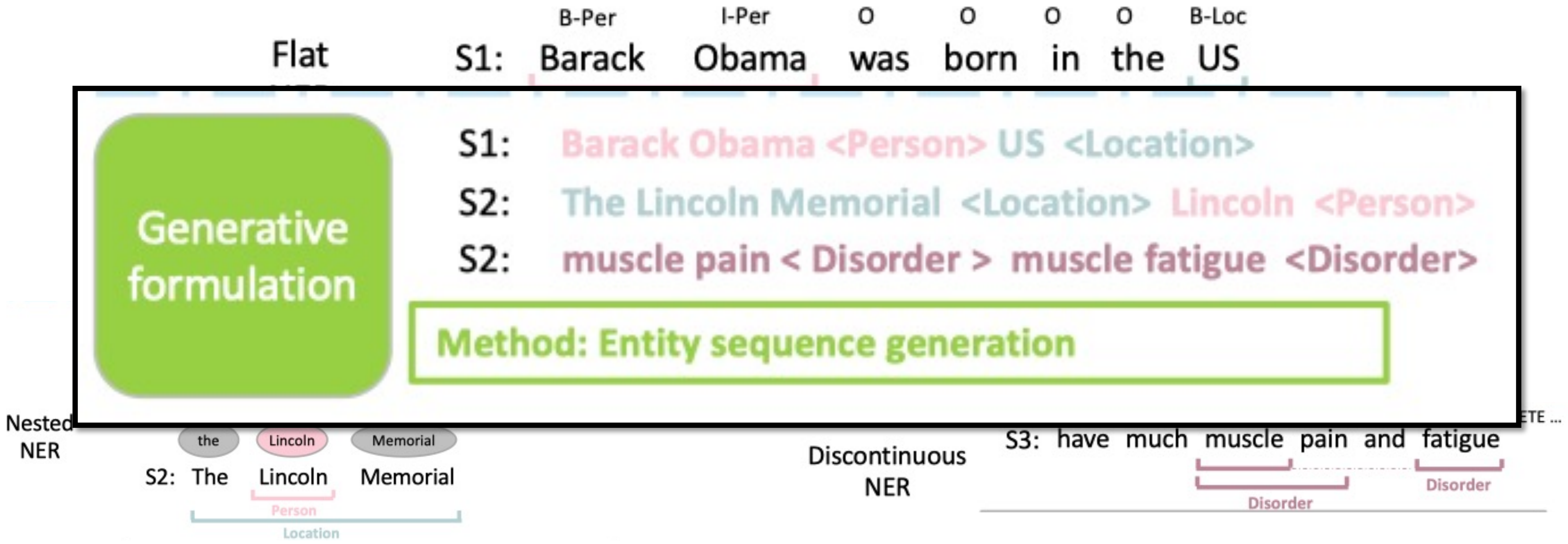


The different formulations make it hard to solve all NER tasks in a unified method

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

FNLP

4 Text Modelling – Named Entity Recognition

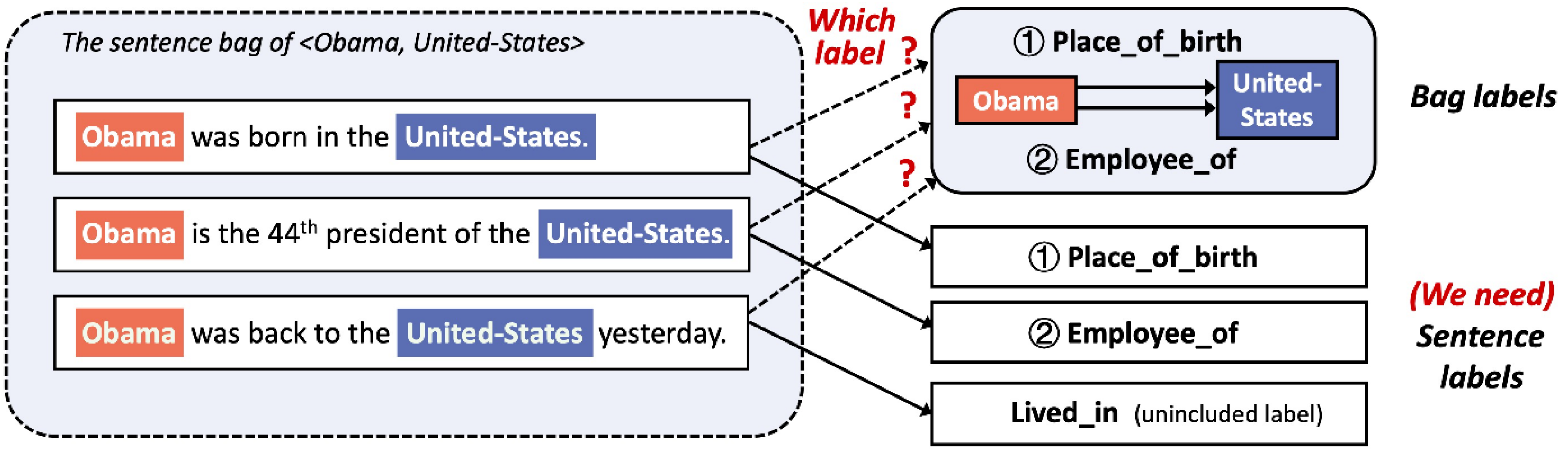


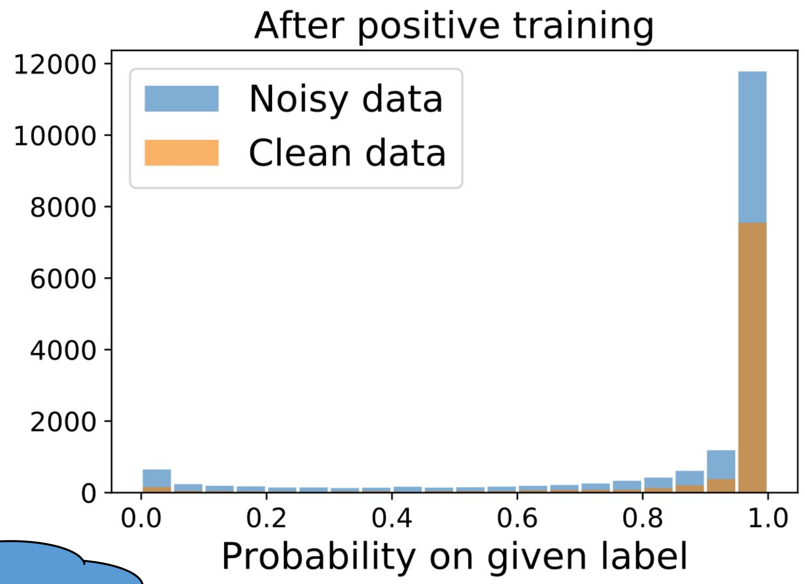
The different formulations make it hard to solve all NER tasks in a unified method

Yan et al., A Unified Generative Framework for Various NER tasks, ACL 2021

4 Noisy Label in Distant Supervision

$$\mathcal{L}_{PT}(f, y^*) = -\sum_{k=1}^{C} y_k \log p_k$$



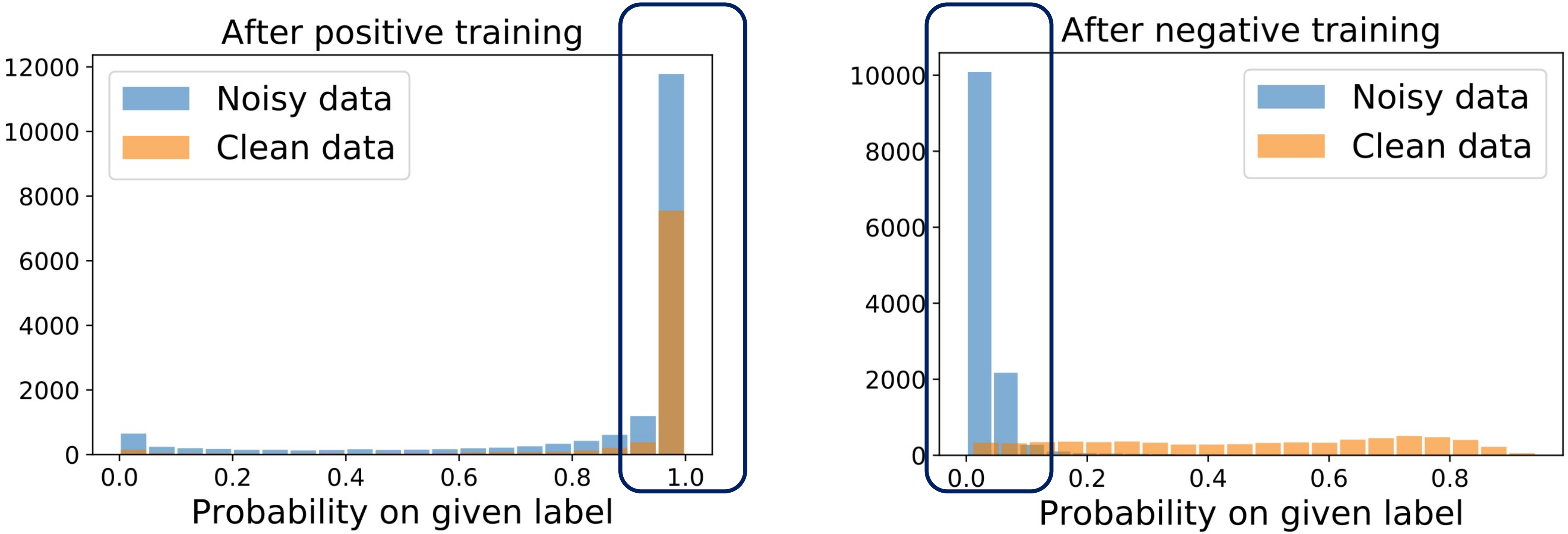


Important for downstream tasks

1. Given bag-level labels, can we obtain sentence-level labels?
2. Sentence bag contains correct labels, incorrect labels, and unincluded labels.
3. Previous positive learning framework cannot distinguish noisy data.

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

FNLP



Comparison between positive and negative training

Ma et al., SENT: Sentence-level Distant Relation Extraction via Negative Training, ACL 2021

1 TextFlint: Robustness Evaluation Toolkit

Watch Star Fork

TextFlint

Thanks for your attention!