BERT-ATTACK: Adversarial Attack Against BERT Using BERT

EMNLP 2020 Long Paper

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, Qipeng Qiu

Linyang Li Fudan-NLP Fudan University linyangli19@fudan.edu.cn

- Adversarial Attack in NLP

Major Problem : Discrete Nature : Cannot Use Gradients;

Solution : Substitution-Based

IMDB	Ori	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ’ s death altered unnecessarily ? to make the story more ‘ horrible ? ’	Negative
	Adv	it is hard for a lover of the novel northanger abbey to sit through this bbc adaptation and to keep from throwing objects at the tv screen... why are so many facts concerning the tilney family and mrs . tilney ’ s death altered unnecessarily ? to make the plot more ‘ horrible ? ’	Positive

SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))	
Premise	Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON)	The boys are in band ***uniforms***.
Adversary (Label: ENT)	The boys are in band ***garment***.
Premise	A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU)	The ***child*** is at the ***beach***.
Adversary (Label: ENT)	The ***youngster*** is at the ***shore***.

Current Methods Summary

- Substitutes-Constraints:
- (1) similar in semantic/grammar/fluency ;
- (2) harmful to NN ;

- Traditional Method:
- Two-Step Algorithm:
- (1) Find places to perturb;
- (2) Replace with similar substitutes;
-

Attack Level {
- Sentence Level
- Word Level
- Char Level

Context: ... commentators had debated whether the figure could be reached as the growth in subscriber numbers elsewhere in Europe flattened.
Original Question: What was happening to subscriber numbers in other areas of Europe? **Prediction:** flattened
Paraphrased Question: What was going on with subscriber numbers in other areas of Europe? **Prediction:** growth

(Sentence-Paraphrase)

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**

(Character change)

Attack Target {
- Gradient — Know the Gradients of the Model
- Score — Know the Score of the classifier
- Decision — Only Know the Label
- Blind — Know nothing

Label : positive

↑

"The movie is great."

↓

Score : positive :0.8; negative:0.2

$(score = Softmax(\overrightarrow{X}))$

Our work : BERT Attack

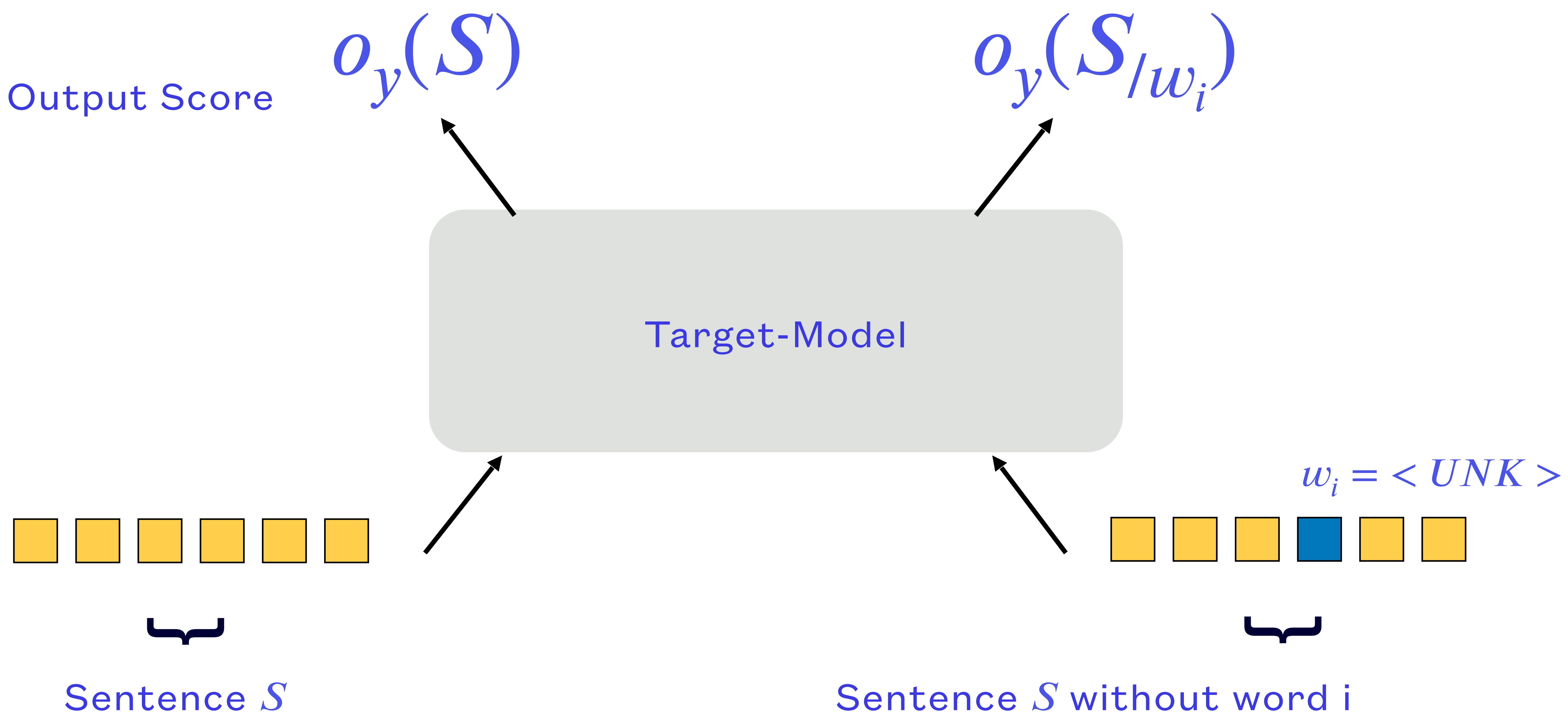
- Major Problem of Substitution-based methods:
- (1): Substitutes are synonyms —-> not context-aware
- (2): Apply Language Models/POS-checking to constrain the perturbations —> inefficient

- Motivation of using Pre-trained Masked-Language Model in Adversarial Attack:
- Fine-tuned Model —-> strong target model ;
- MLM —-> strong LM (substitute generator)

Method of BERT-Attack

- two-steps : (1) finding vulnerable words

Importance of Word: $$I_{w_i} = o_y(S) - o_y(S_{\backslash w_i}),$$



Method of BERT-Attack

- two-steps : (2) using BERT-MLM to generate candidates

deal with sub-words

not mask the word

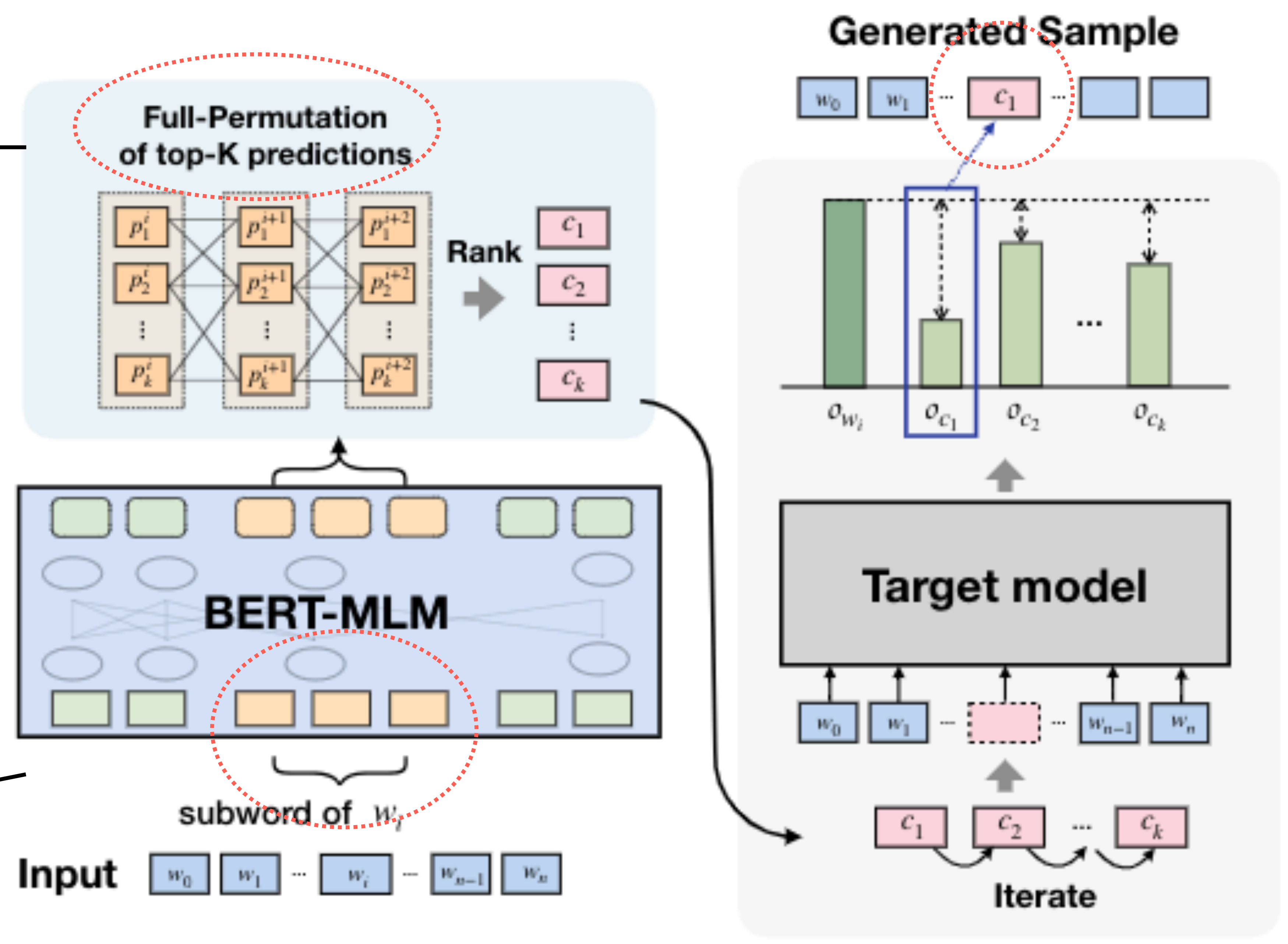


Figure 1: One step of our replacement strategy.

find one substitute

Advantages:

- 1. Using MLM : effective
- context-aware generation of substitutes
- 2. No other constraints :
- efficient
- during the iteration, using LMs is costly

Experiment Result

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/-		-

Experiment Result

Dataset	Model	Ori Acc	Atk Acc	Perturb %
IMDB	Word-LSTM	89.8	10.2	2.7
	BERT-Large	98.2	12.4	2.9
Yelp	Word-LSTM	96.0	1.1	4.7
	BERT-Large	97.9	8.2	4.1
MNLI matched	ESIM	76.2	9.6	21.7
	BERT-Large	86.4	13.2	7.4

Table 3: BERT-Attack against other models.

Dataset	Method	Ori Acc	Atk Acc	Perturb %
MNLI matched	BERT-Atk	85.1	7.9	8.8
	+Adv Train	84.6	23.1	10.5

Table 5: Adversarial training results.

Dataset	Model	LSTM	BERT-base	BERT-large
IMDB	Word-LSTM	-	0.78	0.75
	BERT-base	0.83	-	0.71
	BERT-large	0.87	0.86	-
Dataset	**Model**	ESIM	BERT-base	BERT-large
MNLI	ESIM	-	0.59	0.60
	BERT-base	0.60	-	0.45
	BERT-large	0.59	0.43	-

Table 6: Transferability analysis using attacked accuracy as the evaluation metric. The column is the target model used in attack, and the row is the tested model.

Dataset		Accuracy	Semantic	Grammar
MNLI	Original	0.90	3.9	4.0
	Adversarial	0.70	3.7	3.6
IMDB	Original	0.91	4.1	3.9
	Adversarial	0.85	3.9	3.7

Table 2: Human-Evaluation Results.

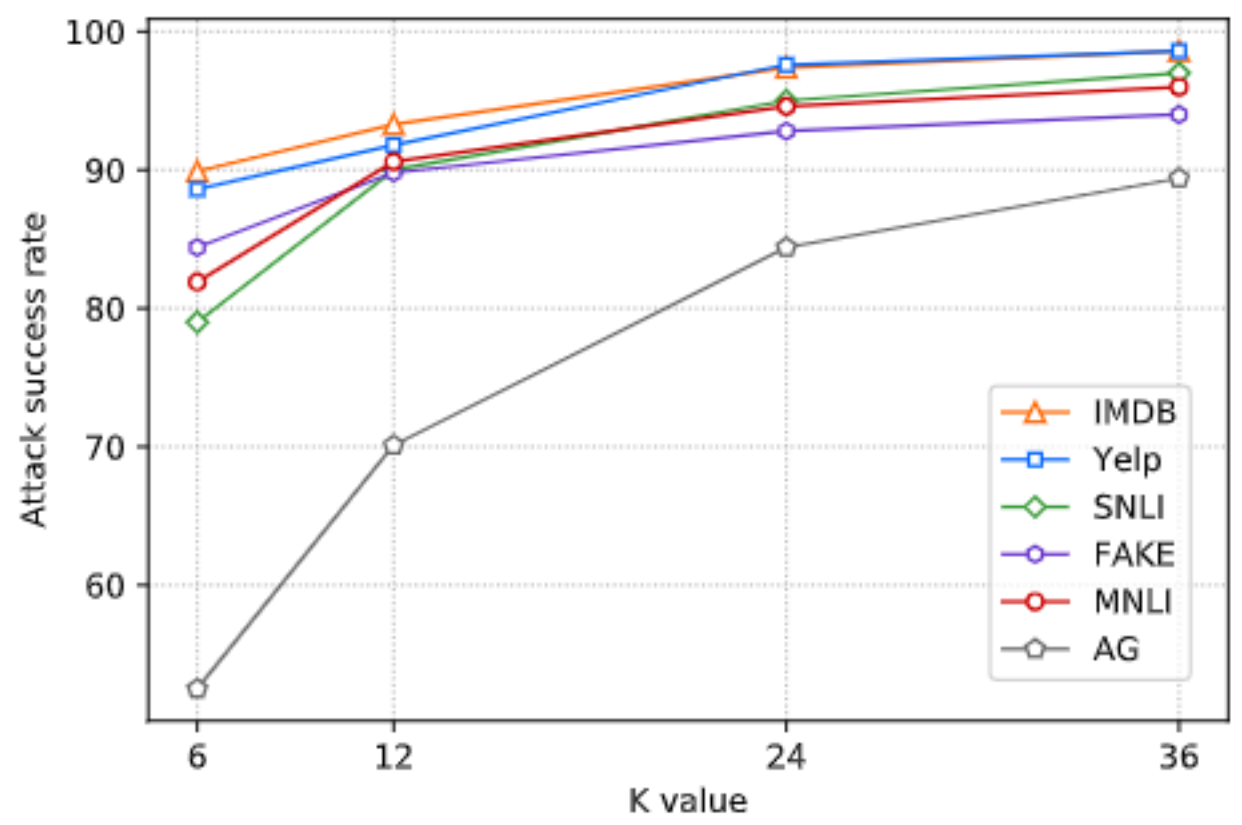
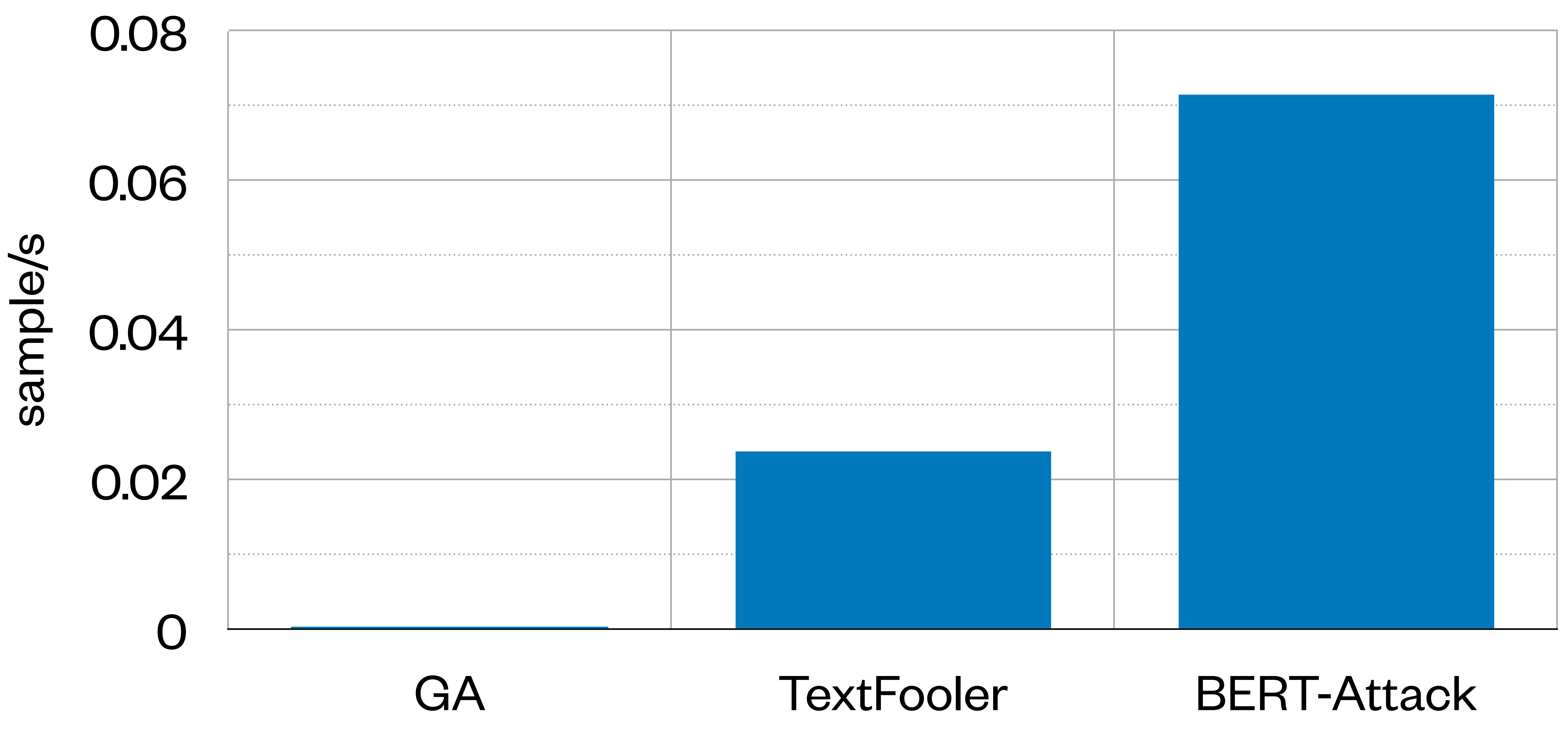


Figure 2: Using different candidate number K in the attacking process.

Runtime

Dataset	Method	Runtime(s/sample)
IMDB	BERT-Attack(w/o BPE)	14.2
	BERT-Attack(w/ BPE)	16.0
	Textfooler(Jin et al., 2019)	42.4
	GA(Alzantot et al., 2018)	2582.0

Table 9: Runtime comparison.



Examples

MNLI	Ori	Some rooms have balconies .	Hypothesis	All of the rooms have balconies off of them .	Contradiction
	Adv	Many rooms have balconies .	Hypothesis	All of the rooms have balconies off of them .	Neutral

IMDB	Ori	i first seen this movie in the early 80s .. it really had nice picture quality too . anyways , i 'm glad i found this movie again ... the part i loved best was when he hijacked the car from this poor guy... this is a movie i could watch over and over again . i highly recommend it .	Positive
	Adv	i first seen this movie in the early 80s .. it really had nice picture quality too . anyways , i 'm glad i found this movie again ... the part i loved best was when he hijacked the car from this poor guy... this is a movie i could watch over and over again . i inordinately recommend it .	Negative

Summary:
We propose a simple, effective and efficient method to craft Adv. samples in NLP.

In textual Adversarial Attack, both effectiveness and efficiency are important.

END

Linyang Li

Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis

Xiaoyu Xing[1*] **Zhijing Jin**[2*] **Di Jin**[3] **Bingning Wang**[4] **Qi Zhang**[1] **Xuanjing Huang**[1]
[1]Fudan University [2]Max Planck Institute, [3]CSAIL, MIT, [4]Sogou Inc.

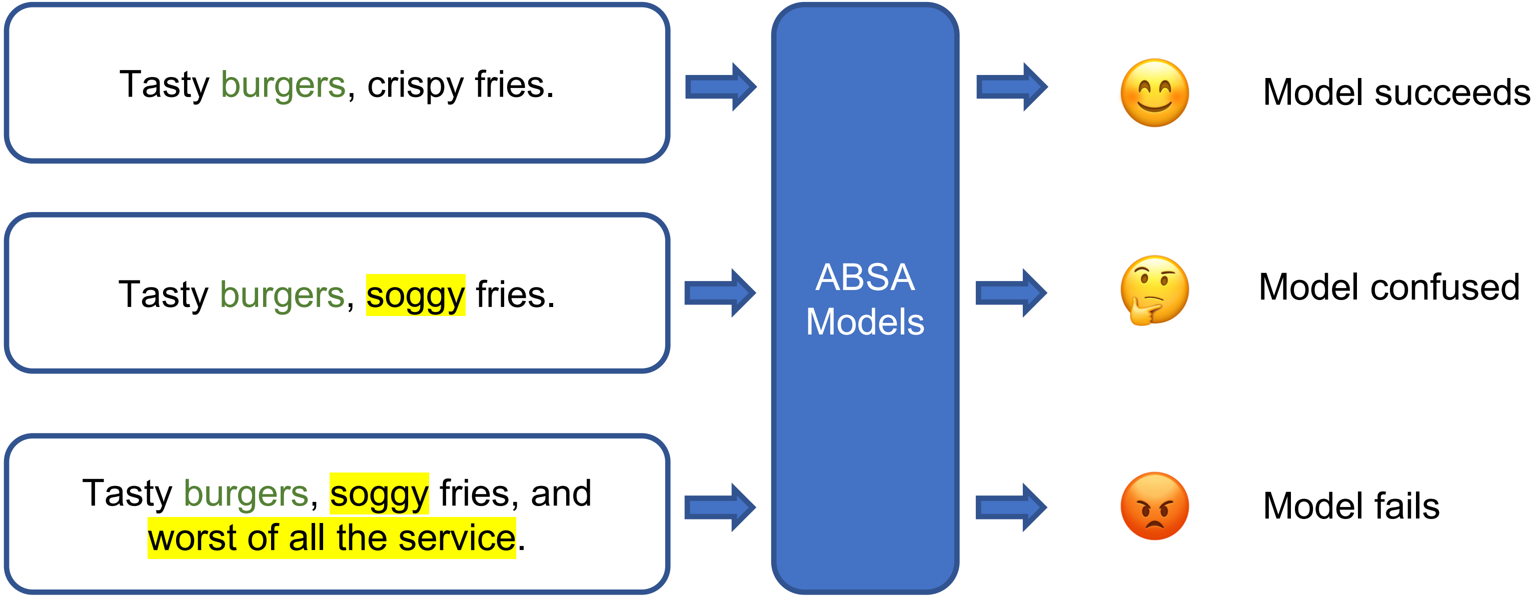
EMNLP 2020

High performance ≠ Strong model

- A strong ABSA model should understand:
  - Aspect
  - Sentiment words
  - Which sentiment words are for the target aspect
- State-of-art models have achieved high accuracy on ABSA tasks.

Do models really understand the correspondence between aspect and sentiment words?

Typical Examples

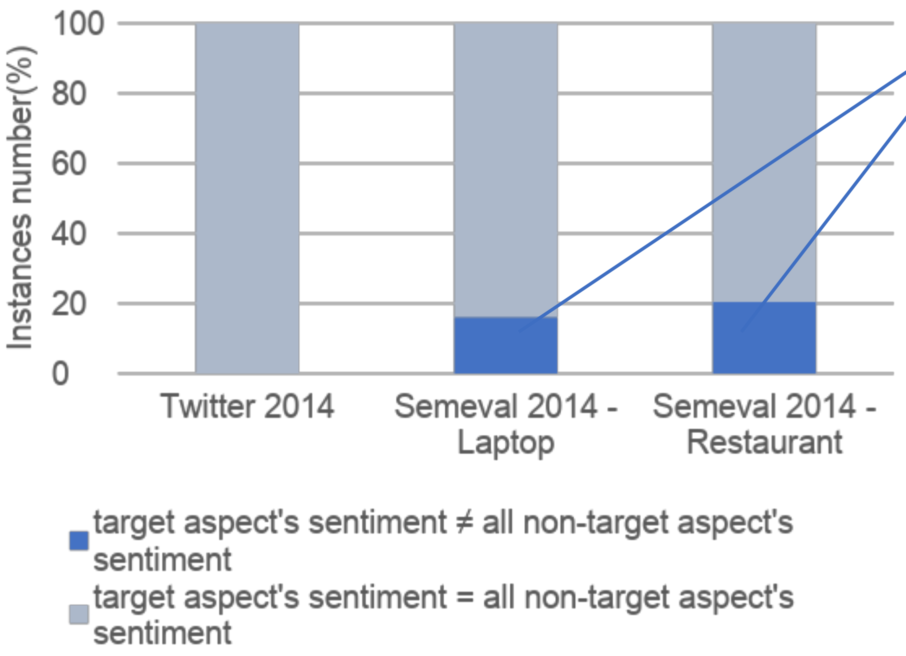


Question about previous models' robustness

A model outputs correct sentiment polarity for the test example

- (Q1) If we **reverse the sentiment polarity of the target aspect**, can the model change its prediction accordingly?
- (Q2) If **the sentiments of all non-target aspects become opposite to the target one**, can the model still make the correct prediction?
- (Q3) If we **add more non-target aspects with sentiments opposite to the target one**, can the model still make the correct prediction?

Existing datasets



Can be used to answer our question

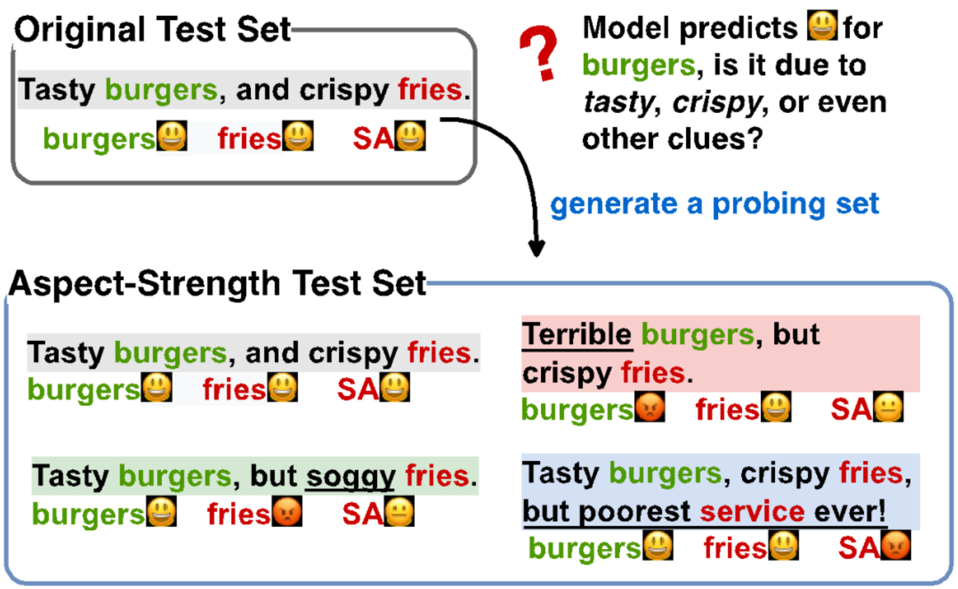
When we test on these subsets,

Laptop: 78.53% ⇩ **59.32%**

Restaurant: 86.70% ⇩ **63.93%**

Over-rely on non-target aspects !

An automatic generation framework



Target aspect: burgers (positive)

Non-target aspect: fries (negative)

- **REVTGT**

tasty -> terrible, positive -> negative

- **REVNON**

crispy -> soggy

- **ADDDIFF**

, but poorest service ever

REVTGT

- It's light and easy to transport. — Get antonyms → It's heavy and difficult to transport.
- The menu changes seasonally. — Add negation → The menu does not change seasonally.
- The food is good, and the décor is nice. — Get antonyms & adjust conjunctions → The food is good, but the décor is nasty.

REVNON

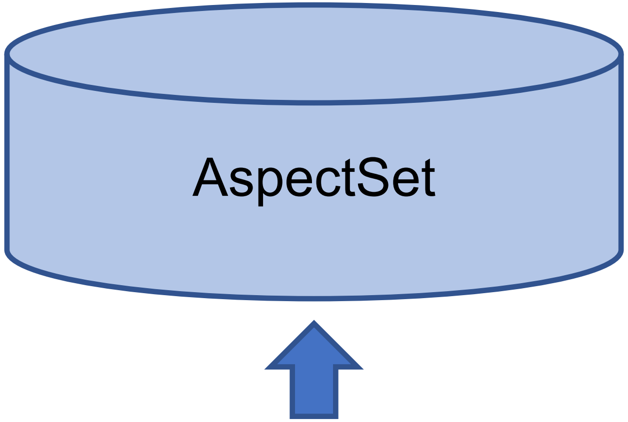
- Flip same-sentiment non-target aspects
- Exaggerate opposite-sentiment non-target aspects

It has great food at a reasonable price, but the service is poor.

but an unreasonable price

and the service is extremely poor

ADDDIFF



Randomly sample 1-3 aspects (different sentiment & not mentioned)

Tasty burgers, crispy fries, but poorest service ever!

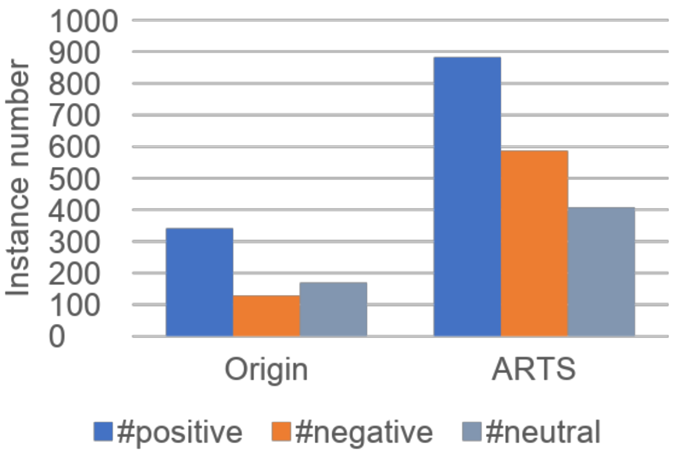
staff is friendly and knowledgeable
desserts are out of this world
texture is a velvety

…

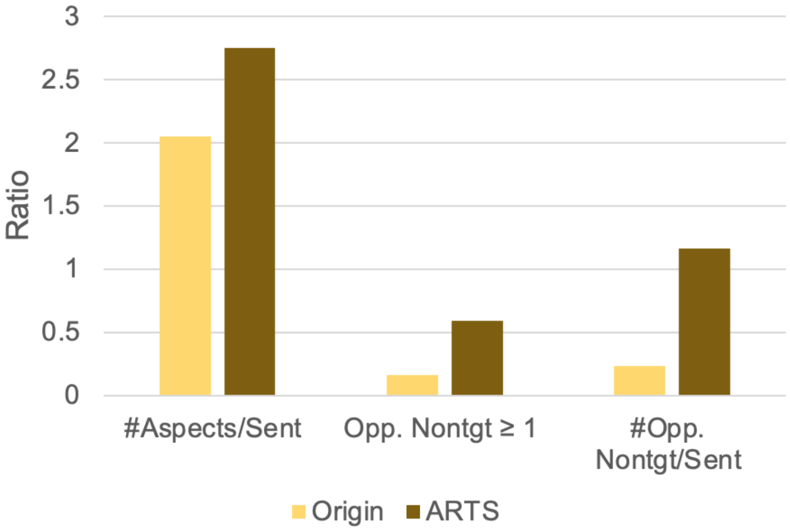
The overall sentiment change from positive to negative.

Dataset Analysis

The dataset is larger and the label is more balanced



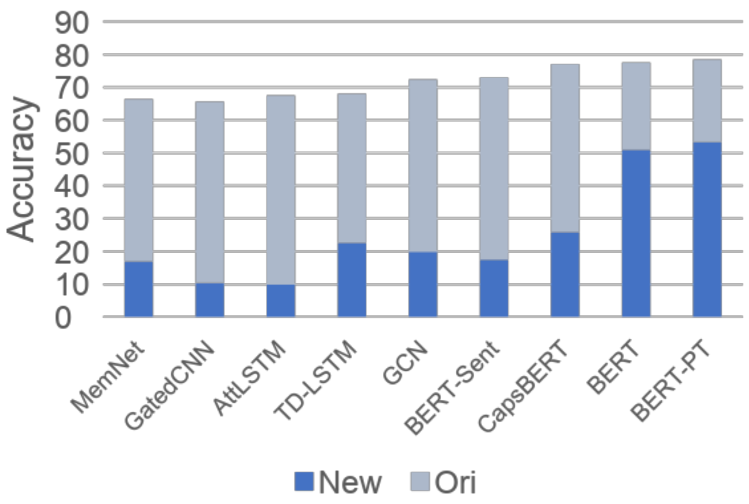
The dataset is more challenging



For restaurant dataset, please refer to our paper.

Experimental Results

$$ARS = \frac{\#\, correct\ units}{\#\, all\ units}$$



Unit

1. Tasty **burgers**, and crispy fries. ✅
2. Terrible **burgers**, but crispy fries. ✅
3. Tasty **burgers**, but soggy fries. ✅
4. Tasty **burgers**, crispy fries, but poorest service ever! ✅

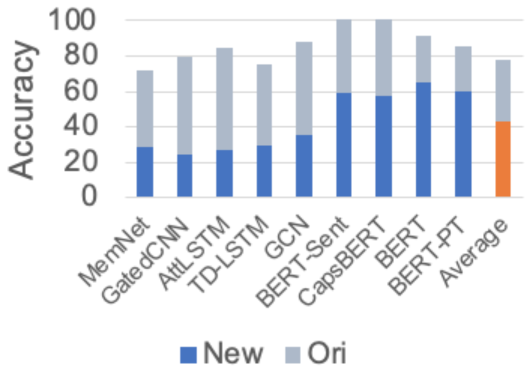
✅

- Overall performance drops dramatically on ARTS.
- BERT-based models are more robust.

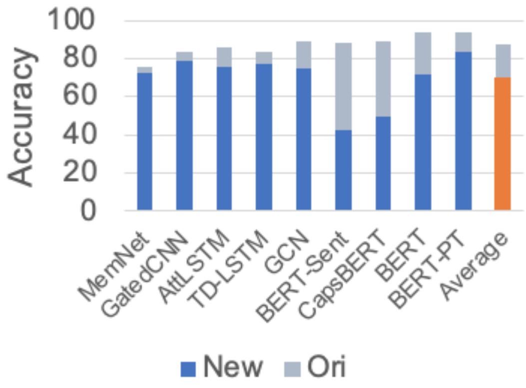
For restaurant dataset, please refer to our paper.

Experimental Results

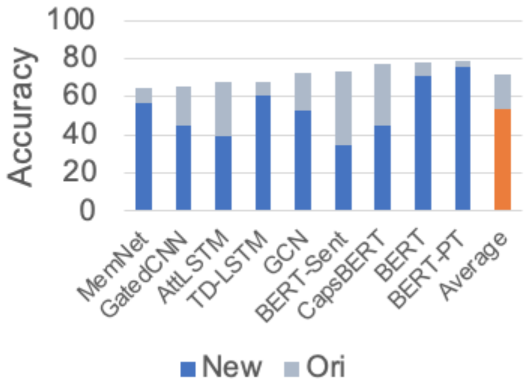
REVTGT



REVNON



ADDDIFF



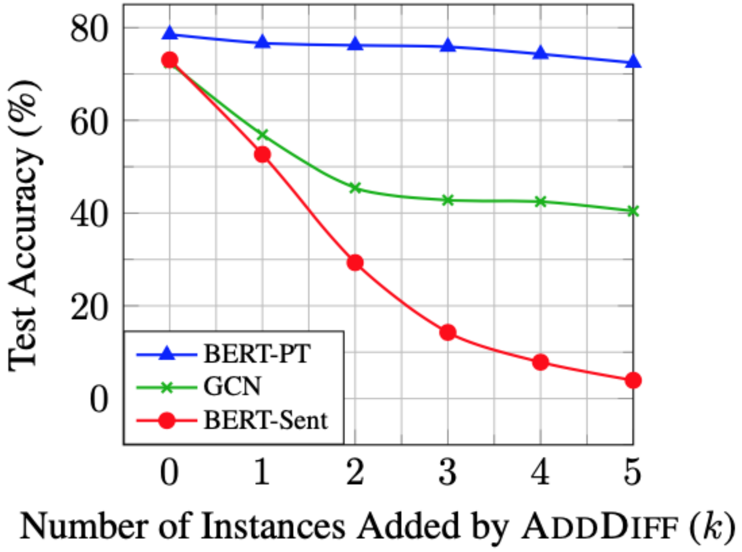
- REVTGT on average induces the most performance drop.
- ADDDIFF causes most non-BERT models to drop significantly .

Variations

Combining multiple strategies

Model	Laptop Ori → New (Change)
MemNet	82.22 → 72.59 (↓09.63)
GatedCNN	84.44 → 59.26 (↓25.18)*
AttLSTM	85.93 → 51.85 (↓34.08)*
TD-LSTM	83.70 → 68.89 (↓14.81)*
GCN	88.89 → 60.74 (↓28.15)*
BERT-Sent	88.15 → 11.85 (↓76.30)*
CapsBERT	90.37 → 24.44 (↓65.93)*
BERT	93.33 → 68.15 (↓25.18)*
BERT-PT	93.33 → 78.52 (↓14.81)*
Average	87.57 → 55.14 (↓32.43) *

ADDDIFF with more aspects



How to effectively model the aspects

Model	Aspect Embedding	Position Aware	Aspect Attention
AttLSTM	✅	❌	✅
GatedCNN	✅	❌	✅
MemNet	❌	❌	✅
GCN	❌	✅	✅
TD-LSTM	❌	✅	❌
CapsBERT	❌	❌	✅
BERT	❌	❌	❌
BERT-PT	❌	❌	❌

[Paper] Tasty Burgers, Soggy Fries: Probing Aspect Robustness in Aspect-Based Sentiment Analysis

Training Strategy

- Train on complex data (MAMS)
- Adversarial Training

Model	Restaurant				Laptop		
	O→O	O→N	MAMS→N	Adv→N	O→O	O→N	Adv→N
MemNet	75.18	21.52	24.02	37.95	64.42	16.93	31.82
GatedCNN	76.96	13.13	18.48	37.50	65.67	10.34	41.85
AttLSTM	75.98	14.64	22.32	48.66	67.55	9.87	42.63
TD-LSTM	78.12	30.18	41.60	62.76	68.03	22.57	54.86
GCN	77.86	24.73	46.51	61.52	72.41	19.91	56.43
BERT-Sent	80.62	10.89	12.95	45.80	73.04	17.40	53.92
CapsBERT	83.66	55.36	61.43	75.80	76.80	25.86	61.23
BERT	83.04	54.82	62.77	74.82	77.59	50.94	65.67
BERT-PT	86.70	59.29	62.77	74.64	78.53	53.29	66.93

Conclusions

- We proposed a simple but effective mechanism to probe the aspect robustness of the models.
- We enhanced the test sets: SemEval 2014 laptop test sets by 294% and restaurant test sets by 315%.
- We probed the aspect robustness of nine ABSA models, and discussed how to improve robustness.

😊 Contact: xyxing18@fudan.edu.cn

Q&A

文本摘要的跨数据集迁移研究

分享者：陈怡然
导师：邱锡鹏教授
复旦大学自然语言处理组

Outline

Outline

CONTENTS

01 Introduction of Text Summarization

02 Motivation

03 Experiment setup and results

04 Conclusions

Introduction of Text Summarization

- **Task description：**

  - A subtask of text generation.

  - shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content.

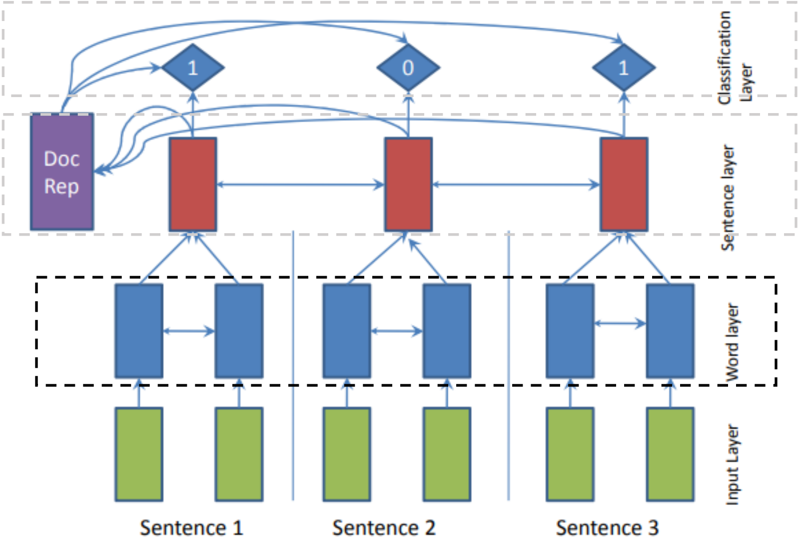
  - Fluent, grammatically correct, repetition, concise, faithfulness, saliency.

- **Main types of summarization systems：**

  - Extractive summarizer (sentence encoder, document encoder, decoder)

  - Abstractive summarizer (encoder decoder)

Introduction of Text Summarization



decoder

document encoder

sentence encoder

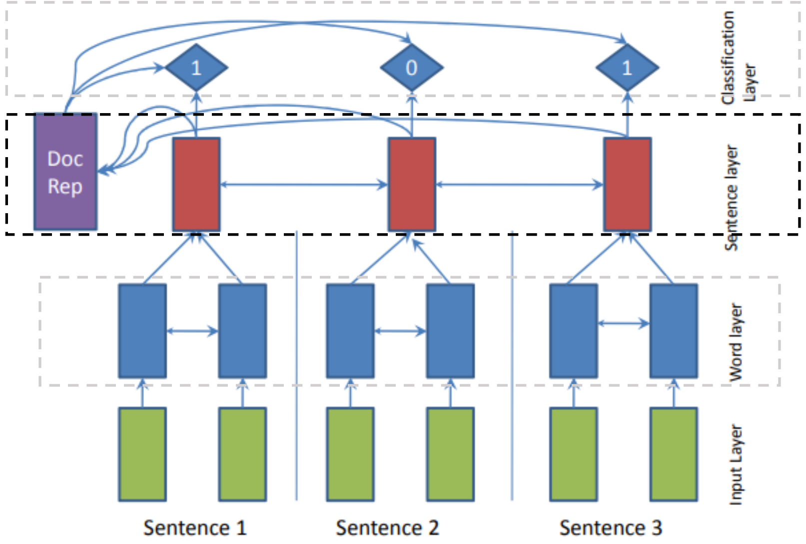
- **Main types of summarization systems：**

  - Extractive summarizer (sentence encoder, document encoder, decoder)

  - Abstractive summarizer (encoder decoder)

Picture: Nallapati, R., Zhai, F., & Zhou, B. (2016). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. arXiv preprint arXiv:1611.04230.

Introduction of Text Summarization



decoder

document encoder

sentence encoder

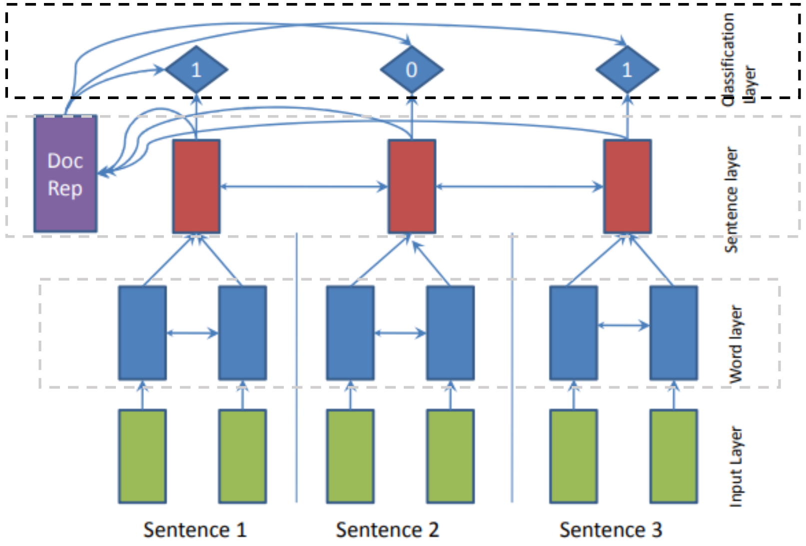
- **Main types of summarization systems：**

  - Extractive summarizer (sentence encoder, document encoder, decoder)

  - Abstractive summarizer (encoder decoder)

Picture: Nallapati, R., Zhai, F., & Zhou, B. (2016). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. arXiv preprint arXiv:1611.04230.

Introduction of Text Summarization



decoder

document encoder

sentence encoder

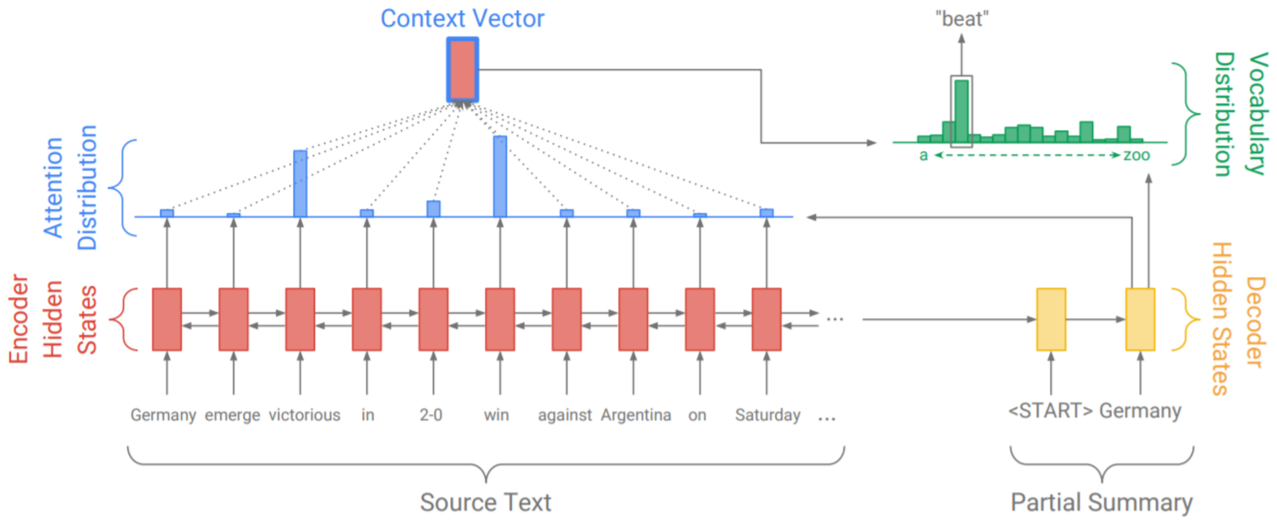
- **Main types of summarization systems：**

  - Extractive summarizer (sentence encoder, document encoder, decoder)

  - Abstractive summarizer (encoder decoder)

Picture: Nallapati, R., Zhai, F., & Zhou, B. (2016). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. arXiv preprint arXiv:1611.04230.

Introduction of Text Summarization



- **Main types of summarization systems：**

  - Extractive summarizer (sentence encoder, document encoder, decoder)

  - Abstractive summarizer (encoder decoder)

Picture: See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Our Work

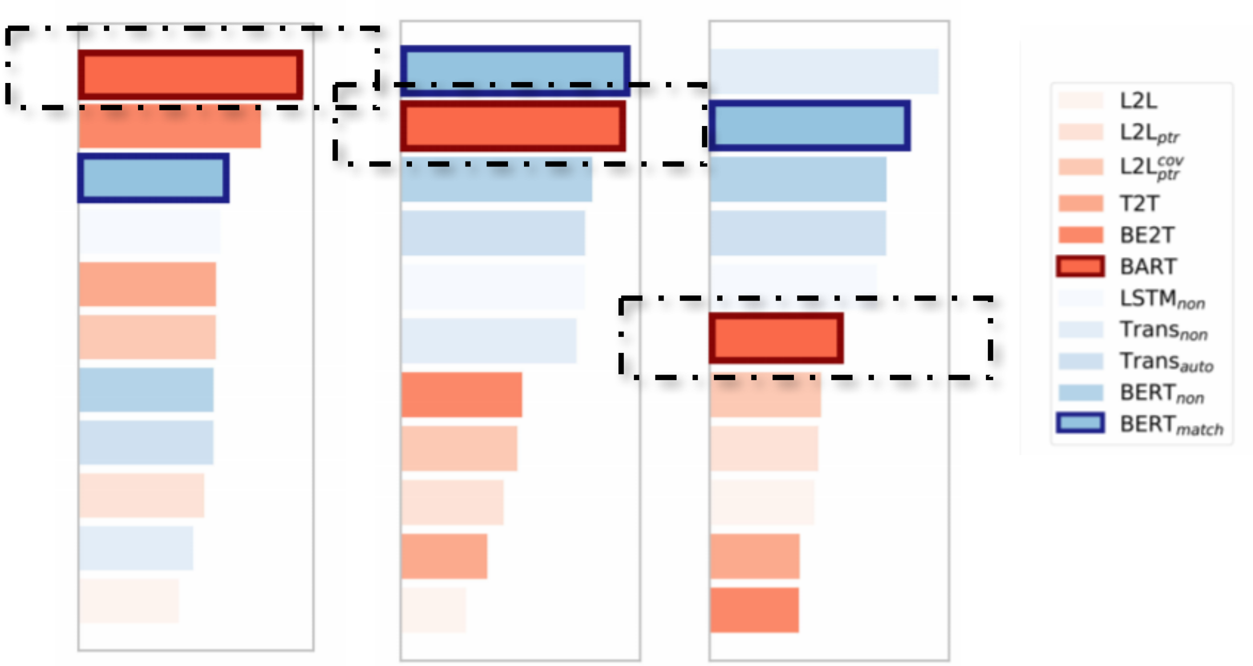
CDEvalSumm: An Empirical Study of _Cross-Dataset Evaluation_ for Neural _Summarization_ Systems

Yiran Chen[*], **Pengfei Liu**[♯*], **Ming Zhong**, **Zi-Yi Dou**[♯], **Danqing Wang**,
Xipeng Qiu[†], **Xuanjing Huang**
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
2005 Songhu Road, Shanghai, China
♯Carnegie Mellon University
{yrchen19,mzhong18,dqwang18,xpqiu,xjhuang}@fudan.edu.cn
{zdou,pliu3}@cs.cmu.edu

Motivation: Ranking Systems based on Different Metrics

In-dataset *New Settings*

BAR
T



Metric: *ROUGE* *Metric1* *Metric2*

- Ranking in a **descending** order
- Each bin -> a system
- Orange -> abstractive systems
- Blue -> extractive systems

Observations

- *The existing SOTA system will not be a SOTA model under CD setting*
- *Abstractive summarizers (in orange) are extremely brittle compared with extractive approaches (larger performance gap)*

Motivation

- **Two questions：**
    - **Q1**: How do different neural architectures of summarizers influence the cross-dataset generalization performances?
    - **Q2**: Do different generation ways (extractive and abstractive) of summarizers influence the cross-dataset generalization ability?

Experiments -- setup

Datasets：

- CNN/DailyMail, Xsum, Pubmed, Bigpatent B, Reddit TIFU

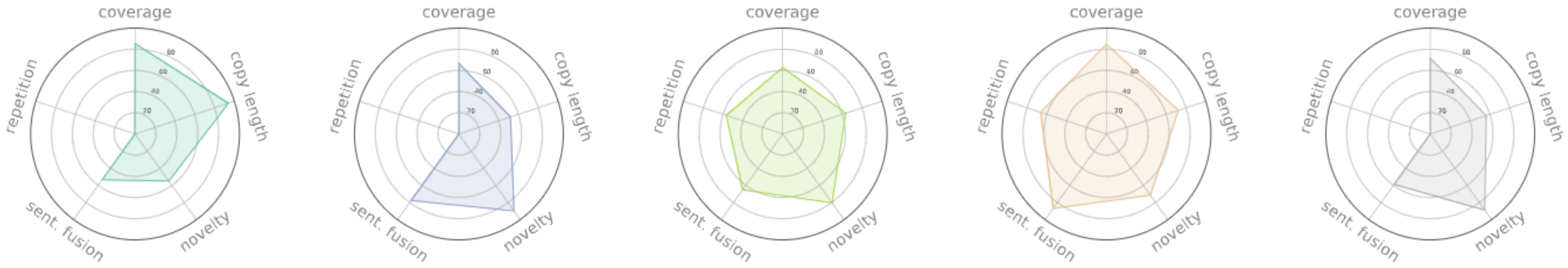
Summarization systems：

- Extractive: $LSTM_{non}$, $Trans_{non}$, $Trans_{auto}$, $BERT_{non}$, $BERT_{match}$
- Abstractive: $L2L$, $L2L_{ptr}$, $L2L_{prt}^{cov}$, $T2T$, $BE2T$, $BART$

Experiments -- setup

Metrics：

- Semantic equivalence: ROUGE
- Factuality: Factcc (Kry´sci´nski et al., 2019)
- Data bias: Coverage, Copy Length, Repetition, Novelty, Sentence fusion score



(a) CNN. (b) Xsum (c) PubMed (d) Bigatent b (e) Reddit

Wojciech Kry´sci´nski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. arXiv, pages arXiv–1910..

Experiments -- setup

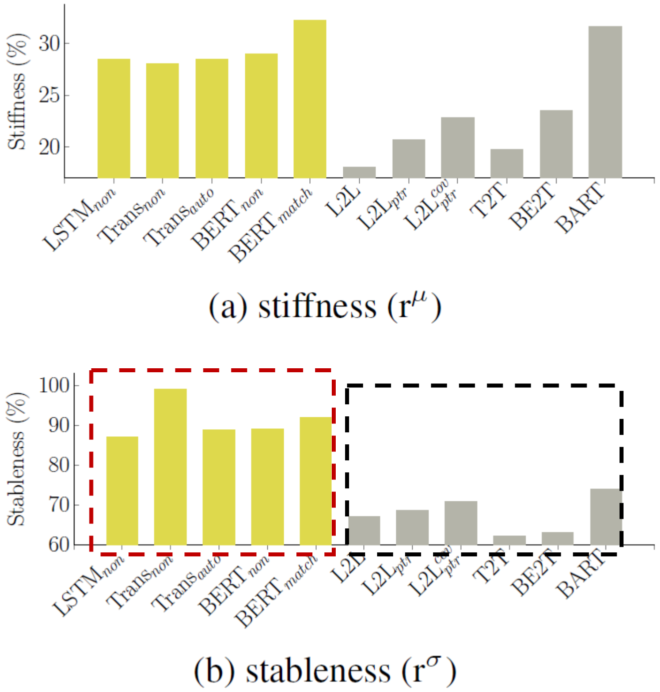
Cross-dataset Measures ：

- Stiffness: $r^{\mu} = \frac{1}{N*N} \Sigma_{i,j} U_{ij}$
- Stableness: $r^{\sigma} = \frac{1}{N*N} \Sigma_{i,j} U_{ij}/U_{jj} \times 100\%$

	$\mathbf{U}_A$ a	$\mathbf{U}_A$ b		$\mathbf{U}_B$ a	$\mathbf{U}_B$ b	Measures	$\mathbf{U}_A$	$\mathbf{U}_B$
a	48	40	a	61	43	*Stiff.*	44	55
b	41	45	b	46	69	*Stable.*	94	84

Table 3: Illustration of two views (*Stiffness*: r^u and *Stableness*: r^{σ}) to characterize the cross-dataset (a and b) generalization based on model A and B. $\mathbf{U_A}$ and $\mathbf{U_B}$ represent two cross-dataset matrix of two models. $r^{\mu}(\mathbf{U_A}) < r^{\mu}(\mathbf{U_B})$ means the model B gains a better cross-dataset absolute performance while $r^{\sigma}(\mathbf{U_A}) > r^{\sigma}(\mathbf{U_B})$ suggests the model A is more robust.

Experiments – ROUGE holistic result



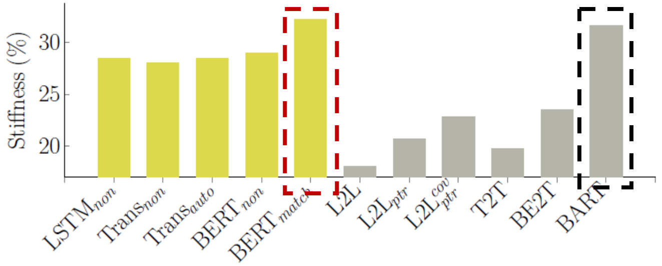
(a) stiffness (r^{μ})

(b) stableness (r^{σ})

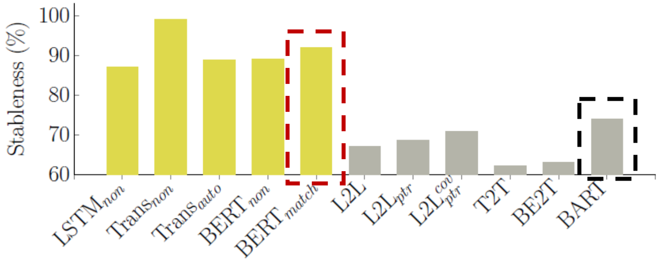
Figure 4: Illustration of stiffness and stableness of ROUGE-1 F1 scores for various models. Yellow bars stand for extractive models and grey bars stand for abstractive models.

- Abstractive models are more brittle compared with extractive models.

Experiments – ROUGE holistic result



(a) stiffness (r^{μ})

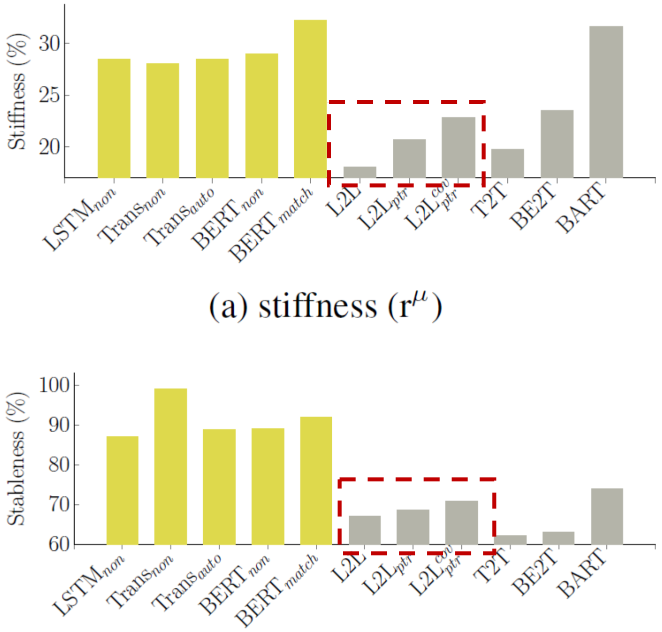


(b) stableness (r^{σ})

Figure 4: Illustration of stiffness and stableness of ROUGE-1 F1 scores for various models. Yellow bars stand for extractive models and grey bars stand for abstractive models.

- Abstractive models are more brittle compared with extractive models.
- $Bart$ is comparable with $Bert_{match}$ in absolute performance. But still lack stableness.

Experiments – ROUGE holistic result



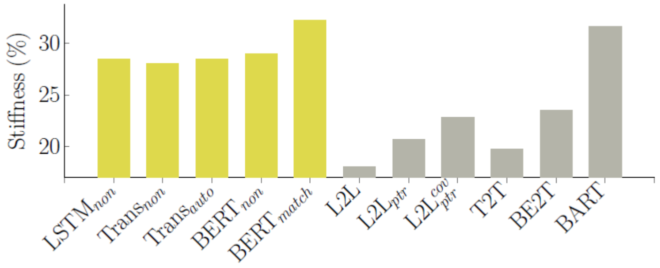
(a) stiffness (r^{μ})

(b) stableness (r^{σ})

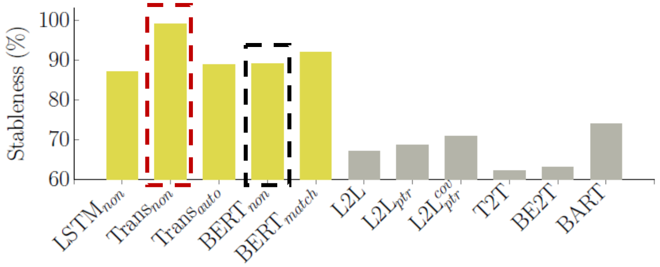
Figure 4: Illustration of stiffness and stableness of ROUGE-1 F1 scores for various models. Yellow bars stand for extractive models and grey bars stand for abstractive models.

- Abstractive models are more brittle compared with extractive models.
- $Bart$ is comparable with $Bert_{match}$ in absolute performance. But still lack stableness.
- Pointer network and coverage mechanism can improve both stiffness and stableness.

Experiments – ROUGE holistic result



(a) stiffness (r^{μ})

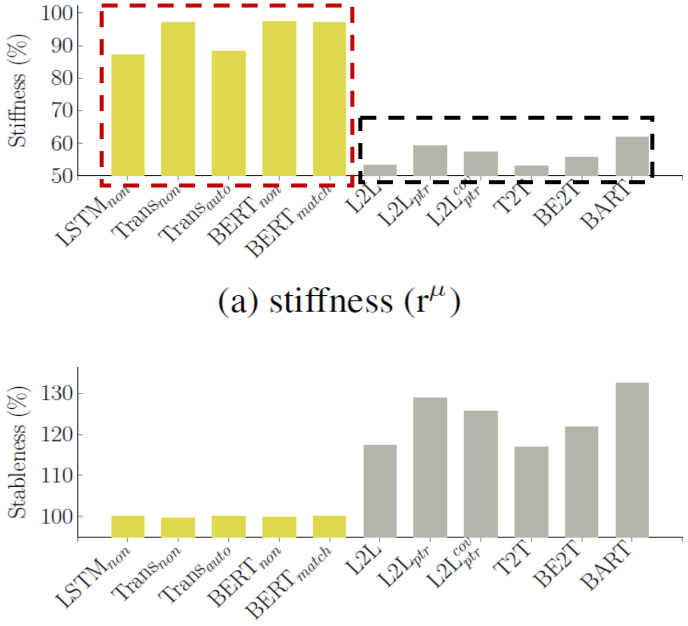


(b) stableness (r^{σ})

Figure 4: Illustration of stiffness and stableness of ROUGE-1 F1 scores for various models. Yellow bars stand for extractive models and grey bars stand for abstractive models.

- Abstractive models are more brittle compared with extractive models.
- $Bart$ is comparable with $Bert_{match}$ in absolute performance. But still lack stableness.
- Pointer network and coverage mechanism can improve both stiffness and stableness.
- $Bert_{non}$ is less stable compared with $Trans_{non}$ though the former equipped with BERT.

Experiments – Factcc holistic result



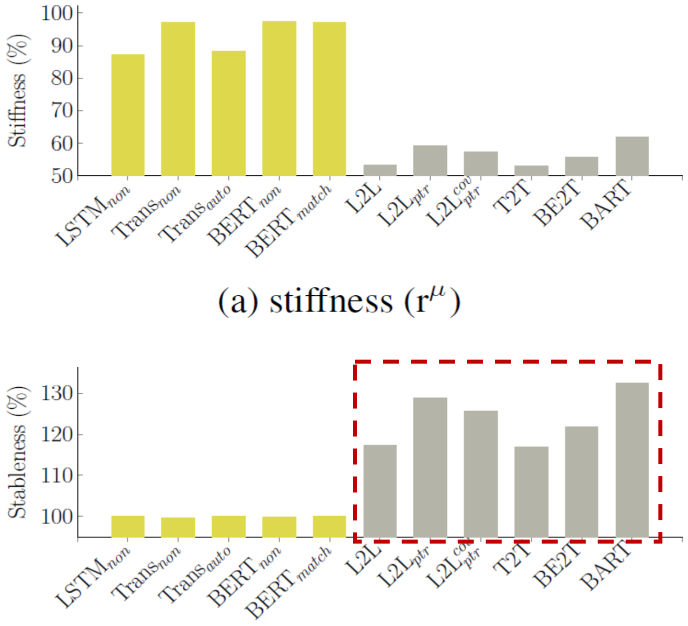
(a) stiffness (r^{μ})

(b) stableness (r^{σ})

Figure 5: Illustration of stiffness and stableness of factuality scores for various models. Yellow bars stand for extractive systems and grey bars stand for abstractive systems.

- Abstractive summarization systems perform extremely worse than extractive summarizers under the metric: factcc.

Experiments – Factcc holistic result



(b) stableness (r^{σ})

Figure 5: Illustration of stiffness and stableness of factuality scores for various models. Yellow bars stand for extractive systems and grey bars stand for abstractive systems.

- Abstractive summarization systems perform extremely worse than extractive summarizers under the metric: factcc.
- Abstractive summarizers possess better cross-dataset performance than in-dataset performance.

Experiments – fine-grained result

analysis aspect	Architecture				Generation way	
model type	EXT		ABS		LSTM	BERTSUM
compare models	$BERT_{match}$ vs. $BERT_{non}$	$BERT_{non}$ vs. $Trans_{non}$	$L2L_{ptr}$ vs. L2L	$L2L^{cov}_{ptr}$ vs. $L2L_{ptr}$	$LSTM_{non}$ vs. L2L	$BERT_{non}$ vs. BE2T
holistic analysis	stiff. : 32.27 vs. 28.98 stable. : 91.98 vs. 88.93	stiff. : 28.98 vs. 28.02 stable. : 88.93 vs. 99.05	stiff. : 20.74 vs. 18.03 stable. : 68.63 vs. 66.93	stiff. : 22.81 vs. 20.74 stable. : 70.71 vs. 68.63	stiff. : 28.51 vs. 18.03 stable. : 87.00 vs. 66.93	stiff. : 28.98 vs. 23.49 stable. : 88.93 vs. 62.93

Fine-grain analysis (ROUGE), $BERT_{match}$ vs. $BERT_{non}$:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	1.6	4.1	4.5	3.0	4.7	3.6
	Xsum	2.9	3.2	3.5	1.6	5.7	3.4
	Pubm.	0.9	4.0	2.4	0.2	8.7	3.3
	Patent b	4.6	3.1	3.5	3.0	3.7	3.6
	Red.	3.3	4.2	3.5	-1.4	3.5	2.6
	avg	2.6	3.7	3.5	1.3	5.3	3.3
		(a)					
normali.	CNN.	0.0	5.8	5.3	0.7	6.9	3.7
	Xsum	3.4	0.0	2.8	-2.7	11.5	3.0
	Pubm.	-1.2	6.1	0.0	-6.5	26.5	5.0
	Patent b	7.3	1.8	2.8	0.0	3.3	3.0
	Red.	4.4	6.2	2.9	-10.5	0.0	0.6
	avg	2.8	4.0	2.7	-3.8	9.6	3.1
		(g)					

$BERT_{non}$ vs. $Trans_{non}$:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	1.8	1.2	0.3	0.8	-10.9	-1.3
	Xsum	-0.9	6.0	0.1	-1.6	-0.7	0.6
	Pubm.	2.5	1.4	0.3	0.6	-2.2	0.5
	Patent b	0.5	1.1	0.2	1.4	3.8	1.4
	Red.	8.3	3.0	-0.1	1.6	5.6	3.7
	avg	2.4	2.5	0.2	0.6	-0.9	1.0
		(b)					
normali.	CNN.	0.0	-23.9	0.1	-1.5	-96.6	-24.4
	Xsum	-6.1	0.0	-0.5	-8.3	-31.8	-9.3
	Pubm.	2.0	-21.0	0.0	-2.2	-33.7	-11.0
	Patent b	-2.6	-24.8	-0.2	0.0	-5.5	-6.6
	Red.	16.3	-12.8	-1.0	1.0	0.0	0.7
	avg	1.9	-16.5	-0.3	-2.2	-33.5	-10.1
		(h)					

$L2L_{ptr}$ vs. L2L:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	4.3	0.5	5.3	3.2	1.5	3.0
	Xsum	3.4	1.4	3.4	4.2	0.1	2.5
	Pubm.	10.3	2.3	4.2	3.0	2.6	4.5
	Patent b	1.1	-1.1	2.5	0.6	-0.3	0.5
	Red.	2.2	3.1	2.6	2.9	4.4	3.0
	avg	4.2	1.2	3.6	2.8	1.7	2.7
		(c)					
normali.	CNN.	0.0	-1.0	4.8	8.7	-9.9	0.5
	Xsum	1.8	0.0	0.7	12.2	-13.8	0.2
	Pubm.	23.3	5.6	0.0	8.4	1.6	7.8
	Patent b	-1.6	-5.8	-0.4	0.0	-14.5	-4.4
	Red.	1.9	8.7	3.4	8.4	0.0	4.5
	avg	5.1	1.5	1.7	7.5	-7.3	1.7
		(i)					

$L2L^{cov}_{ptr}$ vs. $L2L_{ptr}$:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	2.9	1.8	6.4	3.4	1.7	3.2
	Xsum	-0.8	-0.8	-4.5	-2.4	-0.1	-1.7
	Pubm.	4.5	1.7	3.2	3.4	2.7	3.1
	Patent b	1.0	2.0	2.2	4.9	0.8	2.2
	Red.	3.3	1.0	6.5	6.9	-0.0	3.5
	avg	2.2	1.1	2.8	3.2	1.0	2.1
		(d)					
normali.	CNN.	0.0	8.1	9.6	-4.1	8.0	4.3
	Xsum	-6.7	0.0	-19.7	-18.0	-0.2	-8.9
	Pubm.	6.7	7.4	0.0	-1.2	12.6	5.1
	Patent b	-0.1	8.1	0.8	0.0	3.7	2.5
	Red.	5.6	4.9	14.8	11.7	0.0	7.4
	avg	1.1	5.7	1.1	-2.3	4.8	2.1
		(j)					

$LSTM_{non}$ vs. L2L:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	8.6	0.1	13.2	4.9	2.0	5.7
	Xsum	13.1	-8.8	18.3	7.1	3.8	6.7
	Pubm.	18.6	4.8	15.1	11.1	9.0	11.7
	Patent b	19.7	2.8	22.8	8.8	5.9	12.0
	Red.	21.4	7.3	30.7	18.0	3.6	16.2
	avg	16.3	1.2	20.0	10.0	4.9	10.5
		(e)					
normali.	CNN.	0.0	28.4	-0.7	-7.9	-4.8	3.0
	Xsum	18.3	0.0	15.8	2.0	6.6	8.5
	Pubm.	36.8	44.3	0.0	12.5	35.1	25.7
	Patent b	39.6	35.2	31.4	0.0	17.8	24.8
	Red.	44.7	52.9	58.4	35.1	0.0	38.2
	avg	27.9	32.2	21.0	8.3	10.9	20.1
		(k)					

$BERT_{non}$ vs. BE2T:

	fine-grain analysis	CNN.	Xsum	Pubm.	Patent b	Red.	avg
origin	CNN.	1.3	-2.0	3.5	-1.8	-1.7	-0.1
	Xsum	12.9	-17.2	18.3	9.9	1.5	5.1
	Pubm.	17.2	2.9	1.6	-0.3	0.3	4.3
	Patent b	21.8	6.7	15.4	-7.2	5.1	8.4
	Red.	17.8	4.6	20.2	11.4	-4.8	9.8
	avg	14.2	-1.0	11.8	2.4	0.1	5.5
		(f)					
normali.	CNN.	0.0	31.5	5.2	11.1	9.0	11.4
	Xsum	28.4	0.0	45.0	37.8	19.9	26.2
	Pubm.	38.7	42.0	0.0	14.5	11.5	21.3
	Patent b	49.9	53.7	37.2	0.0	34.1	35.0
	Red.	40.1	48.4	50.2	41.5	0.0	36.1
	avg	31.4	35.1	27.5	21.0	14.9	26.0
		(l)					

Table 4: The difference of ROUGE-1 F1 scores between different model pairs. Every column of the table represents the compared results of one pair of models. The line of holistic analysis displays the overall stiffness and stableness of compared models. The rest of the table is fine-grained results, the first line of which is the origin compared results ($\mathbf{U_A} - \mathbf{U_B}$ for model pairs A and B) and the second line is the normalized compared results ($\hat{\mathbf{U}}_\mathbf{A} - \hat{\mathbf{U}}_\mathbf{B}$ for model pairs A and B). For all heatmap, 'grey' and 'red' represent positive and negative respectively. Here we only display compared results for limited pairs of models, all other results are displayed in appendix.

Conclusion

- Abstractive summarizers are extremely brittle compared with extractive approaches.
- BART (SOTA system) is superior over other abstractive models and even comparable with extractive models in terms of stiffness (ROUGE).
- The robustness of models can be improved through either equipped the model with ability to copy span from source document or make use of well trained sequence to sequence pre-trained model (BART).
- Simply adding BERT on encoder could improve the stiffness (ROUGE) of model but will cause larger cross-dataset and in-dataset performance gap.
- Existing factuality checker (Factcc) is limited in predictive power of positive samples.

Conclusion

Contribution:

1. Cross-dataset evaluation is orthogonal to other evaluation aspects (e.g., semantic equivalence, factuality)
2. We have design two measures Stiffness and Stableness, which could help us to characterize generalization ability in different views, encouraging us to diagnose the weaknesses of state-of-the-art systems.
3. We conduct dataset bias-aided analysis and suggest that a better understanding of datasets will be helpful for us to interpret systems' behaviours.

Thanks & QA