



ACL2020 Best Paper & Honorable Mention Papers

Changying Hao



- 1. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Best paper)**
2. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks
3. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

CheckList



1 Motivation

1. Measuring held-out accuracy is a primary approach to evaluate generalization, but it often overestimates the performance of NLP models.
2. Alternative approaches for evaluating models either focus on individual tasks or on specific behaviors.
3. Inspired by principles of behavioral testing in software engineering, this paper introduces CheckList, a task-agnostic methodology for testing NLP models.

CheckList



2 Checklist Check what?

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
...			

Check some natural language *capabilities* that are manifested on the task to be test.

Other capabilities such as *Taxonomy*, *Robustness*, *Fairness*, *Temporal*, *Coreference*, *Semantic Role Labeling*, and *Logic*.

CheckList



2 Checklist How to check?

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
...			

Minimum Functionality test (MFT)

Test case	Expected	Predicted	Pass?
A Testing Negation with MFT Template: I {NEGATION} {POS_VERB} the {THING} .	Labels: negative, positive, neutral		
 I can't say I recommend the food.			
I didn't love the flight.			
	neg	pos	X
	neg	neutral	X
...			
Failure rate = 76.4%			

CheckList



2 Checklist How to check?

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
...			

Invariance test (INV)

Test case	Expected	Predicted	Pass?
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%			

CheckList



2 Checklist How to check?

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
...			

Directional Expectation test (DIR)

Test case	Expected	Predicted	Pass?
C Testing Vocabulary with DIR @AmericanAir service wasn't great. You are lame.	Sentiment monotonic decreasing (\downarrow)	neg neutral	x
@JetBlue why won't YOU help them?! Ugh. I dread you.	\downarrow	neg neutral	x
...			
Failure rate = 34.6%			

CheckList



2 Checklist Generate Test Cases at Scale

Use Templates and RoBERTa mask-and-fill suggestions

<https://github.com/marcotcr/checklist>

```
In [27]: editor.visual_suggest('This is {a:mask} movie.')
```

> This is **a:mask** movie .

FILL IN WITH...

- Check All
- a good
- an amazing
- an excellent
- an awful

Preview

No Data

```
In [26]: editor.selectedSuggestions
```

Wordnet



3 Testing SOTA models with CheckList

sentiment analysis (*Sentiment*):

Microsoft Text Analytics, Google Clouds Natural Language, Amazon Comprehend, BERT-base and RoBERTa-base (RoB)

duplicate question (*QQP*):

BERT-base and RoBERTa-base (RoB)

machine comprehension (*MC*):

BERT-large

CheckList



3 Testing SOTA models with CheckList

Labels: positive, negative, or neutral; INV: same pred. (INV) after **removals/ additions**; DIR: sentiment should not decrease (\uparrow) or increase (\downarrow)

Test TYPE and Description	Failure Rate (%)						Example test cases & expected behavior
	Windows	G	a	RoB			
Vocab.+POS	<i>MFT</i> : Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	<i>MFT</i> : Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	<i>INV</i> : Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	<i>DIR</i> : Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	<i>DIR</i> : Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	<i>INV</i> : Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	<i>INV</i> : Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtue I cri INV @SouthwestAir no thanks → thakns INV
NER	<i>INV</i> : Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	<i>INV</i> : Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D, thanks. INV

Table 1: A selection of tests for **sentiment analysis**. All examples (right) are failures of at least one model.

CheckList



3 Testing SOTA models with CheckList

Label: duplicate \equiv , or non-duplicate \neq ; INV: same pred. (INV) after removals/ additions							
	Test TYPE and Description	Failure Rate		Example Test cases & expected behavior			
			RoB				
Vocab.	MFT: Modifiers changes question intent	78.4	78.0	{ Is Mark Wright a photographer? Is Mark Wright an accredited photographer? }	\neq		
	MFT: Synonyms in simple templates	22.8	39.2	{ How can I become more vocal? How can I become more outspoken? }	\equiv		
	INV: Replace words with synonyms in real pairs	13.1	12.7	Is it necessary to follow a religion? Is it necessary to follow an organized → organised religion?	$\}$	INV	
Taxonomy	MFT: More X = Less antonym(X)	69.4	100.0	{ How can I become more optimistic? How can I become less pessimistic? }	\equiv		
	INV: Swap one character with its neighbor (typo)	18.2	12.0	{ Why am I getting → gettnig lazy? Why are we so lazy? }	INV		
	DIR: Paraphrase of question should be duplicate	69.0	25.0	Can I gain weight from not eating enough? Can I → Do you think I can gain weight from not eating enough?	$\}$	=	
Robust.	INV: Change the same name in both questions	11.8	9.4	Why isn't Hillary Clinton → Nicole Perez in jail? Is Hillary Clinton → Nicole Perez going to go to jail?	$\}$	INV	
	DIR: Change names in one question, expect \neq	35.1	30.1	What does India think of Donald Trump? What India thinks about Donald Trump → John Green ?	$\}$	\neq	
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect \neq	30.0	32.8	Will it be difficult to get a US Visa if Donald Trump gets elected? Will the US accept Donald Trump?	$\}$	\neq	
NER	DIR: Change names in one question, expect \neq	35.1	30.1	What does India think of Donald Trump? What India thinks about Donald Trump → John Green ?	$\}$	\neq	
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect \neq	30.0	32.8	Will it be difficult to get a US Visa if Donald Trump gets elected? Will the US accept Donald Trump?	$\}$	\neq	
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect \neq	30.0	32.8	Will it be difficult to get a US Visa if Donald Trump gets elected? Will the US accept Donald Trump?	$\}$	\neq	

Table 2: A selection of tests for **Quora Question Pair**. All examples (right) are failures of at least one model.

CheckList



3 Testing SOTA models with CheckList

	Test TYPE and Description	Failure Rate (⌚)	Example Test cases (with expected behavior and 🎨 prediction)
Vocab	MFT: comparisons	20.0	C: Victoria is younger than Dylan. Q: Who is less young? A: Dylan 🎨: Victoria
	MFT: intensifiers to superlative: most/least	91.3	C: Anna is worried about the project. Matthew is extremely worried about the project. Q: Who is least worried about the project? A: Anna 🎨: Matthew
Taxonomy	MFT: match properties to categories	82.4	C: There is a tiny purple box in the room. Q: What size is the box? A: tiny 🎨: purple
	MFT: nationality vs job	49.4	C: Stephanie is an Indian accountant. Q: What is Stephanie's job? A: accountant 🎨: Indian accountant
	MFT: animal vs vehicles	26.2	C: Jonathan bought a truck. Isabella bought a hamster. Q: Who bought an animal? A: Isabella 🎨: Jonathan
	MFT: comparison to antonym	67.3	C: Jacob is shorter than Kimberly. Q: Who is taller? A: Kimberly 🎨: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	C: Jeremy is more optimistic than Taylor. Q: Who is more pessimistic? A: Taylor 🎨: Jeremy
Robust.	INV: Swap adjacent characters in Q (typo)	11.6	C: ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... Q: What was the ideal duty → udty of a Newcomen engine? A: INV 🎨: 7 million → 5 million
	INV: add irrelevant sentence to C	9.8	(no example)

Table 3: A selection of tests for **Machine Comprehension**.



4 User Evaluation

1. The team of Microsoft Text Analytics state that CheckList is very helpful.
2. Compare the number of test cases created by people with different conditions for testing BERT on the QQP validation dataset.

Unaided	CHECKLIST		
	Cap. only	Cap.+templ.	
#Tests	5.8 ± 1.1	10.2 ± 1.8	13.5 ± 3.4
#Cases/test	7.3 ± 5.6	5.0 ± 1.2	198.0 ± 96
#Capabilities tested	3.2 ± 0.7	7.5 ± 1.9	7.8 ± 1.1
Total severity	10.8 ± 3.8	21.7 ± 5.7	23.7 ± 4.2
#Bugs ($sev \geq 3$)	2.2 ± 1.2	5.5 ± 1.7	6.2 ± 0.9

Table 4: **User Study Results**: first three rows indicate number of tests created, number of test cases per test and number of capabilities tested. Users report the severity of their findings (last two rows).



5 Conclusion

1. This paper proposes a model-agnostic and task-agnostic testing methodology **CheckList** that tests individual *capabilities* of the model using three different test types.
2. It highlights significant problems at multiple levels in the conceptual NLP pipeline for models that have “solved” existing benchmarks on three different tasks.
3. User studies show the helpfulness of the CheckList.



1. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Best paper)
2. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks
3. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Don't stop pretraining



1 Motivation

1. Language models pretrained on text from a wide variety of sources form the foundation of today's NLP.
2. To test whether it is still helpful to tailor a pretrained model (RoBERTa) to the domain of a target task, this paper studies the second phase of pretraining in domain (*domain-adaptive pretraining*) and the influence of adapting to the task's unlabeled data (*task-adaptive pretraining*).

Don't stop pretraining



2 Domain-Adaptive Pretraining (DAPT)

Domain Similarity

Continue pretraining RoBERTa on a large corpus of unlabeled domain-specific text.

PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0
	PT	News	Reviews	BioMed	CS

Figure 1: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to RoBERTa's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Don't stop pretraining



2 Domain-Adaptive Pretraining (DAPT). Mask LM Loss

Continue pretraining the pretrained RoBERTa on each domain for 12.5k steps (a single pass).

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

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

Don't stop pretraining



2 Domain-Adaptive Pretraining (DAPT). Specifications of Datasets

Consider two text classification tasks under each domain.

Domain	Task	Label Type	Train (Lab.)	Train (Unl.)	Dev.	Test	Classes
BIOMED	CHEMPROT	relation classification	4169	-	2427	3469	13
	[†] RCT	abstract sent. roles	18040	-	30212	30135	5
CS	ACL-ARC	citation intent	1688	-	114	139	6
	SCIERC	relation classification	3219	-	455	974	7
NEWS	HYPERPARTISAN	partisanship	515	5000	65	65	2
	[†] AGNEWS	topic	115000	-	5000	7600	4
REVIEWS	[†] HELPFULNESS	review helpfulness	115251	-	5000	25000	2
	[†] IMDB	review sentiment	20000	50000	5000	25000	2

Table 2: Specifications of the various target task datasets. [†] indicates high-resource settings.

Don't stop pretraining



2 Domain-Adaptive Pretraining (DAPT). Main test Results

Dom.	Task	RoBa.	DAPT	\neg DAPT
BM	CHEMPROT	$81.9_{1.0}$	$84.2_{0.2}$	$79.4_{1.3}$
	\dagger RCT	$87.2_{0.1}$	$87.6_{0.1}$	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$66.4_{4.1}$
	SCIERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.2_{0.9}$
NEWS	HYP.	$86.6_{0.9}$	$88.2_{5.9}$	$76.4_{4.9}$
	\dagger AGNEWS	$93.9_{0.2}$	$93.9_{0.2}$	$93.5_{0.2}$
REV.	\dagger HELPFUL.	$65.1_{3.4}$	$66.5_{1.4}$	$65.1_{2.8}$
	\dagger IMDB	$95.0_{0.2}$	$95.4_{0.2}$	$94.1_{0.4}$

Table 3: Comparison of RoBERTa (RoBa.) and DAPT to adaptation to an *irrelevant* domain (\neg DAPT). Reported results are test F1. \dagger indicates high-resource settings. Best task performance is boldfaced.

Don't stop pretraining



2 Domain-Adaptive Pretraining (DAPT). Results Analysis

Dom.	Task	RoBA.	DAPT	\neg DAPT
BM	CHEMPROT	81.9 _{1.0}	84.2 _{0.2}	79.4 _{1.3}
	[†] RCT	87.2 _{0.1}	87.6 _{0.1}	86.9 _{0.1}
CS	ACL-ARC	63.0 _{5.8}	75.4 _{2.5}	66.4 _{4.1}
	SCIERC	77.3 _{1.9}	80.8 _{1.5}	79.2 _{0.9}
NEWS	HYP.	86.6 _{0.9}	88.2 _{5.9}	76.4 _{4.9}
	[†] AGNEWS	93.9 _{0.2}	93.9 _{0.2}	93.5 _{0.2}
REV.	[†] HELPFUL.	65.1 _{3.4}	66.5 _{1.4}	65.1 _{2.8}
	[†] IMDB	95.0 _{0.2}	95.4 _{0.2}	94.1 _{0.4}

1. DAPT improves over RoBERTa in all domains.
2. DAPT outperforms adapting to an irrelevant domain.
3. \neg DAPT results in worse performance than even RoBERTa.

Don't stop pretraining



3 Task-Adaptive Pretraining (TAPT). Main Test Results

Pretraining on the unlabeled training set for a given task .

The pretraining corpus is smaller than DAPT, but is much more task-relevant.

Perform TAPT for 100 epochs (randomly mask 15% words across epochs)

Domain	Task	ROBERTA	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BIOMED	CHEMPROT	81.9 _{1.0}	84.2 _{0.2}	82.6 _{0.4}	84.4 _{0.4}
	†RCT	87.2 _{0.1}	87.6 _{0.1}	87.7 _{0.1}	87.8 _{0.1}
CS	ACL-ARC	63.0 _{5.8}	75.4 _{2.5}	67.4 _{1.8}	75.6 _{3.8}
	SCIERC	77.3 _{1.9}	80.8 _{1.5}	79.3 _{1.5}	81.3 _{1.8}
NEWS	HYPERPARTISAN	86.6 _{0.9}	88.2 _{5.9}	90.4 _{5.2}	90.0 _{6.6}
	†AGNEWS	93.9 _{0.2}	93.9 _{0.2}	94.5 _{0.1}	94.6 _{0.1}
REVIEWS	†HELPFULNESS	65.1 _{3.4}	66.5 _{1.4}	68.5 _{1.9}	68.7 _{1.8}
	†IMDB	95.0 _{0.2}	95.4 _{0.1}	95.5 _{0.1}	95.6 _{0.1}

Table 4: Results on different phases of adaptive pretraining compared to the baseline ROBERTA.

Don't stop pretraining



3 Task-Adaptive Pretraining (TAPT). Results Analysis

1. TAPT consistently improves the RoBERTa baseline for all tasks across domains.
2. TAPT even exceed DAPT in some tasks.
3. DAPT followed by TAPT achieves the best.

Domain	Task	ROBERTA	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BIOMED	CHEMPROT	81.9 _{1.0}	84.2 _{0.2}	82.6 _{0.4}	84.4 _{0.4}
	†RCT	87.2 _{0.1}	87.6 _{0.1}	87.7 _{0.1}	87.8 _{0.1}
CS	ACL-ARC	63.0 _{5.8}	75.4 _{2.5}	67.4 _{1.8}	75.6 _{3.8}
	SCIERC	77.3 _{1.9}	80.8 _{1.5}	79.3 _{1.5}	81.3 _{1.8}
NEWS	HYPERPARTISAN	86.6 _{0.9}	88.2 _{5.9}	90.4 _{5.2}	90.0 _{6.6}
	†AGNEWS	93.9 _{0.2}	93.9 _{0.2}	94.5 _{0.1}	94.6 _{0.1}
REVIEWS	†HELPFULNESS	65.1 _{3.4}	66.5 _{1.4}	68.5 _{1.9}	68.7 _{1.8}
	†IMDB	95.0 _{0.2}	95.4 _{0.1}	95.5 _{0.1}	95.6 _{0.1}

Don't stop pretraining



3 Task-Adaptive Pretraining (TAPT). Cross-Task Transfer

Pretrain on the other task, and finetune on this task.

BIOMED	RCT	CHEMPROT
TAPT	$87.7_{0.1}$	$82.6_{0.5}$
Transfer-TAPT	$87.1_{0.4} (\downarrow 0.6)$	$80.4_{0.6} (\downarrow 2.2)$
NEWS	HYPERPARTISAN	AGNEWS
TAPT	$89.9_{9.5}$	$94.5_{0.1}$
Transfer-TAPT	$82.2_{7.7} (\downarrow 7.7)$	$93.9_{0.2} (\downarrow 0.6)$

CS	ACL-ARC	SCIERC
TAPT	$67.4_{1.8}$	$79.3_{1.5}$
Transfer-TAPT	$64.1_{2.7} (\downarrow 3.3)$	$79.1_{2.5} (\downarrow 0.2)$
REVIEWS	HELPFULNESS	IMDB
TAPT	$68.5_{1.9}$	$95.7_{0.1}$
Transfer-TAPT	$65.0_{2.6} (\downarrow 3.5)$	$95.0_{0.1} (\downarrow 0.7)$

Performance becoming worse shows that data distributions of tasks within a given domain might differ.

Don't stop pretraining



3 Task-Adaptive Pretraining (TAPT). Data Augmentation

Use unlabeled data or create unlabeled data to pretrain the LM.

- (1) Use available unlabeled data from the human-curated corpus.
- (2) Retrieve related unlabeled data if human-curated data is unavailable.

Pretraining	BIOMED RCT-500	NEWS HYP.	REVIEWS IMDB †
TAPT	79.8 _{1.4}	90.4 _{5.2}	95.5 _{0.1}
DAPT + TAPT	83.0 _{0.3}	90.0 _{6.6}	95.6 _{0.1}
Curated-TAPT	83.4 _{0.3}	89.9 _{9.5}	95.7 _{0.1}
DAPT + Curated-TAPT	83.8 _{0.5}	92.1 _{3.6}	95.8 _{0.1}

Table 6: Test set F1, † indicates high-resource settings.

Pretraining	BIOMED		CS
	CHEMPROT	RCT-500	ACL-ARC
RoBERTA	81.9 _{1.0}	79.3 _{0.6}	63.0 _{5.8}
TAPT	82.6 _{0.4}	79.8 _{1.4}	67.4 _{1.8}
RAND-TAPT	81.9 _{0.6}	80.6 _{0.4}	69.7 _{3.4}
50NN-TAPT	83.3 _{0.7}	80.8 _{0.6}	70.7 _{2.8}
150NN-TAPT	83.2 _{0.6}	81.2 _{0.8}	73.3 _{2.7}
500NN-TAPT	83.3 _{0.7}	81.7 _{0.4}	75.5 _{1.9}
DAPT	84.2 _{0.2}	82.5 _{0.5}	75.4 _{2.5}

Table 7: Test set F1, comparing Rand-TAPT (with 50 candidates) and kNN-TAPT selection.

Don't stop pretraining



4 Computational Requirements

Pretraining	Steps	Docs.	Storage	F_1
ROBERTA	-	-	-	79.3_{0.6}
TAPT	0.2K	500	80KB	79.8 _{1.4}
50NN-TAPT	1.1K	24K	3MB	80.8 _{0.6}
150NN-TAPT	3.2K	66K	8MB	81.2 _{0.8}
500NN-TAPT	9.0K	185K	24MB	81.7 _{0.4}
Curated-TAPT	8.8K	180K	27MB	83.4_{0.3}
DAPT	12.5K	25M	47GB	82.5 _{0.5}
DAPT + TAPT	12.6K	25M	47GB	83.0 _{0.3}

Table 8: Computational requirements for adapting to the RCT-500 task, comparing DAPT and the various TAPT modifications

Don't stop pretraining



5 Conclusion

1. RoBERTa struggles to encode the complexity of a single textual domain, let alone all of language.
2. Domain-adaptive pretraining and task-adaptive pretraining are helpful.
3. Adapting to a task corpus augmented using simple data selection strategies is an effective alternative, especially when resources for domain-adaptive pretraining might be unavailable.



1. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (Best paper)
2. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks
3. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics

Tangled up in BLEU



1 Motivation

1. Automatic metrics are fundamental for the development and evaluation of machine translation systems. Measuring how well automatic metrics match with human judgements of translation quality is important.
2. Previous works have conflict findings on the evaluation of MT metrics, which raise important questions as to the reliability of the accepted best-practices for ranking metrics, and cast doubt over these metrics' utility for tuning high-quality systems.

Tangled up in BLEU



2 Q1: Are metrics unreliable when evaluating high-quality MT systems?

Human evaluation: direct assessment (DA) scores

Automatic metrics: BLEU, TER, CHRF, YISI-1, ESIM and YISI-2

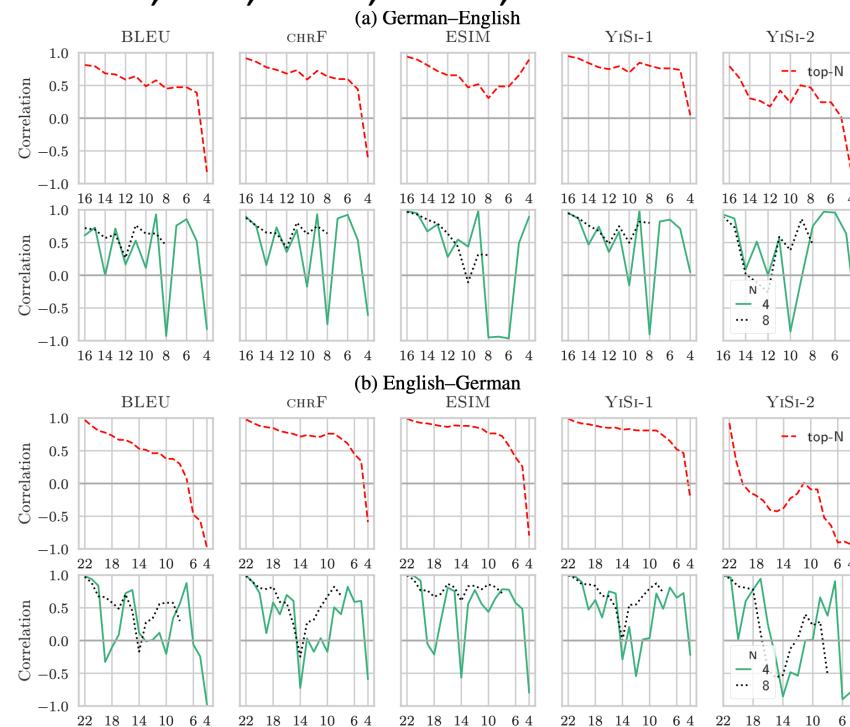


Figure 1: Pearson correlation coefficient computed over the top-N systems (top row), or over a rolling window of 4 or 8 systems (bottom row). The x axis shows the index of the starting system, and systems are sorted by DA quality score.

Tangled up in BLEU



3 Q2: How do outliers affect the correlation of MT evaluation metrics?

A method of detecting outlier systems using human score:

1. Compute Median Absolute Deviation (MAD) , which is the median of all absolute deviations from the median

$$\text{MAD} = 1.483 \times \text{median}(|s - \text{median}(s)|)$$

2. Compute robust scores:

$$z = (s - \text{median}(s)) / \text{MAD}$$

3. Discard systems where the magnitude of z exceeds a cutoff(2.5 in this paper)

Tangled up in BLEU



3 Q2: How do outliers affect the correlation of MT evaluation metrics?

Correlation of metrics with and without outliers (“All” and “–out”, resp.)

#sys	de-en		gu-en		kk-en		lt-en		ru-en		zh-en	
	All	–out										
BLEU	0.81	0.79	0.83	0.97	0.95	0.91	0.96	0.97	0.87	0.81	0.90	0.81
TER	0.87	0.81	0.89	0.95	0.80	0.57	0.96	0.98	0.92	0.90	0.84	0.72
chrF	0.92	0.86	0.95	0.96	0.98	0.77	0.94	0.93	0.94	0.88	0.96	0.84
ESIM	0.94	0.90	0.88	0.99	0.99	0.95	0.99	0.99	0.97	0.95	0.99	0.96
YiSi-1	0.95	0.91	0.92	1.00	0.99	0.92	0.98	0.98	0.98	0.95	0.98	0.90
YiSi-2	0.80	0.61	-0.57	0.82	-0.32	0.66	0.44	0.35	-0.34	0.71	0.94	0.62

Table 1: for the to-English language pairs that contain outlier systems.

#sys	de-cs		en-de		en-fi		en-kk		en-ru		fr-de	
	All	–out	All	–out	All	–out	All	–out	All	–out	All	–out
BLEU	0.87	0.74	0.97	0.81	0.97	0.94	0.85	0.58	0.98	0.95	0.87	0.85
TER	0.89	0.79	0.97	0.84	0.98	0.96	0.94	0.55	0.99	0.98	0.89	0.67
chrF	0.97	0.97	0.98	0.88	0.99	0.97	0.97	0.90	0.94	0.97	0.86	0.80
ESIM	0.98	0.99	0.99	0.93	0.96	0.93	0.98	0.90	0.99	0.99	0.94	0.83
YiSi-1	0.97	0.98	0.99	0.92	0.97	0.94	0.99	0.89	0.99	0.98	0.91	0.85
YiSi-2	0.61	0.12	0.92	-0.01	0.70	0.48	0.34	0.69	-0.77	0.13	-0.53	0.07

Table 2: for the language pairs into languages other than English that contain outlier systems.

Tangled up in BLEU

4 Q3: Can these metrics be relied upon for comparing two systems?

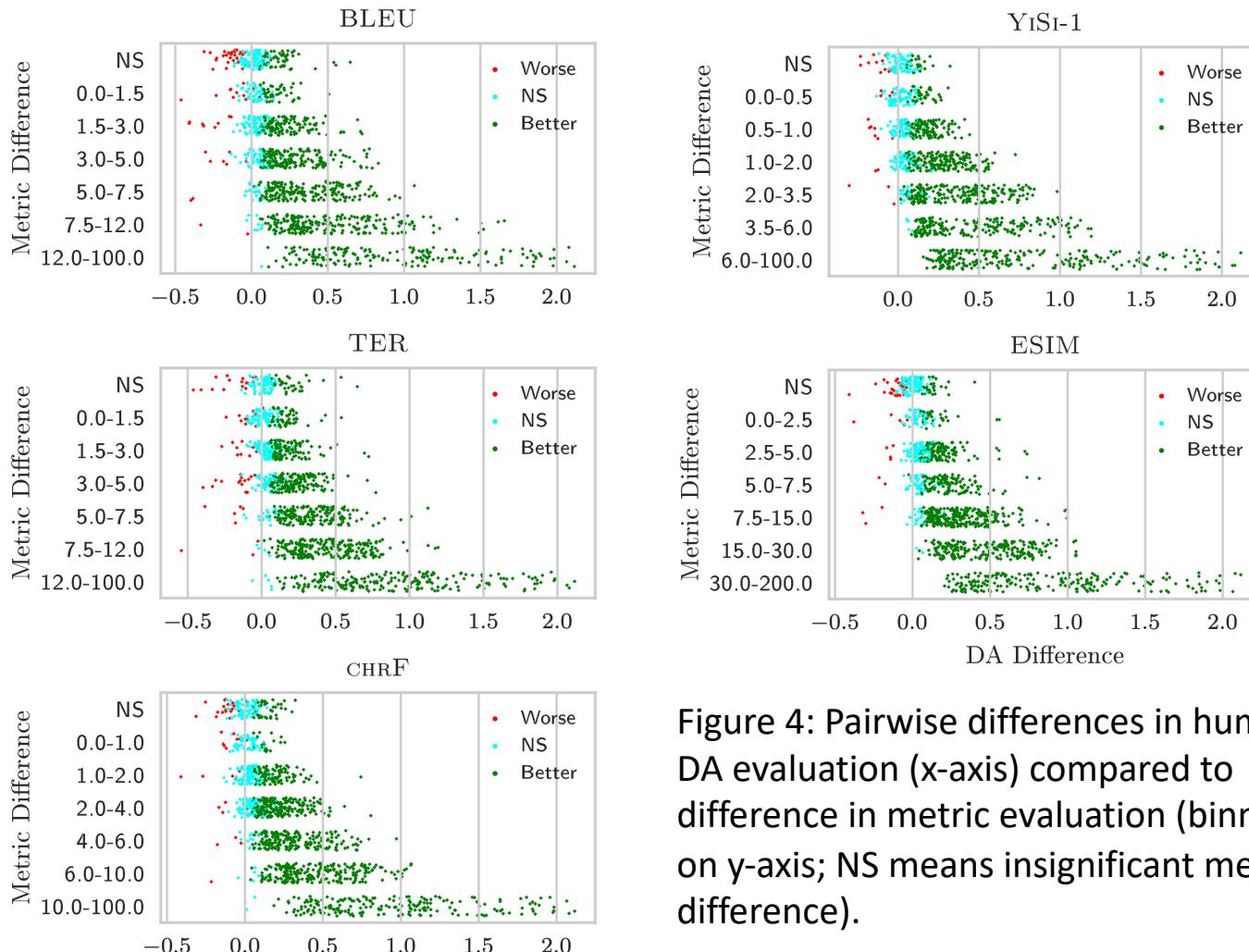


Figure 4: Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference).

Tangled up in BLEU



4 Q3: Can these metrics be relied upon for comparing two systems?

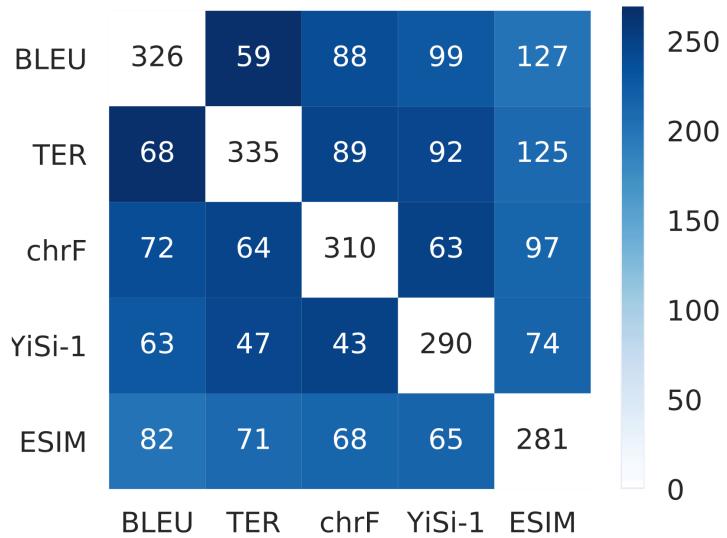


Figure 5: The agreement between metric errors over all 1362 system comparisons. The values in the diagonal indicate the total number of Type 1 and Type 2 errors for the metric. The off-diagonal cells show the total number of errors made by the row-metric where the column-metric is correct.



5 Conclusion

1. It shows that current MT evaluation methods are sensitive to the translations used for assessment.
2. It reveals that BLEU can be misleading when comparing high quality systems.
3. It proposes a new method for identifying outliers, and gives a comparison of BLEU with embedding-based measures.
4. Recommendations:
 - 1) Use the method in this paper to remove outliers before evaluating MT systems.
 - 2) Stop using BLEU or TER, and instead use CHRF, YISI-1, or ESIM
 - 3) Stop using small changes in evaluation metrics as the sole basis to draw important empirical conclusions, and make sure these are supported by manual evaluation.



Tanks!