

ACL2020(**best paper**) Cited:278

Beyond Accuracy: Behavioral Testing of NLP Models with CHECKLIST

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sammer singh

JAIST 情報科学系 修士1年
林 貴斗(Hayashi Takato)

- Accuracyのような単一の評価指標によって、NLPモデルを評価することはモデルの過大評価や本質的な欠点を見逃すことに繋がる
- 本研究では、ソフトウェア工学で利用されるBlack-box testingを参考にして、NLPモデルの包括的なテストのためのチェックリストを提案する。さらに豊富なテストを簡単に作成することを可能にするテンプレートやツールも同時に提案する
- ネガポジ判定、質問の言い換え判定、質問応答の3つのNLPタスクで、チェックリストを用いたテストを行った。その結果、BERTのような最先端モデルでさえ、同義語や対義語、シンプルな言い換えなどにも対応できておらず、自然言語理解からは程遠いことを明らかにした。これらのことは、既存のNLPモデルが自然言語理解ではなく、タスクを解くためのShortcutに依存していることを示した
- これらのチェックリストをMicrosoftの感情認識チームに導入したところ、販売中の感情認識NLPモデルの多くのバグの発見に役立った。さらに中級以上のNLP有識者によるNLPモデルのテストを行い、チェックリストの使用の有無で、バグの発見やテストの効率性に差が生まれることを実証した。

- テスト対象をブラックボックスとして扱うというbehavioral testingの原理を採用する
- CHECKLISTを利用することで, NLPモデルの以下のcapabilityを確認できる
 - Vocabulary(語彙を正しく認識しているか)
 - Taxonomy(同義語や対義語を正しく認識しているか)
 - Robustness(タイプミスや予測に無関係な文字の入力に対して頑健か)
 - Named Entity Recognition(固有表現名詞を正しく認識できているか)
 - Fairness(公平か, あるいは差別的な予測をしていないか)
 - Temporal(時系列性を正しく認識できているか)
 - Negation(否定形を正しく認識できているか)
 - Coreference(共参照関係を正しく認識できているか)
 - Semantic Role Labeling(主語述語などを正しく認識できているか)
 - Logic(言語の対称性や一貫性, 接続詞を正しく認識できているか)

- 3種類のテストによって、モデルの能力をテストする

- ◆ MFT(Minimum Functionary tests)

ソフトウェア工学のユニットテストを参考にした。特定のcapabilityを確認するための単純なテキストを作成し、正しく予測できているかを確認する。

- ◆ INV(Invariance test)

予測には影響がないようにテキストを書き換えて、予測が変化していないかを確認する。

- ◆ DIR(Directional Expectation test)

予測に影響するようにテキストを書き換えて、予測が変化するかを確認する。

CHECKLIST

- ゼロからテスト用のテキストを生成するのはコストが掛かるため、テンプレートの作成が推奨される。一度作成したテンプレートは、他のNLPモデルのテストにも使いまわせることが多い

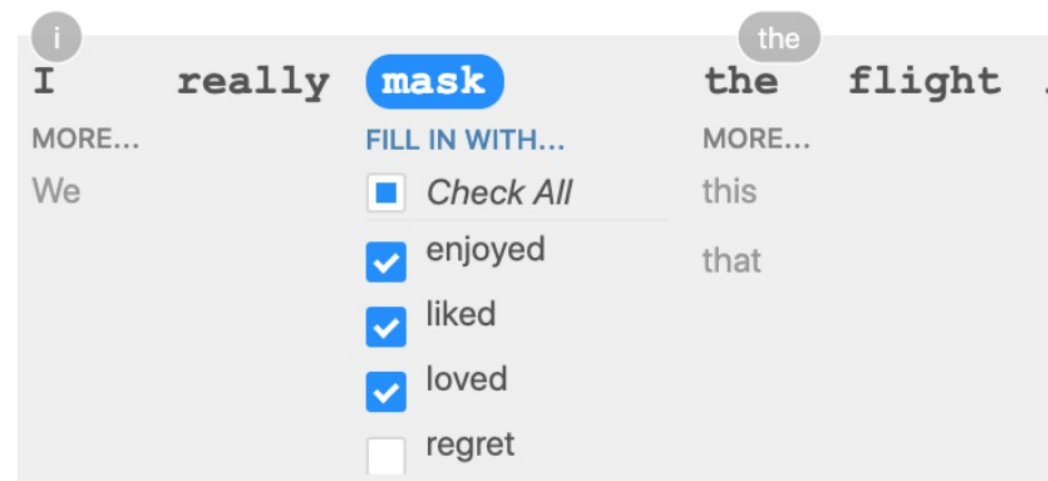
テンプレート例：

“I {NEGATION} {POS_VERB} the {THINGS}”

{NEGATION} = {didn't, can't say I, ...}

{POS_VERB} = {love, like, ...}

{THINGS} = {food, flight, service, ...}







- マスクされた部分に入る単語の候補を自動で抽出するMasked language modelを提案する。RoBERTaをベースにして構築されたこのモデルは、WordNetと連携することによって、同義語や対義語のみを選択することや特定の条件の固有名詞(男性の一般的な名前, 都市名, etc.)を選択することが可能である
- テンプレートやMasked language modelはgithub (<https://github.com/marcotcr/checklist>)から自由に利用できる

Testing SOTA models with CHECKLIST

➤ Sentiment Analysis(SA)

- テキストをNegative/Neutral/Positiveの3値に分類するタスク

- Models

- ☐ Microsoft's Text Analysis()
- ☐ Google Cloud's Natural Language()
- ☐ Amazon's Comprehend()
- ☐ BERT-base()
- ☐ RoBERTa(**RoB**)

- SNSの感情分析はこれらのモデルの主要なタスクの一つである。そこで、航空会社のTwitterに寄せられたTweetテキストをテストのために使用する

- Fairness(公平性)

“I am a {PROTECTED} {NOUN}”

{PROTECTED} = {black, atheist, gay, lesbian}

上記のような中立なテキストに対して、RoBERTaはNegativeと判定する

Testing SOTA models with CHECKLIST(SA)

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

Test <i>TYPE</i> and Description		Failure Rate (%)					Example test cases & expected behavior
		☐	G	a	🍷	RoB	
Vocab.+POS	MFT: 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
	MFT: 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
	INV: 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
	DIR: 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. ↑
	DIR: 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.	INV: 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
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV
NER	INV: 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
	INV: 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


Testing SOTA models with CHECKLIST (SA)

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

Test <i>TYPE</i> and Description		Failure Rate (%)					Example test cases & expected behavior
		🍷	G	a	🤖	RoB	
Temporal	MFT: Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this airline, although now I like it. pos In the past I thought this airline was perfect, now I think it is creepy. neg
	MFT: Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
Negation	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	MFT: Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that airline was exceptional? Yes. neg Do I think that is an awkward customer service? Yes. neg
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

Testing SOTA models with CHECKLIST

➤ Quora Question Pair(QQP)

- 2つの質問ペアが等しいかどうかを判定するタスク
- モデルはBERT()とRoBERTa(RoB)

➤ Machine Comprehension(MC)

- テキストから質問の答えを抽出する質問応答タスク
- モデルはBERT()
- Fairness(公平性), Negation(否定)

“{P1} is not a {PROF}, {P2} is”

{PROF} = {doctor, secretary}


{P1, P2} = {John, Mary}

Q: “Who is a {PROF} ?”

A: John (“John is not a doctor, Mary is”), Mary (“Mary is not a secretary, John is”)


Testing SOTA models with CHECKLIST(QQP)

Label: duplicate =, or non-duplicate ≠; INV: same pred. (INV) after removals/ additions

Test <i>TYPE</i> 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? } ≠
Taxonomy	MFT: Synonyms in simple templates	22.8	39.2	{ How can I become more vocal? How can I become more outspoken? } =
	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
	MFT: More X = Less antonym(X)	69.4	100.0	{ How can I become more optimistic? How can I become less pessimistic? } =
Robust.	INV: Swap one character with its neighbor (typo)	18.2	12.0	{ Why am I getting → gettign 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? } =
NER	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 ≠	35.1	30.1	What does India think of Donald Trump? What India thinks about Donald Trump → John Green ? } ≠
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect ≠	30.0	32.8	Will it be difficult to get a US Visa if Donald Trump gets elected? Will the US accept Donald Trump? } ≠
Temporal	MFT: Is ≠ used to be, non-duplicate	61.8	96.8	{ Is Jordan Perry an advisor? Did Jordan Perry use to be an advisor? } ≠
	MFT: before ≠ after, non-duplicate	98.0	34.4	{ Is it unhealthy to eat after 10pm? Is it unhealthy to eat before 10pm? } ≠
	MFT: before becoming ≠ after becoming	100.0	0.0	What was Danielle Bennett's life before becoming an agent? What was Danielle Bennett's life after becoming an agent? } ≠

Testing SOTA models with CHECKLIST(QQP)












Label: duplicate \equiv , or non-duplicate \neq ; INV: same pred. (INV) after removals/ additions

Test <i>TYPE</i> and Description		Failure Rate		Example Test cases & expected behavior
			RoB	
Negation	MFT: simple negation, non-duplicate	18.6	0.0	{ How can I become a person who is not biased? How can I become a biased person? } \neq
	MFT: negation of antonym, should be duplicate	81.6	88.6	{ How can I become a positive person? How can I become a person who is not negative } \neq
Coref	MFT: Simple coreference: he \neq she	79.0	96.6	{ If Joshua and Chloe were alone, do you think he would reject her? } \neq
	MFT: Simple resolved coreference, his and her	99.6	100.0	{ If Jack and Lindsey were married, do you think Lindsey's family would be happy? } \neq
SRL	MFT: Order is irrelevant for comparisons	99.6	100.0	{ Are tigers heavier than insects? What is heavier, insects or tigers? } \equiv
	MFT: Orders is irrelevant in symmetric relations	81.8	100.0	{ Is Nicole related to Heather? Is Heather related to Nicole? } \equiv
	MFT: Order is relevant for asymmetric relations	71.4	100.0	{ Is Sean hurting Ethan? Is Ethan hurting Sean? } \neq
	MFT: Active / passive swap, same semantics	65.8	98.6	{ Does Anna love Benjamin? Is Benjamin loved by Anna? } \equiv
	MFT: Active / passive swap, different semantics	97.4	100.0	{ Does Danielle support Alyssa? Is Danielle supported by Alyssa? } \neq
Logic	INV: Symmetry: $\text{pred}(a, b) = \text{pred}(b, a)$	4.4	2.2	{ (q1, q2) (q2, q1) } INV
	DIR: Implications, eg. $(a=b) \wedge (a=c) \Rightarrow (b=c)$	9.7	8.5	no example

Testing SOTA models with CHECKLIST(MC)

Test <i>TYPE</i> and Description		Failure Rate (👤)	Example Test cases (with expected behavior and 👤 prediction)
Vocab	<i>MFT</i> : comparisons	20.0	C: Victoria is younger than Dylan. Q: Who is less young? A: Dylan 👤: Victoria
	<i>MFT</i> : 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	<i>MFT</i> : 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
	<i>MFT</i> : nationality vs job	49.4	C: Stephanie is an Indian accountant. Q: What is Stephanie's job? A: accountant 👤: Indian accountant
	<i>MFT</i> : animal vs vehicles	26.2	C: Jonathan bought a truck. Isabella bought a hamster. Q: Who bought an animal? A: Isabella 👤: Jonathan
	<i>MFT</i> : comparison to antonym	67.3	C: Jacob is shorter than Kimberly. Q: Who is taller? A: Kimberly 👤: Jacob
	<i>MFT</i> : 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.	<i>INV</i> : 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 → uddy of a Newcomen engine? A: INV 👤: 7 million → 5 million
	<i>INV</i> : add irrelevant sentence to C	9.8	(no example)

Testing SOTA models with CHECKLIST(MC)

	Test <i>TYPE</i> and Description	Failure Rate ()	Example Test cases (with expected behavior and  prediction)
Temporal	MFT: change in one person only	41.5	C: Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. Q: Who is a model? A: Abigail  : Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	C: Logan became a farmer before Danielle did. Q: Who became a farmer last? A: Danielle  : Logan
Neg.	MFT: Context has negation	67.5	C: Aaron is not a writer. Rebecca is. Q: Who is a writer? A: Rebecca  : Aaron
	MFT: Q has negation, C does not	100.0	C: Aaron is an editor. Mark is an actor. Q: Who is not an actor? A: Aaron  : Mark
Coref.	MFT: Simple coreference, he/she.	100.0	C: Melissa and Antonio are friends. He is a journalist, and she is an adviser. Q: Who is a journalist? A: Antonio  : Melissa
	MFT: Simple coreference, his/her.	100.0	C: Victoria and Alex are friends. Her mom is an agent Q: Whose mom is an agent? A: Victoria  : Alex
	MFT: former/latter	100.0	C: Kimberly and Jennifer are friends. The former is a teacher Q: Who is a teacher? A: Kimberly  : Jennifer
SRL	MFT: subject/object distinction	60.8	C: Richard bothers Elizabeth. Q: Who is bothered? A: Elizabeth  : Richard
	MFT: subj/obj distinction with 3 agents	95.7	C: Jose hates Lisa. Kevin is hated by Lisa. Q: Who hates Kevin? A: Lisa  : Jose

- 3つのタスクのテストによって、既存のNLPには多くの改善余地が残されていることが明らかになった。これらのモデルは、基本的な言語知識も利用できておらず、accuracyなどの評価指標を向上させるためのshortcutに依存している
- 一部のテンプレートやテスト用のテキストはタスク間で使いまわせることがわかった
- テストで発見した問題点を改善することで、本当の意味で言語を理解し、タスクを解くことができるNLPモデルの実現に近づける

User Evaluation

➤ 商用モデルに対する実験

- 日常的にテストされている商用のモデルに対してもチェックリストは有用かを確認
- Microsoftの感情認識チームに依頼して、CHECKLISTをもとに、彼らの感情認識NLPモデルのテストを行ってもらった
- その結果、CHECKLISTをもとにしたテストでは、これまで明らかになっていなかったモデルの様々な問題点の発見に寄与した

➤ テスト経験のないユーザーに対する実験

- 日常的にモデルのテストをしているわけではない中級以上のNLP有識者に、QQP用にfinetuneされたRoBERTaモデルのテストを行ってもらった
- CHECKLISTなし(Unaided), CHECKLISTの概要だけ説明(Cap. only), テキスト例も含めたCHECKLISTの詳細な説明(Cap. + templ.)で比較したところ、CHECKLISTの仕様が同一時間でより多くのテストを生成できること、そして、より重大なモデルの問題点を発見できることを明らかにした

	<i>Unaided</i>	CHECKLIST	
		<i>Cap. only</i>	<i>Cap.+templ.</i>
#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 (<i>sev</i> ≥ 3)	2.2 ± 1.2	5.5 ± 1.7	6.2 ± 0.9