Beyond Accuracy: Behavioral Testing of NLP Models with Checklist

Marco Tulio Ribeiro  
Microsoft Research  
marcotcr@microsoft.com

Tongshuang Wu  
Univ. of Washington  
wtshuang@cs.uw.edu

Carlos Guestrin  
Univ. of Washington  
guestrin@cs.uw.edu

Sameer Singh  
Univ. of California, Irvine  
sameer@uci.edu

Abstract

Although measuring held-out accuracy has been the primary approach to evaluate generalization, it often overestimates the performance of NLP models, while alternative approaches for evaluating models either focus on individual tasks or on specific behaviors. Inspired by principles of behavioral testing in software engineering, we introduce Checklist, a task-agnostic methodology for testing NLP models. Checklist includes a matrix of general linguistic capabilities and test types that facilitate comprehensive test ideation, as well as a software tool to generate a large and diverse number of test cases quickly. We illustrate the utility of Checklist with tests for three tasks, identifying critical failures in both commercial and state-of-art models. In a user study, a team responsible for a commercial sentiment analysis model found new and actionable bugs in an extensively tested model. In another user study, NLP practitioners with Checklist created twice as many tests, and found almost three times as many bugs as users without it.

1 Introduction

One of the primary goals of training NLP models is generalization. Since testing “in the wild” is expensive and does not allow for fast iterations, the standard paradigm for evaluation is using train-validation-test splits to estimate the accuracy of the model, including the use of leader boards to track progress on a task (Rajpurkar et al., 2016). While performance on held-out data is a useful indicator, held-out datasets are often not comprehensive, and contain the same biases as the training data (Rajpurkar et al., 2018), such that real-world performance may be overestimated (Patel et al., 2008; Recht et al., 2019). Further, by summarizing the performance as a single aggregate statistic, it becomes difficult to figure out where the model is failing, and how to fix it (Wu et al., 2019).

A number of additional evaluation approaches have been proposed, such as evaluating robustness to noise (Belinkov and Bisk, 2018; Rychalska et al., 2019) or adversarial changes (Ribeiro et al., 2018; Iyyer et al., 2018), fairness (Prabhakaran et al., 2019), logical consistency (Ribeiro et al., 2019), explanations (Ribeiro et al., 2016), diagnostic datasets (Wang et al., 2019b), and interactive error analysis (Wu et al., 2019). However, these approaches focus either on individual tasks such as Question Answering or Natural Language Inference, or on a few capabilities (e.g. robustness), and thus do not provide comprehensive guidance on how to evaluate models. Software engineering research, on the other hand, has proposed a variety of paradigms and tools for testing complex software systems. In particular, “behavioral testing” (also known as black-box testing) is concerned with testing different capabilities of a system by validating the input-output behavior, without any knowledge of the internal structure (Beizer, 1995). While there are clear similarities, many insights from software engineering are yet to be applied to NLP models.

In this work, we propose Checklist, a new evaluation methodology and accompanying tool for comprehensive behavioral testing of NLP models. Checklist guides users in what to test, by providing a list of linguistic capabilities, which are applicable to most tasks. To break down potential capability failures into specific behaviors, Checklist introduces different test types, such as prediction invariance in the presence of certain perturbations, or performance on a set of “sanity checks.” Finally, our implementation of Checklist includes multiple abstractions that help users generate large numbers of test cases easily, such as templates, lexicons, general-purpose perturbations, visualizations, and context-aware suggestions.

1https://github.com/marcotcr/checklist
As an example, we CheckList a commercial sentiment analysis model in Figure 1. Potential tests are structured as a conceptual matrix with capabilities as rows and test types as columns (examples of each type in A, B and C).

We demonstrate the usefulness and generality of CheckList via instantiation on three NLP tasks: sentiment analysis (Sentiment), duplicate question detection (QQP; Wang et al., 2019b), and machine comprehension (MC; Rajpurkar et al., 2016). While traditional benchmarks indicate that models on these tasks are as accurate as humans, CheckList reveals a variety of severe bugs, where commercial and research models do not effectively handle basic linguistic phenomena such as negation, named entities, coreferences, semantic role labeling, etc., as they pertain to each task. Further, CheckList is easy to use and provides immediate value – in a user study, the team responsible for a commercial sentiment analysis model discovered many new and actionable bugs in their own model, even though it had been extensively tested and used by customers. In an additional user study, we found that NLP practitioners with CheckList generated more than twice as many tests (each test containing an order of magnitude more examples), and uncovered almost three times as many bugs, compared to users without CheckList.

2 CheckList

Conceptually, users “CheckList” a model by filling out cells in a matrix (Figure 1), each cell potentially containing multiple tests. In this section, we go into more detail on the rows (capabilities), columns (test types), and how to fill the cells (tests). CheckList applies the behavioral testing principle of “decoupling testing from implementation” by treating the model as a black box, which allows for comparison of different models trained on different data, or third-party models where access to training data or model structure is not granted.

2.1 Capabilities

While testing individual components is a common practice in software engineering, modern NLP models are rarely built one component at a time. Instead, CheckList encourages users to consider how different natural language capabilities are manifested on the task at hand, and to create tests to evaluate the model on each of these capabilities. For example, the Vocabulary+POS capability pertains to whether a model has the necessary vocabulary, and whether it can appropriately handle the impact of words with different parts of speech on the task. For Sentiment, we may want to check if the model is able to identify words that carry positive, negative, or neutral sentiment, by verifying how it behaves on examples like “This was a good flight.” For QQP, we might want the model to...
understand when modifiers differentiate questions, e.g. accredited in (“Is John a teacher?”, “Is John an accredited teacher?”). For MC, the model should be able to relate comparatives and superlatives, e.g. (Context: “Mary is smarter than John.”, Q: “Who is the smartest kid?”, A: “Mary”).

We suggest that users consider at least the following capabilities: Vocabulary+POS (important words or word types for the task), Taxonomy (synonyms, antonyms, etc), Robustness (to typos, irrelevant changes, etc), NER (appropriately understanding named entities), Fairness, Temporal (understanding order of events), Negation, Coreference, Semantic Role Labeling (understanding roles such as agent, object, etc), and Logic (ability to handle symmetry, consistency, and conjunctions). We will provide examples of how these capabilities can be tested in Section 3 (Tables 1, 2, and 3). This listing of capabilities is not exhaustive, but a starting point for users, who should also come up with additional capabilities that are specific to their task or domain.

2.2 Test Types

We prompt users to evaluate each capability with three different test types (when possible): Minimum Functionality tests, Invariance, and Directional Expectation tests (the columns in the matrix).

A Minimum Functionality test (MFT), inspired by unit tests in software engineering, is a collection of simple examples (and labels) to check a behavior within a capability. MFTs are similar to creating small and focused testing datasets, and are particularly useful for detecting when models use shortcuts to handle complex inputs without actually mastering the capability. The Vocabulary+POS examples in the previous section are all MFTs.

We also introduce two additional test types inspired by software metamorphic tests (Segura et al., 2016). An Invariance test (INV) is when we apply label-preserving perturbations to inputs and expect the model prediction to remain the same. Different perturbation functions are needed for different capabilities, e.g. changing location names for the NER capability for Sentiment (Figure 1B), or introducing typos to test the Robustness capability. A Directional Expectation test (DIR) is similar, except that the label is expected to change in a certain way. For example, we expect that sentiment will not become more positive if we add “You are lame.” to the end of tweets directed at an airline (Figure 1C). The expectation may also be a target label, e.g. replacing locations in only one of the questions in QQP, such as (“How many people are there in England?”, “What is the population of England - Turkey?”), ensures that the questions are not duplicates. INVs and DIRs allow us to test models on unlabeled data – they test behaviors that do not rely on ground truth labels, but rather on relationships between predictions after perturbations are applied (invariance, monotonicity, etc).

2.3 Generating Test Cases at Scale

Users can create test cases from scratch, or by perturbing an existing dataset. Starting from scratch makes it easier to create a small number of high-quality test cases for specific phenomena that may be underrepresented or confounded in the original dataset. Writing from scratch, however, requires significant creativity and effort, often leading to tests that have low coverage or are expensive and time-consuming to produce. Perturbation functions are harder to craft, but generate many test cases at once. To support both these cases, we provide a variety of abstractions that scale up test creation from scratch and make perturbations easier to craft.

Templates Test cases and perturbations can often be generalized into a template, to test the model on a more diverse set of inputs. In Figure 1 we generalized “I didn’t love the food.” with the template “I {NEGATION} {POS_VERB} the {THING}.”, where {NEGATION} = {didn’t, can’t say I, ...}, {POS_VERB} = {love, like, ...}, {THING} = {food, flight, service, ...}, and generated all test cases with a Cartesian product. A more diverse set of inputs is particularly helpful when a small set of test cases could miss a failure, e.g. if a model works for some forms of negation but not others.

Expanding Templates While templates help scale up test case generation, they still rely on the user’s creativity to create fill-in values for each
We release an implementation of CheckList at https://github.com/marcotcr/checklist. In addition to templating features and mask language model suggestions, it contains various visualizations, abstractions for writing test expectations (e.g. monotonicity) and perturbations, saving/sharing tests and test suites such that tests can be reused with different models and by different teams, and general-purpose perturbations such as char swaps (simulating typos), contractions, name and location changes (for NER tests), etc.

### S3 Testing SOTA models with CheckList

We CheckList the following commercial Sentiment analysis models via their paid APIs:

- Microsoft’s Text Analytics (A)
- Google Cloud’s Natural Language (G)
- Amazon’s Comprehend (A)

We also CheckList BERT-base (A) and RoBERTa-base (RoB) (Liu et al., 2019) finetuned on SST-2 (acc: 92.7% and 94.8%) and on the QQP dataset (3/1, 3/2) range are considered neutral.
and

Hillary Clinton

John Green?

Nicole Perez in jail?

gettin lazy?

Nicole Perez going to go to jail?

expected behavior

We create tests for a broad range of capabilities, where we expect models to appropriately handle

MC

considerably when clearly positive phrases (e.g. “You

/uniF1A0

/windows

the DIR tests, the sentiment scores predicted by

neutral sanity checks (e.g. “I like this seat.”). In

trained on binary labels only). Surprisingly, RoB do poorly on neutral predictions (they were

Vocab.

and present subset with high failure rates in Ta-

table tweets for INV

on that domain and use a dataset of unlabeled air-

as a use case for these commercial models, we test

Since social media is listed

et al., 2019), achieving 93.2 F1. All the tests pre-

sented here are part of the open-source release, and

can be easily replicated and applied to new models.

Sentiment Analysis  Since social media is listed as a use case for these commercial models, we test on that domain and use a dataset of unlabeled airline tweets for INV4 and DIR perturbation tests. We create tests for a broad range of capabilities, and present subset with high failure rates in Table 1. The Vocab. +POS MFTs are sanity checks, where we expect models to appropriately handle common neutral or sentiment-laden words. A and RoB do poorly on neutral predictions (they were trained on binary labels only). Surprisingly, G and A fail (7.6% and 4.8%) on sentences that are clearly neutral, with G also failing (15%) on non-neutral sanity checks (e.g. “I like this seat.”). In the DIR tests, the sentiment scores predicted by A and G frequently (12.6% and 12.4%) go down considerably when clearly positive phrases (e.g. “You

are extraordinary.”) are added, or up (G: 34.6%) for negative phrases (e.g. “You are lame.”).

All models are sensitive to addition of random (not adversarial) shortened URLs or Twitter handles (e.g. 24.8% of A predictions change), and to name changes, such as locations (G: 20.8%, A: 14.8%) or person names (G: 15.1%, A: 9.1%). None of the models do well in tests for the Temporal, Negation, and SRL capabilities. Failures on negations as simple as “The food is not poor.” are particularly notable, e.g. G (54.2%) and A (29.4%). The failure rate is near 100% for all commercial models when the negation comes at the end of the sentence (e.g “I thought the plane would be awful, but it wasn’t.”), or with neutral content between the negation and the sentiment-laden word.

Commercial models do not fail simple Fairness sanity checks such as “I am a black woman,” (template: “I am a {PROTECTED}. {NOUN}.”), always predicting them as neutral. Similar to software engineering, absence of test failure does not imply that these models are fair – just that they are not unfair enough to fail these simple tests. On the other hand, A always predicts negative when

4For all the INV tests, models fail whenever their prediction changes and the probability changes by more than 0.1.
<table>
<thead>
<tr>
<th>Test TYPE and Description</th>
<th>Failure Rate (%)</th>
<th>Example Test cases (with expected behavior and ( \hat{A} ) prediction)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vocab</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} comparisons</td>
<td>20.0</td>
<td>C: Victoria is younger than Dylan. ( Q: ) Who is less young? ( A: ) ( \hat{A} ): Dylan ( \hat{A} ): Victoria</td>
</tr>
<tr>
<td>\textit{MFT:} intensifiers to superlative: most/least</td>
<td>91.3</td>
<td>C: Anna is worried about the project. Matthew is extremely worried about the project. ( Q: ) Who is least worried about the project? ( A: ) Anna ( \hat{A} ): Matthew</td>
</tr>
<tr>
<td><strong>Texts</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} match properties to categories</td>
<td>82.4</td>
<td>C: There is a tiny purple box in the room. ( Q: ) What size is the box? ( A: ) tiny ( \hat{A} ): purple</td>
</tr>
<tr>
<td>\textit{MFT:} nationality vs job</td>
<td>49.4</td>
<td>C: Stephanie is an Indian accountant. ( Q: ) What is Stephanie’s job? ( A: ) accountant ( \hat{A} ): Indian accountant</td>
</tr>
<tr>
<td>\textit{MFT:} animal vs vehicles</td>
<td>26.2</td>
<td>C: Jonathan bought a truck. Isabella bought a hamster. ( Q: ) Who bought an animal? ( A: ) Isabella ( \hat{A} ): Jonathan</td>
</tr>
<tr>
<td>\textit{MFT:} comparison to antonym</td>
<td>67.3</td>
<td>C: Jacob is shorter than Kimberly. ( Q: ) Who is taller? ( A: ) Kimberly ( \hat{A} ): Jacob</td>
</tr>
<tr>
<td>\textit{MFT:} more/less in context, more/less synonym in question</td>
<td>100.0</td>
<td>C: Jeremy is more optimistic than Taylor. ( Q: ) Who is more pessimistic? ( A: ) Taylor ( \hat{A} ): Jeremy</td>
</tr>
<tr>
<td><strong>Robust</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{INV:} Swap adjacent characters in ( Q ) (type)</td>
<td>11.6</td>
<td>C: ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... ( Q: ) What was the ideal duty - duty of a Newcomen engine? ( A: ) ( \hat{A} ): 7 million ( \hat{A} ): 5 million</td>
</tr>
<tr>
<td>\textit{INV:} add irrelevant sentence to ( C )</td>
<td>9.8</td>
<td>(no example)</td>
</tr>
<tr>
<td><strong>Temporal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} change in one person only</td>
<td>41.5</td>
<td>C: Both Jason and Abigail were journalists, but there was a change in Abigail, who is now a model. ( Q: ) Who is a model? ( A: ) Abigail ( \hat{A} ): Abigail were journalists, but there was a change in Abigail</td>
</tr>
<tr>
<td>\textit{MFT:} Understanding before/after, last/first</td>
<td>82.9</td>
<td>C: Logan became a farmer before Danielle did. ( Q: ) Who became a farmer last? ( A: ) Danielle ( \hat{A} ): Logan</td>
</tr>
<tr>
<td><strong>New</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} Contests has negation</td>
<td>67.5</td>
<td>C: Aaron is not a writer. Rebecca is. ( Q: ) Who is a writer? ( A: ) Rebecca ( \hat{A} ): Aaron</td>
</tr>
<tr>
<td>\textit{MFT:} ( Q ) has negation, ( C ) does not</td>
<td>100.0</td>
<td>C: Aaron is an editor. Mark is an actor. ( Q: ) Who is not an actor? ( A: ) Aaron ( \hat{A} ): Mark</td>
</tr>
<tr>
<td><strong>Cross</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} Simple coreference, he/she.</td>
<td>100.0</td>
<td>C: Melissa and Antonio are friends. He is a journalist, and she is an adviser. ( Q: ) Who is a journalist? ( A: ) Antonio ( \hat{A} ): Melissa</td>
</tr>
<tr>
<td>\textit{MFT:} Simple coreference, his/her.</td>
<td>100.0</td>
<td>C: Victoria and Alex are friends. Her mom is an agent. ( Q: ) Whose mom is an agent? ( A: ) Victoria ( \hat{A} ): Alex</td>
</tr>
<tr>
<td>\textit{MFT:} former/latter</td>
<td>100.0</td>
<td>C: Kimberly and Jennifer are friends. The former is a teacher. ( Q: ) Who is a teacher? ( A: ) Kimberly ( \hat{A} ): Jennifer</td>
</tr>
<tr>
<td><strong>SRL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\textit{MFT:} subject/object distinction</td>
<td>60.8</td>
<td>C: Richard bothers Elizabeth. ( Q: ) Who is bothered? ( A: ) Elizabeth ( \hat{A} ): Richard</td>
</tr>
<tr>
<td>\textit{MFT:} subject/predicate distinction with 3 agents</td>
<td>95.7</td>
<td>C: Jose hates Lisa. Kevin is hated by Lisa. ( Q: ) Who hates Kevin? ( A: ) Lisa ( \hat{A} ): Jose</td>
</tr>
</tbody>
</table>

Table 3: A selection of tests for Machine Comprehension.

\{\textbf{PROTECTED}\} is black, atheist, gay, and lesbian, while predicting positive for Asian, straight, etc.

With the exception of tests that depend on predicting “neutral”, \( \hat{A} \) and RoB did better than all commercial models on almost every other test. This is a surprising result, since the commercial models list social media as a use case, and are under regular testing and improvement with customer feedback, while \( \hat{A} \) and RoB are research models trained on the SST-2 dataset (movie reviews). Finally, \( \hat{A} \) and RoB fail simple negation MFTs, even though they are fairly accurate (91.5%, 93.9%, respectively) on the subset of the SST-2 validation set that contains negation in some form (18% of instances). By isolating behaviors like this, our tests are thus able to evaluate capabilities more precisely, whereas performance on the original dataset can be misleading.

**Quora Question Pair** While \( \hat{A} \) and RoB surpass human accuracy on \textit{QQP} in benchmarks (Wang et al., 2019a), the subset of tests in Table 2 indicate that these models are far from solving the question paraphrase problem, and are likely relying on shortcuts for their high accuracy.

Both models lack what seems to be crucial skills for the task: ignoring important modifiers on the \textit{Vocab}, test, and lacking basic \textit{Taxonomy} understanding, e.g. synonyms and antonyms of common words. Further, neither is robust to typos or simple paraphrases. The failure rates for the \textit{NER} tests indicate that these models are relying on shortcuts such as anchoring on named entities too strongly instead of understanding named entities and their impact on whether questions are duplicates.

Surprisingly, the models often fail to make simple \textit{Temporal} distinctions (e.g. is\# used to be and before\#after), and to distinguish between simple \textit{Coreferences} (he\#she). In \textit{SRL} tests, neither model is able to handle agent/predicate changes, or active/passive swaps. Finally, \( \hat{A} \) and RoB change predictions 4.4% and 2.2% of the time when the question order is flipped, failing a basic task requirement (if \( q_1 \) is a duplicate of \( q_2 \), so is \( q_2 \) of \( q_1 \)). They are also not consistent with \textit{Logical} implications of their predictions, such as transitivity.
Machine Comprehension  Vocab+POS tests in Table 3 show that the model often fails to properly grasp intensity modifiers and comparisons/superlatives. It also fails on simple Taxonomy tests, such as matching properties (size, color, shape) to adjectives, distinguishing between animals-vehicles or jobs-nationalities, or comparisons involving antonyms.

The model does not seem capable of handling short instances with Temporal concepts such as before, after, last, and first, or with simple examples of Negation, either in the question or in the context. It also does not seem to resolve basic Coreferences, and grasp simple subject/object or active/passive distinctions (SRL), all of which are critical to true comprehension. Finally, the model seems to have certain biases, e.g. for the simple negation template “\{P1\} is not a \{PROF\}, \{P2\} is.” as context, and “Who is a \{PROF\}?” as question, if we set \{PROF\} = doctor, \{P1\} to male names and \{P2\} to female names (e.g. “John is not a doctor, Mary is.”; “Who is a doctor?”), the model fails (picks the man as the doctor) 89.1% of the time. If the situation is reversed, the failure rate is only 3.2% (woman predicted as doctor). If \{PROF\} = secretary, it wrongly picks the man only 4.0% of the time, and the woman 60.5% of the time.

Discussion We applied the same process to very different tasks, and found that tests reveal interesting failures on a variety of task-relevant linguistic capabilities. While some tests are task specific (e.g. positive adjectives), the capabilities and test types are general; many can be applied across tasks, as is (e.g. testing Robustness with typos) or with minor variation (changing named entities yields different expectations depending on the task). This small selection of tests illustrates the benefits of systematic testing in addition to standard evaluation. These tasks may be considered “solved” based on benchmark accuracy results, but the tests highlight various areas of improvement – in particular, failure to demonstrate basic skills that are de facto needs for the task at hand (e.g. basic negation, agent/object distinction, etc). Even though some of these failures have been observed by others, such as typos (Belinkov and Bisk, 2018; Rychalska et al., 2019) and sensitivity to name changes (Prabhakaran et al., 2019), we believe the majority are not known to the community, and that comprehensive and structured testing will lead to avenues of improvement in these and other tasks.

4 User Evaluation

The failures discovered in the previous section demonstrate the usefulness and flexibility of CheckList. In this section, we further verify that CheckList leads to insights both for users who already test their models carefully and for users with little or no experience in a task.

4.1 CheckListing a Commercial System

We approached the team responsible for the general purpose sentiment analysis model sold as a service by Microsoft (\( \square \) on Table 1). Since it is a public-facing system, the model’s evaluation procedure is more comprehensive than research systems, including publicly available benchmark datasets as well as focused benchmarks built in-house (e.g. negations, emojis). Further, since the service is mature with a wide customer base, it has gone through many cycles of bug discovery (either internally or through customers) and subsequent fixes, after which new examples are added to the benchmarks. Our goal was to verify if CheckList would add value even in a situation like this, where models are already tested extensively with current practices.

We invited the team for a CheckList session lasting approximately 5 hours. We presented CheckList (without presenting the tests we had already created), and asked them to use the methodology to test their own model. We helped them implement their tests, to reduce the additional cognitive burden of having to learn the software components of CheckList. The team brainstormed roughly 30 tests covering all capabilities, half of which were MFTs and the rest divided roughly equally between INVs and DIRs. Due to time constraints, we implemented about 20 of those tests. The tests covered many of the same functionalities we had tested ourselves (Section 3), often with different templates, but also ones we had not thought of. For example, they tested if the model handled sentiment coming from camel-cased twitter hashtags correctly (e.g. “#IHateYou”, “#ILoveYou”), implicit negation (e.g. “I wish it was good”), and others. Further, they proposed new capabilities for testing, e.g. handling different lengths (sentences vs paragraphs) and sentiment that depends on implicit expectations (e.g. “There was no \{AC\}” when \{AC\} is expected).

Qualitatively, the team stated that CheckList was very helpful: (1) they tested capabilities they had not considered, (2) they tested capabilities that they had considered but are not in the benchmarks,
and (3) even capabilities for which they had benchmarks (e.g. negation) were tested much more thoroughly and systematically with CheckList. They discovered many previously unknown bugs, which they plan to fix in the next model iteration. Finally, they indicated that they would definitely incorporate CheckList into their development cycle, and requested access to our implementation. This session, coupled with the variety of bugs we found for three separate commercial models in Table 1, indicates that CheckList is useful even in pipelines that are stress-tested and used in production.

4.2 User Study: CheckList MFTs

We conduct a user study to further evaluate different subsets of CheckList in a more controlled environment, and to verify if even users with no previous experience in a task can gain insights and find bugs in a model. We recruit 18 participants (8 from industry, 10 from academia) who have at least intermediate NLP experience, and task them with testing a finetuned QQP for a period of two hours (including instructions), using Jupyter notebooks. Participants had access to the QQP validation dataset, and are instructed to create tests that explore different capabilities of the model. We separate participants equally into three conditions: In Unaided, we give them no further instructions, simulating the current status-quo for commercial systems (even the practice of writing additional tests beyond benchmark datasets is not common for research models). In Cap. only, we provide short descriptions of the capabilities listed in Section 2.1 as suggestions to test, while in Cap. + templ., we further provide them with the template and fill-in tools described in Section 2.3. Only one participant (in Unaided) had prior experience with QQP. Due to the short study duration, we only asked users to write MFTs in all conditions; thus, even Cap. + templ. is a subset of CheckList.

We present the results in Table 4. Even though users had to parse more instructions and learn a new tool when using CheckList, they created many more tests for the model in the same time. Further, templates and masked language model suggestions helped users generate many more test cases per test in Cap. + templ. than in the other two conditions – although users could use arbitrary Python code rather than write examples by hand, only one user in Unaided did (and only for one test).

<table>
<thead>
<tr>
<th></th>
<th>Unaided</th>
<th>CheckList Cap. only</th>
<th>CheckList Cap. + templ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Tests</td>
<td>5.8 ± 1.1</td>
<td>10.2 ± 1.8</td>
<td>13.5 ± 3.4</td>
</tr>
<tr>
<td>#Cases/test</td>
<td>7.3 ± 5.6</td>
<td>5.0 ± 1.2</td>
<td>198.0 ± 96</td>
</tr>
<tr>
<td>#Capabilities tested</td>
<td>3.2 ± 0.7</td>
<td>7.5 ± 1.9</td>
<td>7.8 ± 1.1</td>
</tr>
<tr>
<td>Total severity</td>
<td>10.8 ± 3.8</td>
<td>21.7 ± 5.7</td>
<td>23.7 ± 4.2</td>
</tr>
<tr>
<td>#Bugs (sev ≥ 3)</td>
<td>2.2 ± 1.2</td>
<td>5.5 ± 1.7</td>
<td>6.2 ± 0.9</td>
</tr>
</tbody>
</table>

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).

Users explored many more capabilities on Cap. only and Cap. + templ. (we annotate tests with capabilities post-hoc); participants in Unaided only tested Robustness, Vocabulary+POS, Taxonomy, and few instances of SRL, while participants in the other conditions covered all capabilities. Users in Cap. only and Cap. + templ. collectively came up with tests equivalent to almost all MFTs in Table 2, and more that we had not contemplated. Users in Unaided and Cap. only often did not find more bugs because they lacked test case variety even when testing the right concepts (e.g. negation).

At the end of the experiment, we ask users to evaluate the severity of the failures they observe on each particular test, on a 5 point scale. While there is no “ground truth”, these severity ratings provide each user’s perception on the magnitude of the discovered bugs. We report the severity sum of discovered bugs (for tests with severity at least 2), in Table 4, as well as the number of tests for which severity was greater or equal to 3 (which filters out minor bugs). We note that users with CheckList (Cap. only and Cap. + templ.) discovered much more severe problems in the model (measured by total severity or # bugs) than users in the control condition (Unaided). We ran a separate round of severity evaluation of these bugs with a new user (who did not create any tests), and obtain nearly identical aggregate results to self-reported severity.

The study results are encouraging: with a subset of CheckList, users without prior experience are able to find significant bugs in a SOTA model in only 2 hours. Further, when asked to rate different aspects of CheckList (on a scale of 1-5), users indicated the testing session helped them learn more about the model (4.7 ± 0.5), capabilities helped them test the model more thoroughly (4.5 ± 0.4), and so did templates (4.3 ± 1.1).

i.e. have taken a graduate NLP course or equivalent.

5 Related Work

One approach to evaluate specific linguistic capabilities is to create challenge datasets. Belinkov and Glass (2019) note benefits of this approach, such as systematic control over data, as well as drawbacks, such as small scale and lack of resemblance to “real” data. Further, they note that the majority of challenge sets are for Natural Language Inference. We do not aim for CheckList to replace challenge or benchmark datasets, but to complement them. We believe CheckList maintains many of the benefits of challenge sets while mitigating their drawbacks: authoring examples from scratch with templates provides systematic control, while perturbation-based INV and DIR tests allow for testing behavior in unlabeled, naturally-occurring data. While many challenge sets focus on extreme or difficult cases (Naik et al., 2018), MFTs also focus on what should be easy cases given a capability, uncovering severe bugs. Finally, the user study demonstrates that CheckList can be used effectively for a variety of tasks with low effort: users created a complete test suite for sentiment analysis in a day, and MFTs for QQP in two hours, both revealing previously unknown, severe bugs.

With the increase in popularity of end-to-end deep models, the community has turned to “probes”, where a probing model for linguistic phenomena of interest (e.g. NER) is trained on intermediate representations of the encoder (Tenney et al., 2019; Kim et al., 2019). Along similar lines, previous work on word embeddings looked for correlations between properties of the embeddings and downstream task performance (Tsvetkov et al., 2016; Rogers et al., 2018). While interesting as analysis methods, these do not give users an understanding of how a fine-tuned (or end-to-end) model can handle linguistic phenomena for the end-task. For example, while Tenney et al. (2019) found that very accurate NER models can be trained using BERT (96.7%), we show BERT finetuned on QQP or SST-2 displays severe NER issues.

There are existing perturbation techniques meant to evaluate specific behavioral capabilities of NLP models such as logical consistency (Ribeiro et al., 2019) and robustness to noise (Belinkov and Bisk, 2018), name changes (Prabhakaran et al., 2019), or adversaries (Ribeiro et al., 2018). CheckList provides a framework for such techniques to systematically evaluate these alongside a variety of other capabilities. However, CheckList cannot be directly used for non-behavioral issues such as data versioning problems (Amershi et al., 2019), labeling errors, annotator biases (Geva et al., 2019), worst-case security issues (Wallace et al., 2019), or lack of interpretability (Ribeiro et al., 2016).

6 Conclusion

While useful, accuracy on benchmarks is not sufficient for evaluating NLP models. Adopting principles from behavioral testing in software engineering, we propose CheckList, a model-agnostic and task-agnostic testing methodology that tests individual capabilities of the model using three different test types. To illustrate its utility, we highlight significant problems at multiple levels in the conceptual NLP pipeline for models that have “solved” existing benchmarks on three different tasks. Further, CheckList reveals critical bugs in commercial systems developed by large software companies, indicating that it complements current practices well. Tests created with CheckList can be applied to any model, making it easy to incorporate in current benchmarks or evaluation pipelines.

Our user studies indicate that CheckList is easy to learn and use, and helpful both for expert users who have tested their models at length as well as for practitioners with little experience in a task. The tests presented in this paper are part of CheckList’s open source release, and can easily be incorporated into existing benchmarks. More importantly, the abstractions and tools in CheckList can be used to collectively create more exhaustive test suites for a variety of tasks. Since many tests can be applied across tasks as is (e.g. typos) or with minor variations (e.g. changing names), we expect that collaborative test creation will result in evaluation of NLP models that is much more robust and detailed, beyond just accuracy on held-out data. CheckList is open source, and available at https://github.com/marcotcr/checklist.

Acknowledgments

We would like to thank Sara Ribeiro, Scott Lundberg, Matt Gardner, Julian Michael, and Ece Kamar for helpful discussions and feedback. Sameer was funded in part by the NSF award #IIS-1756023, and in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office of Naval Research.
References


Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In


