Deriving Behavioral Tests from Common Sense Knowledge Graphs

Yasaman Razeghi Robert L. Logan IV Sameer Singh
University of California, Irvine
{yrazeghi, rlogan, sameer}@uci.edu

Abstract

Although NLP models have demonstrated "superhuman" performance on common sense reasoning tasks, it is unclear whether these models truly have common sense knowledge. Constructing evaluation datasets to test this knowledge is expensive due to the manual effort involved, and is also limited in scope. Meanwhile, common sense knowledge graphs (CSKGs) aim for a wide coverage of structured common sense knowledge, but can not be directly used for testing purposes. In this work, we introduce a semi-automated approach that leverages CSKGs to construct out-of-domain evaluation sets for NLP tasks that are more scalable than purely manual approaches. Using this procedure, we create test cases from two popular CSKGs—ConceptNet and ATOMIC—to test the common sense reasoning capability of models trained for natural language inference (NLI) and question answering (QA). These tests reveal interesting differences in failure modes of these models; models trained on NLI tend to perform better on tests of ontological knowledge, e.g. 'is a' and 'used for' relations, failing on tests that require understanding 'desires', 'needs', and 'wants', while QA models perform better on tests that involve 'wants', and 'desires'.

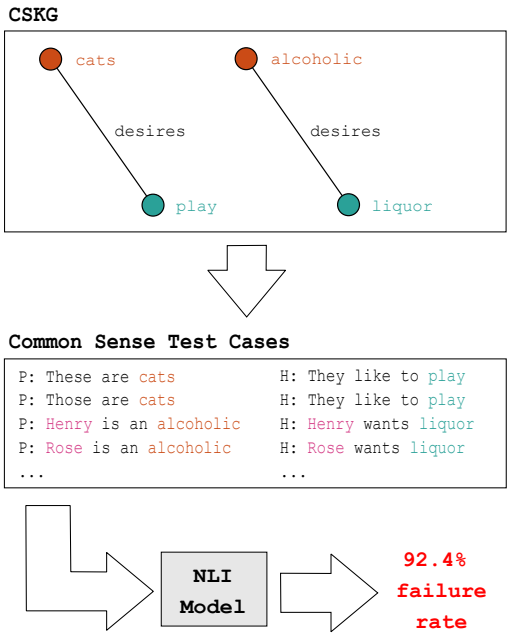

Figure 1: **Application of the Proposed Pipeline**. Using a common sense knowledge graph, we create a set of test cases to evaluate NLI models (P: premise, H: hypothesis). Models using common sense should always predict *entailment*, however BART-large fails on 92.4% of examples.

1 Introduction

Evaluating common sense reasoning capabilities of NLP models is an important yet complex procedure. Previous approaches have introduced test sets to probe *pretrained language models* on their factual (Petroni et al. 2019) or common sense knowledge (Zhou et al. 2020b). Supervised models, on the other hand, are usually evaluated on held-out sets that closely resemble the training data (Bhagavatula et al. 2020; Bisk et al. 2020), which can cause difficulty discerning whether a model with good performance has truly learned to solve the task or is just exploiting biases in the dataset (Gururangan et al. 2018; Gardner et al. 2020). An alternative approach is to use *behavioral* testing (Ribeiro et al. 2020) on these models where we measure the model capabilities by generating behavior-specific inputs for the model and comparing its output to expected behavior.

Behavioral testing for common sense comes with its own challenges. Notably, generating test cases and, in general, generating datasets is a costly and time consuming procedure. In the case of common sense reasoning, collecting common sense knowledge and converting it to a form that is suitable for a given task (e.g., premise-hypothesis pairs in natural language inference) requires a lot of effort. To decrease the cost of test case generation, previous works have proposed methods that partially automate certain aspects of the process. Zhou et al. (2020a) introduce a systematic procedure leveraging first order logic rules to prepare a benchmark to evaluate a natural language inference (NLI) model's common sense inference capabilities. Another example is CheckList (Ribeiro et al. 2020) which provides a platform for quickly creating numerous test cases to evaluate an NLP model's linguistic capabilities, but does not cover the cre-

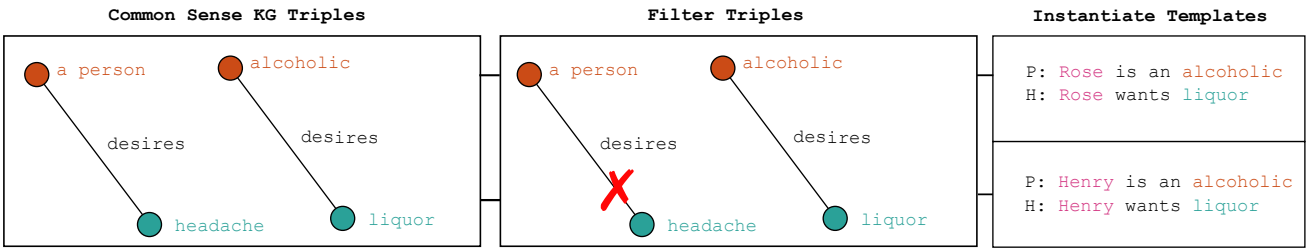

Figure 2: **Test Case Construction Pipeline**. We first *filter* the triples from common sense knowledge graphs to get semantically meaningful triples. The head (orange) and tail (teal) are used to *instantiate* relation-specific templates. The templates may contain additional slots for things such as names (magenta) which do not affect the outcome of the test case. (P: premise, H: hypothesis)

ation of tests that specifically target common sense.

In this work, we introduce a pipeline to create simple, scalable test cases to evaluate the common sense reasoning capabilities of NLI and question answering (QA) models. Our test cases are constructed from common sense knowledge graphs (CSKGs) such as ConceptNet (Liu and Singh 2004) and ATOMIC (Sap et al. 2019). However, we cannot directly test models using this structured information; information from CSKGs needs to be rendered in natural language to form valid model inputs. To create test cases for NLP models from these common sense knowledge graphs, we first select which knowledge triples we want to use. Then, using task and relation-specific templates, we generate test cases in the same format as the model input, and introduce variations on each template to increase the linguistic diversity of the tests.

An application of our proposed pipeline is illustrated in Figure 1. In this example, we create test cases to evaluate whether a BART (Lewis et al. 2020) model finetuned on the MNLI dataset (Williams, Nangia, and Bowman 2018) has learned to perform inference on instances requiring an understanding of desires (as captured in ConceptNet). These test cases are constructed so that the model should always predict *entailment*. We find this model makes an incorrect prediction 92.4% of the time, indicating that it's ability to solve instances requiring this understanding is limited. By studying performance across a variety of relations, we obtain detailed insights about the commonsense reasoning capabilities of different models.

2 Constructing Test Cases with CSKGs

In the following section we describe our method for constructing test cases from a common sense knowledge graph. An overview of our approach is provided in Figure 2.

Common Sense Knowledge Graphs A knowledge graph consists of a collection of triples *(head, relation, tail)* indicating the relations between the head and tail entities. While most knowledge graphs capture factual information about specific entities, e.g., *(Barack Obama, married to, Michelle Obama)*, common sense knowledge graphs instead try to capture generally agreed upon knowledge about more generic entities, e.g., *(alcoholic, desires, liquor)*. Our goal is to use this structured common sense knowledge to create high quality, simple test cases for a given task, such as NLI.

Filtering Facts Unfortunately, common sense knowledge graphs can contain triples that are not universally agreed upon, e.g., ConceptNet contains the triple *(a person, desires, a headache)*, and ATOMIC contains the triple *(personX reads six books, PersonX wants, never read again)*. Because the validity of these triples is subjective, they are not suitable for creating our test cases, so we filter them out. While there is some potential to perform this filtering using annotated confidence scores (if they are available), we have found that many problematic triples still have a high associated confidence, e.g., the ConceptNet example above. Thus, to ensure the quality of our test cases, we additionally perform a manual filtering step. While this limits the scalability of our approach, it only involves *simple* acceptability judgements (i.e., "Does this sentence make sense or not?") which can be scaled easily using crowd-sourcing, as annotation can be performed with little to no training.

Instantiating Templates Because triples are not stated in natural language, we introduce a number of templates to convert triples into properly formatted inputs for each relation and task. These templates specify where the *head* and *tail* of the triple should occur in the text, as well as placeholders for other fields like *names* that can vary across instances without affecting the expected label. For example, in Figure 2 we provide instantiated templates for the *desires* relation on NLI.

3 Evaluation Setup

3.1 Common Sense Knowledge Graphs

We use the following CSKGs to create our test cases (see the Appendix for details):

ConceptNet (Liu and Singh 2004): A structured database for common sense knowledge containing examples like *(a person, desires, feel happy)*.

ATOMIC (Sap et al. 2019): A knowledge graph capturing social event-based common sense knowledge , such as *(PersonX quits the job, PersonX wants, search for a new job)*.

3.2 Tasks

We create common sense test cases for natural language inference and common sense question answering as these tasks are critical in natural language understanding benchmarks (Williams, Nangia, and Bowman 2018). Example test cases for this task can be found in the Appendix.

Natural Language Inference We evaluate RoBERTa-large (Liu et al. 2019) and BART-large fine-tuned on the MNLI dataset. In MNLI, the goal is to classify the relationship between a pair of sentences, *(premise, hypothesis)*, as *entailment*, *neutral* or *contradiction*. Here, our goal is to create test cases evaluating common sense knowledge of these fine-tuned models. An example of these test cases is shown in Figure 2, with premise *Rose is an alcoholic* and hypothesis *Rose wants liquor*. These test cases are constructed so that, if the NLI model uses common sense knowledge while doing inference, the model should always predict *entailment*.We evaluate performance in both a 2- and 3-way setting. In the 3-way setting, evaluation is performed using all of the model's labels. Since model's tend to overwhelmingly prefer the *neutral* label on these cases, we additionally perform a 2-way evaluation where the model is forced to choose between the *entailment* and *contradiction* labels.

Common Sense Question Answering We evaluate on the Physical Interaction Question Answering (PIQA) (Bisk et al. 2020) dataset. For this task, given a *goal* sentence, the task is to choose between two possible solutions. This *goal* can be a question like *how do you put eyelashes on?* or a phrase e.g, *ice box*. To create our test cases, our templates for this task need to specify a goal, a correct and a wrong solution. To construct wrong solutions, for each correct triple, *(head, relation, tail)*, we negatively sample 5 tail nodes from the set of tail nodes that do not have the same relation with the head. We also check all the wrong answers manually to make sure that the test cases are semantically correct and are in high quality before instantiating the templates. We evaluate two models, RoBERTa-large and BERT-large (Devlin et al. 2019), fine-tuned on PIQA dataset on our generated test cases. These fine-tuned models have the accuracy of 76.3% and 66.2% on the PIQA development set, respectively.

4 Results

Natural Language Inference Results are provided in Table 1. We observe a large discrepancy in performance between the 2-way and 3-way settings, indicating that NLI models often predict *neutral* on these instances, but are capable of picking correctly between the other labels when forced. Generally, RoBERTa appears to perform better than BART. The failure rates across categories are also informative and show that models trained on NLI tend to perform better on tests involving ontological knowledge, e.g., "isa" and "usedfor", while they struggle at tests about "wants", "needs" and "desires." Specifically on the "desires" tests, the models have high failure rates even in the 2-way setting.

Common Sense Question Answering Results for the question answering task are provided in Table 2. With one exception, failure rates are lower for RoBERTa compared to BERT, and both models perform better on the tests than the NLI models from the previous section. These results are expected given that these models are fine-tuned on the PIQA dataset which is a commonsense reasoning task itself.

Rel	**3-way MNLI**		**2-way MNLI**	
	RoBERTa	BART	RoBERTa	BART
ConceptNet				
capableof	55.9	67.6	15.8	20.6
desires	93.0	92.4	50.3	54.4
createdby	73.8	80.7	8.0	19.0
usedfor	53.3	62.5	1.1	2.8
isa	10.0	15.3	4.7	5.9
madeof	58.4	76.1	16.8	28.3
hasa	71.4	81.0	19.6	25.6
ATOMIC				
xAttr	60.3	68.3	5.3	3.8
xNeed	100.0	99.4	36.2	26.3
xIntent	84.9	86.1	5.5	6.0
xEffect	86.4	91.4	17.3	31.4
xWant	95.6	98.0	21.6	36.4

Table 1: Failure rates for **NLI** model on our test cases.

ConceptNet	**PIQA**	
	RoBERTa	BERT
capableof	7.4	14.7
desires	5.4	17.9
createdby	4.3	14.0
usedfor	1.3	14.2
isa	4.1	26.0
madeof	9.1	17.3
hasa	5.3	18.9

ATOMIC	**PIQA**	
	RoBERTa	BERT
xAttr	11.6	33.2
xNeed	19.2	25.2
xIntent	15.2	24.2
xEffect	29.0	26.7
xWant	11.3	22.9

Table 2: Failure rates for **PIQA** model on our test cases.

5 Conclusions and Future Work

We introduce a pipeline to create common sense behavioural evaluation sets using common sense knowledge graphs. The pipeline consists of simple steps, however the resulting evaluation sets provide complex insights concerning model capabilities to perform common sense reasoning. While this semi-automated pipeline reduces the overhead of test case generation, it still requires manual effort mostly because of the low quality triples in the CSKGs. Applying our pipeline to other tasks and pursuing automated methods for denoising common sense knowledge graphs, are interesting avenues for future research.

Acknowledgements

We would like to thank the our reviewers and the members of UCI NLP for valuable feedback. This material is based upon work sponsored in part by NSF award #IIS-1817183 and in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research.

References

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; tau Yih, W.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=Byg1v1HKDB.

Bisk, Y.; Zellers, R.; LeBras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7432–7439. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/6239.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Gardner, M.; Artzi, Y.; Basmov, V.; Berant, J.; Bogin, B.; Chen, S.; Dasigi, P.; Dua, D.; Elazar, Y.; Gottumukkala, A.; Gupta, N.; Hajishirzi, H.; Ilharco, G.; Khashabi, D.; Lin, K.; Liu, J.; Liu, N. F.; Mulcaire, P.; Ning, Q.; Singh, S.; Smith, N. A.; Subramanian, S.; Tsarfaty, R.; Wallace, E.; Zhang, A.; and Zhou, B. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1307–1323. Online: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.117.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324* .

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.703. URL https://www.aclweb.org/anthology/2020.acl-main.703.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22(4): 211–226.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases?

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.442. URL https://www.aclweb.org/anthology/2020.acl-main.442.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics. URL http://aclweb.org/anthology/N18-1101.

Zhou, P.; Khanna, R.; Lin, B. Y.; Ho, D.; Ren, X.; and Pujara, J. 2020a. RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms. *arXiv preprint arXiv:2005.00782* .

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020b. Evaluating Commonsense in Pre-Trained Language Models. In *AAAI*, 9733–9740.

A Supplementary Material

The relation types, number of templates for each relation and number of test cases are represented in Table 3. The examples of test cases for both tasks and datasets are shown in Table 4.

Rel	triples	MNLI		PIQA	
		Templates	Tests	Templates	Tests
ConceptNet					
capableof	89	8	272	10	700
desires	44	8	171	100	223
createdby	81	16	336	10	600
usedfor	182	2	360	10	1482
isa	84	8	170	10	716
madeof	58	4	226	10	450
hasa	82	8	168	10	628
ATOMIC					
xAttr	31	15	451	15	399
xNeed	58	15	847	15	802
xIntent	41	15	598	15	539
xEffect	45	15	653	15	660
xWant	39	15	407	15	548

Table 3: **Dataset Statistics**

Task	CSKG	Relation	Example Test Case
NLI	ConceptNet	desires	P: These are cats H: They like to play
NLI	ATOMIC	xAttr	P: Dylan has always made good grades H: Dylan seems intelligent
PIQA	ConceptNet	desires	goal: These are dogs, sol1: they like to catch frisbees, sol2: they like to hear stories
PIQA	ATOMIC	xAttr	goal: Judy has passed the test, sol1: Judy is intelligent, sol2: Judy is supportive

Table 4: **Example Test Cases** for the tasks of NLI and PIQA from the common sense knowledge graphs