Non-entailed subsequences as a challenge for natural language inference

R. Thomas McCoy
Department of Cognitive Science
Johns Hopkins University
tom.mccoy@jhu.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Introduction: Natural language inference (NLI) — the task of determining whether a premise entails a hypothesis — is a central challenge for natural language understanding systems (Condoravdi et al., 2003; Dagan et al., 2006; Bowman et al., 2015). The availability of large sets of premises and hypotheses generated through crowdsourcing has made it possible to train neural networks without explicit logical representations to perform this task; such systems have reached considerable accuracy on these data sets (Radford et al., 2018; Kim et al., 2018). Recent studies have identified biases in these data sets which complicate the interpretation of these successes; for instance, statistical regularities in crowdsourced hypotheses make it possible to reach substantial accuracy without even considering the premise (Gururangan et al., 2018; Poliak et al., 2018). Since neural networks excel at capturing such statistical regularities, success on biased data sets may reflect fallible heuristics rather than deep language understanding, underscoring the need for a controlled experimental approach for evaluating NLI systems. To this end, we introduce a challenge set that targets the following possible heuristic:

(1) **The subsequence heuristic:** Assume that a sentence entails all of its subsequences.

This heuristic is attractive to a statistical learner because it often yields the correct answer for NLI sentence pairs:

(2) a. John likes Baltimore a lot. →
John likes Baltimore.
b. Roses are red, and violets are blue →
Violets are blue.

The subsequence heuristic is not a generally valid inference strategy, however; for example, it incorrectly predicts that the following sentence pairs are instances of entailment:

(3) Alice believes Mary is lying. ↛
Alice believes Mary.

(4) The book on the table is blue. ↛
The table is blue.

(5) The student sent the gift by Max yawned. ↛
The student sent the gift.

We conjecture that pairs such as (3)-(5), in which the hypothesis is a *nonentailed, nonconstituent* subsequence of the premise, are highly unlikely to be generated as potential contradictions by untrained annotators; consequently, they will not be available when training the model and will not be reflected in standard accuracy metrics.

We propose to create a challenge set that leverages the syntactic constructions illustrated in (3)-(5), as well as other constructions, to generate sentence pairs in which the hypothesis is a nonentailed nonconstituent subsequence of the premise. We demonstrate the viability of our approach with a set of sentences modeled after (3). These sentence are referred to in psycholinguistics as NP/S sentences (e.g., Pritchett 1988), because the verb (*believe*) can take either a direct object noun phrase (NP) or a sentence (S) as its complement; the hypothesis *Alice believes Mary* is the result of incorrectly assuming that the complement of the verb is the noun phrase *Mary* instead of the sentence *Mary is lying*. We evaluate a number of competitive NLI models on this challenge set. To anticipate our results, the accuracy of these models was close to 0% (when chance performance is 50%), supporting the hypothesis that they rely on the subsequence heuristic.

Models: We assess the performance of five neural-network NLI models. All models consisted of bidirectional LSTMs trained in two stages, following Wang et al. (2018): first, on one of the pre-training tasks described below, and then on NLI (with a classifier predicting the labels *entail-*

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 358-360.
New York City, New York, January 3-6, 2019

ment, *contradiction* and *neutral*), using the MNLI data set (Williams et al., 2018). Our pre-training tasks were: NLI using the MNLI corpus, combinatory categorial grammar (CCG) supertagging using tags from CCGbank (derived from the Penn Treebank) (Hockenmaier and Steedman, 2007), image generation from captions using the MS COCO data set (Lin et al., 2014), and language modeling (LM) using the WikiText-103 corpus (Merity et al., 2016). We also tested a model without pre-training, in which the encoder had random weights but the classifier was still trained on MNLI.

Data set creation: We generated premises using the template NP_1 V_1 S_1, where (i) NP_1 appeared as the subject of V_1 in the MNLI training corpus, (ii) the subject of S_1 appeared as the direct object of V_1 in the corpus, and (iii) S_1 appeared in the corpus (not necessarily as a complement of V_1). These conditions ensured that our examples were in the domain on which the models were trained, and that the models had been exposed to all words and dependencies in our examples. For example, based on the sentences in (6) from the MNLI training corpus, we generated the example in (7):

(6) a. The Knights believed that their goal was justified, however they would succumb to infighting.
b. No one believed the story that Miss Howard has made up.
c. San'doro said the story was awful.

(7) The Knights believed the story was awful. $\nrightarrow$ The Knights believed the story.

We built our examples around the verbs *heard*, *believed*, *felt*, and *claimed*. We generated 200 sentence pairs and had each one annotated by three workers on Amazon Mechanical Turk. We kept the 88 examples for which two of the annotators agreed that the example made sense and that the correct label was *not entailment*. Some premises from our data set shown are in (8)-(10), with the associated non-entailed hypotheses underlined:

(8) They claimed the cinema is in a steel sphere.

(9) The committee felt the pressure was applied by oversight entities.

(10) They heard the miners were prepared to fight.

	MNLI	NP/S	NP/S (no neg.)
MNLI	0.75	0.08	0.01
CCG	0.67	0.17	0.03
MSCOCO	0.61	0.24	0.03
LM	0.72	0.06	0.00
Random	0.73	0.03	0.01
Chance	0.33	0.50	0.50

Table 1: Accuracies on MNLI, our unmodified NP/S set, and our NP/S set with negation words removed.

Results: Table 1 reports accuracies on the MNLI development set and our NP/S set. All models performed reasonably well on MNLI but substantially below chance on the NP/S set. Closer inspection revealed that most examples that the models correctly labeled *not entailment* had a negation word in the premise but not the hypothesis:

(11) They heard the tapes are of **no** importance $\nrightarrow$ They heard the tapes.

(12) The young American believed the statistician is **not** involved. $\nrightarrow$ The young American believed the statistician.

This observation suggests that even when the models correctly labeled an NP/S example as *not entailment* they may have done so using a heuristic that relied heavily on irrelevant negation words. To test whether this was the case, we removed all negation words from the NP/S examples; as shown in Table 1, this caused the accuracy of all models to fall to nearly 0, suggesting that the models were indeed using a negation-word-based heuristic. Thus, even when the models provided the correct label on the NP/S evaluation set, they generally did so for the wrong reason.

Conclusions: All models perform poorly on the NP/S evaluation set, especially when irrelevant negation words are removed. These results indicate that standard neural models trained on crowdsourced NLI data sets are prone to heuristics based on subsequences and negation and suggest that there is substantial room for improving the sophistication of NLI models. The clear and interpretable results of our evaluation strategy motivate expanding our data set to include additional constructions with similar properties, some of which are illustrated in (3)-(5), to create an ambitious standard for measuring progress in NLI. In future

work, we will also expand this data set into a more general test suite for evaluating which heuristics a model has learned. This test suite will include the subsequence heuristic and the negation heuristic from the current work, as well as other heuristics based on properties such as lexical overlap between the premise and the hypothesis. We will also investigate other types of models trained on NLI, such as non-neural models and tree-based neural models, to test whether reliance on the subsequence heuristic arises from the the NLI task or from the sequential nature of standard RNNs, or both.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information. *arXiv preprint arXiv:1805.11360*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191. Association for Computational Linguistics.

Bradley L. Pritchett. 1988. Garden path phenomena and the grammatical basis of language processing. *Language*, 64(3):539–576.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122. Association for Computational Linguistics.