Final Project: Unsound Heuristic Detection via Dataset Transformations

Submitted by: Robert Dickerson

1 Abstract

Natural language inference (NLI) is the problem of deciding whether one piece of text entails another. In a recent paper [4], McCoy et al. look at the problem of NLI models that perform well over some training set due only to learned shallow heuristics that do not actually scale to wider varieties of data sets. They construct a dataset called HANS over which NLI models using certain superficial heuristics are shown to be inaccurate.

This project proposes to investigate another way of obtaining datasets that expose the use of bad heuristics, namely by perturbing sentence pairs that already exist in training and development sets. If this approach proves to be a sufficient way of detecting heuristic use, it may scale well to other models, heuristics, or domains without requiring the creation of bespoke templates for specific heuristics.

2 Introduction

Textual Entailment is a relation that holds between two sentences P (the premise) and H (the hypothesis) when H logically follows from P. For example, ”The lawyer drove his car to visit a client” entails ”The lawyer drove a car,” but not ”The client drove a car.” Natural Language Inference (NLI, also referred to as Recognizing Textual Entailment or RTE) is the problem of modeling and classifying textual entailment.

In [4], McCoy et al. look at the problem of NLI models that perform well over some training set due only to unsound heuristics that do not actually scale to wider varieties of data sets. They construct a set of 30,000 sentence pairs called HANS over which NLI models using three specific bad heuristics are shown to be inaccurate. The HANS data set was constructed from inference templates built by hand to expose models using these three heuristics.

The goal of this project is to experiment with exposing the presence of similar heuristics by making transformations to the datasets used in training of the NLI models themselves. These transformations may include swapping or shuffling words, replacing words with synonyms or antonyms, and adding or removing words from sentences. The idea is related to adversarial input perterbations for neural classifier [5], data noising for neural network smoothing [9]. Some of the specific transformations were inspired by [7].

2.1 Novel Aspects

Although this project tries to achieve the same end goal as [4], it differs in its approach. Instead of creating datasets from bespoke templates written per individual heuristic, the approach here is to randomly perturb existing datasets in a way that exposes unsound reasoning. This project also bears similarities to [7] and [9], although these works apply transformations to training data in order to prevent overfitting and speed up convergence during the training itself. In contrast, this project seeks to perturb training datasets to expose undesirable model behavior after training is complete.

3 Problem Definition

The basic problem of natural language inference is, given two sentences called the *premise* and the *hypothesis*, to determine whether or not the premise entails the hypothesis. This project uses the MNLI dataset [8], a collection of premise / hypothesis pairs labeled as *entailment* if the premise entails the hypothesis, *contradiction* if the premise contradicts the hypothesis, or *neutral* if the premise does not entail anything about the hypothesis one way or the other. (As in [4], this project collapses *contradiction* and *neutral* into a single *non-entailment* label during evaluation.) An NLI classifier, then, must correctly characterizes the entailment relation between two input sentences.

This project does not attempt to build an NLI classifier, but rather attempts to assess the presence of unsound heuristics in existing NLI classifiers. An unsound heuristic is taken to be any behavior that classifies based on shallow syntactic characteristics of the sentences rather than deeper semantic meaning. For example, classifying entailment whenever the hypothesis is any subsequence of the premise may often work, but is not sound in general. Given an NLI classifier and corpus, this project seeks to automatically generate a dataset that yields low accuracy from the classifier only if the classifier employs an unsound heuristic.[1]

4 Technical Approach

The approach taken in this project is to generate novel datasets by perturbing data from the model's training corpus. The idea is that, if an NLI model reasons solely on superficial characteristics of the input sentences, then making superficial changes to the training data should be able to confuse the model. The different perturbations considered in this project are given in Figure 1.

One of the main problems with this approach is maintaining or predictably altering the gold label in the labeled training data. Replacing a word, reordering a sentence, etc. may or may not have an effect on the entailment relation between the two sentences. For example, *premsubseq*, which replaces the hypothesis with a random subsequence of the premise, could produce the valid entailment "I ate a cheeseburger" / "I ate" or the nonsensical non-entailment "I at a cheeseburger" / "a cheeseburger." In this case, what should happen to the gold label depends on the chosen subsequence.

The approach adopted here is to set the gold label according to what a given transformation will *usually* do; in this case, *premsubseq* sets the gold label to non-entailment, as a random subsequence of the premise will not be a valid entailment in most cases. (See Figure 1 for the label behavior of each transform.) Although this approach is imperfect, having the wrong gold label on some examples is acceptable as long as the overall trend of the model's accuracy on the perturbation remains visible. In other words, if incorrect gold labels push a model's accuracy from 5% to 15%, there is still a strong signal that the model is using a bad heuristic. Whether or not these gold label inaccuracies are noisy enough to drown out this signal is a central question of this project.

[1] The definition of what constitutes a "superficial" or unsound heuristic is not made precise in [4] or in this project, and is arguably a subjective distinction. It may be the case that a classifier employing seemingly shallow reasoning but which nevertheless performs well in its target context should not be considered undesirable or incorrect. The rest of this paper will set aside the issue of what constitutes an unacceptable heuristic and focus only on generating datasets that cause NLI models to perform poorly.

Name	**Entailment**	**Description**
shuffle	non-entailment	Arbitrarily reorder the words in both the premise and hypothesis.
shuffleprem	non-entailment	Sets the hypothesis to an arbitrary reordering of the words in the premise.
premsubseq	non-entailment	Sets the hypothesis to an arbitrary subsequence of the premise.
neghyp	flip	Negates the hypothesis by prepending the clause ”it is not the case that.”
negprem	non-entailment	Negates the premise by prepending it with the clause ”it is not the case that.”
hXsyn	keep	*(X = verb, adv, or noun).* Randomly replaces each word with X part of speech in the hypothesis with a synonym.
hXant	flip	*(X = verb, adv, or noun).* Randomly replaces each word with X part of speech in the hypothesis with an antonym.

Figure 1: Dataset perturbations used to transform datasets in ways that might expose unsound heuristics in NLI models. Models marked as “keep” do not change the entailment label. Models marked as “non-entailment” change every label to non-entailment. Models marked as “flip” change entailment to non-entailment, contradiction to entailment, and leave neutral labels untouched.

5 Evaluation

5.1 Rationale

The main question this project tries to answer is whether simple syntactic transformations on NLI datasets are sufficient to expose the presence of unsound or shallow heuristics in NLI models.

5.2 Experimental Settings

Code used in the evaluation is available on Github at `https://github.com/rcdickerson/fritz`.

To generate the datasets for evaluation, the transforms described in the previous section were applied to the development sets of the MNLI [8] corpus. Synonym and antonym based replacements were carried out using WordNet [2]. The MNLI corpus has two separate development sets with similar accuracy characteristics (see Figure 2); the two sets were concatenated together to form a single base dataset for the perturbation transforms.

Evaluation was done on a Decomposable Attention model (DA) as described in [6] and an Enhanced Sequential Inference Model (ESIM) as described in [1]. The DA model is conceptually a bag-of-words approach, while ESIM is an LSTM-based approach. Both models were trained on the MNLI corpus [8] using the stock implementations available in the AllenNLP library [3]. The exact training configuration is available in the Github repository for this project, linked above. Both of these models were also evaluated in [4], where poor performance on the non-entailment examples in HANS indicated that both models make use of unsound heuristics in their classifications.

Each evaluation dataset was run through each model to determine the model’s overall accuracy for each kind of perturbation. Note that these accuracy results, presented in Figure 2, are relative to the gold label as set by the perturbation transforms as described in the previous section.

5.3 Results

Accuracy results are listed in Figure 2. The first three rows of each graph are the MNLI development sets, which indicate the baseline performance of the models over stock MNLI. The fourth row is the performance on the HANS dataset, and the following rows are the performance on the perturbation transforms described in the previous section.

The accuracy profile characteristic for HANS [4] is a high entailment accuracy with a very low non-entailment accuracy, indicating that the models are overeager to classify entailment on examples generated from HANS's shallow heuristic templates. This behavior was reproduced here, as seen in Figure 2. The perturbations closest to the heuristics tested for in HANS are *shuffleprem* and *premsubseq*, which similarly have low non-entailment accuracy. As these perturbations set the gold label to non-entailment across the board, there are no entailment examples to consider with the perturbation-based approach. However, it seems likely that the low performance on these perturbation types are due to the same shallow heuristics described in [4].

The various synonym and antonym based transformations are interesting cases, as there are multiple ways for models to end up misclassifying. Some classification errors are actual mistakes by the NLI model. For example, DA correctly classifies a pair with the hypothesis"there is an old woman in the village that I have been visiting". When the hypothesis is changed to "there is an old *man* in the village that I have been visiting" by *hnounant*, DA misclassifies the result. Other misclassifications, however, are caused by word replacements that do not affect the entailment relation in the desired way. One (somewhat gruesome) hypothesis "Jon stabbed the man's throat multiple times" is changed to "Jon stabbed the civilian's throat multiple times" by *hnounant*. Here, WordNet on input "man" used the synonym "serviceman" to generate the antonym "civilian". DA incorrectly classifies the result, although it is not clear the word replacement was strong enough to justify the switch in gold label in this case. Similar questionable antonym switches, for example "... forced to enter the country for work" into "... forced to enter the urban area for work" lead to similar questionable gold label changes.

For this reason, is currently unclear whether the synonym and antonym replacement based transforms are providing a valid signal about the presence of shallow heuristics in each model. Answering this question would require a more careful quantification of the source of misclassifications, which outside the scope of this project. It may be possible to mitigate some of the misclassifications due to inaccurate gold labels with smarter synonym / antonym selection, which is also left to future work.

6 Summary

Although it remains unclear how much signal some dataset transformations are yielding, it appears that *shuffleprem* and *premsubseq* is likely to be catching the heuristics described in [4]. Synonym and antonym replacements look promising, but more work is needed to judge their overall efficacy. Other potential future work could include evaluating the transformed datasets over more NLI models (the original proposal was to evaluate over two additional models, but this was cut for time), investigating incorporating perturbations back into the model training, checking for undesirable heuristics not described in [4], and training or perturbing corpora other than MNLI.

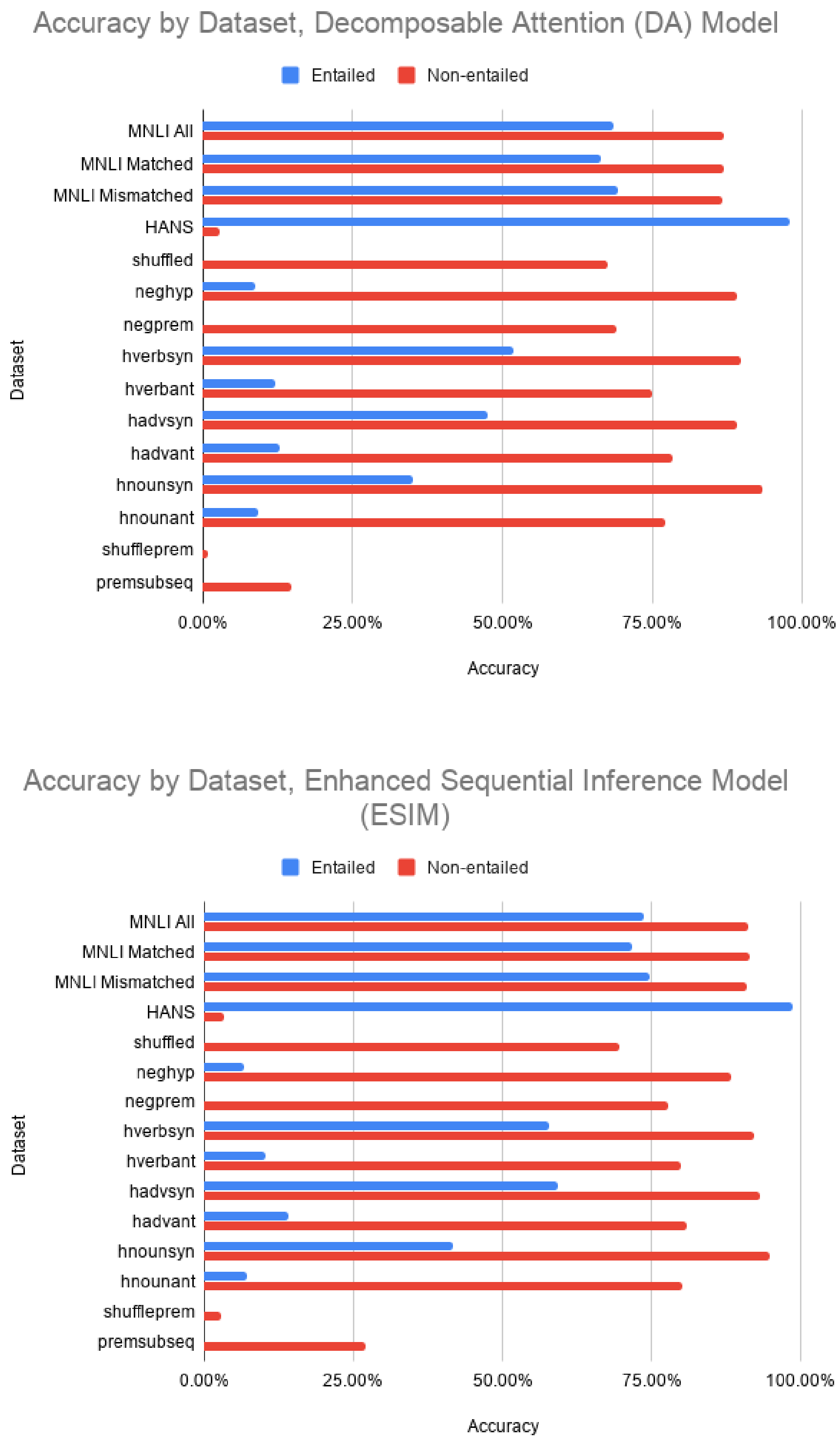

Figure 2: Accuracy of the Decomposable Attention (DA) model and Enhanced Sequential Inference Model (ESIM) over various datasets.

References

[1] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[3] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. E. Peters, M. Schmitz, and L. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018.

[4] T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019.

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[7] J. W. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[8] A. Williams, N. Nangia, and S. Bowman. 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, 2018.

[9] Z. Xie, S. I. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, and A. Y. Ng. Data noising as smoothing in neural network language models. *CoRR*, abs/1703.02573, 2017.