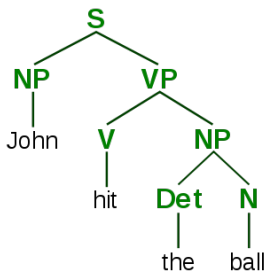
# Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

## I- Highlight and explanation:

This paper tackles the accuracy of heuristics in Natural Language Inference. It shows that even though heuristics succeed in the majority of the time and that the loss function gives good results, the models would fail to predict more challenging cases that require to understand underlying generalizations. A heuristic is a scientific method consisting in solving a problem in an approximative way, by choosing a certain limit for the loss function.
Natural Language Inference consists in determining if a sentence entails another. For instance, NLI would determine whether the sentence « The doctor was paid by the actor » entails the sentence « The actor paid the doctor ». This issue was tackled by determining whether a premise sentence entails a hypothesis. To show the limits generated by heuristics on this model, the authors used HANS (Heuristic Analysis for NLI Systems) as a training set. This datasets contains examples where the heuristic failed. This dataset is going to be used on models that do use these heuristic but still obtain good results. The goal is to consequently drop their accuracy scores.
This paper focuses on three main heuristics:

- Lexical overlap heuristic: with this heuristic, a premise entails a sentence if they are both based on the same words.
- Subsequence heuristic: if the hypothesis contains a subsequence of the premise, then there is entailment
- Constituent heuristic: if the hypothesis is a subtree of the parse tree generated from the premise, then the premise entails the hypothesis. A constituent parse tree divides the phrase into Noun (N), Verbal Phrase (VP), Verb (V), Noun phrase (NP), or D (determiner, definite article « the » for instance), and constitutes a tree. The following parse tree is for the sentence « John hit the ball »:




To do so, the model was trained on a dataset supporting these heuristic, such as MNLI or SNLI, and the test is going to be done on HANS. The HANS dataset is based on phrase templates, each template being accompanied by many couples of premises/hypothesis labelled with Entailment (E) or Non-entailment (N).
Four models were applied to HANS:

- the Decomposable Attention model, a sort of bag of words model which does not include word order as a parameter.
- The Enhanced Sequential Inference Model (ESIM), a sequential model (i.e the order of the words matters) using a bidirectional LSTM to encode sentences
- Stack-augmented Parser-Interpreter Neural Network (SPINN) which is based on parse trees

- Bidirectional Encoder Representations form Transformers model (BERT), a bidirectional model using the whole sentence at once rather than reading it from left to right. This is a state of the art model that has been developed by Google

Without any surprise, these models achieved very well on the MNLI dataset, with very high accuracies. However, on the HANS dataset, the accuracies were much lower, showing that there is indeed a heuristic problem in the construction of these datasets. In cases of entailment, the models showed very good results as the answers were in line with the hypothesized heuristics. The problem was with the non-entailment phrases, with accuracies lower than 10% for every model. The figure below, taken trom the paper, summarize the results of each model.

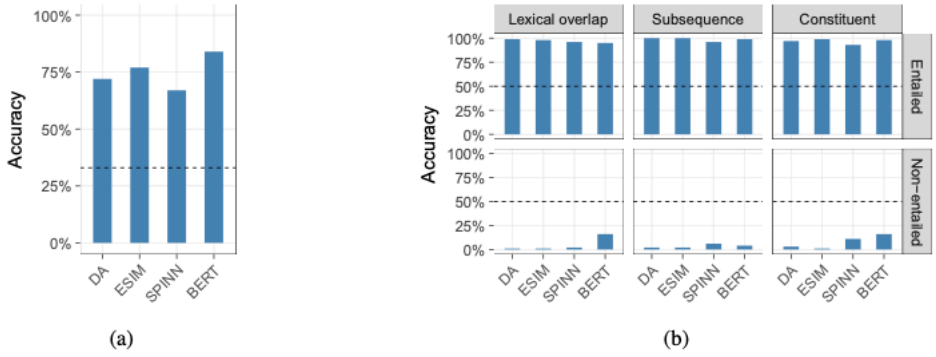



Figure 1: (a) Accuracy on the MNLI test set. (b) Accuracies on the HANS evaluation set, which has six subcomponents, each defined by its correct label and the heuristic it addresses. Dashed lines show chance performance. All models behaved as we would expect them to if they had adopted the heuristics targeted by HANS. That is, they nearly always predicted *entailment* for the examples in HANS, leading to near-perfect accuracy when the true label is *entailment*, and near-zero accuracy when the true label is *non-entailment*. Exact results are in Appendix G.

Although all the models performed poorly, the models did not exactly behave the same way:

- The DA model performed very poorly. This model is also called bag of words, meaning that it does not take into account the order of the words when processing the sentences. The bad result thus was very predictable. The results show that the model does not make any distinction between the three heuristics, and considers that they all follow the same heuristic : lexical overlap. Indeed, it is the only one which does not involve word order.
- The ESIM model too makes hardly no distinction between the heuristics, leading us to think that it does not use the word order to make a prediction even though it has access to that information.
- The SPINN model performs the best for the subsequence overlap, which might be due to its tree structure. It also performed very well for the constituent overlap. The tree structure is also believed to be the reason why the model has learned the importance of specific constituents on the meaning of a sentence. However, we notice that its results were slightly lower on cases where the correct answer is entailment. We can also see that the SPINN model has the worst accuracy compared to the three other models on the MNLI test set.
- BERT is, as the authors assumed, the best model. This is first shown on the MNLI test set, where its accuracy is the best. Then, it performed relatively well on all kinds of heuristics. It was the best model for constituent overlap, and especially for lexical overlap where we see a big difference with the other models. It was close to SPINN on

the subsequence heuristic. These results suggest that BERT takes more often into account word order as a parameter for its prediction than DA, ESIM or SPINN.

Another important thing to highlight is that a model do not behave the same on subcategories of heuristics. For instance, BERT achieved 39% accuracy on conjunction, but 0% accuracy on subject-object interversion.

## II- Analysis of advantages and issues:

The HANS dataset was made of 3 categories of heuristics : the constituent heuristic, the subsequence heuristic and the lexical overlap heuristic. All these categories are linked and even are nested: the constituent heuristic is a particular case of subsequence heuristic, which is a subcase of lexical overlap. The authors of the paper took this issue into consideration to make sure that each element of each category only belongs to one specific heuristic. This allowed to show that even though the inclusion between the different categories, models can behave worse on a subcategory than it does on a more general category. However, this phenomenon raises questions on the level of difficulty set for each category: are the categories all made of examples of the same difficulty?

Then, the authors compared the results of the models with the results of actual human beings. It showed that humans performed worse on the HANS dataset than they did on MNLI, which is similar to the results of the models. However, the accuracy was the same whether the label was Entailment or Non Entailment for humans, whereas the models performed significantly worse on Non Entailment. This permitted to rule out the possibility that the human errors could be affected by heuristics.

The aim of this paper was to show that even with high accuracies, there are flaws in the datasets used to make NLI. They succeeded in doing so by showing on different models that the performances are very poor for more challenging premises, and that this is caused by poor heuristics and a learning method that is not suited for more general phrase structures. The advantage of comparing 4 different models and seeing that they all performed poorly on the HANS dataset is that the authors could exclude the possibility that the problem came from the models. This highlights the fact that the training dataset if problematic and should be changed to achieve better results.

Finally, the authors do propose a solution to the problem: they created a dataset made of MNLI plus examples of each category included in HANS. The results were outstanding for most models. This proves that a dataset is less likely to learn a heuristic if the training dataset does not support them. The authors then compared the BERT model to other challenging datasets, while training them on the MNLI and the MNLI augmented testing set. The MNLI+ does improve the results: not very much for long phrases but dramatically for short premises.

## II- Personal take on the paper:

This paper is very interesting because it raises a problem that cannot be seen easily. It points out the limits of NLI by tackling the limits of the method used. The training dataset and the test dataset are usually the same when we test a model. However, this approach could hide deficiencies in the predictions, just as we have seen with the heuristics in this

paper. Not only have they proved this point, but they also came up with a solution to the problem with their augmented dataset.

One problem I noticed in this paper is the examples included in the HANS dataset. The premises are quite difficult and are not used on an everyday basis. Plus, some premises make no sense, such as « If the actor slept, the judge saw the artist ».

Finally, I think that this paper can have much impact in the data science and NLP community. Indeed, it reminds us the importance of choosing the right dataset if we want to make good predictions, but also models can always be improved in a way or another. This work is still open for improvements (such as the amount of data to add to maximize accuracy), and the authors admit it.