Benjamin LeBrun
benjamin.lebrun@mail.mcgill.ca

Final Results
MAIS 202

09/03/20

Natural language inference (NLI) is the task of determining whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) given a premise. For instance, the following sequence is tagged as *entailment*.[1]

(1) *Premise: A soccer game with multiple males playing.*
Hypothesis: Some men are playing a sport.

1 Results

In my previous deliverable, I mentioned that the results achieved on the HANS dataset were not consistent with those outlined by McCoy et al. Thus, I decided to re-implement my BERT model. To do so, I used huggingface[2] (formerly known as pytorch-transformers), which provides general-purpose architectures for NLU tasks. In doing so, I achieved results consistent with those outlined in the paper. The following deliverable will outline these results.

Recalling from previous deliverables, HANS is an NLI dataset that tests specific hypotheses about invalid heuristics that NLI models may adopt.[3] These syntactic heuristics are split into three categories: lexical overlap, subsequence, and constituent. It is important to note that even though these heuristics often lead to the correct answer, they are not valid inference strategies as they fail on numerous examples.[4] These three categories are defined by McCoy et al. in the following ways.[5] Note that the $\not\rightarrow$ symbol means "does not entail".

- **Lexical overlap**: Assume that a premise entails all hypotheses constructed from words in the premise.
  - e.g. **The doctor** was **paid** by **the actor**. $\not\rightarrow$ The doctor paid the actor.
- **Subsequence**: Assume that a premise entails all of its contiguous subsequences.
  - e.g. The doctor near **the actor danced**. $\not\rightarrow$ The actor danced.
- **Constituent**: Assume that a premise entails all complete subtrees in its parse tree
  - e.g. If **the artist slept**, the actor ran. $\not\rightarrow$ The actor slept.

Tables 1 and 2 show BERT's results on HANS. Table 1 shows the model's accuracy on examples where the correct label was `entailed`, whereas Table 2 shows the model's accuracy on examples where the correct label was `non-entailed`.

Heuristic	Accuracy
Lexical overlap	0.9588
Subsequence	0.9852
Constituent	0.992

Table 1: Heuristic entailed results

Heuristic	Accuracy
Lexical overlap	0.4464
Subsequence	0.0958
Constituent	0.1524

Table 2: Heuristic non-entailed results

One can easily see that BERT does not perform well on examples where the correct label is `non-entailed`. As a comparison, if the model was simply guessing, it would achieve results close to 50%. Humans, for their part, can achieve an 76% to 97% accuracy.[6] These results suggest that BERT does indeed adopt syntactic heuristics when it comes to natural language inference.

To explore these results in depth, we can assess the model's performance on the 30 sub cases provided by HANS. Table 3 shows the sub cases in which BERT failed to achieve more than 5% accuracy. Note that the HANS test dataset contains 1,000 examples from each of these sub cases.

[1] http://nlpprogress.com/english/natural_language_inference.html
[2] https://huggingface.co/transformers/
[3] https://arxiv.org/pdf/1902.01007v4.pdf
[4] Ibid.
[5] Ibid.
[6] Ibid.

Sub case	Accuracy	Template	Example
Non-entailment: Subject-object swap	0.022	The N_1 V the $N_2 \not\rightarrow$ The N_2 V the N_1	The senators mentioned the artist. $\not\rightarrow$ The artist mentioned the senators.
Non-entailment: NP/S	0.028	The N_1 V_1 the N_2 V_2 the $N_3 \not\rightarrow$ The N_1 V_1 the N_2	The managers heard the secretary encouraged the author. $\not\rightarrow$ The managers heard the secretary.
Non-entailment: Past-participle	0.004	The N_1 V_1 P the N_1 $V_2 \not\rightarrow$ The N_1 V_1 P the N_2	The senators paid in the office danced. $\not\rightarrow$ The senators paid in the office.
Non-entailment: Outside embedded clause	0.03	P the N_1 V_1 the N_2, the N_3 V_3 the $N_4 \not\rightarrow$ The N_3 V_1 N_4	Unless the authors saw students, the doctors helped the bankers. $\not\rightarrow$ The doctors helped the bankers.
Non-entailment: Disjunction	0.022	The N_1 V_1, or the N_2 V_2 the $N_3 \not\rightarrow$ The N_2 V_2 the N_3	The judges resigned, or the athletes mentioned the author. $\not\rightarrow$ The athletes mentioned the author.

Table 3: Sub cases in which BERT achieved < 0.05 accuracy

These examples offer us a glimpse into which types of sentence structures the model has difficulty with. As well, they shed light on the difficulty of generalizing to more challenging cases - ones that the model has never seen before.

Indeed, we can further assess how the model develops these heuristics by augmenting the training data with HANS examples from each heuristic. In others words, we can feed the model examples in attempt to see if this will increase accuracy. To do so, we re-train BERT on both the MNLI data set and 30,000 novel examples from HANS. After just one epoch, BERT achieves the following results.

Heuristic	Accuracy
Lexical overlap	1.0
Subsequence	1.0
Constituent	1.0

Table 4: Heuristic entailed results

Heuristic	Accuracy
Lexical overlap	0.9996
Subsequence	1.0
Constituent	1.0

Table 5: Heuristic non-entailed results

The stark contrast between these results and those of the pre-augmented training set suggest that while BERT is efficient when it comes to forgetting the heuristics. Indeed, after "seeing" 10,000 examples from each heuristic only once, BERT achieves almost 100% accuracy on HANS.

To test just how little the number of examples needs to be, I trained BERT on augmented data sets of varying sizes. To do so, I began by randomly selecting 100 examples from each sub-case (1000 from each heuristic) with at least one from each template. Unfortunately, the idea of doing so only came to me the night this deliverable is due. As a result, the results below show only a dataset of 3,000 examples (100 from sub-case, 1,000 from each heuristic). More will be added shortly.

Heuristic	Accuracy
Lexical overlap	0.9528
Subsequence	0.9938
Constituent	0.9898

Table 6: Heuristic entailed results

Heuristic	Accuracy
Lexical overlap	0.9752
Subsequence	0.9536
Constituent	0.9982

Table 7: Heuristic non-entailed results

Surprisingly, BERT can achieve impressive accuracy having only seen 3,000 examples, as opposed to the 30,000 provided in the dataset. As a reminder, the MNLI dataset contains almost 400,000 examples.

Smaller amounts will be added shortly.

2 Final demonstration proposal

My poster presentation will contain the following sections:

- **Introduction**: Introduce NLI by offering an introduction similar to the one at the beginning of this deliverable. Additionally, I will introduce the MNLI and HANS datasets with a few short sentences.
- **Hypothesis**: This will depend on which direction I take. (see results and discussion)
- **Model**: Introduce BERT and its architecture. (If I have space)
- **Method**: Brief explanation of the steps I took to produce these results.
- **Results and Discussion**: In this section, I will present the results shown above. Although, I feel as though I need to narrow my scope. In my opinion, there are two directions I could take. The first of which would be discussing BERT's syntactic heuristics, the second a discussion on BERT's ability to correct these heuristics.

  The first would entail presenting the examples that BERT struggled with, similar to what is done in Table 3. This would be close to what was done in the paper introducing HANS.

  The second direction would entail presenting the results obtained on HANS, and then presenting results obtained after augmenting the MNLI dataset in various ways (adding various amounts of examples from HANS). This was not done in the paper introducing HANS.

  There might also be a solution that combines both of these directions.
- **Conclusion**: Conclude, again, based on the direction I take.
- **References**: Self-explanatory