**Reviewing: Overcoming the Lexical Overlap Bias Using Predicate-Argument Structures**

Starting with, this is a must read for anyone using transformer models (Bert, Xlnet) for solving NLP problems. It is no big secret that the transformer babies are kicking ass in pretty much all NLP scoreboards. But does kicking ass in nlp scoreboards equate to real life applicability. For last year or so, I have been here and there testing transformers on various other domains. And I will be honest, I was impressed but not all the time. Model makes a lot of mistakes that it seems counterintuitive to what they should have learned. This was a real problem and I searched for a solution and that’s when I came upon this paper. This paper claims it will help me with my problems. So, let’s dig deep.

In the “Introduction” itself the paper states this obvious fact-

*“For instance, given the premise “Neil Armstrong was the first man who landed on the Moon”, the model may recognize the sentence “Moon was the first man who landed on the Neil Armstrong” as an entailing hypothesis or a plausible ending because it has a high lexical overlap with the premise.”*

I understood it wasn’t just me alone getting transferability issues with Textual Entailment models. That was a ‘whew’! :P

Paper goes to compare itself with existing solutions which tries to address this problem and they all require a few things-

*“new models or training procedures and the model’s complexity remains unchanged.”*

**Experimental Setup**

Let’s discuss on the experimental setup which they used to check if they could overcome the lexical Bias.

They are evaluating 2 NLP Tasks-

1. SWAG- “*Given a premise that is a partial description about a situation, GCI is the task of reasoning about what is happening and predicting what might come next. SWAG models this task as a multiple-choice answer selection, in which the premise is given and the correct and three incorrect endings are presented as candidate answers.”*

” *We created three different adversarial datasets based on the SWAG development set for evaluating the lexical overlap bias. These datasets evaluate the model’s understanding of (1) syntactic variations, (2) antonym relations, and (3) named entities in the presence of high lexical overlap*”

Let me be honest with you, I haven’t done much work with Swag and the task seems very interesting from a NLP perspective. However, I haven’t been able to visualize, how any pretrained models from SWAG can be used for solving upstream business problems. But I am always open to new ideas that challenges my beliefs and in search for it.

1. Natural Language Inference. Given a premise and a hypothesis, NLI is the task of determining whether the hypothesis entails, contradicts, or is neutral to the premise.

Now to overcome the Lexical Bias they wanted an evaluation dataset. Both the original SWAG & MultiNLI dataset have these biases, so evaluating on their test datasets won’t be useful.

So, for SWAG, they created 3 evaluation datasets. According to them-

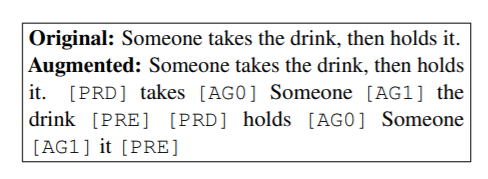
“*These datasets evaluate the model’s understanding of (1) syntactic variations, (2) antonym relations, and (3) named entities in the presence of high lexical overlap.”*

For NLI they are using an existing available dataset. Heuristic Analysis for NLI Systems (HANS) dataset. This was a welcome surprise to me.

**Their Approach**

If you have noticed earlier, I had mentioned that their approach does not involve any change in the model or any other complexity. So, the only way they could solve this is using data augmentation. And if this problem can be solved like this, it’s a big win for all of us as this can be applied for any dataset of our choice.

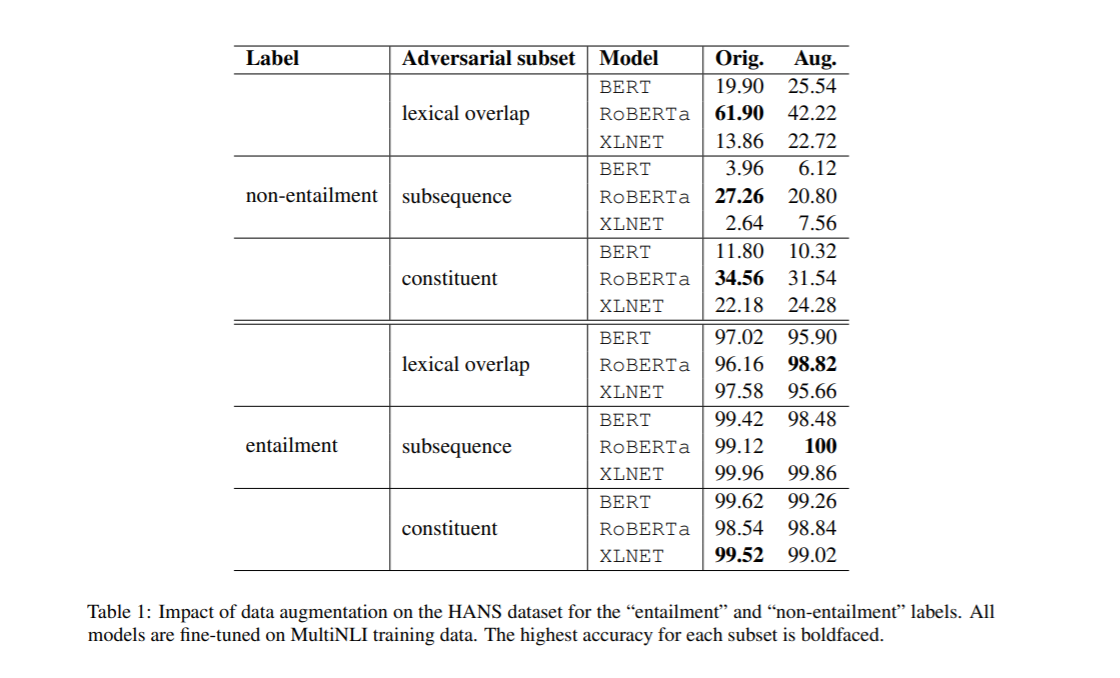
They augment the training set by introducing predicate-argument structure using the ProbBank-style semantic role labelling model of Shi and Lin (2019) , which has the state-of-the-art results on the CoNLL-2009 dataset, to get predicate-argument structures.. This is how data example will look after augmentation-



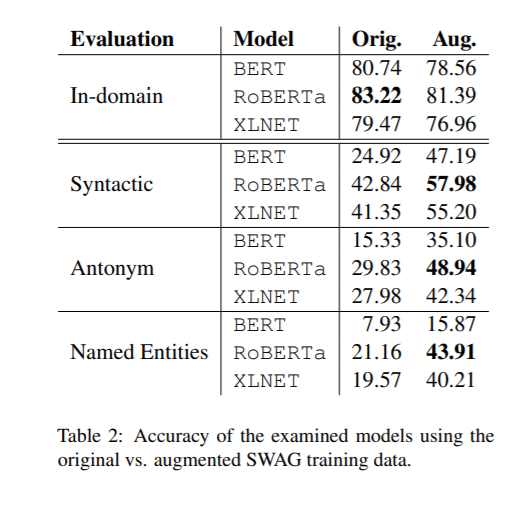
**Models Used for Setup**

They use bert, xlnet and Roberta models available from huggingface and then finetuned the models on Swag & MultiNLI Dataset for the GCI & NLI task respectively.

**Results**



1. Data augmentation is causing a substantial increase in “non-entailment” case for BERT & XLNET. for Roberta accuracy seems to have gone down a bit.
2. Clearly Roberta seems to be the most robust in case of MNLI with & without data augmentation. Time to use Roberta more.



1. The results are more clear in the case of Swag Dataset. Augmentation has increased accuracy for all 3 evaluation datasets. Here, also RobertA is dominating the pack.
2. It is to be concluded that models trained on MNLI are tending to have less issues when compared to the Swag ones. So, it seems that finetuning can have significant impact on the model’s understanding of natural language.
3. The results on the HANS dataset can vary by a large margin using different random seeds during fine-tuning, e.g., BERT accuracy on the lexical overlap subset can vary from 6% to 54%. This is a bit disappointing that random seeds can have such a significant impact. It qualifies the need for more dissection of BERT models.
4. They tried SRL using a different model (OPENIE which is less accurate). In this case also SRL resulted in an improvement in accuracy.
5. They alse tested the augmentation on Roberta-large. Here also, there was an improvement when they checked on Adversial SWAG datasets.
6. Luckily adding augmentation on test data did not yield any difference. It would have been a pain if it was the alternative.

**Key Takeaways-**

1. We need to get more understanding of Transformer Models. We now seem to have research papers conflicting each other on their understanding of Transformer Models
2. Better training and evaluation datasets are needed for NLP. SuperGlue is a big win, but we need more.
3. For now, I am of the opinion that transformer models will need significant finetuning before they can adapt to a business scenario.