The Impacts of Premise Length on NLI Models

Matthew Pelz
University of Texas at Austin, Department of Statistics and Data Sciences
mattpelz@pm.me

Abstract

Natural language inference classifies relationships between premise-hypothesis pairs. NLI models require training examples and the SNLI is one of the most commonly used sets. Previous studies found that SNLI examples may be suboptimal due to brevity of some premises. We explore impacts of premise length on NLI model accuracy by using previously published and author-generated NLI examples. We find that the inclusion of long-premise training examples may make models more resilient and that short-premise examples may lead to a skewed labeling distribution.

1 Introduction

Natural language inference (NLI) is a fundamental project within the field of natural language processing. Over the past two decades, researchers have studied numerous approaches for NLI data processing and modeling. While many language modeling tasks can be quite complex, the task of NLI is relatively simple: given a *premise* statement and a *hypothesis* statement, classify the premise-to-hypothesis relationship as either entailment, neutral, or contradiction. Put another way, if we assume a premise to be true, what can we sat about the hypothesis? Can we be certain that it is also true (entailment), can we safely reject it (contradiction), or are we can we do neither (neutral)?[1]

Given the three possible labels {E, N, C} for each premise-hypothesis pair, NLI amounts to a straightforward classification problem. Still, achieving high accuracy in the task requires the development of language models capable of learning complicated interrelationships involving syntax and vocabulary. NLI research has exploded in recent years after the introduction of such model frameworks. While these state-of-the-art models have achieved breakthrough accuracy scores, they have led researchers to scrutinize the nature of the actual underlying learning processes. High accuracy scores are less meaningful if they reflect some kind of statistical trick within the model rather than true linguistic comprehension. With this in mind, some research has focused on the character of training data and the search for *artifacts* in the data that a model might leverage to minimize loss without gaining any true, generalizable comprehension.[2]

For examples, Gururangan et al found three common artifacts that allow models to boost accuracy by recognizing patterns in hypotheses. Examining the SNLI and Multi-Genre Natural Language Inference (MNLI) datasets, the authors found that entailment hypotheses tend to contain gender-neutral references, that neutral hypotheses tend to include purpose clauses, and that contradiction hypotheses tend to include negation. The authors showed that models can correctly classify many examples by looking for these cues without considering any information in the premise. Based on goals of NLI, such inferences should not be possible.[3]

One framework for identifying data artifacts is the use of *adversarial* datasets, which punish the model's reliance on spurious information. Such research is conducted in the hopes that adversarial data, once demonstrated to weaken model performance, could be used to train or pretrain new, more robust models. Such models may not necessarily achieve higher accuracy, but they

[1] Gururangan et al (2018); McCoy et al (2019).
[2] McCoy et al (2019).
[3] Gururangan et al (2018); McCoy et al (2019).

would ideally be more successful against further adversarial challenges and thus be more generalizable.[4]

The Stanford Natural Language Inference (SNLI) corpus includes 570,000 premise-hypothesis pairs written and labeled by humans. The SNLI dataset is widely used in NLI modeling, and its familiarity and accessibility has made it a useful subject for researchers analyzing data artifacts.[5] Among those taking the adversarial approach to the SNLI data are Nie et al, who in 2019 introduced the appropriately named "Adversarial Natural Language Inference" (ANLI) corpus. Seeking to construct a dataset that was more difficult to classify, the researchers focused on creating premises longer than those found in the SNLI. They explain that the ANLI "remedies a problem with SNLI, namely that its contexts (or premises) are very short, because they were selected from the image captioning domain. We believe longer contexts should naturally lead to harder examples, and so we constructed ANLI contexts from longer, multisentence source material."[6] Indeed, premises in the ANLI dataset (average length of about 325 characters) tend to be much longer than those in the SNLI dataset (average length of about 72 characters).

The authors in Nie et al demonstrate that augmenting SNLI data with ANLI data in training does not hinder performance on NLI tasks, suggesting that such a procedure might help to generate more robust models. Yet they do not fully explain the rationale for identifying premise length as a key characteristic determining hardness. In the present study, we explore the impact of premise length on model performance by training different versions of an ELECTRA-small model through selective inclusion and exclusion of SNLI and ANLI data. We compare the premise lengths for correctlu and incorrectly labeled examples to see if patterns emerge. We verify the conclusions from Nie et al that the inclusion of ANLI examples, with longer premise lengths, does not negatively impact model performance. Lastly, we explore the possibility that premise length may have the potential to function as a kind of data artifact.

2 Procedure

To explore the potential impacts that premise length can have on accuracy and robustness, we generate two models for natural language inference and evaluate each on four distinct datasets. Both models use the ELECTRA-small framework, created by Clark et al in 2020 and provided with CS388 course materials.[7] Model 1 trains this model using only SNLI training data. Model 2 trains using SNLI training data as well as the 'train_r3' split of the ANLI data.[8]

To test the author's claim in Nie et al that augmentation of SNLI data with ANLI does not hinder performance, we evaluate the SNLI validation data using both models. Next, to explore the adversarial effects of the ANLI data, we evaluate both models using the ANLI 'dev_r1' split.

Having recognized the presence of artifacts in the SNLI corpus, we must also recognize the potential for similar bias in the ANLI corpus. Therefore, to further test the impacts of training on ANLI data, we evaluate Model 1 and Model 2 on two datasets annually generated by the author.

2.1 ANLI splits

The published ANLI corpus includes nine splits: three training splits (train_r1, train_r2, train_r3) three development splits (dev_r1, dev_r2, dev_r3), and three test splits (test_r1, test_r2, test_r3). As explained by the authors in Nie et al, the splits increase in hardness as they proceed from the r1 splits to the r3 splits. Here, we use the hardest training data, train_r3, and the easiest development data, dev_r1.[9] This decision stems from the fact that our goal is not to pressure test the difficulty of the ANLI splits; we assume that the claims in Nie et al regarding the relative difficulty of the splits is accurate. Rather, we train with the hardest data to give Model 2 an opportunity to learn from the most challenging ANLI examples before we examine impacts on evaluation. Conventiently, choices of ANLI splits have little effect on our analysis of premise length, as the average premise length for each split is close to 325 characters.

[4] Bartolo et al (2020); Liu et al (2020).
[5] Bowman et al (2015).
[6] Nie et al (2019).
[7] Clark et al (2020).
[8] The augmentation is a concatenation of the two datasets using only the 'premise', 'hypothesis', and 'label' features from the ANLI data.
[9] The train_r3 split has 100,459 examples and the dev_r1 split has 1000.

Premise	Hypothesis	Label
Tourism in the Maldives began in 1972. A United Nations mission on development which visited the Maldives Islands in the 1960s did not recommend tourism, claiming that the islands were not suitable. Ever since the launch of the first the resort in Maldives in 1972, however, tourism in Maldives has flourished. The arrival of the first tourist's group is estimated to have occurred in February 1972.	The Maldives has a thriving tourism industry.	0: entailment
All the trawlers were found to have problems including leaking, hogging, and severe stability issues. Eleven of the trawlers, after modifications, made the Atlantic crossing. After one foundered in a storm all the trawlers were relieved of escort duties, had one gun and some ballast removed to reduce draft to one more suitable for the work.	Skiffs are easier to operate than trawlers.	1: neutral
Walid Adel Shour is a Lebanese footballer who plays as a midfielder for Lebanese Premier League club Shabab Sahel, on loan from Lebanese Premier League club Ahed, and the Lebanon national team. Starting his senior career at Ahed, he was loaned out to Racing Beirut for one season in 2018. Upon his return to Ahed, he helped them win the 2019 AFC Cup.	Walid Adel Shour has never been to Lebanon.	2: contradiction

Table 1 - Samples from Pelz-1 dataset, which features examples with long premises. Premise texts comes from random Wikipedia pages. Hypotheses generated by author.

Premise	Hypothesis	Label
Trains first arrived in 1881 in Castelldefels with the introduction of the Vilanova line.	The Vilanova line, which included a connection allowing trains to arrive in Castelldefels, was introduced back in the nineteenth century.	0: entailment
The by-election was held due to the death of the incumbent Conservative MP, Joseph Leech.	Joseph Leech died of a heart attack at the age 78 and is survived by his wife and three daughters.	1: neutral
Born in Bucks County in the Province of Pennsylvania, Bennet attended the common schools.	Bennet was in fact born in international waters and his parents sent him to elite educational institutions around the world.	2: contradiction
Since 2008 Gedaschko is state minister of Economic Affairs and Employment in Hamburg.	Not only does Gedaschko work as a state minister in Hamburg, he has maintained this particular position within the government since 2008.	0: entailment
The land has undergone extensive open-cast mining and subsequent restoration.	Open-hearted restoration and subsequent mining have brought changes to this extensive land.	1: neutral
The security protocol protecting many Wi-Fi devices can essentially be bypassed.	The tremendous value of various types security protocols lies in the fact that no one can bypass them.	2: contradiction

Table 2 - Samples from Pelz-2 dataset, which features examples with hypothesis character length greater than premise character length. Premise texts comes from random Wikipedia pages. Hypotheses generated by author.

2.2 Author-generated datasets

To facilitate exploration of the impacts of premise length, the author has manually generated two small datasets. The first, denoted Pelz-1, features 51 examples with equally distributed labels. The premises range in length from 300 to 360 characters. This range was chosen to mimic the ANLI corpus. The second dataset, Pelz-2, features 30 examples, also with equally distributed labels. In this case, the premise lengths are chosen to mimic a subset of the SNLI corpus. Samples from the two datasets are presented in tables 1 and 2.

For both author-generated datasets, text for the premises comes from random Wikipedia pages.[10] For Pelz-1, hypotheses range in difficulty but none are intended to be overly vague for most human annotators. For Pelz-2, hypotheses are intended to be more uniformly easy, but each is generated so that the hypothesis has a longer character length

[10] https://en.wikipedia.org/wiki/Special:Random

than its associated premise. This choice was made to investigate a potential data artifact, as described in the results section.

Neither the Pelz-1 nor Pelz-2 datasets have been reviewed by secondary annotators to verify their clarity or accuracy. Therefore, there exists the potential for errors within both sets. The hope here is that the author-generated datasets will provide means for exploration rather than firm proof regarding the existence of premise-length-related data artifacts.

3 Results

After training Model 1 with SNLI data and training Model 2 using SNLI data augmented with the train_r3 split from the ANLI data, we evaluate each model on the evaluation sets described above and find noteworthy differences in accuracy between the two models.

3.1 Evaluation using SNLI data

Evaluating on the SNLI validation split, Model 1, trained only on SNLI training data, achieves an accuracy of 0.8978. This is in line with expectations provided with CS388 course materials. On the same evaluation data, Model 2, trained with SNLI and ANLI data, has an accuracy of 0.8971. The similarity between the two scores supports the conclusion made by Nie et al that augmentation of SNLI data with ANLI data does not hinder performance.

3.2 Evaluation using ANLI data

Using the dev_r1 split of the ANLI data, Model 1 had an evaluation accuracy of 0.3190 and Model 2 had an evaluation accuracy of 0.4650. Clearly, the ANLI data was much more difficult to classify, just as its authors intended. We note that Model 2 performed much better on this task, perhaps due to the fact that the SNLI + ANLI training data includes more examples with long premises matching the lengths of those found in the ANLI evaluation set. We also consider the possibility that the ANLI training and evaluation splits themselves share biases or artifacts that a model might learn, so comparing results on a third-party dataset may be illustrative.

3.3 Evaluation using author-generated data

Eavluating on the Pelz-1 datset, Model 1 has an evaluation accuracy of 0.6275, Model 2 substantially outperforms this mark with an evaluation accuracy of 0.7843. This finding, along with the results from 3.1 and 3.2, suggest that the inclusion of examples with longer premises can help models evaluate similar examples without decreasing total performance. This alone, however, does not necessarily mean that characteristics of premise length allow models to find shortcuts in the data. In order to explore this possibility, we utilize the Pelz-3 dataset to scrutinize examples from the SNLI training set in which the hypothesis has more characters than the premise. There are 1161 such examples found in SNLI evaluation data, and only 222 (19.1 percent) have entailment labels. Model 1, trained only on SNLI data, predicts only 197 entailment labels.

To investigate the idea that Model 1 may have learned during training that examples with hypotheses longer than their premises are less likely to have entailment labels, the Pelz-2 dataset includes only such examples, all intended to be fairly easy for most human annotators. Evaluated on this dataset, Model 1 has an accuracy of 0.6667, but more importantly it predicts only five entailments out of the 30 examples, suggesting the model may indeed have a bias against predicting entailment in these cases.

The Pelz-2 dataset is not evaluated using Model 2 as model comparison is not relevant to this aspect of the study.

4 Analysis

Through examination of the evaluation results for both models across the various datasets, we find evidence that (1) training models with long premises may improve model performance better when predicting labels for other examples with long premises and (2) that examples from the SNLI training split that have very short premises may prompt models to draw incorrect correct conclusions due to the unequal labeling distribution in these examples.

Eval Set	Model 1	Model 2
SNLI	0.8978	0.8971
ANLI	0.3190	0.4650
Pelz-1	0.6275	0.7843

Table 3 – Comparing model accuracy using different evaluation sets. Model 1 trains only on SNLI data while Model 2 trains on both SNLI and ANLI data.

4.1 Long premises

We find that Model 1, trained on the SNLI corpus and relying on its relatively brief premises, struggles to evaluate examples with longer premises from the ANLI data. Evaluation accuracy figures are presented in Table 3. Model 2, trained using SNLI and ANLI data, is not substantially worse when evaluated using only SNLI data, and it strongly outperforms Model 2 on the evaluation sets that use longer premises.

To see this trend from another perspective, we can look at the premise lengths of examples that the models label incorrectly. Table 4 lists the average premise lengths for incorrect labels when evaluating on the SNLI and Pelz-1 datasets. These figures, combined with the accuracy differential from Table 3, support the notion that Model 2 performs better at predicting the longest examples from the SNLI set and that some of the advantage Model 2 has in predicting the Pelz-1 data stems from its ability to predict the longest examples.

Eval Set	Model 1	Model 2
SNLI	74.39	73.95
Pelz-1	329.5	326.6

Table 4 – Average premise length of incorrectly labeled examples from the SNLI and Pelz-1 evaluation sets. Model 1's mistakes tend to have longer premises.

4.2 Short premises

While Model 1 may struggle with long-premise examples, it excels when dealing with short premises. For examples from the SNLI dataset with a shorter-than-mean premise length, Model 1 has a .9050 evaluation accuracy. Our last task is to explore possible spurious reasons behind this high performance. We consider the results from evaluation of the models on the Pelz-2 data. As Table 5 shows, among the SNLI examples with premises shorter than their hypotheses, the actual number of entailment examples is disproportionately small (0.1912). If Model 1 has recognized such a trend and consequently limited the number of entailment predictions, we would expect that the distribution of predictions to reflect a bias. This appears to be the case, as Model 1 predicts an even smaller proportion of entailments (0.1697) for examples in this category. Certainly, there could be many reasons for this small proportion of predicted entailments. This could easily be a case of a smart model predicting each example well without any systematic bias.

To further investigate the issue, we use the Pelz-2 dataset, which includes only premise-longer-than-hypothesis examples. Unlike the SNLI subset, however, the distribution of gold labels for the Pelz-2 dataset is uniform. In this case, we observe tendency for Model 1 to avoid entailment labels, predicting only 5 entailments out of 30 examples.

Why might Model 1 have fewer entailment predictions? A clear understanding of the mechanism at work would require a deeper analysis of the underlying distributions. Still, the practice of generating the Pelz-2 examples inspired some ideas regarding how human authors may play a role.

If we consider each character as a unit of information, then a hypothesis with more characters than its premise consequently has the capacity to contain more information. This capacity is naturally well suited to contradictory or neutral examples, in which the "extra" information can be incongruent or superfluous to the premise. To create entailment examples that fit the pattern, however, the hypothesis in essence becomes a more verbose version of the premise (or of a segment of the premise). Such verbosity may seem unnatural for human authors tasked with writing hypotheses for natural language inference datasets, given that writing and composition education typically promotes a style based on concision and clarity. Of course, there could be other factors in play, but this theory may explain the small proportion of such examples in the SNLI data.

Eval Set	Actual proportion of entailments	Predicted proportion of entailments
SNLI (short-premise subset)	0.1912	0.1697
Pelz-2	0.3333	0.1667

Table 5 – Proportions of entailments among actual and predicted labels when evaluating Model 1 on a subset of SNLI data and an author-generated dataset. Both datasets feature only examples in which the premise has more characters than the hypothesis.

5 Conclusion

In this paper, we explore the impact of premise length on model performance. We created two ELECTRA-small models: one model trained using only data from the SNLI corpus, which has an average premise length of about 72 characters, and a second model trained using the same SNLI data as well as ANLI data, which has an average premise length of about 325 characters. Using multiple datasets and categorizing them according to premise length, we find that the inclusion of the ANLI data does not harm overall model performance even when evaluating on only short-premise examples. Furthermore, we find that including the long-premise ANLI examples in training enables much higher model performance on long-premise evaluation sets. Lastly, focusing on examples in which the hypothesis has more characters than the associated premise, we present evidence that models trained on SNLI data can learn to favor the labels "neutral" or "contradiction" over "entailment" due to label distributions found in a subset of the SNLI data.

References

Max Bartolo, Alastair Roberts, Johannes Weibl, Sebastian Reidel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotation. In Transactions of the Association for Computational Linguistics, Volume 8, 2020 pages 662-678. arXiv:2002.00293

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. EMNLP 2015. arXiv:1508.05326

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR). arXiv:2003.10555

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In Proceedings of NAACL. arXiv:1803.02324

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics. arXiv:1707.07328

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial Training for Large Neural Language Models. arXiv:2004.08994

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019.Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. arXiv:1902.01007

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. arXiv:1910.14599

Eric Wallace, Shi Feng, Nikhil Kandpa, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. EMNLP 2019. arXiv:1908.07125