Analyzing and Improving ELECTRA Performance on Neutrally Labeled Premise/Hypothesis Pairs

Single Author – mwl759
The University of Texas at Austin, MSDS

Abstract

Overreliance on popular annotated datasets like SNLI and MNLI can hinder model generalization as they can be fraught with annotation artifacts. Artifacts can stem from crowd worker bias, lexical choice, and even spurious sentence length correlations making them nontrivial to address. This study investigates the impact of artifacts outlined by Glockner et al. and Gururangan et al. on an ELECTRA model trained on the canonical SNLI dataset. Through a detailed analysis of model error types, error categories, and sentence length, we reveal that the initial model underperforms on neutrally labeled premise/hypothesis pairs. Focusing on this challenging data subset, we find that fine-tuning on many diverse data sources improves model generalization. Despite these promising findings, model validation on additional out of domain test sets is necessary for a production ready model. In future work we recommend incorporating lexical knowledge, similar to Chen et al.'s Knowledge based Inference Model, as it may also improve model performance. Overall, this study highlights the importance of rigorous and diverse data collection for model training.

1 Introduction

NLI (Natural Language Inference), also known as RTE (Recognizing Textual Entailment) tries to identify if a premise is entailed by a hypothesis. SNLI (Stanford Natural Language Inference) (Bowman et al., 2015), and other commonly used NLI benchmarks are often used for training, testing, and validation of NLI models. Training and validating only a few datasets can often lead to poor performance when used in production because of "annotation artifacts" (Gururangan et al. 2018).

Sources of these artifacts are numerous. This includes sentence length, lexical choice, and general crowd worker bias. It has also been shown that a model's ability to predict a test example can depend on the number of similar examples in the training set (Glockner et al. 2018). Despite these implications we will set out to verify that the assertions put forth by Glockner et al., Gururangan et al., and others, still hold true with an ELECTRA model further trained on the standard SNLI dataset.

After we identify some of these artifacts, we will also try to incrementally improve the model's performance by fine-tuning on training data from diverse genres and sources. This single concatenated training set will be referred to as "Custom Fine-Tune Training Set" or CF-TTS. Our hope is that each data source provides the model information not covered by the other datasets. If, for example, SNLI skews labels because the entailed hypothesis sentences are often very short, other datasets who have greater diversity of sentence length for entailed hypotheses may help remove this spurious artifact. By adding more data diversity this may help remove artifacts such as sentence length, label imbalances, annotator bias, etc.

We will then measure performance on the training data sources' respective test sets. These test sets will be concatenated into a single validation dataset, which we'll call "Custom Fine-Tune Validation Set" or CF-TVS in this study. Our measure of success will be incremental improvements on categories the model performed poorly on or improvements to overall model evaluation accuracy on CF-TVS.

2 Model, Data Sources, and Analysis Tools

The model we used was the ELECTRA-small (Clark et al., 2020) implementation trained on the standard Hugging Face SNLI dataset for 3 epochs and batch size 64. We refer to this as the "Base ELECTRA Model" or "Base" Model. We did not change the standard training dataset distributed by Hugging Face. We also did not alter the standard run.py training procedure provided.

Fine-tune training sets used in CF-TTS were minimally processed via concatenation only and were sourced from Hugging Face. The datasets used to fine-tune the model include MNLI (Williams et al., 2018), "Breaking NLI" (Glockner et al.) validation set, AmericasNLI (Ebrahimi et al., 2021) validation set, ANLI (Nie et al. 2019), SciTaiL (Khot et al. 2019), WANLI (Liu et al. 2022), and FEVER-NLI (Nie et al. 2019). Each fine-tune training set, except SciTaiL which contains neutral and entailment only, contains premise/hypothesis pairs labeled with entailment, contradiction, or neutral.

We leaned on the "Breaking NLI" dataset from Glockner et al.[1], SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and our CF-TVS dataset for our error analysis. Notably, the "Breaking NLI" premise is taken directly from the SNLI corpus. However, it replaces one word or phrase from the premise to generate the hypothesis. These premise/hypothesis pairs are semantically grouped into 14 unique "categories" based on the entity replaced, lending this useful for analyzing incorrect predictions. MNLI and CF-TVS are better measures of generalization than SNLI alone given they have many more genres and variations in validation data.

Data, tables, and charts, were generated and analyzed using Docker, Python, Apache Superset, and DuckDB.

3 Analyzing Incorrect Predictions

To guide where to start improving the model, we first identified places the Base model failed. We analyzed error type, error category for the datasets that have it, and sentence length. We also include general and specific error examples. Recall, 0: Entailment, 1: Neutral, 2: Contradiction for all tables in this paper.

Dataset	**Eval Accuracy: Base Model**
SNLI	89.85%
Glockner "Breaking NLI"	94.96%
MNLI	70.80%
Custom Combined Test Set CF-TVS	67.98%

Table 1. Overall Evaluation Accuracy – Base Model

3.1 Error Type

When trained on the SNLI training dataset, the model correctly identified 8,802 of 9,842 (89.43% accuracy) examples. Broken down by error types on the incorrect predictions, the model was less likely to falsely predict contradiction when the gold label is entailed (8.3% of incorrect predictions) and vice versa (6.4% of incorrect predictions). The balance (~85% of incorrect predictions) involve the neutral label, or neutral prediction, in some combination. This indicates that the model struggles with neutrally labeled premise/hypothesis pairs. Neutral pairs tend to be more ambiguous and complex than contradictions and entailments as we will see in section 3.4 and 3.5.

Error Type	**Count**	**Percent of Total**
Label: **1**, Predicted: 2	230	22.11%
Label: 2, Predicted: **1**	220	21.15%
Label: **1**, Predicted: 0	219	21.06%
Label: 0, Predicted: **1**	218	20.96%
Label: 0, Predicted: 2	86	8.27%
Label: 2, Predicted: 0	67	6.44%

Table 2. Base Model - SNLI Error Type Categorization

The model struggles with neutral pairs and predictions on the MNLI test dataset as well. However, this error occurs more frequently when the gold label is neutral as opposed to the model incorrectly predicting neutral for the other gold labels. Notably, it also predicts neutral incorrectly to the same extent it did with the SNLI dataset.

[1] https://github.com/BIU-NLP/Breaking_NLI/tree/master

75.14% of the error types involve neutral labels, or predictions, in some combination.

Error Type	**Count**	**Percent of Total**
Label: **1**, Predicted: 2	1196	21.30%
Label: **1**, Predicted: 0	1185	21.10%
Label: 0, Predicted: **1**	929	16.54%
Label: 2, Predicted: **1**	910	16.20%
Label: 2, Predicted: 0	816	14.53%
Label: 0, Predicted: 2	580	10.33%

Table 3. Base Model - MNLI Error Type Categorization

Similar analysis on the "Breaking NLI" test dataset from Glockner et al. reveals the model struggles with the contradiction gold label. This is to be expected, to some extent, because there are many more contradiction instances for the model to mis-predict: 7,164 of the 8,193 (87%).

Error Type	**Count**	**Percent of Total**
Label: **2**, Predicted: 1	160	42.22%
Label: **2**, Predicted: 0	142	37.47%
Label: 1, Predicted: **2**	32	8.44%
Label: 0, Predicted: **2**	31	8.18%
Label: 0, Predicted: 1	13	3.43%
Label: 1, Predicted: 0	1	0.26%

Table 4. Base Model - "Breaking NLI" Error Type Categorization

For our large custom dataset CF-TVS, we see that the model still can't predict neutrally labeled pairs very well. CF-TVS data is more diverse, larger, and has better label balancing making this a stronger indicator of how well our Base model will generalize. This highlights that the model still struggles with neutral labels, just as it did with the MNLI and SNLI. 72.54% of the error types involve neutral labels, or predictions, in some combination.

Error Type	**Count**	**Percent of Total**
Actual: 1, Predicted: 0	4705	26.49%
Actual: 1, Predicted: 2	3801	21.40%
Actual: 2, Predicted: 0	2858	16.09%
Actual: 0, Predicted: 1	2263	12.74%
Actual: 2, Predicted: 1	2115	11.91%
Actual: 0, Predicted: 2	2019	11.37%

Table 5. Base Model - "CF-TVS" Dataset Error Type Categorization

3.2 Error by "Breaking NLI" Category

Recall that Glockner et al.'s "Breaking NLI" premise/hypothesis pairs are semantically grouped into 14 unique "categories" based on the single word or phrase they replaced. Surprisingly, the model performed very well on the "Breaking NLI" test dataset from Glockner et al. It predicted 7,814 of 8,193 (95.4%) with no additional training or fine-tuning. We do not have a working hypothesis as to why our Base model performs so well despite their claims. It may be possible modern architectures such as ELECTRA are able to catch this specific artifact type. However, this claim is questionable as ELECTRA has the same architecture as BERT.

Despite this, of the 4.6% that were incorrectly classified, we see the category the model performed the worst in was in "Antonyms" and "Antonyms Wordnet". This accounted for 50% of the incorrect predictions, or 190 of 379 absolute incorrect counts. Notably, other incorrect categories include "rooms" accounting for 13% of incorrect predictions, followed by "synonyms" at 8%, and "drinks" at 6%.

Category	**Incorrect Count**	**Percent of Total**
Antonyms wordnet	109	28.76%
antonyms	81	21.37%
rooms	50	13.19%
synonyms	31	8.18%
drinks	23	6.07%
vegetables	20	5.28%
ordinals	18	4.75%
cardinals	16	4.22%
colors	11	2.90%
materials	6	1.58%
nationalities	5	1.32%
planets	5	1.32%
instruments	3	0.79%
countries	1	0.26%

Table 6. Base Model - "Breaking NLI" Incorrect by Category

3.3 Sentence Length

It has been suggested that hypothesis sentence length can be an artifact (Gururangan et al. 2018). Figure A.4 and A.7 show significant right skew indicating there are many shorter hypotheses in the SNLI dataset – specifically entailed examples compared to neutral ones. Given the baseline model is trained on this, it is unsurprising that Figure A.1 demonstrates that the model incorrectly predicts entailment on longer sentence lengths (42 chars avg. length of incorrect examples vs. 34 characters length of correct examples). This corroborates Gururangan et al.'s claims. Furthermore, Figure A.9 demonstrates that MNLI has a balanced blend of labels across sentence lengths. Training on diverse datasets such as this may help reduce the model's dependence on these types of artifacts.

3.4 General Class: Errors

In general, the model struggles with neutral labels. Neutral pairs can be difficult for human annotators as well. The "Breaking NLI" dataset by Glockner et al., also contains an "annotator labels" column capturing each individual annotators score. All 47 neutral examples were non-unanimous. In comparison 88% of entailment examples and 82% of contradiction examples were consistently labeled by all three annotators in Figure A.10.

The model also generally struggles with antonyms and synonyms. We see this in the Table 5. category breakdown for "Breaking NLI". This could be due to lack of lexical context and that antonyms/synonyms are more complex than the other categories.

3.5 Specific Class: Errors

Appendix tables A.11 and A.12 both demonstrate some specific errors from the base model. This includes both annotation errors and model errors. This is a small subset of only 5 examples from SNLI, however these specific errors shed more light on why the Base model may not generalize.

Annotation errors are often driven by nuance in specificity. This can be when an annotator labels something as neutral when the correct label is entailed. An example pair being: premise - "A group of people are standing on a platform and waiting for the subway train.", hypothesis - "People gathering in a group" with gold label 1 (neutral) and predicted label 0 (entails). In this case, the annotator should have used gold label 0 (entails), but mis-labeled the data instead. These mistakes can be driven by simple human error or nuance in how the annotator understood the pair.

A specific case of the model's failure to generalize well can be seen with small perturbations to the hypothesis. An example being: premise - "The greyhounds are running quickly in this race.", hypothesis - "The big greyhounds are running quickly in this race" gold label 0 (entails) and predicted label 1 (neutral). Even misspellings, such as "caine" instead of "cane" in table A.11, can lead the model to incorrectly predict 2 (contradiction) as the label. These examples corroborate Glockner et al. in that a single word change can trigger an incorrect prediction.

4 Approach

Based on our analysis, we hypothesize the model can generalize better on premise/hypothesis pairs with neutral labels. Our first theory was that by exposing the model to more diverse datasets filtered to *neutral-only* labeled premise/hypothesis pairs, the model will generalize better to this subset.

Our second theory is that training on a variety of different data sources collected by many different organizations will allow the model to generalize better *regardless of whether we filter to neutral-only labels*. We believe doing so may help reduce the model's dependence on artifacts such as sentence length, label imbalances, annotator bias, etc. by letting the model see artifacts from different angles in different contexts.

For our first theory, we processed and concatenated the datasets used to train the model, MNLI (Williams et al., 2018), "Breaking NLI" (Glockner et al.), AmericasNLI (Ebrahimi et al., 2021), ANLI (Nie et al. 2019), etc. into a single Json file: CF-TTS. *All non-neutral gold labels were removed from this dataset.* Once processed, the base ELECTRA model previously trained on SNLI was further fine-tuned with a batch size of 64 with "task" nli.

Our second theory followed the same procedure as the first generating CF-TTS, however we kept all of the original gold labels. Once processed, we still fine-tuned on top of our base ELECTRA model trained on SNLI with a batch size of 64 and "task" nli but for 5 epochs.

5 Results and Discussion

After fine-tuning on our first theory, the dataset with neutral only gold-labels, the model predicted neutral for *every* validation example. This was unintended and only biased the model to always predict the majority label. This approach did not work as expected and so our first theory was quickly invalidated.

Continuing with our second theory, we kept all of the CF-TTS gold labels intact during fine-tuning. This led to more compelling results. Though this second approach worsened performance on the standard SNLI dataset, performance increased for all other test sets. This is perhaps unsurprising as the Base model itself was only trained on the SNLI dataset. It had "overfit" to the artifacts in the SNLI dataset and therefore did not generalize to the other datasets.

Dataset	Eval Accuracy: Base Model	Eval Accuracy: Fine-Tuned Model
SNLI	89.85%	84.34%
Glockner "Breaking NLI"	94.96%	99.21%
MNLI	70.80%	81.81%
Custom Combined Validation Set "CF-TVS"	67.98%	78.88%

Table 7. Overall Evaluation Accuracy – Base and Fine-tuned Models

The new Fine-Tuned model is much more adept at handling neutrally labeled premise/hypothesis pairs as well. Of the % incorrect neutrally labeled premise/hypothesis pairs, each test set, except Glockner et al.'s "Breaking NLI", performed ~8-12% better.

Dataset	% *Incorrect* Neutral Gold Label: Base Model	% *Incorrect* Neutral Gold Label: Fine-Tuned Model
SNLI	43.17%	33.48%
MNLI	42.40%	34.20%
Glockner "Breaking NLI"	8.71%	46.15%
Custom Combined Validation Set "CF-TVS"	47.89%	35.36%

Table 8. Overall Evaluation Accuracy – Base and Fine-tuned Models

Returning to the SNLI test set with our fine-tuned model, we see that despite the lower overall performance the model has shifted its top error type category to predicting neutral for contradiction pairs. This is double the nearest error type category (~36% vs ~18%). Though SNLI is a flawed dataset, this indicates the model is more familiar with neutrally labeled premise/hypothesis pairs having been exposed to more data.

Error Type	Count	Percent of Total
Actual: 2, Predicted: 1	592	36.10%
Actual: 1, Predicted: 0	307	18.72%
Actual: 0, Predicted: 1	287	17.50%
Actual: 1, Predicted: 2	242	14.76%
Actual: 0, Predicted: 2	114	6.95%
Actual: 2, Predicted: 0	98	5.98%

Table 9. Fine-Tuned Model - SNLI Error Type Categorization

We see again that the Fine-Tuned model is more apt to predict a neutral label whereas the Base model often missed neutral gold label pairs. The distribution of error types has remained balanced, if only shifted. Despite the balance shift, overall evaluation accuracy is much higher.

Error Type	Count	Percent of Total
Actual: 0, Predicted: 1	859	24.04%
Actual: 2, Predicted: 1	813	22.75%
Actual: 1, Predicted: 2	667	18.67%
Actual: 1, Predicted: 0	555	15.53%
Actual: 2, Predicted: 0	347	9.71%
Actual: 0, Predicted: 2	332	9.29%

Table 10. Fine-Tuned Model - MNLI Error Type Categorization

With 99% of Glockner et al. "Breaking NLI" predicted accurately, analysis is more skewed. We believe this is why there is such a large discrepancy between the two incorrect neutral gold label percentages.

Revisiting Glockner et al.'s error categories, the fine-tuned model shows improvement in the general antonyms class, but still misses out on the "antonyms wordnet" category. The fine-tuned model has significantly improved on synonyms as well, indicating that it may have made some progress on "understanding" more complex concepts. Curiously, "simpler" categories such as drinks and vegetables are still quite high.

Error Type	Count	Percent of Total
Actual: 1, Predicted: 2	29	44.62%
Actual: 2, Predicted: 1	21	32.31%
Actual: 0, Predicted: 2	8	12.31%
Actual: 2, Predicted: 0	4	6.15%
Actual: 0, Predicted: 1	2	3.08%
Actual: 1, Predicted: 0	1	1.54%

Table 11. Fine-Tuned Model - “Breaking NLI” Error Type Categorization

Category	Incorrect Count	Percent of Total
antonyms_wordnet	16	24.62%
cardinals	11	16.92%
drinks	9	13.85%
antonyms	7	10.77%
vegetables	7	10.77%
colors	6	9.23%
ordinals	4	6.15%
nationalities	1	1.54%
materials	1	1.54%
synonyms	1	1.54%
rooms	1	1.54%
planets	1	1.54%

Table 13. Fine-Tuned Model - “Breaking NLI” Incorrect by Category

Finally, our fine-tuned model exhibits the same tendency to predict neutral as it did on the MNLI test set. In this case, however, the error types are more balanced. To some degree, the fine-tuned model is better spreading its errors over several types.

Error Type	Count	Percent of Total
Actual: 0, Predicted: 1	2442	20.84%
Actual: 2, Predicted: 1	2434	20.77%
Actual: 1, Predicted: 0	2376	20.27%
Actual: 1, Predicted: 2	1768	15.09%
Actual: 2, Predicted: 0	1647	14.05%
Actual: 0, Predicted: 2	1052	8.98%

Table 12. Fine-Tuned Model - “CF-TVS” Dataset Error Type Categorization

5.1 Evaluating the Improvement

Training the model on diverse data has proven to be very effective in increasing the Base model’s ability to generalize. Not only has there been improvements to overall evaluation accuracy, but also to the sub-group we wanted to improve: neutral gold labeled premise/hypothesis pairs both in total counts and percent mixes. The fix generally improved performance. However, validation on Glockner et al. “Breaking NLI” revealed there were simple categories with only a single word from that category replaced, such as vegetables or drinks, that we believe should have been correctly predicted.

6 Next Steps and Further Investigation

Other areas to investigate include using external lexical knowledge to improve the new fine-tuned model. Glockner et al. and Chen et al. both demonstrate this concept with KIM (Knowledge-based Inference Model). This may help the model generalize even better. Additional models, other than ELECTRA or BERT, may also yield better results.

More robust categorization of common NLI datasets could help researchers better understand the types of errors models are making. This could be done automatically or manually during the data collection process. Regardless, there are countless papers that stress the importance of proper data collection. As can be seen from the results here, diverse and properly labeled datasets are paramount.

It would be interesting to see how this model performs against other natural language problems such as QA. Some open questions that can be further investigated include: why did the Base model do so well on Glockner et al. “Breaking NLI”? This dataset was supposed to exemplify issues with data artifacts in SNLI but did not. Additionally, how can we verify that more data truly addresses the artifacts in question? Training on more data may only help the model within the domains it was trained on. Additionally, the model may unintentionally be learning the artifacts of all of those datasets as well. Lastly, incorporating more in-production validation or trying to test on new out of domain datasets may expose weaknesses in the model and should be further investigated.

7 Conclusions

Our initial goal was to identify artifacts and corroborate the work of Gururangan et al. and Glockner et al. We started with analyzing the specific and general errors of our ELECTRA Base model on test sets from SNLI, MNLI, Glockner et al.'s "Breaking NLI", and a custom dataset of CF-TVS. We found the model generally struggled with neutrally labeled premise/hypothesis pairs and somewhat with lexical relations. We then fine-tuned on CF-TTS to try and improve model performance on these challenging subsets of NLI data.

With renewed focus on improving this subset, we had two theories. One theory is that the Base model could be fine-tuned with data filtered to *neutral-only* gold labels from the CF-TTS dataset we constructed. This method proved ineffective as the model would then only predict neutral as a label. Our second theory was that training on a variety of different data sources, with *all CF-TTS gold labels intact*, would allow the model to generalize better. These results were more compelling, and we found that this drastically improved model evaluation accuracy on our chosen test datasets. It also greatly improved the model's ability to correctly identify neutral gold labels, shifting the distribution of error types to be more balanced. Our second theory was confirmed and we ultimately met our success criteria.

References

[Bowman et al., 2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and D. Christopher Manning. 2015. A large annotated corpus for learning natural language inference. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

[Chen et al., 2018] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL).* Melbourne, Australia

[Clark et al., 2020] 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).*

[Ebrahimi et al., 2021] Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. In *arXiv:2104.08726 [cs], CoRR, volume abs/2104.08726. https://arxiv.org/abs/2104.08726*

[Glockner et al., 2018] Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In Proceedings of the 56th Annual Meeting of the Association for *Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

[Gururangan et al., 2018] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. New Orleans, Louisiana

[Khot et al., 2018] Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A Textual Entailment Dataset from Science Question Answering. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI).*

[Liu et al., 2022] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In Proceedings of the Association for Computational Linguistics (ACL), January 2022. URL: https://arxiv.org/pdf/2201.05955

[Nie et al., 2019] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics.*

[Nie et al., 2019] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI).*

[Williams et al.,2018] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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, New Orleans, Louisiana, June. Association for Computational Linguistics.

A Appendices

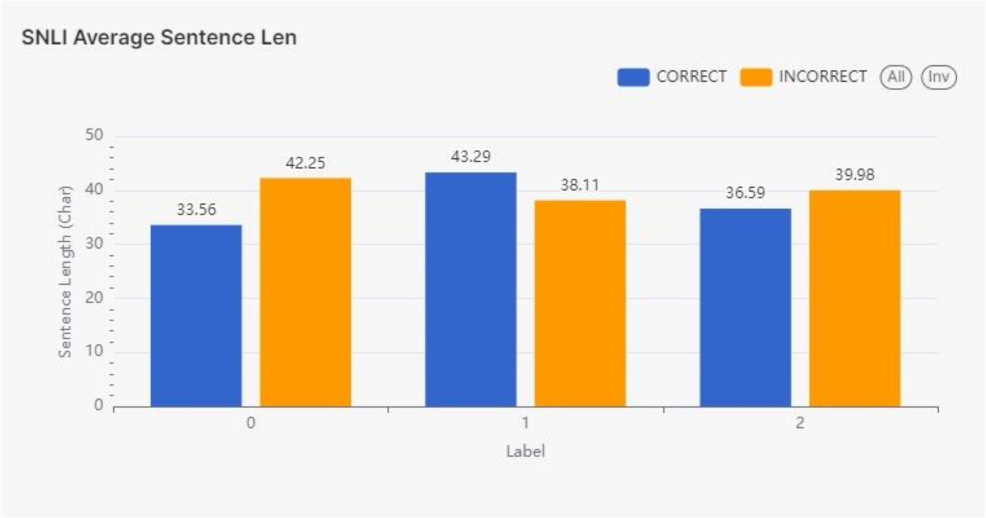

Figure A.1: SNLI Sentence Length by Label

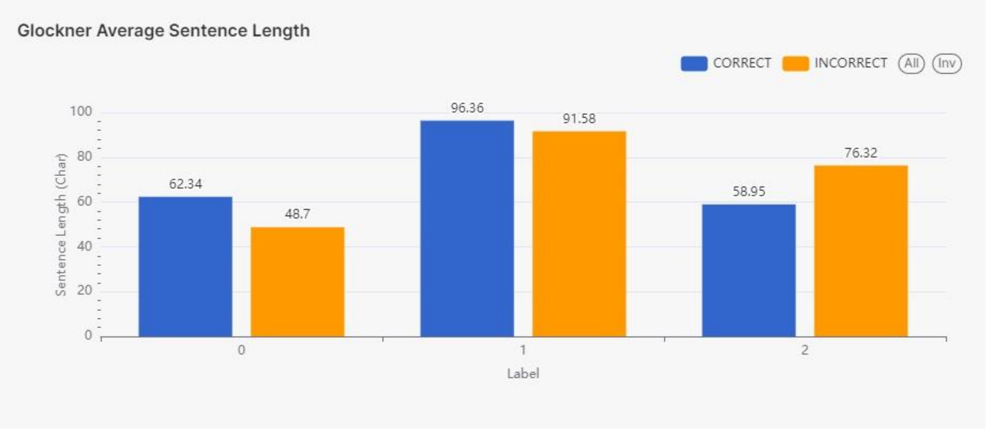

Figure A.2: Glockner et al. "Breaking NLI" Sentence Length by Label

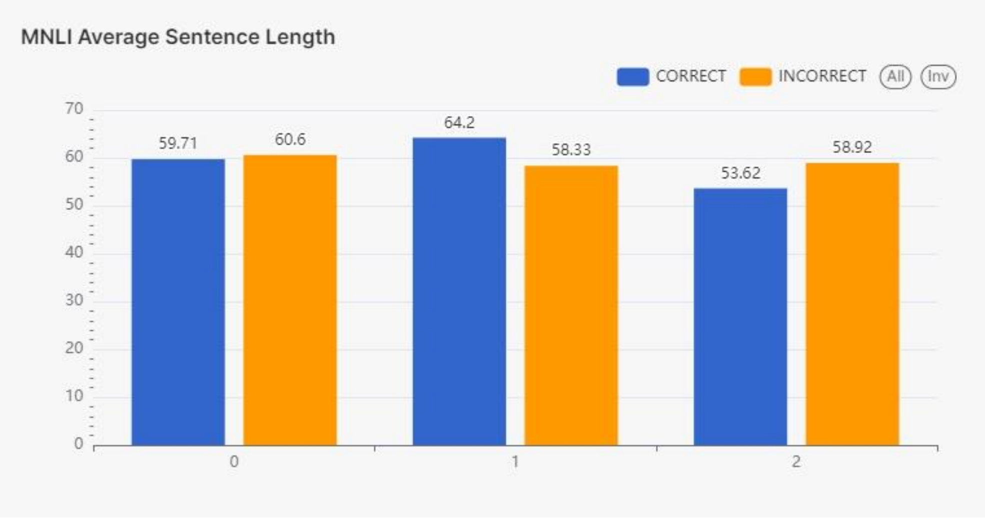

Figure A.3: MNLI Sentence Length by Label

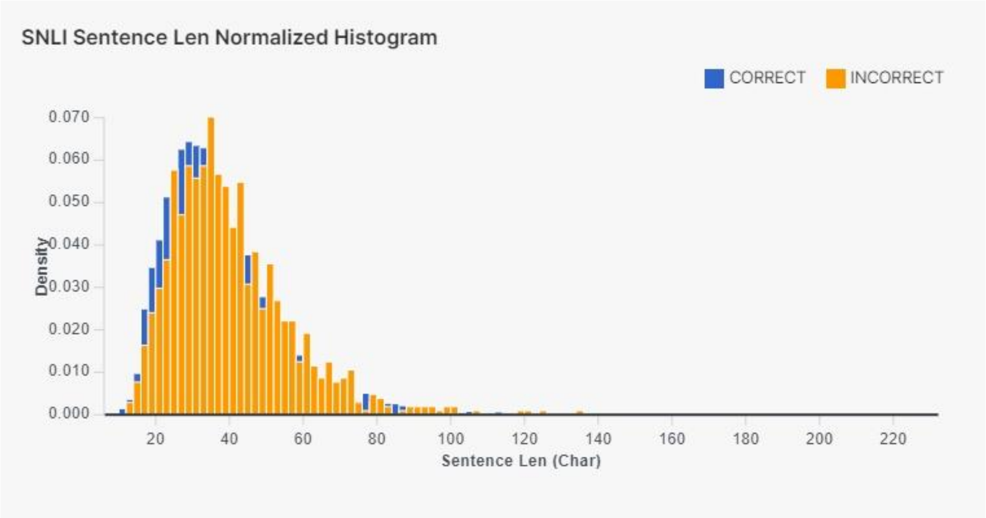

Figure A.4: SNLI Sentence Length Normalized Histogram

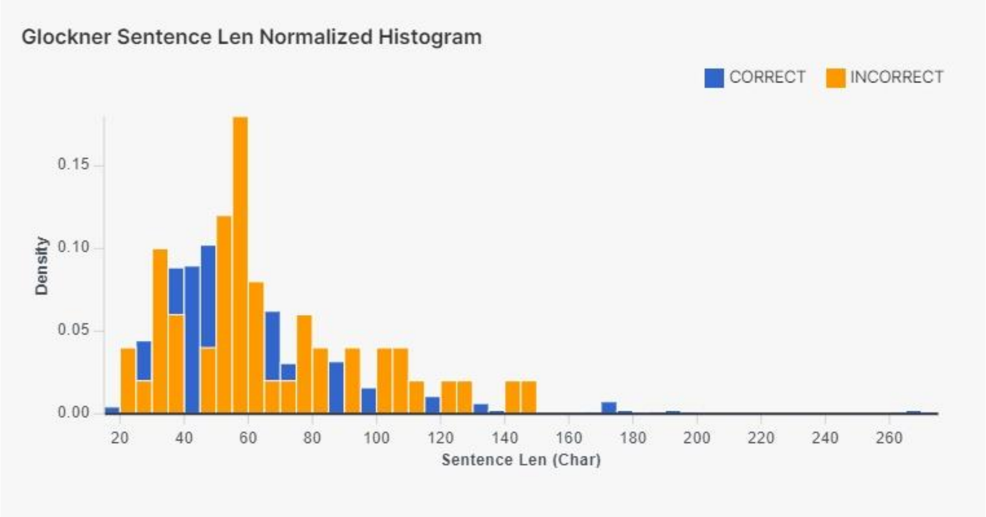

Figure A.5: Glockner et al. “Breaking NLI” Sentence Length Normalized Histogram

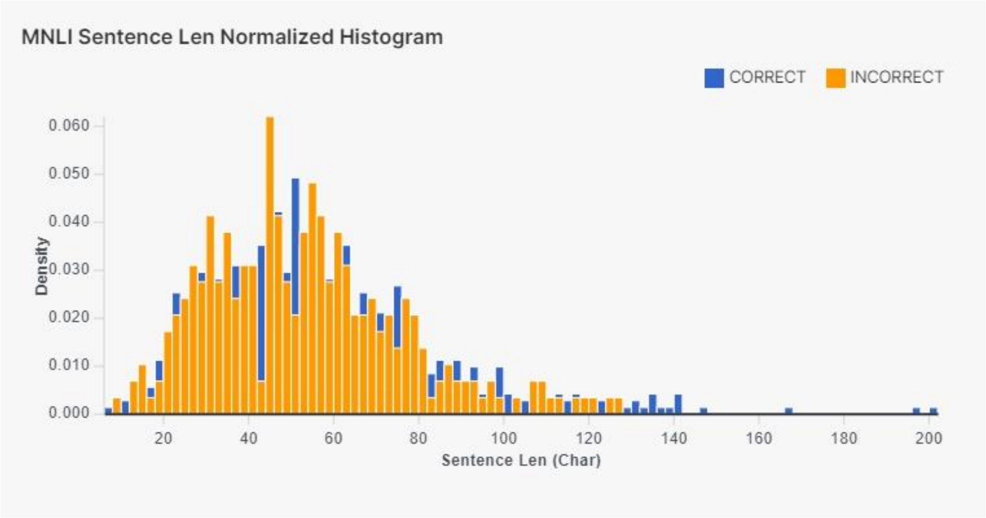

Figure A.6: MNLI Sentence Length Normalized Histogram

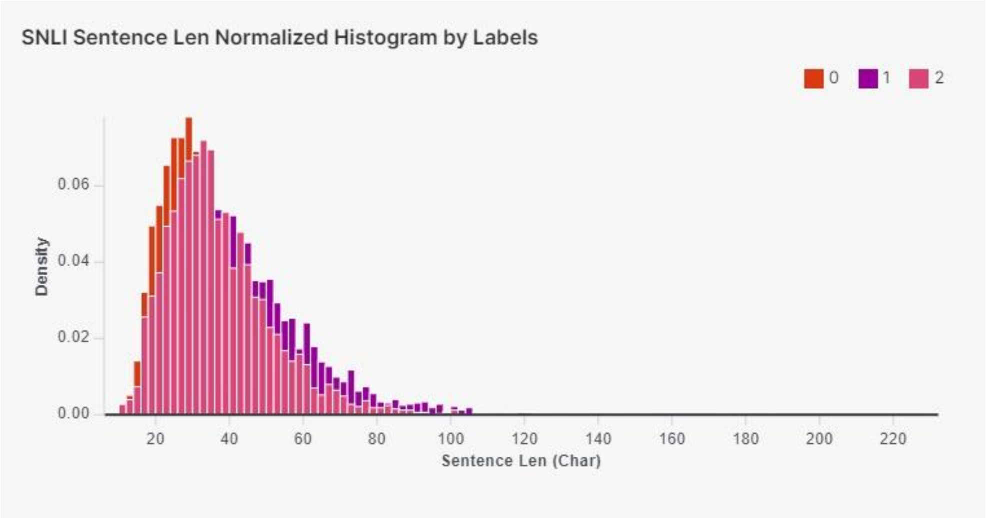

Figure A.7: SNLI Sentence Length Normalized Histogram

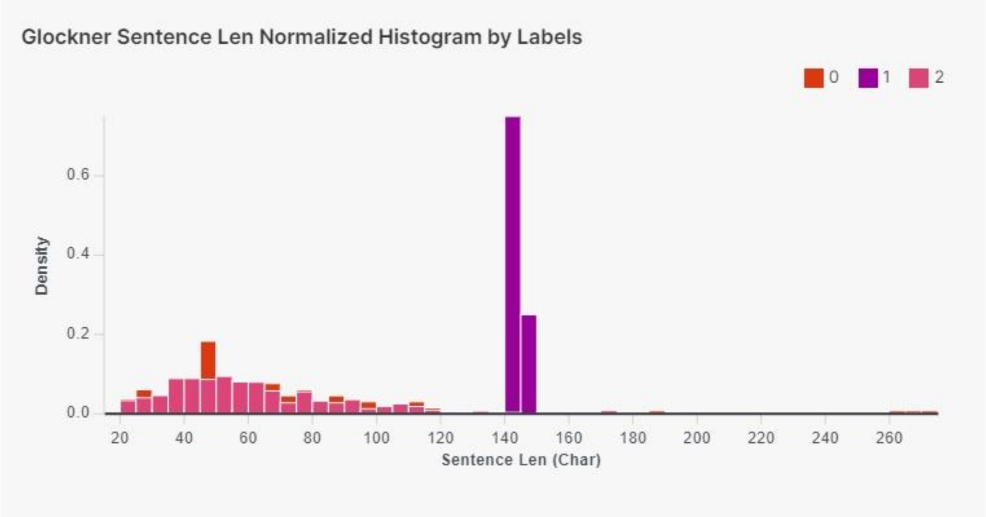

Figure A.8: SNLI Sentence Length Normalized Histogram

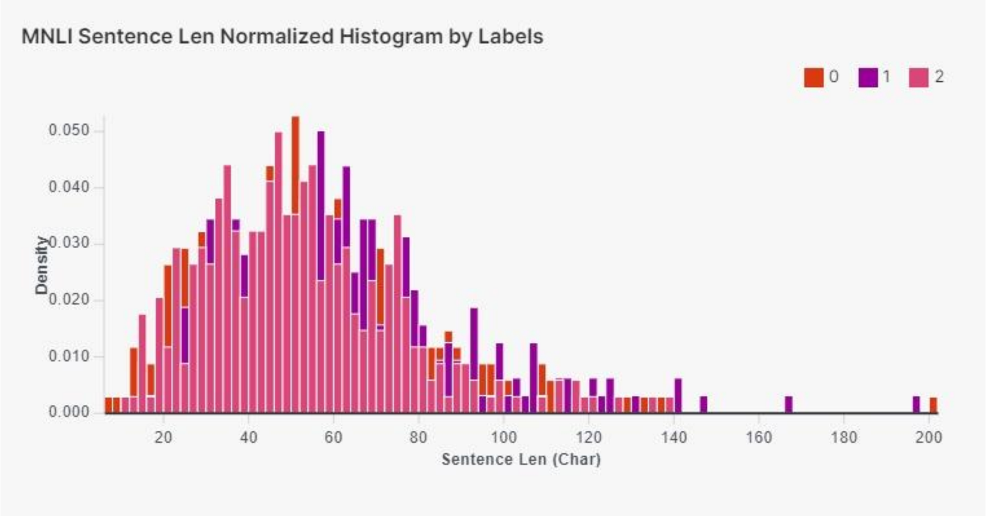

Figure A.9: SNLI Sentence Length Normalized Histogram

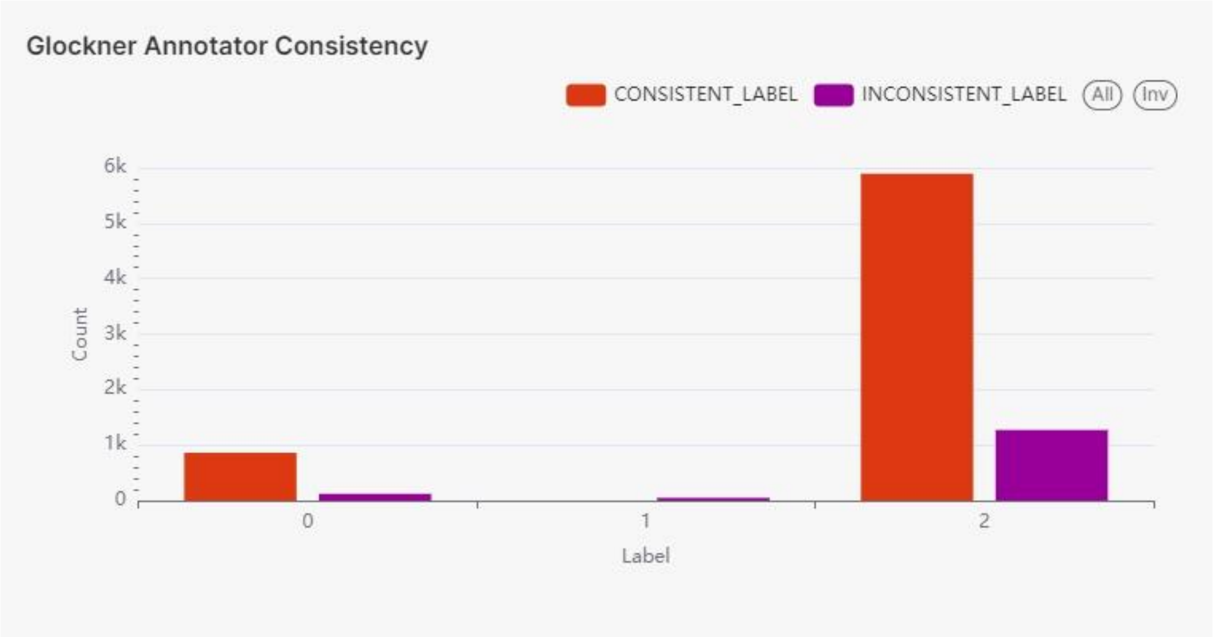

Figure A.10: Glockner Annotator Consistency

Error Type	Premise	Hypothesis	Error Category
Label: 0, Predicted: 1	The greyhounds are running quickly in this race.	The big greyhounds are running quickly in this race.	Model Error – Single Word
Label: 2, Predicted: 1	A man cooks a large amount of shellfish in a wok outdoors.	the water looks cold to swim in.	Incorrect Label
Label: 1, Predicted: 0	An elderly man sleeping in a outdoor chair while another woman sits next to him.	Two people are outside.	Incorrect Label - Specificity
Label: 1, Predicted: 0	A group of people are standing on a platform and waiting for the subway train.	People gathering in a group	Incorrect Label - Specificity
Label: 1, Predicted: 0	An older gentleman speaking at a podium.	A man giving a speech	Incorrect Label - Specificity
Label: 0, Predicted: 2	A man in a black long-sleeved tee-shirt is walking down the street with a cane.	A man walking with a caine.	Misspelled

Table A.11. Specific SNLI Type Categorization Examples

(0: Entailment, 1: Neutral, 2: Contradiction)

Error Type	Premise	Hypothesis
Label: 1, Predicted: 0	Enter the realm of shopping malls, where everything you're looking for is available without moving your car.	Everything can be found inside a shopping mall.
Label: 1, Predicted: 0	Then he is very sure.	He is very sure of himself.
Label: 1, Predicted: 0	Others love to see it in the middle of the heaviest monsoon, its marble translucent, its image blurred in the rain-stippled water channels of its gardens.	It is especially beautiful during the monsoon season.
Label: 1, Predicted: 0	High Crimes is painfully shoddy, even for a book rushed to press.	Books that are rushed to press are usually shoddy.
Label: 1, Predicted: 0	It was other-worldly.	It was a spiritual event.

Table A12. MNLI Type Categorization – Incorrect Labels

(0: Entailment, 1: Neutral, 2: Contradiction)