A consensus-powered approach to NLI corpus cleaning

December 6, 2021

Abstract

Corpus construction for the teaching of NLI models is an expensive process that necessitates compromises during data collection. Well-known corpora, such as the Stanford Natural Language Inference (SNLI) corpus, uses single annotator judgments on the majority of its premise-hypothesis pairs. An approach to identify errant judgments is proposed and evaluated using the ELECTRA-small-discriminator model. While accuracy does not seem to increase at face value, deeper analysis of model prediction scores show an increased ability for models to discern incorrect gold labels.

1 Background

Language is fundamentally a consensus driven, communicative model for information exchange, where meaning is prescribed not by a single speaker, but by multiple. To create a high-quality, annotated corpus for the purposes of teaching natural language inference (NLI), annotations must either be derived from an objective set of rules or by the consensus of qualified judges.

The use of singly annotated examples, while enabling less expensive corpus construction, allows the possibility of incorrect gold labels being included in a training corpus, whereby a model may learn incorrect associations and negatively impact its predictive capabilities.

A corpus commonly used to train models for NLI is the SNLI corpus[1] which constitutes a collection of 570k premise-hypothesis pairs written by 2500 workers. 90% of the corpus is annotated by a single worker, while 10% of the corpus underwent a further validation phase using an additional four workers. 58% of the validated examples show unanimous agreement with the original worker's judgment, while 98% show majority agreement with the original worker.

Though the results of the SNLI validation phase suggest high quality of the non-validated portion of the corpus, it is by no means a guarantee. Several examples highlight the challenges of relying on possibly singly-annotated pairs from the SNLI corpus.

In one example, the premise "Two

monks are visiting a big city." followed by the hypothesis "The monks are dressed in their robes enjoying the big city." is annotated as Entailment. Following the rules of logical entailment described by MacCartney et al.[2] and used in the creation of the SNLI corpus[1], this example should be annotated as Neutral, as it is not definitely true that the monks are wearing their robes, nor that they are enjoying the city.

In another example, the premise "A man reading the paper at a cafe" followed by the hypothesis "A man reading the paper while drinking cafe" is annotated as Entailment, when this pair should be again labeled as Neutral, as the drinking of coffee might be true, but is not definitely true.

This type of error is best described as a worker performing textual entailment rather than logical entailment. The differences in label assignment between logical and textual entailment is described in the following table.

Table 1: Annotation given an inference

Inference	**Logical**	**Textual**
True	0	0
Maybe True	1	0
Unknowable	-	1
Maybe False	-	2
False	2	2

where 0 = entails, 1 = neutral, and 2 = contradicts

Given that the purpose of the SNLI corpus is to provide data for logical entailment tasks, workers who make textual entailments, contribute incorrect gold labels if textual entailment does not match the logical entailment. If incorrectly annotated examples are allowed into the training corpus, then trained models may learn spurious correlations that ultimately impact its performance in NLI tasks.

To determine the presence of logical-textual entailment mismatches, ELECTRA-small was trained for three epochs on the SNLI training set and evaluated against the test set. This trained model is referred to as "Model 0". Mismatches made by Model 0 were split into categories depending on the type of label mismatch.

This results in six categories of mismatches, i.e. the label predicted was 1, the gold label was 2. Mismatch types are abbreviated PXGY, where X is the value of the label predicted by the trained model, and Y is the value of the gold label given by the annotator(s). The number of each mismatch type is shown in Fig. 1.

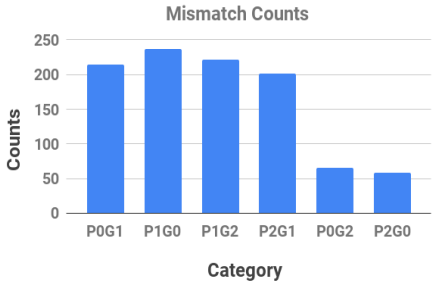

Fig. 1: Number of examples in each mismatch category

For this study, the mismatch type P1G0 was of primary focus. 237 P1G0 mismatches were examined and re-annotated to investigate possible metrics for annotator error identification. The vast majority of P1G0 mismatches, where the annotator was incorrect, was a result of the annotator

assigning Entails to a hypothesis with more specificity than the premise.

The absolute difference between the model scores of the predicted and gold labels was identified as an interesting metric to judge annotator accuracy. The magnitude of this difference can be taken to represent the model's confidence in its prediction versus the annotator gold label and will be referred to as "model confidence" henceforth.

Sorting P1G0 mismatches by model confidence reveals a potential strategy for identifying and cleaning corpora used for NLI. Fig. 2 shows the percentage of incorrectly annotated P1G0 mismatches for a given model confidence interval.

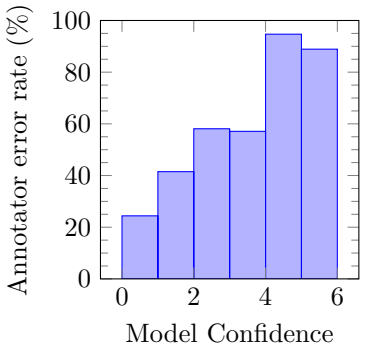

Fig. 2: Fraction of annotator mistakes versus model confidence in its own prediction

As seen in Fig. 2, there is a marked increase in the percentage of annotator errors with increasing model confidence. Model 0, having trained on the SNLI train set, represents a student who has learned from a collective of 2500 teachers. For very high model confidences, given that the collective is far larger than any single annotator, it is highly likely that the single annotator is mistaken. Above a model confidence of 2.0, approximately 58% of P1G0 mismatches were incorrectly annotated by the original author, and above a model confidence of 4.0, 95% of examples were incorrectly annotated.

One example where the model confidence was greater than 4.0 is the following premise-hypothesis pair: "Two wrestlers jump in a ring while an official watches." followed by "Two wrestlers were just tagged in on a tag team match.". This example was incorrectly labeled Entails by the original annotator, whereas Model 0 correctly predicted Neutral.

It is hypothesized that by using a trained model to identify possible, incorrectly annotated examples and prune said examples from the training corpus, a resultant model trained on the pruned corpus may perform better on the task of logical entailment since the new model has not learned on incorrect examples. The methodology and results of this approach are now discussed.

2 Methods

ELECTRA-small was trained for three epochs on the SNLI training corpus to create Model 0. Model 0 was evaluated against the SNLI test set to collect baseline performance. Mismatches made by the trained model were recorded according to the type of mismatch, i.e. P0G1, P1G2, etc.

237 P1G0 mismatches generated by Model 0 were newly annotated and model confidence for each mismatch was determined by taking the absolute difference in prediction scores of

the model's predicted label and the original annotator's gold label. The new annotations were only used for the purposes of classifying which examples were incorrectly labeled by the original annotator, and not used in any subsequent model training.

A threshold value for model confidence was determined and used to prune training examples from the SNLI training set. The model confidence threshold was determined as the value above which incorrect logical entailments by SNLI annotators comprise more than 50% of the mismatches in a given score range. For this study, a model confidence threshold of 2.0 was used.

To perform the pruning, the SNLI training set was divided into ten folds. A pretrained ELECTRA-small model was trained for three epochs on nine training folds and evaluated against one test fold. Ten models were created using different permutations of training and test folds for full coverage of the SNLI training set. These models are referred to as "Model F#" where # represents the number of the test fold used for evaluation.

Using the model confidence threshold and evaluating test folds with respective Model F#s, P1G0 mismatches with model confidence scores above the threshold were identified and pruned from each fold. The pruned folds were merged to form a new training corpus, upon which ELECTRA-small was trained for three epochs and evaluated against the original SNLI test set. This trained ELECTRA-small model is referred to as "Model 1". Model 1 accuracy was compared against Model 0 accuracy using the SNLI test set. Model confidence scores from Model 0 and Model 1 were also examined to measure the effects of using a pruned corpus.

3 Results

Baseline metrics for Model 0 and Model F#s are shown in Table 2. It can be seen that Model F#s have high accuracy numbers which indicates that despite training on 90% the size of the SNLI training set, the Model F# series still have high predictive capability.

Table 2: Baseline metrics for ELECTRA-small

Name	**Test Set**	**Accuracy**
Model 0	SNLI Test	89.86
Model F0	Fold 0	87.84
Model F1	Fold 1	90.27
Model F2	Fold 2	92.46
Model F3	Fold 3	94.25
Model F4	Fold 4	95.50
Model F5	Fold 5	87.79
Model F6	Fold 6	90.59
Model F7	Fold 7	92.32
Model F8	Fold 8	94.24
Model F9	Fold 9	95.51

Of the 237 P1G0 mismatches generated by Model 0, most errors made by the original annotators were due to textual entailments where a hypothesis that is more specific than the premise was deemed Entailment. In one such example, the premise "A skateboarder does a trick at a skate park." followed by the hypothesis "The skateboarder is performing a heelie kick flip." is incorrectly labeled Entailment. Examples containing typos were deemed as annotator errors for the purposes of determining a model confidence threshold.

P1G0 mismatches with model confidence above the chosen model confidence threshold of 2.0 were tallied from the evaluation of each of the ten test folds of the SNLI training set with their respective Model F#. A total of 4689 examples were identified and automatically pruned from the SNLI training set. This number represents less than 1% of the size of the original SNLI training set.

After Model 1 was trained on the pruned corpus, performance was measured against the SNLI test set. The accuracy of Model 0 and Model 1 is compared in Table 3.

Table 3: Accuracy comparison between models

Model Name	**Accuracy**
Model 0	89.87
Model 1	89.65

From the accuracy results above, it seems that there is no improvement in performance by pruning possibly incorrectly annotated examples. However, upon close examination of P1G0 mismatches generated by Model 0 and Model 1, several interesting features appear.

For further evaluation of the pruning strategy, P1G0 examples can be divided into three types: Type 1 consists of P1G0 mismatches generated by Model 0 alone, Type 2 consists of P1G0 mismatches generated by Model 1 alone, Type 3 consists of P1G0 mismatches generated by both Model 0 and Model 1.

Looking at a Type 1 example, the premise "A man drives a motorcycle, with a woman in an orange sari and a young girl in a blue dress riding in the backseat." followed by the hypothesis "A woman in an orange sari is a passenger on a motorcycle." was incorrectly predicted Neutral by Model 0. The model confidence in this example was 1.48 for Model 0. Model 1 correctly labels this example as Entailment and thus does not generate this P1G0 example. There are 47 Type 1 examples, of which 85% were true errors made by Model 0 that were correctly labeled by Model 1.

Looking at a Type 2 example, the premise "Two girls play with origami" and the hypothesis "The girls hold paper" is a new P1G0 mismatch generated by Model 1, which Model 0 correctly labeled. The model confidence for this example was 2.18. 46 Type 2 examples were recorded.

Of these 46 mismatches, 71% had a model confidence of 1.0 or below, suggesting the effect of pruning the original SNLI training set perturbed examples whose prediction scores between the predicted and gold label were very close for Model 0.

Since the number of Type 1 and Type 2 examples are roughly equal, this may explain why overall accuracy has not improved between Model 0 and Model 1, along with the fact that only one of the mismatch categories was investigated for this study.

Lastly, Type 3 examples were closely examined due to the ability to directly measure the impact of pruning the SNLI training set. 186 Type 3 P1G0 mismatches were recorded. One such example is the following premise-hypothesis pair: "Four people in a kitchen" followed by "Four people in a kitchen cooking" had a Model 0 confidence of 1.40 and a Model 1 confidence

of 3.73. This example was incorrectly labeled Entailment by the original annotator.

Such an example suggests that after pruning P1G0 examples with incorrect gold labels, Model 1 has become more confident in its prediction of incorrectly annotated P1G0 pairs. Fig. 2 shows more clearly the effect of pruning on the classification strength of the two models. Change in confidence was calculated by taking the difference between Model 1 and Model 0's confidence values.

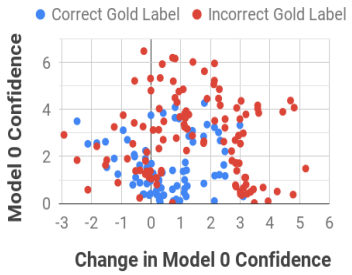

Fig. 2: Effect of pruning on model confidence

As seen in Fig. 2, the vast majority of premise-hypothesis pairs above a Model 0 confidence of 4.0 were incorrectly annotated by the SNLI workers. This trend was previously seen in Fig. 1. Of note however, is that for examples where confidence scores increased by more than 2.5 from Model 0 to Model 1, the vast majority are also incorrectly annotated by the original workers.

Many of these examples were previously below the model confidence threshold used for pruning and creating the training corpus for Model 1. This suggests that by removing the ability of incorrect annotations to influence the learning of a model, the model becomes more adept in making correct predictions.

It is noted that there are some examples with correct gold labels, where Model 1 is more confident in its own incorrect prediction. It is hypothesized that by using a model confidence threshold of 2.0 for pruning, correctly annotated examples were removed from the training set which may lead to increased model confidence in an incorrect prediction, along with the appearance of more Type 2 P1G0 mismatches.

The following premise-hypothesis pair is an example where Model 1 has become more confident in its incorrect prediction. The premise "A customer places his order at Quiznos while another customer looks on." followed by "Someone orders food at Quiznos." is incorrectly predicted as Neutral by Model 0 and Model 1. Model 0 confidence is 0.85 while Model 1 confidence is 2.62.

Using a more strict model confidence threshold of 4.0 for pruning to remove examples much more likely to be incorrectly annotated, around 95% as seen in Fig. 1, may provide the benefits of increased classification strength of incorrect annotations, while reducing Type 2 P1G0 mismatches with correct gold labels.

Future Approaches Potential future strategies in cleaning the SNLI corpus and refining the pruning method are now discussed. Given the harmful effect of including incorrect annotations on the ability of a trained model to perform NLI tasks, annotators identified by a trained model to have contributed incorrect gold labels

should be subjected to a review of all provided examples, up to and including removal of their examples from the training corpus.

As the SNLI corpus does not provide annotator IDs with examples, the task of removing examples belonging to an individual annotator was not possible, but nonetheless, remains a viable tool for corpus cleaning, provided the decision to remove an annotator's contributions is determined solely on the annotator's ability to perform logical entailment, and not any other bias.

As previously mentioned, the use of a model confidence threshold of 2.0 for pruning, where approximately 58% of examples were incorrectly annotated, may be too low. The use of a more strict threshold of 4.0, where approximately 95% of examples were incorrectly annotated, may reduce Type 2 P0G1 mismatches while maintaining the improvements in classification of incorrectly annotated examples. This may lead to an improvement in overall accuracy of a trained model to perform NLI compared to the baseline Model 0.

While this study only investigated P1G0 mismatches, similar pruning methods may be employed in other mismatch categories, such as P1G2, P2G1, etc. However, any pruning strategy must take into account the reason why an example may be incorrectly annotated. In the case of P1G0 mismatches, nearly all incorrectly annotated examples were due to textual entailment errors, whereas the SNLI task called for logical entailment.

A cursory glance at the P2G0 mismatch category shows why pruning may not be a good strategy for all categories. The example "A woman hanging the laundry outside" followed by "A woman is putting her clothes out to dry." is predicted by Model 0 as "contradicts" versus the annotator's correct judgment of "entails". Another example "Two girls and a guy are involved in a pie eating contest" followed by "Three people are stuffing their faces." is incorrectly predicted as "contradicts". Both examples rely on idiomatic expressions which may not be present in large numbers in a genre-unrestricted corpus such as SNLI. Rather than pruning, augmentation of the training corpus to include more world knowledge would be a more appropriate strategy for dealing with P2G0 mismatches.

4 Summary

A strategy for leveraging corpus size to identify incorrect gold labels has been proposed and relies on training models on subsets of the entire training corpus, to provide predictions on the subsets used for testing.

Investigating and validating examples where the baseline model (Model 0) predicted "Neutral" and the original annotator labeled "Entails", showed that above a certain difference in model prediction scores for each label, the model was more likely to be correct despite not matching the annotator gold label.

This difference was used as a threshold to prune examples of the mismatch type P1G0 from each of the test subsets to create a pruned training corpus which was used to train a new ELECTRA-small model (Model 1). While there is no increase in objective accuracy when Model 1 is evalu-

ated against the SNLI test set, analysis of the mismatches generated by Model 1 showed an increased ability to identify incorrectly annotated examples in the SNLI test set.

By learning on the judgments of a large number of annotators, a model's confidence in its prediction can be equated to the collective confidence of the corpus annotators. For examples where the model has very high confidence in its prediction, despite a mismatch with the annotator's gold label, giving deference to the model to prune appears to increase its ability to discover more incorrectly labeled examples.

Since the learning of any student is directly impacted by the quality of the teachings they study, any tool to improve the quality of training corpora, especially for large corpora such as SNLI, is valuable. The general applicability of the pruning stratagem remains unknown, and methods to identify incorrect annotations from different mismatch categories will likely involve the use of multiple strategies. Preliminary results of the P1G0 pruning strategy however, are very encouraging towards the development of highly accurate classifiers for incorrectly labeled logical entailment examples.

References

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

[2] Bill MacCartney, Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*