Evaluating Baseline NLI models on Two Distinct Modes of SNLI Dataset Segmentation

Griffin Tarpenning `gritarp@stanford.edu`

Abstract

The development of increasingly complex natural language inference models have seen dramatic performance gains on common datasets in recent years. However, these models also have been shown to perform poorly when tested on slightly different NLI tasks and datasets, suggesting that perhaps the models aren't truly gaining as much semantic understanding as previously thought. To better understand this issue, we explore different categorization and segmentation techniques on the SNLI dataset, comparing the baseline model performance on these different sections. This paper implements semi supervised learning with KMeans as a new way of segmenting the SNLI dataset, and while KMeans clusters did prove to have common similarities, this form of segmentation did not prove more successful for error analysis than previous methods. However, a strong foundation is laid for more extensive semi-supervised clustering attempts in the future.

1 Introduction

Using machine and deep learning on text data has been a hot area of analysis for decades due in part by its canonical simplicity on the small scale and infinite complexity on the large scale. One of the premier tasks within this genre of textual analysis is Natural Language Inference (NLI), a classification task where a machine predicts whether a given premise and hypothesis pair are an entailment, neutral, or contradiction. Although NLI has been used as a metric for evaluating the success of Natural Language Processing models for decades, with the publishing of large robust datasets like the Stanford Natural Language Inference dataset (Bowman et al., 2015) and later the MultiNLI dataset (an identically formatted dataset to SNLI but with wider variation of relationships) in the last five years, there has been a renewed focus and seemingly fantastic advancement in performance.

However, with the ease of access these datasets provide, comes an arms race of data scientists intent on besting the leaderboard; while the spirit of developing models that learn semantic features of text, understanding, is largely lost. Recently, models with state of the art performance have been shown to produce lackluster or even disastrous results when tested on slightly different data (Glockner et al. 2018, Naik et al. 2018). Moreover, when models are cross-trained and tested models boasting strong 85+% performances on the SNLI leaderboard freefall up to 30% (Talman and Chatzikyriakidis 2018). These findings, which will be discussed in the following section, point to the need for a more robust framework for testing models, and more importantly, a better understanding of the semantic features current models are able to learn.

Amidst the barrage of newer and larger models, one avenue that is increasingly being investigated is changing the testing data to provide more transparency on the limitations of NLI models. Authors are harnessing the many different pre-categorized genres in the MultiNLI dataset (Geva et al. 2019), creating new datasets focused on breadth of reasoning rather than genre (Poliak et al, 2018), and making adjustments/modifications to the existing SNLI and MultiNLI datasets (Glockner et al. 2018, Naik et al. 2018). The following section discusses some of the most interesting scholarship falling into the latter category.

2 Related Work

One recent work with dramatic findings comes from Talman and Chatzikyriakidis, who demonstrate that many NLI models are highly specific to the dataset they are trained on. Rather than truly

understanding language and modeling inference, the top models are essentially gaming the training/test data by finding patterns that don't actually translate to semantic meaning. They tested their own implementation of a Hierarchical BiLSTM with max pooling, the now ubiquitous ESIM model developed by Chen et al. 2017, the KIM model (Chen et al., 2018) , and an ELMo modified ESIM model (Peters et al., 2018). When trained on MultiNLI and tested on SNLI, the models saw performance drops of 10% on average. Although this drop is not as large as some of their results, it show that the models generalize poorly, especially when considering a performance drop even half that size on the SNLI leaderboard would be massive. The authors suggest that NLI is genre dependent to explain performance drops, and rather than address the model construction, their analysis focuses on how datasets could be constructed in superior ways to better promote generalization.

While cross training and testing yields somewhat interesting results, the large number of changing variables with different datasets highlight the need for a more nuanced approach. Multiple recent papers have focused on the creation of testing sets that have an entirely shared vocabulary with SNLI training data, with small key differences (Glockner et al. 2018, Naik et al. 2018). Maintaining the same vocabulary as the training set is a crucial advantage over dataset swapping, as natural models can be tested alongside models that have access to external knowledge like the KIM model (Chen et al., 2018). This adversarial sample generation proves catastrophic for some models, with Glockner et al. reporting SNLI, MultiNLI + SNLI, and SciTail + SNLI trained ESIM model performance dropping nearly 20% on average. However, a big winner was the KIM model, its external knowledge only allowing a 5.1% slide from the normal SNLI test set. Although significant, the reliance on swapping words for antonyms limits their findings, especially when considering the plethora of diverse meanings each word in English has. Perhaps adversarial samples contained words that, while present in the training set, hadn't been used with the same exact meaning and thus were impossible to understand. Another shortcoming is the size of the test set, which is a mere 8,000 samples, less than 1% of the size of SNLI.

A more complete suite of adversarial samples comes from (Naik et al. 2018), where based on a manual analysis of misclassified samples, a completely new test set was generated through manual and programmatic effort. They found that the following prominent features were the source of the majority of errors in their classifier: antonyms, numerical reasoning, word overlap, strong negation, length mismatch. The baseline BiLSTM model, while achieving 70% on the MultiNLI test set, performs in the range: 49-65% accuracy on the adversarial test categories. This analysis is significantly more extensive than all previously mentioned, but fundamentally they are relying on human taggers to identify which sections, and for what reasons, models performed poorly. That concession becomes the basis for this paper.

3 Segmenting the SNLI Dataset

3.1 Overview

The SNLI dataset is made up of 570,000 premise, hypothesis pairs to be classified between entailment, neutral, or contradiction. For our analysis the SNLI dataset is segmented in two fundamentally different ways. The first follows in a similar vein as Naik et al. 2018, by tagging sentences with features that might be especially easy or challenging for a model to predict. The second way that SNLI data is segmented is by semi-supervised KMeans clustering. Unlike the papers previously discussed, no new data is generated for either method, rather test data is partitioned directly from the SNLI and then sorted into categories. 90% of the SNLI dataset is exclusively for training, while the other 10% is partitioned for testing. Additionally, while the KMeans segmentation ensures that 100% of the testing data is accounted for split into n clusters, the categorical analysis merely tags premise-hypothesis pairs of interest. Moreover, while it is impossible for an example to appear in multiple KMeans clusters, it is possible that a single premise-hypothesis pair could appear in multiple different categories, like sentiment and average word length for example.

3.2 Hand Chosen Categories

The following features serve as flags for potentially more difficult premise-hypothesis pairs. For each feature, a high-difference and a low-difference category is created. For example, a category exists for premise-hypothesis pairs with a large difference in average worth length, as well as

a different category for premise-hypothesis pairs with a small difference in average worth length

- High/Low average word length difference between premise and hypothesis
- High/Low difference in total number of words
- High/Low percentage of words that overlap between premise and hypothesis
- High/Low difference in average sentiment between premise and hypothesis

3.3 KMeans Segmentation

For the KMeans clustering, test data is first separated from the SNLI dataset. The data is then fed into a KMeans classifier, using TF-IDF vector representations, and segmented into different groups. By the semi-supervised nature of this classifier, it is impossible to predict how many different clusters would lead to an optimal split of the data, or even what features will be prominent. Thus, KMeans was run 6 different times, each with a different number of clusters in the following: [2, 3, 4, 5, 6, 10]. This allowed for a substantial spread of different features defining the characters as seen in the examples below.

4 Results

4.1 Models

For evaluation, this paper used five distinct baseline models. The non-neural net models included: Bag of words, Logistic Regression, and Random Forest Classifier. The fourth model chosen was a simple BiLSTM. The implementation of the BiLSTM follows from the implementation from Chen et al. 2017. Premises and Hypothesis are encoded with separate recurrent neural networks (RNNs) from a tree representation structure, concatenated, and then classified at the top with a bidirectional RNN. A hidden dimension of 200 was used. Finally, the last model included was a variant of a simple sentence encoding classifier. Premise-hypothesis pairs were encoded word for word with pre-trained GloVe embeddings, concatenated, encoded through a single direction RNN, and then finally sent through two fully connected layers with a dropout.

BOW	LogReg	RanFor	BiLSTM	RNN
65.7%	72.1%	70.9%	74.4%	68.2%

Table 1: Baseline SNLI Results.

```
Category *highSentimentDif*: 628
[{'hypo': 'an older gentleman is enjoying his orange juice at a new caf
e.', 'prem': 'an elderly man is drinking orange juice at a cafe.'}, {'hyp
o': 'cheerleaders cheer for the football team.', 'prem': 'cheerleaders ar
e on the field cheering.'}, {'hypo': 'a small human in the water.', 'pre
m': 'a child is paddleboarding outdoors'}, {'hypo': 'three people went to
the planetarium.', 'prem': 'man in white t-shirt and white beard plays el
ectric guitar with a fiddler in a band.'}, {'hypo': 'a man is riding a bi
cycle up a graffiti covered wall.', 'prem': 'a man on a small bicycle per
forming a trick on a wall covered with graffiti.'}, {'hypo': 'an elderly
alien and another younger alien', 'prem': 'an elderly woman and another y
ounger woman greeting each other.'}, {'hypo': 'an older couple sit and wa
tch a young woman dance.', 'prem': 'a happy older couple share a dance wh
ile a younger dancing woman looks on.'}, {'hypo': 'a lad fires a weapo
n.', 'prem': 'a boy at a gun range aims and shoots.'}, {'hypo': 'a young
boy holds a crying baby.', 'prem': 'a young boy holds an infant which app
ears to be in distraught.'}, {'hypo': 'a boy learns to play football.',
'prem': 'a young child wearing a yellow wetsuit rides on a surfboard.'}]
```

Figure 1: 10 random examples from the high sentiment difference category.

4.2 Baseline SNLI Results

The baseline results are relatively consistent with current baselines, but trend lower. This could be because of the slight reduction in training data (from 100% to 90%), or a lack of competitive hyper-tuning. Either way, their purpose is to provide a canvas to see how different training data sections affect accuracy, so the overall performance is not imperative to significant findings. The BiLSTM performs the best, which is to be expected considering its complexity and widespread adoption from NLI tasks. However, it is worth noting that the Logistic Regression is quite close with only a 2.3% decrease in performance. Because of their vastly different structure yet nearly comparable performance, looking to these two models will be important in further analysis sections.

4.3 Hand Selected Categories

Here, the hand selected categories show consistent but marginal changes on the performance of the models. We see some expected results, as well as some unexpected. Because of the novelty of the sentiment category, and importance of understanding how examples in each category differ, Figure: 1 shows a few examples randomly selected from the sentiment category.

4.4 KMeans Clustering

The clustering of the SNLI dataset by KMeans proved somewhat lackluster in terms of performance differentiations by cluster, but not necessarily in whole. The results are reported below, showing each model's performance on each clus-

Category	BOW	LR	RF	BiL	RNN
word len.					
$\geq$ high	0.614	0.696	0.699	0.641	0.733
$\leq$ low	0.658	0.723	0.712	0.682	0.746
# words					
$\geq$ high	0.610	0.671	0.682	0.666	0.704
$\leq$ low	0.668	0.737	0.711	0.683	0.750
overlap					
$\geq$ high	0.682	0.750	0.719	0.692	0.759
$\leq$ low	0.617	0.687	0.698	0.640	0.725
sentiment					
$\geq$ high	0.649	0.729	0.712	0.684	0.741
$\leq$ low	0.657	0.723	0.704	0.678	0.745

Table 2: Categorical Model Results.

tering regime. Nearly all of the model results show performance within the realm of just random variation, however there are a few slight outliers. For example, on the 5 cluster run, cluster 1 caused worse performance over all 5 models by a relatively significant delta. Although the absolute change wasn't much, it shows that clustering can have a statistically significant affect. The Logistic Regression saw the most significant drop, with the other linear models following, and the neural network models seeing a smaller impact. We must look at examples from that individual cluster to see what might have caused the dip in performance. Looking at the 5930 entries in this cluster, we can see the following interesting attributes:

- **Average word length difference** : 0.94
- **Average word overlap %**: 0.4

These features, explained earlier in the categorical analysis section, are higher than the average over the SNLI testing set used (0.89 and 0.35 respectively), which might explain the drop in performance. What features KMeans used to determine this cluster is hard to tell without digging into a few examples. So, the following is a random sample of 5 hypothesis from the cluster: 'a few people in a restaurant setting...", "people waiting at a light on bikes", "the people are on skateboards...", "people on bicycles waiting at an intersection", "there are people just getting on a train.". Immediately, we notice the theme of transportation and time, but this gives us little clues into why this theme might be harder to computer. Perhaps the SNLI training set contained fewer examples pertaining to transportation, or the transportation premise-hypothesis pairs are more challenging to evaluate.

The above analysis into one cluster further highlights the importance of qualitative analysis. Thus, the following lists show randomly chosen example snippets from each cluster in the 4-cluster, and then 10-cluster regimes, ideally showing the progression from broader themes to more specific.

4 Cluster Segmentation

- **cluster 1**: 'A blond man is drinking...' 'an elderly man is drinking...' 'a man with a red shirt is watching...' 'a man is sleeping under a bench.' 'a white horse is pulling a cart while a man stands and watches'
- **cluster 2**: 'two blond women are hugging...' 'two women who just had lunch...' 'a lady in a black and white striped shirt...' 'a lady is about to propose...' 'two ladies are reading through binders' 'two dogs are sleeping while a third eats...'
- **cluster 3**: 'a few people in a restaurant setting...' 'people waiting at a light on bikes' 'the people are on skateboards...' 'people on bicycles waiting at an intersection' 'there are people just getting on a train.'
- **cluster 4**: 'woman in white in forground...' 'a woman ordering pizza' 'a woman wearing all white...' 'a woman in a green jacket and hood over her head...' 'a woman is walking across the street...'

10 Cluster Segmentation

- **cluster 1**: 'the man is working on the computer', 'A man is performing for money', 'The man plays guitar', 'The young man has glasses on his face', 'An old man is standing by a building in downtown'
- **cluster 2**: 'A skier is away from the rail', 'The skier was on the edge of the ramp', 'The bikers are riding Harley's', 'A barber is inside his shop standing in the...', 'Firemen walking outside'
- **cluster 3**: 'The people are sitting at desks in school', 'People wearing pink and purple get ready for...', 'The speaker is the people's boss', 'The scene is in the evening', 'A group sits together outside and talk over drinks'

Cluster #	BOW	LR	RF	BiL	RNN
2					
c1: 14263	0.654	0.720	0.709	0.688	0.742
c2: 42737	0.658	0.721	0.709	0.683	0.745
3					
c1: 36233	0.676	0.720	0.710	0.682	0.744
c2: 12878	0.654	0.720	0.709	0.688	0.742
c3: 7889	0.638	0.723	0.708	0.688	0.745
4					
c1: 29446	0.667	0.719	0.707	0.682	0.744
c2: 12596	0.654	0.720	0.709	0.688	0.742
c3: 7598	0.658	0.723	0.708	0.688	0.745
c4: 7360	0.662	0.729	0.711	0.690	0.745
5					
c1: 5930	0.630	0.704	0.701	0.679	0.739
c2: 12252	0.654	0.720	0.709	0.688	0.742
c3: 7250	0.662	0.729	0.711	0.690	0.745
c4: 7517	0.658	0.723	0.708	0.688	0.745
c5: 24051	0.656	0.717	0.708	0.681	0.743
6					
c1: 5226	0.650	0.707	0.704	0.682	0.740
c2: 7278	0.672	0.729	0.711	0.890	0.745
c3: 6635	0.658	0.723	0.708	0.688	0.745
c4: 5315	0.636	0.708	0.705	0.684	0.744
c5: 11682	0.652	0.720	0.709	0.689	0.742
c6: 20864	0.655	0.718	0.708	0.681	0.744
10					
c1: 1729	0.687	0.730	0.707	0.690	0.746
c2: 19595	0.639	0.718	0.708	0.681	0.744
c3: 4426	0.659	0.708	0.705	0.684	0.744
c4: 4457	0.650	0.709	0.704	0.685	0.744
c5: 9146	0.653	0.722	0.709	0.689	0.743
c6: 1339	0.689	0.732	0.712	0.690	0.746
c7: 5438	0.662	0.710	0.709	0.690	0.745
c8: 4187	0.658	0.723	0.708	0.688	0.745
c9: 1018	0.633	0.701	0.698	0.679	0.740
c10: 3191	0.658	0.723	0.708	0.688	0.742

Table 3: KMeans Model Results.

- **cluster 4**: 'A lady is kneeling before the priest at church', 'A lady wearing a cover', 'A young girl wears sandals and walks on hopsco...', 'A lady wearing a yellow top is standing on the...', 'Two women are looking for something'

5 Discussion

When considering splitting the SNLI dataset, or any NLI dataset into subsections for targeting testing/training, some categories might be more obvious choices than others. For example, focusing on examples with strong negation might provide insight into where a model can be improved, or where more training data is necessary. These types of categories clearly have immediate use. What we saw from KMeans was a way of breaking down the dataset in a different manner. Because of KMeans reliance on TF-IDF, clusters were largely broken down by different themes, with potentially surprising consistency. However, there are little to no discernable differences between all models when tested on these different clusters.

Why might this be the case? Although genre analysis of SNLI and especially MultiNLI is currently a hot topic in the NLI community, the cluster differentiation from KMeans was not quite the same. The most common words in premise-hypothesis pairs appear prominently as the differentiating features for each cluster, commonly "man," "woman," "children," among others. That being said, there are clearly some themes being picked up by the cluster analysis. For example in cluster 4 in the 10-cluster example list we can see that the core theme is not just surrounding one word, but rather being female. Keywords in this cluster do not just include "woman," but also "lady" and "girl," which strong suggests that the clusters are at least beginning to make basic work associations. All of the clusters had some obviously theme, but the translation from this association to performance gains on the SNLI is less clear.

I believe that the idea to use semi-supervised learning methods in NLI data processing could be a very useful tool for improving generalization of models. Although the results from this analysis are lackluster, there is enough positive feedback through the variations in performance by cluster and qualitative data to validate the need for future research. The impact of having an algorithm automatically filter the dataset into groups, and then cross-train models on these groups, could be quite large. In conclusion, the results shows a promise that semi-supervised learning methods can potentially be useful and accurate at segmenting existing datasets for better error analysis and testing. Unfortunately, there just weren't enough distinct categories for obvious compelling results to come to fruition.

6 Future Work

The next step for research into how KMeans, and other semi-supervised segmentation algorithms, compares to manual breakdown of categories in the SNLI dataset is to experiment with categories as training data rather than testing data. Instead of using the clusters and categories as testing data for models trained on the remaining 90% of the SNLI, segment 90% of the dataset and train models based on these categories. There are a few clear ways of reaching interesting conclusions with this flipped method. The first is simply to see if models trained in one section perform better on other samples from that section. Moreover, a more compelling extrapolation from there is to use these categories to more decisively create batches in the training process. Experimentation surrounding the inclusion and exclusion of certain categories at different moments in the training process could lead to interesting performance differences. Additionally, it might prove interesting to train each category one at a time, plotting segmented test performance as more categories are fed through the network. This might give further insight into how each category is processed and internalized by the models, as well as how differently examples from each category appear to a given model.

8 References

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In ACL.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In Proceedings of ACL.

Aarne Talman, Stergios Chatzikyriakidis. 2018. WIP.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In Proceedings of ACL.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Penstein Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In Proc. of COLING.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proc. of EMNLP.

Mor Geva, Nadav Schor, Meishar Meiri. 2019. Evaluating Domain Adversarial Neural Networks on Multi-Genre Natural Language Inference. Tel-Aviv University. https://www.cs.tau.ac.il/ joberant/teaching/nlp_spring_2019/past_projects/domain_adversarial.pd

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, Benjamin Van Durme. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In Proc. of EMNLP.