Kathleen Xue (kxue@usc)
James Ke (jianxink@usc)
Guohao Li (guohaol@usc)
Yichen Wu (wuyichen@usc)

CSCI 544 Project

CSCI 544
December 5, 2019

Submission

We chose to use Github to submit our code. Here is the repository.

a. Task Description, Problem Formulation

The task we decided to complete for this course project was Natural Language Inference, a task in which the model tries to determine whether a hypothesis is true, false, or undetermined given a premise. The way this problem is formulated is as a classification problem, where each hypothesis is classified as neutral, contradiction, or entailment based on the premise. For instance, the following table gives a good example of the natural language inference task and classifying hypotheses.

Text	Judgment	Hypothesis
A woman walks into the department store.	contradiction	The woman is sleeping.
A girl is eating ice cream.	neutral	Two girls are talking and eating.
A lady meets up with her friends.	contradiction	The lady is by herself.
Two men argue about stocks	entailment	The men are talking.

b. Baseline Algorithm

The baseline algorithm our group decided to use for this task was the Bowman model, which is a simple lexicalized classifier that implements 6 features types:

1. BLEU score of the hypothesis in comparison with the premise (with n-gram length between 1 and 4)
2. Difference in length between hypothesis and premise
3. Amount of overlap between the words used in the hypothesis and the words used in the premise
4. An indicator for every unigram and bigram in the hypothesis
5. Indicators for cross-unigrams between the premise and the hypothesis
6. Indicators for cross-bigrams between the premise and the hypothesis

The reason we chose to use a simple lexicalized classifier is because it actually performs quite well on the SNLI dataset, which is also used for our improved algorithm (in the Bowman paper, the lexicalized classifier was able to correctly classify 78 percent of the test dataset).

c. Approach

In our approach, we decided to take inspiration from the paper DR-BiLSTM: Dependent Reading BiLSTM for Natural Language Inference by Reza Ghaeini et al. The DR-BiLSTM model has four major components: input encoding, attention, inference, and finally classification.

In terms of input encoding, because RNNs are commonly used in this step, the paper chose to utilize a bidirectional LSTM to complete this. The intuition for using the bidirectional LSTM instead of RNN is that this encoding method gives a more informative encoding by taking into account the history of both the premise and the hypothesis, which an RNN is incapable of doing.

In the attention step of the model, the paper utilizes a soft-alignment to connect relevant sub-components between the premise and the hypothesis. We then take these vectors and concatenate them with the difference and element-wise product vectors, before feeding them into a feed-forward neural layer with a ReLU activation function.

During the inference step of the model, the paper chooses to use another bidirectional LSTM to combine the two vectors computed from the attention step. The bidirectional LSTM is similar to the one used in the encoding step, but instead of only using dependent reading information, the inference steps passes in both dependent reading information as well as independent reading information into a max-pooling layer (which allows us to maximize the inferencing ability of the model because we now consider both independent and dependent readings).

In classification, the final step of the model, the paper takes the vectors aggregated from the inference stage and feeds them into a multilayer perceptron classifier with a tanh activation and softmax output layer.

Our group decided to build this model from end to end in order to replicate its results and see for ourselves the improvement that this model has on natural language inference as compared with a simple lexicalized classifier. We used boilerplate preprocessing, embedding, and training code from Github user Aurelien Coet's repository. All the final results of both models can be viewed in our team's project repository.

d. Results Analysis

We found from running both the baseline algorithm and the paper algorithm on training and test datasets that the paper algorithm achieved a significantly higher accuracy than the baseline algorithm. The below table compares the training and validation accuracies of

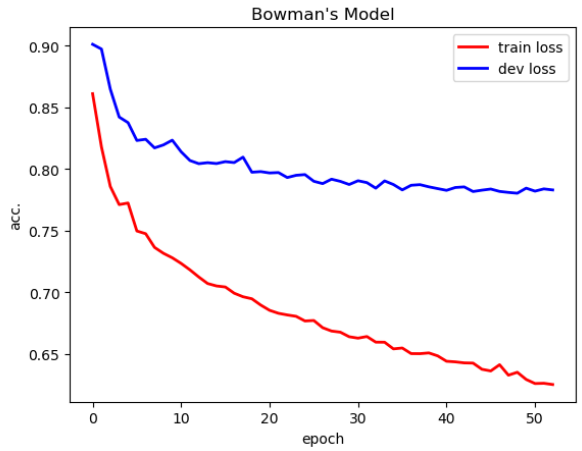
the paper algorithm (with both the preset hyperparameters and with some hyperparameter tuning) with the training and validation accuracies of the baseline algorithm.

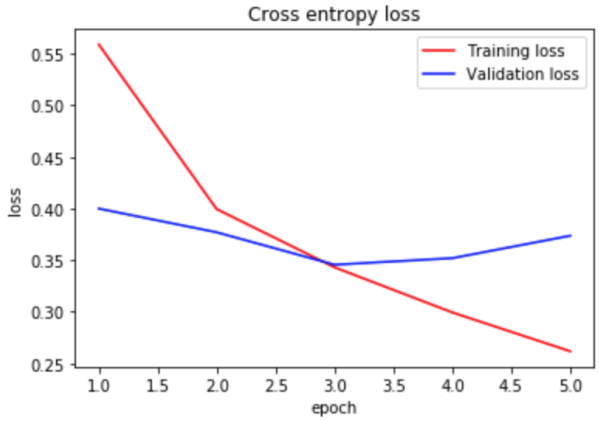
	Training	Validation
Bowman model	68.50%	64.09%
DR-BiLSTM model (preset hyperparameters)*	90.81%	87.18%
DR-BiLSTM model (with batch=64)	89.29%	87.07%

* hidden=450, batch=32, learning rate=0.0004, dropout=0.4

Moreover, looking at the cross-entropy loss over multiple epochs, we can immediately tell that the cross entropy loss in the paper algorithm becomes lower earlier than the cross entropy loss of the baseline algorithm.

As seen below, the cross-entropy loss of the Bowman model (left graph) approaches a minimum of around 0.60 after 50 epochs, but the DR-BiLSTM cross entropy loss (right graph) hits a minimum of around 0.25 after only 5 epochs. This means that the DR-BiLSTM model has a lower cross entropy loss, or a lower divergance between predicted judgments and actual judgments, than the Bowman model.





Beyond basic analyses, our group decided to analyze other more specific features of the model to understand the DR-BiLSTM and Bowman models' "rationales" better. We decided to analyze results based on three types of features: overlap between premise and hypothesis, sentence length, and whether the sentence contains specific types of words (we chose to analyze based on whether the sentence contains negation, quantifier, and belief words). The following table displays the results we obtained by dividing the model results by feature.

In terms of overlap, we defined a premise and hypothesis pair as having high overlap if the sentences share more than 70% of tokens; regular overlap if the sentences share between 30 and 70% tokens; and low overlap if the sentences share fewer than 30% of tokens.

In terms of sentence length, we defined a premise and hypothesis pair as being long if either sentence contains more than 20 tokens; regular if either sentence is between 5 and 20 tokens; and short if either sentence is fewer than 5 tokens.

For negation, we constructed a negation set from an online dictionary: {no, not, none, no one, nobody, nothing, neither, nowhere, never, hardly, scarcely, barely, doesn't, isn't, wasn't, shouldn't, wouldn't, couldn't, won't, can't, don't}. If either of the sentences contains any of the words in the negation set, we defined the premise-hypothesis pair as containing negation. The same was done for quantifiers and beliefs; for quantifiers, we used the set {much, enough, more, most, less, least, no, none, some, any, many, few, several, almost, nearly}; for beliefs, we used the set {know, believe, understand, doubt, think, suppose, recognize, forget, remember, imagine, mean, agree, disagree, deny, promise}.

Surprisingly, the Bowman model did poorly with high-overlap sentences while the DR-BiLSTM model did well on that exact same category. This might be because the Bowman model does not take into account sequence for sentences, and may in fact have missed some key semantic details that the DR-BiLSTM model was able to catch for these high-overlap sentences.

	Bowman	**DR-BiLSTM**
Overlap: High	0.57	0.92
Overlap: Reg	0.64	0.88
Overlap: Low	0.69	0.86
Length: Long	0.68	0.86
Length: Reg	0.67	0.87
Length: Short	0.66	0.90
Contains Negation	0.78	0.88
Contains Quantifier	0.69	0.85
Contains Belief	0.71	0.86

We then looked at accuracy based on label, and found the test accuracy results displayed in the following table.

	Bowman	**DR-BiLSTM**
Entailment Accuracy	0.70	0.89
Neutral Accuracy	0.63	0.81
Contradiction Accuracy	0.68	0.90

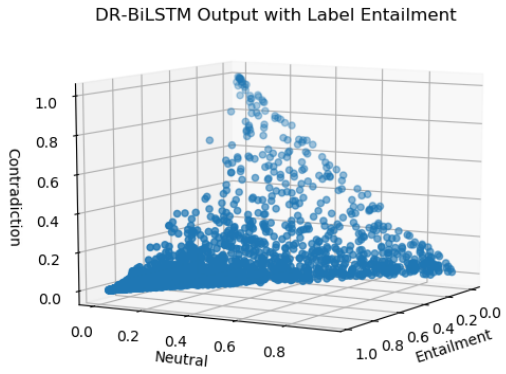
Because both models' neutral statement accuracies were consistently lower than their entailment and contradiction statement accuracies, we wanted to further delve into how the models might be labeling neutral as opposed to entailment and contradiction hypotheses. We took the DR-BiLSTM's predicted softmax probabilities and plotted them on a 3D graph to better visualize the model's perspective of each of the three labels.

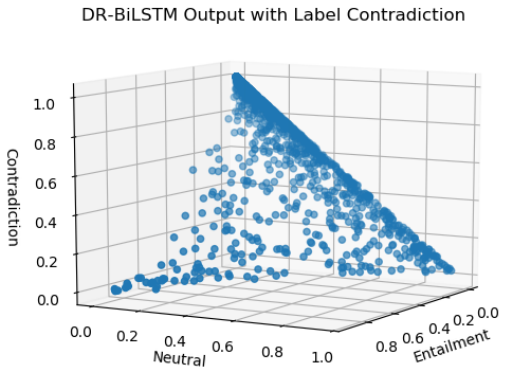
Kathleen Xue (kxue@usc)
James Ke (jianxink@usc)
Guohao Li (guohaol@usc)
Yichen Wu (wuyichen@usc)

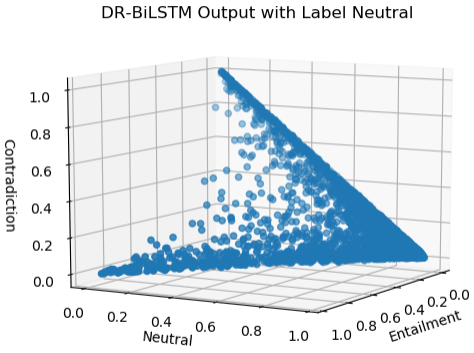
CSCI 544 Project

We can interpret each of the datapoints in the graphs below as how much the DR-BiLSTM model is "leaning" toward each of the labels. This is possible because the output probabilities are gotten from a softmax layer, which weights the probabilities and has the property that the sum of the probabilities sum to 1. So, for instance, given a datapoint [0.1, 0.2, 0.7] (where index i indicates the predicted probability that the hypothesis is of label i, and index 0=Entailment, index 1=Neutral, and index 2=Contradiction), when this datapoint is plotted onto the graph, we can see that the datapoint is majority-weighted in contradiction.

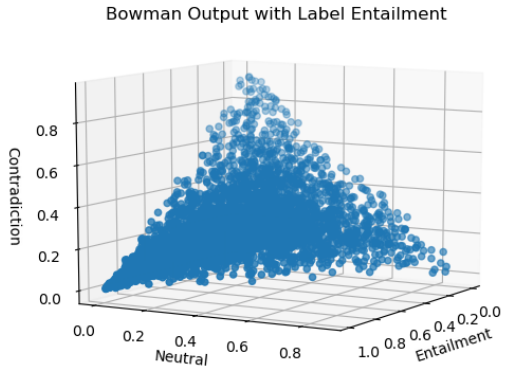
As such, we can easily tell from the graphs below that for ground truth entailment and contradiction hypotheses, the model very heavily leans toward the correct labels. For entailment and contradiction, in fact, the model very clearly knows that a ground truth entailment hypothesis is not a contradiction hypothesis, and vice-versa. But for ground truth neutral hypotheses, the model is actually visibly "confused" and has heavy collections of datapoints leaning towards both entailment and contradiction, meaning that the model in general can discern a neutral hypothesis correctly, but that the model sees the neutral hypothesis as close to both entailment and contradiction.

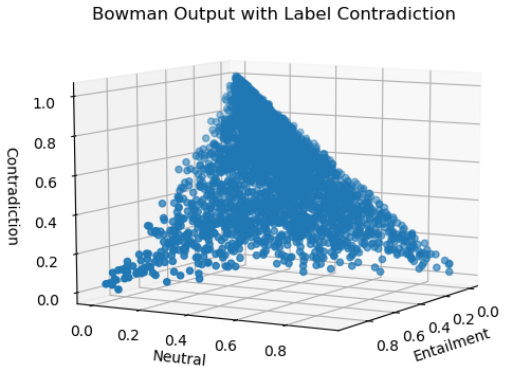


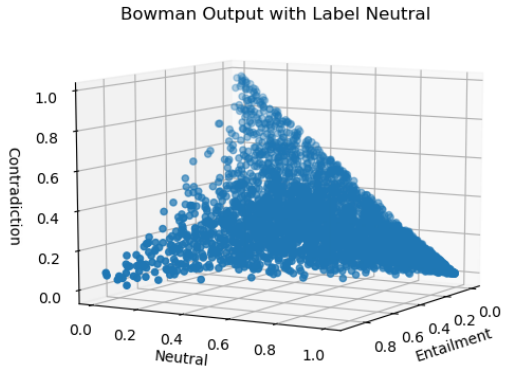




Contrasting the DR-BiLSTM model with the Bowman model, it becomes immediately clear that the DR-BiLSTM model's output probabilities are far more certain than the probabilities outputted by the Bowman model. Although the Bowman model only performs on average about 20% worse than the DR-BiLSTM model, the output probabilities it generates from its softmax layer have more noise, as we can see from the fact that the datapoints in the graphs below are not clustered tightly near the ground truth label but are rather clustered more generally around the ground truth label. Whereas for the DR-BiLSTM model it is much more clear-cut if a hypothesis belongs to entailment as opposed to contradiction, for the Bowman model each label is a fuzzy region and the Bowman model does not seem to see the neutral label as a middle ground between entailment and contradiction as the DR-BiLSTM model does.







e. Qualitative Analysis

Given a list of test sentences from the SNLI test dataset (below), we ran both the Bowman model and the DR-BiLSTM model on these five sentences and outputted their labels as com-

pared with the ground truth.

1. **Text:** This church choir sings to the masses as they sing joyous songs from the book at a church.
   Hypothesis: The church has cracks in the ceiling.
2. **Text:** This church choir sings to the masses as they sing joyous songs from the book at a church.
   Hypothesis: The church is filled with song.
3. **Text:** This church choir sings to the masses as they sing joyous songs from the book at a church.
   Hypothesis: A choir singing at a baseball game.
4. **Text:** A woman with a green headscarf, blue shirt and a very big grin.
   Hypothesis: The woman is young.
5. **Text:** A woman with a green headscarf, blue shirt and a very big grin.
   Hypothesis: The woman is very happy.

The following table displays the 5 judgments each model made along with the ground truth judgments.

	1	**2**	**3**	**4**	**5**
Ground truth	Neutral	Entailment	Contradiction	Neutral	Entailment
Bowman	Contradiction	Contradiction	Contradiction	Neutral	Neutral
DR-BiLSTM	Contradiction	Entailment	Contradiction	Neutral	Neutral

The DR-BiLSTM model predicted 3 labels correctly out of 5 examples. The judgment for the first sentence should be neutral, while the model predicted contradiction; the judgment for the last sentence should be entailment, while the model predicted neutral. The Bowman model, on the other hand, predicted 2 labels correctly out of 5 examples. Both models predicted correctly for the third and the fourth pairs; both models predicted the incorrect "contradiction" for the first pair, and the incorrect "neutral" for the last pair.

Perhaps the reason that both models incorrectly labeled the first sentence is that none of the words in the hypothesis are synonyms of the words in the text, but that the word "cracks" might be considered similar to the word "books" in some models. In terms of the last sentence, it is unclear why both models predicted the judgment would be neutral rather than entailment, as "happy" should be a very close synonym to "grin". One factor that might have helped cause the models' incorrect judgments is that the models might not have assigned "a very big grin" to "a woman" because the given text is a phrase rather than a complete sentence. If the models failed to connect the fact that it was the woman who was grinning, then it makes sense that both of them saw the hypothesis of the woman being happy as neutral rather than entailment.