Kathleen Xue, James Ke,
Guohao Li, Yichen Wu

CSCI 544 Project

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%
DRLSTM model (preset hyperparameters)*	90.81%	87.18%
DRLSTM 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 0.8 after 400 epochs, but the DRLSTM cross entropy loss (right graph) hits a minimum of around 0.25 after only 5 epochs. This means that the DRLSTM model has a lower cross entropy loss, or a lower divergance between predicted judgments and actual judgments, than the Bowman model.
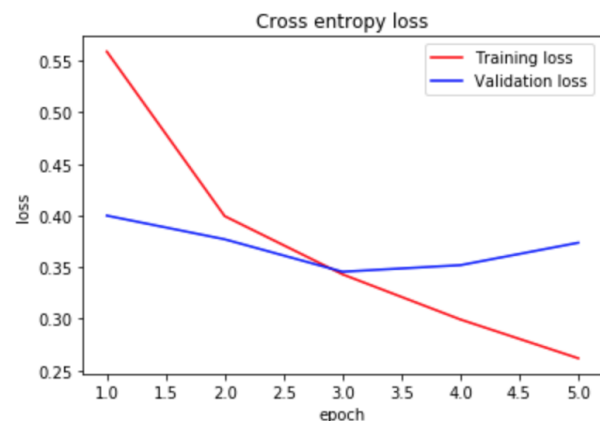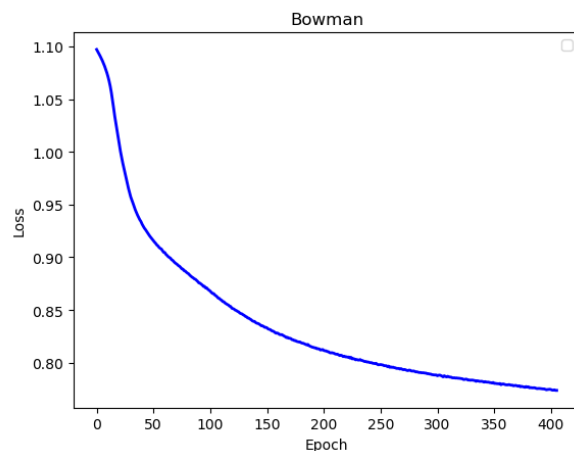
e. Qualitative Analysis

Given a list of test sentences from the SNLI test dataset (below), we ran both the Bowman model and the DRLSTM model on these five sentences and outputted their labels as compared 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	Entailment	Entailment	Contradiction	Entailment	Neutral
DRLSTM	Contradiction	Entailment	Contradiction	Neutral	Neutral

The DRLSTM model predicts 3 labels correctly out of 5 examples. The judgment for the first sentence should be neutral, while the model predicts contradiction; the judgment for the last sentence be entailment, while the model predicts neutral. The Bowman model, on the other hand, predicts 2 labels correctly out of 5 examples. Both models predict correctly for the second and the third pairs; both models predict 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 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 had the grin, then it makes sense that both of them saw the hypothesis of the woman being happy as neutral rather than entailment.