1 Human vs. Machine Translation Classifier

The answer for this problem is divided into the following 3 parts:

A. Classifier design

B. Additional features considered for evaluation.

C. Evaluation of Classifier performance.

D. Citations

Part A

Classifier Design

I'm using libsvm based Support Vector Machines (SVM)[1] as the classifier to predict the labels based on the features listed below. I use the 10% of the tail-end of labeled data as the development set and the rest of the 90% as the training set. Based on my evaluation described in Part C, I picked the following features to be part of my final model:

1. gleu_scores[3]

GLEU score for the candidate sentence

2. ratio_num_tokens_source_candidate

Ratio of the number of tokens in the source sentence to the candidate sentence.

3. ratio_mean_token_length_source_candidate

Ratio of mean token length in the source sentence to the candidate sentence.

4. ratio_common_bigrams_candidate_reference

Ratio of the common bigrams in the candidate and reference sentence to the total number of bigrams in the candidate and reference sentence.

The performance evaluation of the model is discussed later in Part C.

Part B

Additional features considered for evaluation

Initially, I considered and computed the following additional features for the Classifier to predict based upon:

1. ratio_num_char_source_candidate
Ratio of the number of characters in the source sentence to the number of characters in candidate sentence.

2. ratio_num_token_candidate_reference
Ratio of the number of tokens in the candidate sentence to the number of tokens in the reference sentence.

3. ratio_tree_height_candidate_reference
Ratio of the parse tree height for the candidate sentence to the parse tree height for the reference sentence. The trees were constructed using the Stanford Parser library[2].

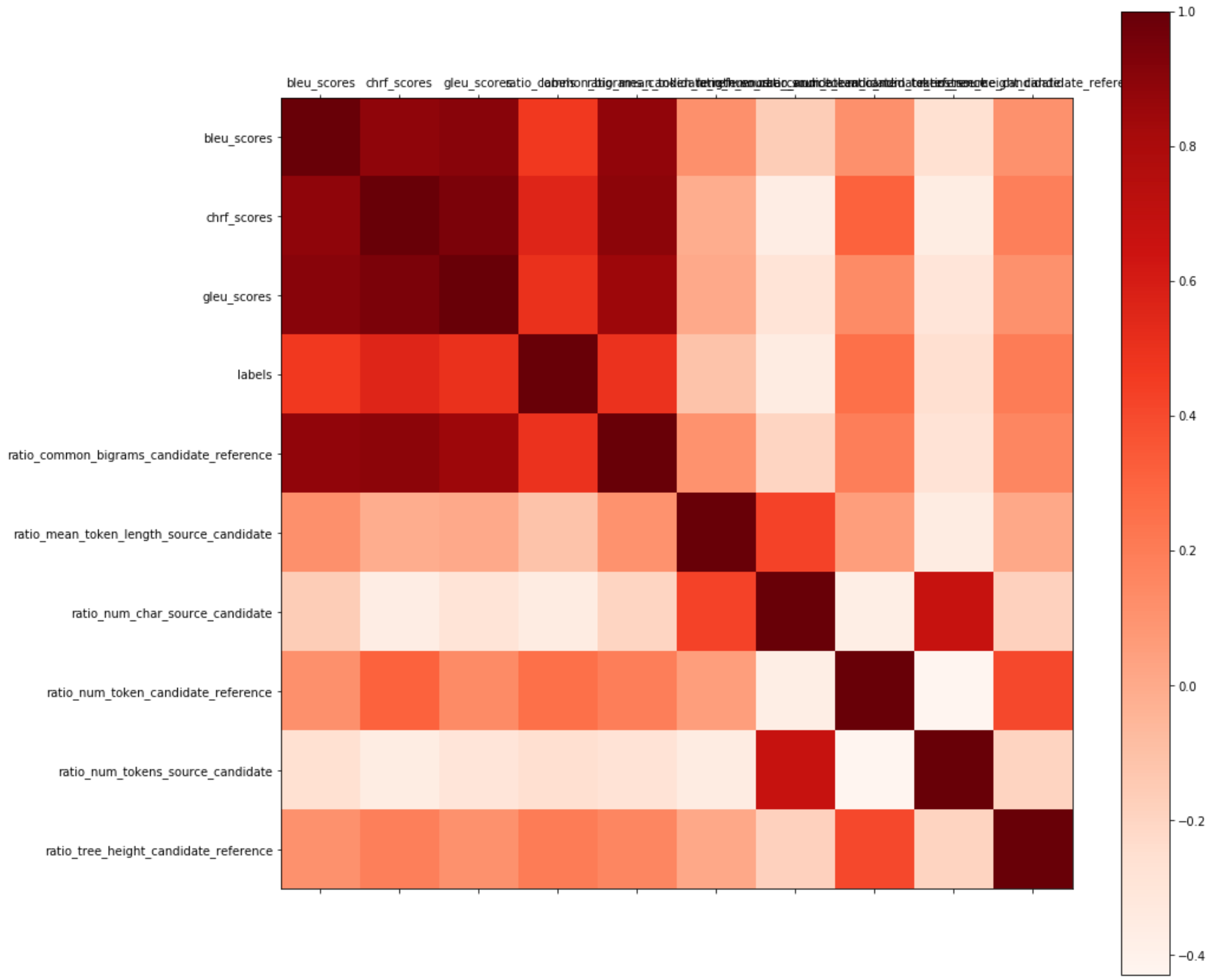
4. chrf_scores[4]
CHRF score for the candidate sentence.

5. bleu_scores
BLEU score for the candidate sentence.

However, based on my evaluation described in the following section, I ended up using only 4 features listed in Part A.

Part C

I ran a test with the training and development dataset split of 80% and 20% respectively and generated a correlation matrix for the 9 features. Initially, I decided to peruse the Correlation matrix to shortlist the features by determining what features are most suitable for this Classifier. The intuition I was following here was that features that are most correlated to the Label are the best features.



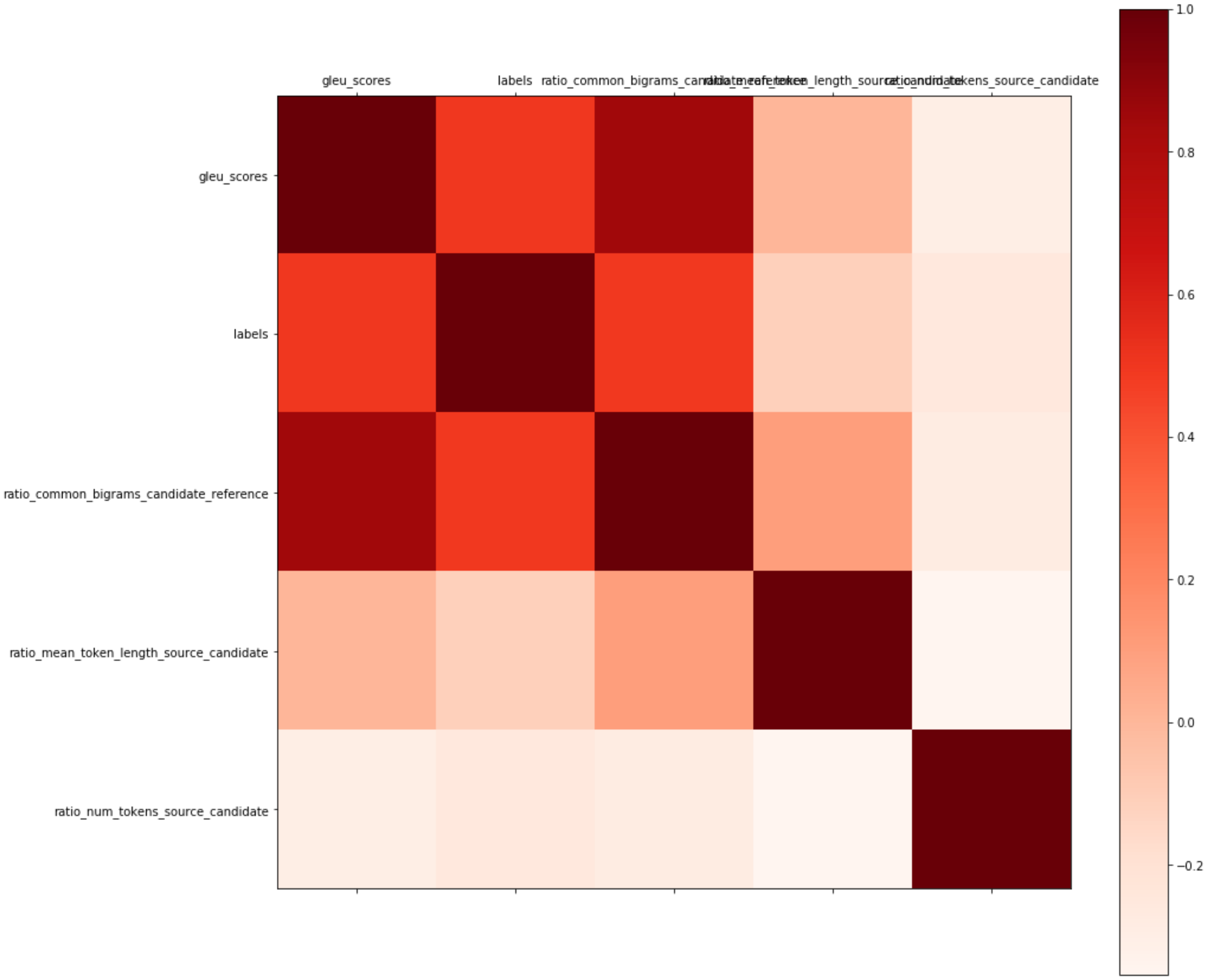
However, this is not true since the features that are most correlated to the Label are also bound to be correlated to each other (as shown in the above diagram).

Hence, I decided to discard this approach.

Next, I ran over 450 tests on the training and development dataset split of 90% and 10% respectively to test the accuracy of the Classifier with different subsets of the above 9 features. I then re-ran these 450+ tests for 80% and 20% split of the training and development dataset.

I complied the accuracy and F1 scores for these 900+ tests together and hand-picked the feature-set with the highest accuracy and F1 score. This feature set was: [gleu_scores, ratio_num_tokens_source_candidate, ratio_mean_token_length_source_candidate, ratio_common_bigrams_candidate_reference].

Here is the correlation matrix for the selected feature set:



The accuracy for the resulting Classifier on the development data is 82.67% and the F1 score is 0.8354.

Here is the confusion matrix for the above run:

	Predicted: Machine	Predicted: Human
Actual: Machine	29	8
Actual: Human	5	33

Part D: Citations

[1] "Sklearn.Svm.SVC — Scikit-Learn 0.18.1 Documentation". *Scikit-learn.org*. N. p., 2017. Web. 30 May 2017.

[2] "The Stanford Natural Language Processing Group ". *Nlp.stanford.edu*. N. p., 2017. Web. 30 May 2017.

[3] Wu, Yonghui, et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv preprint arXiv:1609.08144* (2016).

[4] Popovic, Maja. "chrF: character n-gram F-score for automatic MT evaluation." *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*. 2015.

[5] "Natural Language Toolkit — NLTK 3.2.4 Documentation". *Nltk.org*. N. p., 2017. Web. 30 May 2017.