NLP Research Application
Machine Translation - Binary Classification

Zihao (Robin) Lin
Cornell University, College of Engineering, B.S in Electrical and Computer Engineering
Email: zl755@cornell.edu
Github Repository: NLP Application

In this classification task, the goal is to determine whether a translation from a Chinese source text was created by a human or by a machine. When I first started, it was not entirely clear how I could approach the problem. The data given to us consists of the source text in Chinese, reference translation in English, candidate translation in English, the Bilingual Evaluation Understudy Score, and the ground truth label. Given all of these parameters, how do we process the data so that it could be useful in a classification model? After thinking about how a human would approach this task, I realized that we tend to evaluate the translations based on how natural and similar they are to human language. Thus, I reasoned that we could use the Bilingual Evaluation Understudy Score and additional features related to text similarity between the reference and candidate as features for a statistical or machine learning model.

For the final solution, I compared the performance of two classification models: **Support Vector Machine and Logistic Regression** (Github Repository). An overview of the training and validation process for the two models are shown below.

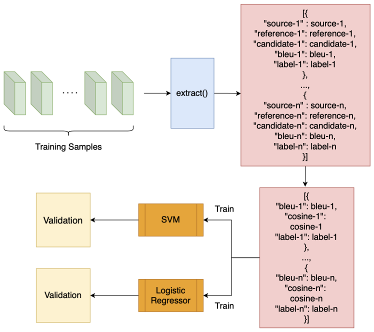

Fig. 1. SVM and Logistic Regression Training Process

For training, the training samples are parsed from train.txt into a list of dictionaries using the function extract() within extract.py, with each dictionary representing a single sample. Then, using the candidate and reference texts, the Cosine Similarity Score for each sample is computed. The NLTK library is used to tokenize the input text. The standard equation for Cosine Similarity is used for term-frequency vectors of the reference and candidate.

After the Cosine Similarity Score has been computed for each sample, the 3-tuple (Bleu_score, Cosine_similarity_score, label) for each sample are inputted for training. A label of ”0” corresponds to human translation and a label of ”1” corresponds to machine translation. The two classification models used are: SVM and Logistic Regression. To implement the models, the Sklearn library was used. Overall, Logistic Regression (LR) performed slightly better than SVM. The final testing accuracy and F1-Score for both models are shown below.

TABLE I
PERFORMANCE OF MODELS

Model	Testing Accuracy	F1-Score
SVM	74.14%	0.717
LR	75.29%	0.736

The logistic regression decision boundaries fitted on the training data as well as its confusion matrix are also shown below. Overall, the model classified a total of 131 testing samples correctly and 43 samples incorrectly.

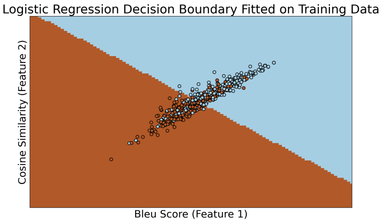

Fig. 2. LR Decision Boundary (Orange: Machine, Blue: Human)

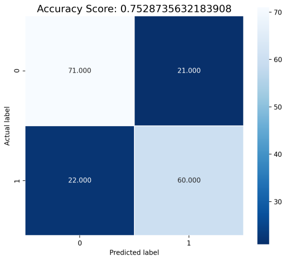

Fig. 3. Confusion Matrix