Feature Selection

Features consist of 3 parts: translation metrics, linguistic features of the candidate sentence and the sentence length. It is assumed that the machine translations differ from human translations in sentence structure. So I used Stanford CoreNLP to build a constituency parsing tree to extract features. There are 10 features in total, including:

- Translation metrics: Unigram Bleu score and METEOR score.
- Linguistic features: number of **S** (sentence), **NP** (noun phrase), **VP** (verb phrase), **NN** (singular noun), **SBAR** (sbar), **DT** (determiner) and **JJ** (adjective) in the constituency parsing tree. All features are normalized by the length of the candidate sentence.
- The ratio of the length of the candidate to the length of the reference: $\frac{L_{cand}}{L_{ref}}$

Classification

Support Vector Machine with RBF kernel is used for this translation classification task. To avoid overfitting on the test dataset, the original training text is split into a training dataset (467 samples) and a validation dataset (117 samples). The final test dataset is only evaluated once.

Results

The following tables shows the performance:

	Accuracy	F1 Score
Training	0.675	0.706
Validation	0.735	0.774
Testing	0.735	0.758

The features extracted from the constituency parsing tree are not expressive enough to show the differences between human translation and machine translation. In this mini-project, original Chinese sentences were not used as the input of the model.

References

[1] Stanford CoreNLP.
`https://stanfordnlp.github.io/CoreNLP/`

[2] Yitong Li, Rui Wang and Hai Zhai. *A Machine Learning Method to Distinguish Machine Translation from Human Translation*.

[3] Roee Aharoni, Moshe Koppel and Yoav Goldberg. *Automatic Detection of Machine Translated Text and Translation Quality Estimation*.