Pura Peetathawatchai (pp448)

Human vs. Machine Translation Classification Task

Features

Bleu score – We hypothesize that humans are generally better translators than machines, so the Bleu scores of human translations should be generally higher than that of machine translations.

Structural similarity to reference – Human translators have an intuition for rearranging sentence structure to make the sentence sound idiomatic both in context of the topic and also in relation to surrounding text, whereas machine translators most likely follow an algorithm that preserves the original syntax tree as much as possible, without regards to idiomaticness. Assuming that human translators are likely to think similarly, we therefore hypothesize that human translations will generally have greater structural similarity to the reference, while machine translations generally have less.

Structural similarity to original text – Chinese follows an SVO syntactic structure similarly to English, so the syntactic structure of the original and directly translated sentences should not be too different. For similar reasons to the previous feature, we therefore hypothesize that machine translations will generally have greater structural similarity to the original text, while human translations generally have less.

Vocabulary similarity to reference – Each translated word has many synonyms. Human translators pick the synonym that is most appropriate to the context of sentence and perhaps even the surrounding choice of words. On the other hand, machine translators follow a different heuristic for choosing synonyms that may not be able to account for idiomaticness and context as well. Once again, assuming that human translators are likely to think similarly given a particular context, we hypothesize that human translations will generally use more similar vocabulary to the reference, while machine translations will use generally more different vocabulary.

Feature Engineering

Differences in syntactic structure between sentences – We measured this by using POS taggers to translate both sentences into sequences of POS tags. We then used the edit distance divided by the average of the lengths of the two sequences (to account for sequence length) as a rough measure of the difference between the two sequences. For this task, we used NLTK for POS-tagging English sentences and Stanford's for Chinese sentences. However, NLTK uses the Penn tagset while Stanford's uses the LDC Chinese Treebank tagset, so we had to translate both to Universal Dependencies tagset when measuring structural similarity of the candidate to the original text.

Vocabulary similarity – We used the bag-of-words approach to translate each sentence to a vector with each entry representing a unique word that exists in at least one of the two sentences. If a sentence has that word in it, that entry has value 1 in its vector, otherwise it has value 0. We then measured the difference in vocabulary used in the two sentences by taking the cosine difference between the two vectors.

Scaling – each feature was normalized via standardization before training.

Model Selection

Based on our earlier hypotheses, we believe that the classes should be fairly easily separable in the feature space (i.e. a hyperplane or simple smooth surface will do). We therefore opted for a simple model like the Random Forest and the linear soft-margin SVM, both of which make few assumptions about the data. We performed hyperparameter search using a 70-30 training-validation split on the provided training data, but did not yield significant results (i.e. similar validation scores for different parameter values). For the Random Forest, we therefore used max_depth = 2 (since we hypothesize a fairly simple separating surface) and 10 trees (since the dataset is small). For the SVM, we used C = 1 (default value) and no kernels (since we hypothesize a fairly simple separating surface).

Results

Both models performed similarly. The mean F_1 scores were 0.78 for the Random Forest and 0.76 for the SVM.

GitHub Link: https://github.com/poonpura/translation