Failed Attempts:

I tried to use synsets from wordnet to find and replace synonyms for each of those hyp1 and hyp2 based on reference sentence and replacing words in h1 and h2 if they have similar word(similarity>85%) in ref sentence. This helped slightly and sometimes didn't improve accuracy when it comes to be combined with other features.

I've tried different evaluation metrics like CHRF, GLEU and RIBES score, they were mostly like similar to simple meteor and sometimes had lower scores.

Then I tried to set a classifier for learning and predicting labels based on different features based on this paper(https://pdfs.semanticscholar.org/17ca/3a00f70899534611d94b5d79af7190efddf5.pdf) as follow:

BLEU, GLEU, CHRF, RIBES, log of ngram precision(n=1..4) , ratio of hypothesis and reference length, ngram precision (n=1..4), hypothesis and reference length and tried to set Random forest, SVM and Neural network ranking classifier to learn. I used grid search for trying different values for parameters(for tuning classifiers) and also features importance to see what feature had most impact in finding results. After running for a long time I got my results and I decided to remove some of these features since they hadn't that much impact. Features like(log of ngram precision(n=1..4), ratio of hypothesis and reference length, hypothesis and reference length). Again I tried to train a classifier with 30,40 and 50 percent of data for training(avoiding to be overfitted for unseen data) and different combination of remaining features, I got results about 56% on my training data and about 52.5 on my test data. After looking at results I realized that samples like 1 and -1 have been repeated nearly as twice as 0. So I decided to use oversampling(SMOTE library) for my data to increase zeros in my dataset by generating synthetic data. After doing that I trained classifier again and I saw a drop in scores. Maybe because those features are highly sometimes correlation and also being closed together(different parameters for different sentences) that method caused a higher overlap between those scores and features so it mislead trainer. I also tried different algorithms for training like Decision trees and XGBoost but it didn't help. I also used word2vec method, for comparing each given words in hyps and calculating differences between those words to ref's words and then summing them up those scores for each hyp, it needed very high computational power to be run in reasonable time. I also tried CREG(https://github.com/redpony/creg) to do ordered logit for ranking, it didn't show me that much of a an imporovment.

Worked attempts:

After running and learning from those features and trying different combinations of those metrics and methods I decide to add few more features. First of all I created a stemmed version of ref and hyps. Then calculated intersection of same words in hyps and ref. then calculating recall and precision and also harmonic F score for both original and stemmed version. Calculating chunks in both hyps in comparison to ref and giving a penalty score for those that are shared. For calculating chunk I used longest common string algorithm and each time I removed substring from hyp till no one left. Then compounding all those score into one: f_mean * (1 - (gamma * (penalty_for_chunk**beta))). After adding this metric I found out there is a easier way for calculating that word2vec technique, I used spacy library to find out about two sentece similarity and then add that metric to my feature sets. I also added n-gram POS precision, n=1...4

to set of my features since those POS tagger will be a good sign for finding out if our hyp is close to ref or not in terms of fluency and human readability.

After all I had 12 different feautres: simple METEOR, POS Precision, sentecene similarity, SCORE,BLEU

Each of those features have two variable since we have two hyp.

Now I trained my classifier and after trying different combination and also using feature importance library I realized that BLEU score wasn't helping that much, I think that it's because of overlaps between BLEU and other metric it didn't have that much impact, so I removed that score from my feature set. Now I have 10 and run the training. I'm getting close to 60% acc for my training and close to 55% for my test data. I again used(25,30,40,50) percent of my data for testing.