Algorithm Description

I implemented the evaluation process based on the METEOR(1) algorithm with the modification of assigning weights to the words matched by different methods.

Given h1, h2 and ref, my method calculates two scores, s(h1, ref) and s(h2, ref) and returns:

1	if s(h1, ref) > s(h2, ref)
0	if s(h1, ref) = s(h2, ref)
-1	if s(h1, ref) < s(h2, ref)

Where

$$s(h, ref) = Fmean * (1 - Penalty)$$

$$Fmean = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 * \frac{\#chunks}{match\_score}$$

$$P = \frac{match\_score}{\#words\ in\ h}$$

$$R = \frac{match\_score}{\#words\ in\ ref}$$

The match_score is calculated in three stages.

Stage 1 exact match: For every word w in ref, search for an unmatched word in h which is exactly the same as w. If find, match_score += 1.0 and tag those two words as matched.

Stage 2 stem match: For every unmatched word w in ref, search for an unmatched word in h which has the same stem as w. If find, match_score += 0.8 and tag those two words as matched. I used wordnet.morphy() to get the stem of a word.

Stage 3 synonym match: For every unmatched word w in ref, search for an unmatched word in h which is w's synonym. If find, match_score += 0.6 and tag

those two words as matched. To identify the synonym relationship, I used wordnet.synets() to check if two words share a common semantic object.
Finally I get the fewest possible number of chunks such that the unigrams in each chunk are in adjacent positions in the hypothesis translation, and are also mapped to unigrams that are in adjacent positions in the reference translation

(1) Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. Vol. 29. 2005.