Sarika Halarnakar - shalarn1
Sindhuula Selvaraju - sselvar4
Machine Translation Assignment 3: Evaluation

Contents of the Repository:

1. data : This directory contains the data files given to us and a smaller file we used for testing
2. results: This directory contains result files for the data given to us. The baseline result file is result_original
3. check : This has not been modified by us
4. compare-with-human-evaluation : This has not been modified by us
5. evaluate_original : This is the original evaluate file given to us
6. evaluate_meteor : This version of the evaluation program has the implementation of METEOR
7. evaluate_keywords : This version of the evaluation program compares keywords from the reference sentence and the hypothesis to find which one is better
8. evaluate_wordnet : This version of the evaluation program compares synonyms for each word of the reference sentence with each word of both hypothesis
9. evaluate_similarity : This version of the evaluation program compares the similarity of the words of the reference sentence with each word of both hypothesis
10. evaluate_stemandlem : This version of the evaluation program compares synonyms for each word of the reference sentence with each word of both hypothesis. But after finding synonyms it stems, lemmatizes and stems and lemmatizes the words and gives 3 output files.

NOTE: Running any of the evaluate_xyz files will automatically add the result to the working directory so remember to delete the old result before running any program.

Part 1: METEOR Implementation

Usage: (In Evaluator)

```
python evaluate_METEOR
python compare-with-human-evaluation < result_METEOR
```

Part 2: Wordnet Implementation

Usage: (In Evaluator)

```
python evaluate_stemandlem
python compare-with-human-evaluation < result_Lem
python compare-with-human-evaluation < result_Stem
python compare-with-human-evaluation < result_LemnStem

python evaluate_similarity
python compare-with-human-evaluation < result_similarity

python evaluate_keywords
python compare-with-human-evaluation < result_keywords


python evaluate_wordnet
python compare-with-human-evaluation < result_wordnet
```

Description:

We decided to use Wordnet to improve accuracy. We implemented Wordnet in 6 ways using the python library nltk.

1. We evaluated the hypotheses by comparing the words of the hypotheses to the synonym set of the reference sentence. We used METEOR with the original reference sentence and hypotheses.

2. Then, within METEOR we found the synonym set of each word in the reference sentence and then checked if each of the hypotheses was in the combined synonym set of the reference sentence(see evaluate_wordnet).

3. We evaluated the hypotheses by comparing keywords from the reference sentence and the hypotheses (see evaluate_keywords). We first extracted the keywords from the reference sentences and the hypotheses. Then, we still used METEOR, but with the keywords and alpha as input and we tested to see if a keyword from the hypotheses was in the synonym set of one of the keywords from the references.

4. We experimented with lemmatizing and stemming (see evaluate_LemnStem). We used METEOR with the original reference

sentence and hypotheses, but within METEOR we found the synonym sets for all the words in the reference sentence and then lemmatized and stemmed each synonym in the synonym set. We then lemmatized and stemmed each word of the hypotheses and checked if it was in the lemmatized and stemmed synonym list.

5. We followed the same process as (3) but we experimented with only lemmatizing the words

6. We followed the same process as (3) but experimented with only stemming the words.

7. In evaluate_similarity we tried to find the similarity between the synonyms of the words in the hypothesis and those in the reference sentence. This algorithm gave an accuracy of 1.0 with just 10 sentences while the original algorithm given to us gave an accuracy of 0.1. This method however takes a very long time run (3 hours for 5000 sentences on grad systems). Also, as the number of sentences are increased the accuracy keeps decreasing.

Results:

	Original	Meteor	Synonym Match	Keyword Extraction	Lemmatize & Stem	Lemmatize	Stem	Synonym Similarity
Accuracy	0.449312	0.504068	0.505124	0.177957	0.506336	0.503051	0.50704	0.449585

The Synonym Match, Lemmatize & Stem, and Stem all reached an accuracy above METEOR. Our Keyword Extraction implementation plummeted. Just lemmatizing did not increase the accuracy as compared to both lemmatizing and stemming , but just stemming increased the accuracy enough. Synonym similarity provided a slightly better score than the original but was still not as good as Stem.