Evaluation

Adam Poliak — apoliak1 David Russell - drusse19

March 25, 2016

All code mentioned can be found at `https://github.com/azpoliak/evalutor`. At time of submission, the leader board was not working so we were unable to submit our results there.

1 METEOR

1.1 METEOR Implementation

As instructed in the assignment, we implemented METEOR in the script `./evaluate` by calculating and using precision and recall as described in the assignment. The implementation is contained in the method `meteor`

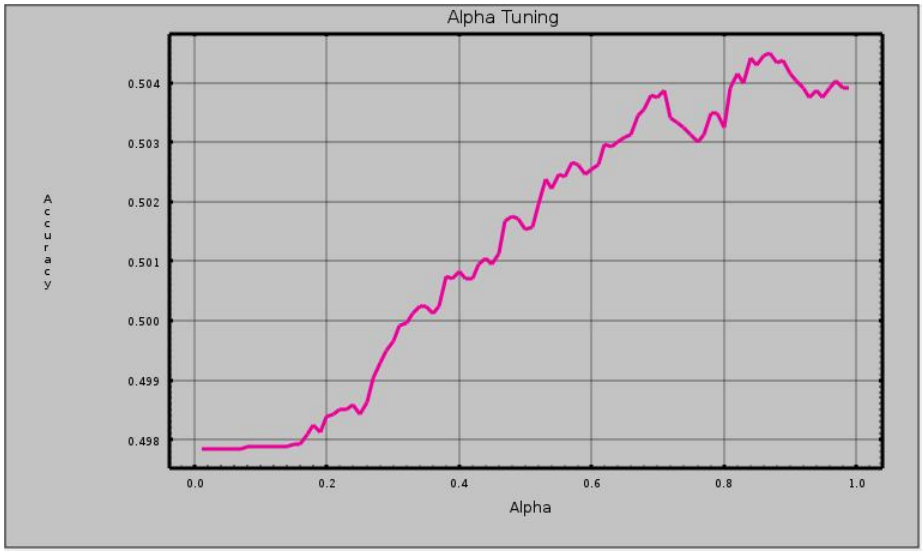
1.2 Running METEOR

To run our implementation of METEOR, `./evaluate` must be called with the α flag. To add an α parameter, use the **-a** flag followed by a decimal number to represent the α. If the **-a** flag is not provided, then the baseline implementation of `./evaluate` provided will be ran instead.

1.3 α tuning

After running `./evaluate` 100 times, we determined that the best alpha on the dev data is $\alpha = .87$ resulted in an accuracy of 0.5045

The graph below shows the results from tuning alpha. The x-axis represents the alphas and the y-axis represents the accuracy for corresponding alphas used in METEOR.



The data in the graph was generated by running the script `test_alpha` to determine the best α. This script takes about 5 minutes to run.

2 WordNet Synonyms + METEOR

2.1 Overview

As discussed in the handout, METEOR uses both precision and recall to calculate the accuracy. Both precision and accuracy are based on the number of common words in sentence `h` and in sentence `e` as

$$Precision = \frac{|h \cap e|}{|h|}$$

and

$$Recall = \frac{|h \cap e|}{|e|}.$$

The numerators above only includes tokens that appear in both sentence `h` and `e` and ignores synonyms. This new approach includes both the number of tokens and high quality synonyms that appear in `h` and `e`. We define high quality synonyms as synonyms in WordNet that have a score above a threshold. We define this new concept as $h \frown e$

2.2 Generating number of synonyms

After calculating $|h \cap e|$, we create new sentences `h'` and `e'` where

$$\texttt{h'} = (\texttt{h} \backslash \texttt{e})$$

$$\texttt{e'} = (\texttt{e} \backslash \texttt{h})$$

Thus, $\texttt{h'} \cap \texttt{e'} = \emptyset$ and $|\texttt{h'} \cup \texttt{e'}| = |\texttt{h'}| + |\texttt{e'}|$.
For every word in `e'`, we use `NLTK`'s `wn.synsets` to a set of all `synsets` of each word in `e'`. Next, for every word in `h'`, we generate each word's `synsets` and

call `NLTK`'s `lch_similarity` on each `h'` words' `synsets` with each `e'` words' `synsets`. If that new value is greater or equal to the WordNet synonym threshold, $|h \cap e|$ is incremented by 1 and then continue to the next word in `h'`. We define this resulting value, i.e. the sum of $|h \cap e|$ and the number of synonyms in `h'` and `e'`, as $|h \frown e|$.
Thus in the WordNet synonyms and METOER implementation, we define

$$Precision = \frac{|h \frown e|}{|h|}$$

and

$$Recall = \frac{|h \frown e|}{|e|}$$

.

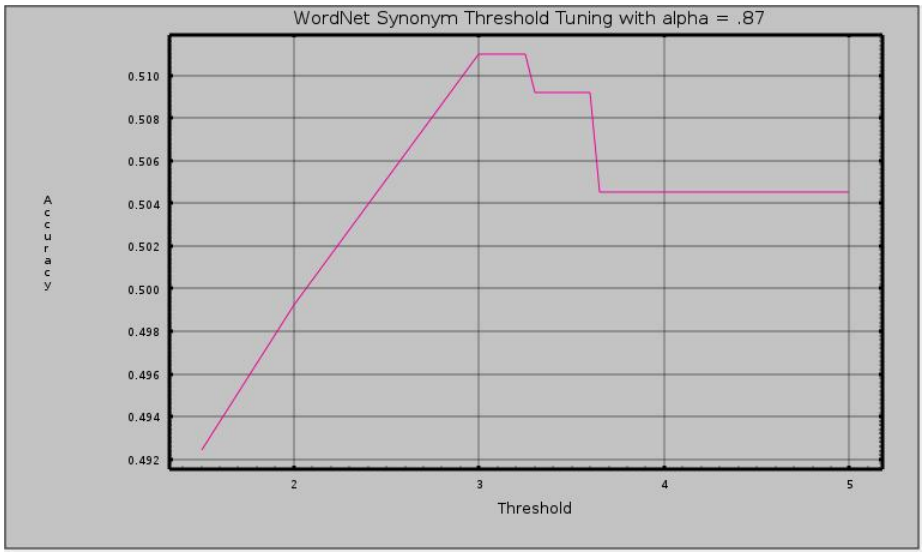
2.3 Running WordNet Synonyms + METEOR

To run our the combo, `./evaluate_word_net` must be called with the `t` flag. To add a WordNet synonym threshold parameter, use the `t` flag followed by a number to represent the WordNet synonym threshold parameter. The default `t` flag is .26.
This script takes roughly $2.5 - 3.5$ hours and uses roughly 900 `MB` when ran on the entire `dev` dataset.

2.4 Synonym threshold tuning

After running `./evaluation` many times, we determined that the best WordNet synonym threshold on the dev data when $\alpha = .87$ is in the range of 3 to 3.25. In such cases, the accuracy was on the dev was 0.510990, which is a slight improvement of the basic METOER implementation.

The graph below shows the results from tuning the WordNet synonym threshold. The x-axis represents the threshold and the y-axis represents the accuracy for corresponding thresholds used.



3 Future Work

3.1 α and synonym threshold tuning combined

We tuned the WordNet synonym threshold only on $\alpha = .87$. It would be interesting to tune both at the same time to determine the best α and WordNet synonym threshold pairing.

3.2 WordNet synonym threshold combined with the default evaluation method.

The default evaluation just calculates the number of tokens that appear in both `h` and `e`. WordNet synonym threshold combinbed with the default evaluation method would include the number of synonyms in `h` and `e` to that count