Word Vector Models for Translation Evaluation

Akshay Srivatsan, Ian Mukherjee, Nathan Smith

March 25, 2016

1 Introduction

Our group incrementally implemented three translation decoder algorithms beginning with the addition of the meteor metric described in the assignment guidelines. After completing the baseline meteor metric implementation, we used Google's word2vec to obtain 200-dimensional vectorizations of the words being evaluated. Firstly, we implemented a simple algorithm that computed the summation of word vectors for each hypothesis to form a sentence vector, and compared their cosine similarities against the reference sentence vector. While this yielded slight improvements over the meteor metric, it did not consider word order when assessing the accuracy of the translations. Thus, we added an addition n-gram evaluation component, where uni through quad-grams in the hypothesis and reference sentences were compared to find the number of shared words, and weighted to create a n-gram accuracy metric. This was then multiplied by the cosine similarity previously calculated for the sentence, and the final metrics were compared against the two hypotheses to determine the more accurate translation. This yielded are best accuracy results, at 0.53.

2 Meteor Model:

By implementing the meteor metric, the new evaluation model takes into account the precision and recall of each hypothesis versus the reference translation. While the initial comparison metric only took into account the number of shared words between the hypothesis and reference sentences, including the precision and recall provide better measures of the translation relevance. The meteor equation is as follows:

$$l(h,e) = \frac{P(h,e) * R(h,e)}{(1-\alpha) * R(h,e) + \alpha * P(h,e)}$$

We looked into previous research regarding the optimal alpha value, and determined that 0.82 would yield the most accurate results.

This new implementation provided the following results:

Accuracy Test for the Meteor Model:

Accuracy	0.504732

This accuracy is identical to the leaderboard baseline in previous years, and we could thus conclude the metric was properly implemented.

3 Vectorized Sentence Model:

After completing the meteor model implementation, we considered the use of word vectorizations, as by comparing word vectors we could determine the semantic similarity between words more precisely than simply observing if they are equivalent. Using Google's word2vec, we generated a list of the 50,000 most common words in their vectorized forms. For each sentence, we iterated through each word and summated their word vectors. The cosine similarity between the hypothesis and reference sentence was then computed, with a 1 being printed if h1 yielded a higher cosine similarity. Accuracy for the Vectorized Sentence Model is as follows:

Accuracy	0.523662

4 N-Gram/Vectorized Sentence Model:

While the sentence model yielded higher accuracies by considering word meaning and similarity, simply computing the vector summation ignored word order for the sentences. To address this, we developed an additional metric that slightly modified the meteor implementation, looking at the number of words shared for 1 through 4-grams in the hypothesis and reference sentences. For each n-gram length, we iterate through the reference sentence, and build a set of all n-grams that occur. We then iterate through the hypothesis sentence, and increment a counter each time an n-gram in the hypothesis is found in the reference set. We then divide this total score by the summation of n-gram set lengths. Finally, we multiply this by our previously calculated vectorized sentence to produce a final metric to compare between the two hypotheses.

Accuracy for the N-Gram Sentence Model is as follows:

Accuracy	0.533

5 Bibliography:

References

[1] Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, Philipp Koehn Monte Carlo inference and maximization for phrase-based translation

[2] Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, Hinrich Schutze The operation Sequence Model

[3] Daniel Ortiz Martinez, Ismael Garcia Varea, Francisco Casacuberta Nolla Generalized Stack Decoding Algorithms for Statistics Machine Translation