CIS 526 Homework 3: Evaluation

James Yang

March 6, 2015

1 Introduction

Evaluating the quality of machine translation models involves the hiring of multi-lingual humans to evaluate entire translated documents, a process that can be very expensive depending on the size of the corpuses. Ongoing research in machine translation evaluation aims for an automated solution that mimics human evaluation criteria. In this homework, we were tasked with implementing and experimenting with different evaluation metrics. In this document, we discuss the inner workings, drawbacks, and results of the METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric, BLEU (Bilingual Evaluation Understudy) metric, and a custom statistical evaluation model.

2 METEOR metric

Compared to the BLEU metric, the METEOR metric is known to provide better sentence-level evaluation scores as it also considers translation recall as well as precision in evaluating translation quality. The evaluation score s of hypothesis sentence h given reference sentence r is given by the following equation:

$$s(h,r) = \left(1 - \gamma \left(\frac{c}{m}\right)^{\beta}\right) \frac{P(h,r) \cdot R(h,r)}{(1-\alpha)R(h,r) + \alpha P(h,r)}$$

where P and R are precision and recall of unigrams in the reference and hypothesis sentences defined as

$$P(h,r) = \frac{|h \cap r|}{|r|} \ , \ R(h,r) = \frac{|h \cap r|}{|h|}$$

c is the number of matched "chunks" in the hypothesis sentence as defined in [1], m is the number of matched unigrams, and α, β, and γ are tunable parameters that weigh precision, recall, and the chunking penalty.

In the evaluation algorithm, two hypothesis sentences are scored separately, providing scores s_1 and s_2 for each hypothesis respectively. If $s_1 > s_2$, then the first model is chosen as the better for that particular sentence. If $s_1 < s_2$, then the second model is chosen as the better for that particular sentence. If both are equal in quality, both are chosen as valid. In our training set, the confusion matrix indicated that the metric should rarely be on the fence about determining the better translation. Also, it was evident from the training data that the second model was typically better. Therefore, when the scores for both models tied, the second model was awarded the translation.

3 BLEU metric

The BLEU metric did not perform nearly as well as METEOR on evaluating sentence-level translation quality. This is mostly due to the formula for calculating the BLEU score s:

$$s(h,r) = p_b \exp\left(\sum_{i=1}^{n} \lambda_i \log\left(P(r,h)_i\right)\right)$$

where p_b is the brevity penalty given by

$$p_b = \min\left(1, \frac{|h|}{|r|}\right)$$

$P(r,h)_i$ is the precision of matched i-grams up to n-grams, and λ_i is the weight for its associated precision.

Typically, the weights are uniform, simplifying the equation to

$$s(h,r) = \min\left(1, \frac{|h|}{|r|}\right) \prod_{i=1}^{n} P(h,r)_i$$

We can immediately see an inherent problem with this metric, where within a sentence, an n-gram is not matched between the reference and hypothesis sentence, ultimately setting the score for that hypothesis to 0. To compensate for this as a sentence-level metric, any n-gram match that results in 0 precision is set to 1, and the resulting product becomes non-zero.

4 Linear Regression metric

As an alternative to a simple formula for measuring the quality of a sentence, I also implemented a linear regression model meant to mimic human judgment in sentence evaluation. Learning was based on n-gram match precision of each hypothesis. More formally, my feature vector $\mathbf{x}$ is of length $mn + 1$ where m is the number of translation models and n is the number of desired matched n-grams, ergo

$$\mathbf{x} = \{P(r, h_j)_i, 1\} \,|\, j = \{1, \cdots, m\}, i = \{1, \cdots, n\}$$

Given a set of D vectors $X \in \mathbb{R}^{D \times (mn+1)}$ and associated labels $Y \in \mathbb{R}^D$, we perform a basic maximum likelihood estimate to obtain the weights for each vector component. We thus predict

$$\hat{y} = \text{sign}(\hat{w}^T \mathbf{x})$$

$$\hat{w} = \left(X^T X\right)^{-1} X^T Y$$

5 Results

Table 1 shows the results of METEOR, BLEU, and the linear regression model using the sample data.

Metric	Grade
METEOR	0.524171
BLEU	0.514784
Linear Regression	0.518773

Table 1: Results on sample data.

We can see from the results that METEOR performed best among the attempted metrics, followed by the learning method and finally the augmented BLEU metric. Strangely enough, a chunking penalty on METEOR seemed to degrade performance. As a result, the following parameters were chosen empirically: $\alpha = 0.82$, $\beta = 0$, and $\gamma = 0$.

It was expected that the linear regression method would have performed better, but it is also noted that the chosen features may not have been the most desirable. Adding more features such as POS matching and n-gram match recall would improve the score marginally. Also noted is that linear regression is not the best choice for discrete classification, and perhaps a more suitable learning method such as support vector machines or neural nets might be preferred.

References

[1] Banerjee, S. and Lavie, A. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005