EVALUATION

Evaluation is a crucial step in this project as the evaluation of text summarization is a challenge. The best way to evaluate summarized documents is by human evaluation. However, this is impractical and not scalable. Automatic evaluation is hard as there is not one correct way to summarize. There may be two summarizes that uses completely different words and sentence structures and yet be equally good summaries. To this end, researchers have come up with comparative metrics that can be used as a heuristic to measure the quality of summarization. We have used BLEU[5] and ROUGE[6] to evaluate the summaries generated by the methods we developed.

BLEU (Bilingual Evaluation Understudy) is an algorithm used for evaluating the quality of the generated text. BLEU is measure of precision. It checks whether the n-grams that are present in the candidate are also present in the reference summaries. It also includes a brevity penalty that are longer than the reference penalty. BLEU is calculated using the following formula:

$$\log BLEU = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} \frac{1}{N} \log p_n$$

Where, $p_n$is the precision score obtained from the summaries. The precision score is calculated on a sentence to sentence basis. We first compute the n-gram matches sentence by sentence. Next, we add the clipped n-gram counts for all the candidate sentences and divide by the number of candidate n-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was developed at the Institute of Information Sciences at USC and is a universal method used to evaluate text summarization. While BLEU is a precision measure, ROGUE is a metric that evaluates the recall of the candidate summary with respect to the reference summaries i.e., it measures the degree to which the candidate summary contains all the n-grams that are present in the reference summary.

The BLEU score is precision measure while ROGUE is a recall measure. We decided that the best way to evaluate the summaries would be to compute F_1score which as the harmonic mean of BLEU and ROUGE scores.

The training set we used contained human-generated summaries of combined articles based on topic. Therefore, we combined the summaries generated by our methods, and used this combined summary for evaluation. The results obtained from the evaluation are tabulated in the results section and discussed in the discussion section.