MOULIN ROUGE: A Method for Automatic Evaluation of Abstractive Summaries

Luke Nelson
5281-4937-09
lukenels@usc.edu

1 – Introduction

The breadth of uses for automatic text summarization is vast and far reaching. Somewhat recent advancements in the area of sequence-to-sequence modeling has paved the way for automatic abstractive text summarization. Prior to these developments, extractive methods stood at the focal point of automatic text summarization. For extractive text summarization, entire phrases and sentences are hoisted from the source text and combined to create a summary. Abstractive summarization is a somewhat more anthropomorphically inspired extension of extractive summarization, whereby novel words and phrases not featured in the source text are generated, similar to how a human written abstract or summary would be. While both methods share the same goal of capturing the central content set forth by the document being summarized, abstractive summarization extends the objective further – the supplementary aim is to produce a summary that is original rather than one that is merely an abbreviated transcription of the source.

Evaluation of automatically produced abstractive summaries is the bottleneck for future progress in the space of abstractive summarization, as headway stems from efficient and accurate ways to evaluate the quality of the summaries being generated. Human judges are unfortunately not a viable option for the purpose of abstractive summary evaluation as a result of both their inefficiency and cost. ROUGE scores - the most widely used metric for automatic summary evaluation - were developed to remedy this barrier for efficient evaluation of extractive text summaries, however these methods are not apt to fully assess the quality of abstractive summaries. ROUGE scores test the content of summaries by comparing the words and phrases within the automatically produced candidates to those within any number of extant reference summaries. Because they neglect to additionally measure a summary's level of abstraction, they are inadequate evaluators of automatically produced abstractive summaries. In this paper, we propose MOULIN ROUGE, a new model for the evaluation of abstractive text summaries that augments the ROUGE-S score in order to accommodate the abstractive dimension of the summaries under consideration. We describe ROUGE-S, which we will be using as our baseline model, in Section 2, and then we introduce our augmentation to this baseline in Section 3. Section 4 discusses the methods employed to evaluate the proposed model. Section 5 concludes the paper with a discussion surrounding the implications of our model's design.

2 – ROUGE-S

While there are four primary ROUGE measures, we have chosen ROUGE-S as our baseline. The DUC designated ROUGE-S, ROUGE-1, and ROUGE-2 as the preferred summary evaluation metrics for DUC 2007. Additionally, Lin et al. (2004) concludes that ROUGE-S maintains a high correlation with human judges for evaluating summary content. ROUGE-S measures the overlap of skip-bigrams between a candidate translation and a set of reference translations, where a skip-bigram is any pair of words in their sentence order, allowing arbitrary gaps. After skip-bigram matches between the reference and candidate are determined, the F1 score is computed and the resulting value is the ROUGE-S score.

3 – MOULIN ROUGE

Our proposed model, MOULIN ROUGE, augments the ROUGE-S in order to reward abstraction while still remaining conscious of content. Like the Moulin Rouge cabaret, think of our proposed model as a new form intended for more mature audiences - namely abstractive summaries. ROUGE-S compares candidate summaries to a reference or set of references. The MOULIN ROUGE will follow suit while additionally comparing candidate summaries to the source documents from which they were constructed. Comparing the candidate summaries to the source document will allow the model to 1. Penalize summaries for large sequences of N-gram similarity and 2. Reward summaries for novel vocabulary.

MOULIN ROUGE is calculated as the weighted harmonic mean between ROUGE-S and the Abstraction Score. As you can see in (1) below, beta is a hyperparameter. Increasing its size allows one to value abstraction over content.

$$MOULIN\ ROUGE = (1+\beta^2)\cdot\frac{ROUGE{-}S \cdot Abstraction\ Score}{(\beta^2 \cdot ROUGE{-}S) + Abstraction\ Score} \tag{1}$$

The Abstraction Score is similar to ROUGE-S in that it takes on a value from 0 to 1. This allows us to combine the scores for content and abstraction into one metric. The Abstraction Score, itself, is the harmonic mean of two inputs, the Novel Phrase metric, and the Novel Word metric. The Novel Phrase metric rewards summaries for not plagiarizing the source text via the following calculation:

$$Novel\ Phrase = 1 - \frac{\#\ of\ 5{-}gram\ matches\ between\ article\ and\ candidate}{\#\ of\ 5{-}grams\ in\ the\ candidate} \tag{2}$$

The 5-gram could be swapped for any other N-gram if the change results in a better evaluator for abstractive summaries. Then, after removing stop-words and words that occur with high frequency, the Novel Word metric is calculated as:

$$Novel\ Word = 1 - \frac{\#\ of\ unique\ 1{-}gram\ matches\ between\ article\ and\ candidate}{\#\ of\ unique\ words\ in\ candidate} \tag{3}$$

This metric rewards summaries for having novel words that are not in the article being summarized.

4 – Method

The data we will be using to evaluate MOULIN ROUGE will be generated at the next TAC/DUC conference. Similar to DUC 2003, the proposed conference will produce single document summaries of about 100 words from news articles. We will be using 500 Los Angeles Times articles published between 2018 and 2020. For each article, four summaries will be produced - three candidates and one reference. The three candidate summaries will consist of an extractive summary, a semi-abstractive summary, and a highly-abstractive summary. The extractive summary will be a lead-3 baseline produced with python scripts. The semi-abstractive summary will be produced by the hybrid pointer-generator sequence-to-sequence model proposed by Liu et al. (2017). We consider the summaries produced by this model to be only slightly abstractive, as the final model that was trained copied whole article sentences 35% of the time, whereas the manually produced reference summaries only copied data 1.3% of the time. We are assuming that this model will be able to train on the same dataset from Liu et al. (2017) and perform similarly, as a result of the model's large training set and the comparability between CNN news articles and LA Times articles. Lastly, the highly-abstractive summary and reference summary will be manually produced by human participants at the DUC. Participants formulating the highly-abstractive summaries will not be able to directly reference the article when composing the summary, so as to encourage abstraction.

After summaries are produced, DUC participants will manually rate them. The participants will be rating each summary by two scores, a content score, and an abstraction score. The content score is the same as the content-centric "readability score" used at DUC 2007. The abstraction score will be a metric for how abstract the given article is. A python script will highlight matching 5-grams or larger that match between the given summary and the source so that the human evaluator has an idea of how often phrases were directly copied. Both scores will be assessed on a scale from 1 to 5, with a 1 signifying "poor content" or "no abstraction" and a 5 representing "good content" or "high abstraction." MOULIN ROUGE and ROUGE-S scores will also be evaluated on all summaries.

To assess the efficacy of MOULIN ROUGE, we calculate the Pearson correlation coefficient between MOULIN ROUGE scores and human assigned scores. Human scores are the ground truth. The intuition is that if MOULIN ROUGE is effective, it will score summaries similar to how a human would. Thus, MOULIN ROUGE will assign high scores to summaries that maintain the important content from the reference and display a high level of abstraction. Alternatively, it will assign low scores to summaries that are entirely extractive with poor content. The model is also expected to do a good job scoring all types of summaries in-between these two extremes. We chose to include various summary types in our dataset in order to gain a wholistic idea of how MOULIN ROUGE performs across a range of different summaries. We will also compute the correlation between ROUGE-S scores and human assigned scores in order to determine whether MOULIN ROUGE more accurately models human judgement regarding abstractive summaries than ROUGE-S. If MOULIN ROUGE has a high correlation with the human scorers, then we can conclude that MOULIN ROUGE is a good evaluation metric for abstractive text summaries. Furthermore, if MOULIN ROUGE has a higher correlation with the human scorers than ROUGE-S does, we can conclude that MOULIN ROUGE is more effective at evaluating abstractive summaries than ROUGE-S.

5 – Discussion

In this paper, I introduced MOULIN ROUGE, an automatic evaluation model for abstractive summarization. This model augments ROUGE-S in order to reward abstraction while still remaining conscious of content. A considerable advantage of MOULIN ROUGE is its tuning parameter, β, which allows the user to specify the importance of content vs. abstraction. The larger the beta, the higher the weight of abstraction in the MOULIN ROUGE score. A potential limitation of MOULIN ROUGE is that ROUGE-S may be a poor metric for evaluating the content of abstractive summaries. Because ROUGE-S is sensitive to word order and word choice, it is possible that given two summaries containing equally good content, the more abstractive of the two would receive a lower ROUGE-S score. This could ultimately result in a lower than deserved MOULIN ROUGE score for a summary that is both highly abstractive and has good content, thus making MOULIN ROUGE ineffective. Albeit less popular, there are alternative summary evaluation metrics that aim to overcome ROUGE's limitations surrounding lexical dependencies. The Pyramid based method assesses a summary by aggregating content units from the summaries of a "wise crowd." This method is better positioned than ROUGE to handle semantic equivalence in evaluation of abstractive summary content. Another metric, known as ParaEval, determines semantic closeness through a domain-independent paraphrase table. Like ROUGE, neither of these alternatives measure for level of abstraction. A follow-up would be to replace ROUGE-S in our model with ParaEval and the Pyramid based method to determine if either would result in a higher correlation with the human judges. Another possibility I would like to investigate builds on neural-network based approaches in which words are "embedded" into a low dimensional space. In these models, each word is represented as a d-dimensional vector, and vectors close to one another are shown to be semantically related. I propose utilizing this approach to judge the content of summaries. We would still need to address an augmentation to measure for abstraction, but this approach might allow us to develop a worthy competitor to MOULIN ROUGE for furthering progress in abstractive text summarization.

References

Document Understanding Conferences. https://www-nlpir.nist.gov/projects/duc/index.html, last accessed 25 Nov 2020.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *In Text Summarization Branches Out: Association for Computational Linguistics Workshop*.

Liu, P.J., C.D. Manning, and A. See. 2017. *Get To The Point: Summarization with Pointer-Generator Networks.* Cambridge University Press.

Nenkova, A., R. Passonneau, and K. McKeown. 2007. *The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation*. ACM Transactions on Speech and Language Processing.

Zhoe, L., C.-Y. Lin, and D.S. Munteanu. 2006. *ParaEval: Using Paraphrases to Evaluate Summaries Automatically*. Association for Computational Linguistics.