Developing New Methods for Evaluation of Document Summaries

by

Sofu, Burak & Gültop, Şükrü Can

Submitted to the Department of Computer
Engineering in partial fulfillment of
the requirements for the degree of
Bachelor of Science

Undergraduate Program in Computer Engineering
Boğaziçi University
Spring 2019

Developing New Methods for Evaluation of Document Summaries

APPROVED BY:

Tunga Güngör .....
(Project Supervisor)

DATE OF APPROVAL:

ABSTRACT

Developing New Methods for Evaluation of Document Summaries

Text summarization is the task of forming a summary of a given document. There are two types of summaries: extracts (just selecting some of the important sentences in the document as the summary) and abstracts (generating new sentences from scratch to form the summary). In this project, we deal only with extract type of summarization. The classical metric used to evaluate the success of a summarization system is the Rouge metric.

By reading the paper "ROUGE: A Package for Automatic Evaluation of Summaries", Lin, C-Y., Workshop On Text Summarization Branches Out, 2004, we used summaries and reference documents to measure goodness of evaluation and implemented a new method for evaluation of document summaries using a specialized version of edit distance method for words and created scores to correlate with Rouge.

ÖZET

Doküman Özetlerinin Değerlendirilmesinde Yeni Yöntemler Geliştirilmesi

Metin özetleme, verilen bir belgenin özetini oluşturma görevidir. İki tür özet vardır: özetler (sadece belgedeki önemli cümlelerin bir kısmını özet olarak seçerek) ve özetler (özeti oluşturmak için sıfırdan yeni cümleler oluşturarak). Bu projede sadece özet türlerini ele alıyoruz. Bir özetleme sisteminin başarısını değerlendirmek için kullanılan klasik ölçüm, Rouge ölçüsüdür.

"ROUGE: A Package for Automatic Evaluation of Summaries" adlı makaleyi okuyarak, Lin, CY., Workshop On Text Summarization Branches Out, 2004 değerlendirmesini ölçmek için özetleri ve referans belgeleri kullandık ve daha iyi ölçüm için "edit distance" yönteminden yararlanarak yeni bir çözüm geliştirdik.

1. INTRODUCTION AND MOTIVATION

In this project we tried to evaluate the goodness of the existing extraction summary evaluation methods, mainly ROUGE which is the main method for many years. It's already a widely accepted and succesful method, developped my many programmers over the years. Since it's a package, there're additions to it which extends the tests to address all the cases.

There has been surveys on the subject, automated summary evaluation, most are considering complex methods to improve quality of evaluation. On the contrary the ROUGE package is quite simple and made simple by purpose as it's the most general way to evaluate. If we'd have ignored the ROUGE and try to develop a base for summary evaluations, we would end up with a package similar to ROUGE as many others would also.

Considering this trying to find out a simple measure to add into ROUGE but also sufficiently different and a new approach is not an easy task. Early thoughts always ends up with somewhat complex algorithm that takes many things into account to get a closer measure to evaluate better.

We wanted to try and find a possible weak point of ROUGE to start with and develop an idea to overcome that part that lacks true evaluation. The thing we focused is to inspect the documents sentence wised. How do the sentences differ from each other? These were the questions we asked.

During this project our main motivation was to create something that will or could be used by others. Better evaluation will improve the area in every way since it would show more certainly that the works are on the right path. While this project has nothing to do with the actual summarization, with the help of this new additions, summarization process will be more accurate if we correctly address missing parts of the existing ROUGE methods. It will help future works of NLP related AI studies.

2. STATE OF THE ART

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an automatically produced summary or translation against a set of reference summaries which are human created.

There're three main measures in the ROUGE package:

(i) ROUGE-N – Measures unigram, bigram, trigram and higher order n-gram overlap

(ii) ROUGE-L – Measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.

(iii) ROUGE-S – Is any pair of word in a sentence in order, allowing for arbitrary gaps. This can also be called skip-gram coocurrence. For example, skip-bigram measures the overlap of word pairs that can have a maximum of two gaps in between words. As an example, for the phrase "cat in the hat" the skip-bigrams would be ***"cat in, cat the, cat hat, in the, in hat, the hat."***

In the technical side we evaluate differences of these methods by looking at the recall and precision calculations. **Precision** tells us *how many of the selected words were correct.* On the other hand, **recall** tells us *how many of the words that should have selected were actually select.*

Precision and recall are two opposite sides of the measure. Weighing one will decrease the other one. In the standart ROUGE 2.0 package that we're using which is implemented in Java defaults 0.5 ratio so that precision and recall are equally important on the measure.

We can also consider extreme cases:

- **Boost precision:** If summarization has picked the nearly all the true words but did this with plenty of unnecessary words the precision will be low thus it will score low. On the other hand, it could get a high score even if it didn't get the whole context but accurate with a few words on the hand.
- **Boost recall:** This is the complete opposite. It will score high only if summarization has got the matching words in the reference document. It will not care the word count.

There's a research that disscuss ROUGE's goodness. It says that "it has been shown to correlate well with human judgements, it is biased towards surface lexical similarities". This means that humans tends to care more about the meaning while ROUGE just measures similarity based on words overlapping or in sequence.

There has been on-going efforts to improve on automatic summarization evaluation measures, such as the Automatically Evaluating Summaries of Peers (AESOP) task in TAC (Dang and Owczarzak, 2009; Owczarzak, 2010; Owczarzak and Dang, 2011). However, ROUGE remains as one of the most popular metric of choice, as it has repeatedly been shown to correlate very well with human judgements (Lin, 2004a; Over and Yen, 2004; Owczarzak and Dang, 2011).

In this project, we tried to find a new measure which also correlate well with human judgements. When we did the gold scoring 100 documents to create a human score we realized that the most important thing humans care when scoring is to match the most important sentences of reference summary in the generated summary. It has to capture the main theme as well as the important sentences or as near as possible. We thought this could be possible if we modify edit distance to use words instead of characters and compare similar sentences of the both documents to measure if they're similar enough.

3. METHODS

To observe the lacking and successful parts of the current ROUGE metrics and experiment with the text summary pairs we need large texts, their reference summaries which are summarized by humans, and extracted summaries which are summarized by summarization tools.

The texts and reference summaries we need for summarization are collected from well-known data sets which are CNN news and BBC News data sets. To give those texts and their reference summaries to summarization tools as inputs we needed to format them in the form of summarization tools and ROUGE evaluation tool wanted. To achieve this we formatted the name and directory structure of the files.

We mainly used BBC news dataset for out observations, because the reference summaries of the BBC news dataset are also summarized as extractive summary. With this reference summaries we could see easily which aspects of the current rouge metrics lacking. We used CNN news dataset as our test and scoring dataset, because it gives us more realistic results.

After preparing our data, we get candidate summaries by our two systems which are gensim and pyteaser. Both uses the text rank algorithm to extract the texts. We tried our system summaries and reference summaries are consistent by fixing system summary lengths to reference summary lengths.

To observe ROUGE metrics and their conclusions, we used ROUGE 2.0 tool, which is re-implementation of original ROUGE tool which is provided by Chin-Yew Lin. To use this tool, we formatted our files and directory format to expected format of ROUGE 2.0 tool. Then, we analyzed our candidate summaries with ROUGE 2.0. As a result of this evaluation we had our recall, precision and f-score metrics for each ROUGE statistics.

After preparing all the data we need to analyze the ROUGE statistics and metrics, we started to determine the success of those statistics and metrics. To do this, we got the texts, reference summaries, and candidate summaries and compare them as humans. We did not classify documents to make process to be more fair. We did not know which summary is reference and which is candidate. With completing this phase, we expect to observe some shortcomings and success of ROUGE and improve them. We completed this phase with a hundred samples from our CNN dataset.

After getting all the necessary statistics of our systems and rouge, we observed some shortcomings of the rouge scoring like mentioned in the Chin-Yew Lin article. We tried to overcome some fallbacks of the rouge scores. In the n gram scores, the orderings of n-grams are not important like in the following example, 2-grams generates the same score for 2nd and 3rd sentence but the meanings are opposite.

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police

In the below example rouge-l the longest common sub sequence is calculated as same but

X: [A B C D E F G]
Y1: [A B C D H I K]
Y2: [A H B K C I D]

Another lacking side of the rouge-l is only counts the main in-sequence words; therefore, other alternative LCSes and shorter sequences are not reflected in the final score. S4.

The above algorithm compares all the sentences of candidate and reference summaries. For each reference summary sentence it chooses the best candidate sentence

Algorithm 1 Edit Distance Score algorithm

procedure ROUGE EDIT DISTANCE(*candidate*,reference) ▷ The g.c.d. of a and b
  $m \leftarrow words(reference)$ ▷ Word count of reference summary
  $n \leftarrow words(candidate)$ ▷ Word count of candidate summary
  $totalEditDistance \leftarrow 0$ ▷ total edit distance
  for x in sentences(candidate) **do** ▷ For each sentence in candidate
    $eds \leftarrow []$ ▷ edit distances list
    for y in sentences(reference) **do** ▷ For each sentence in reference
      $ed \leftarrow LevenshteinDistance(x, y)$ ▷ edit distance of two sentences
      $eds.append(ed)$ ▷ adding to edit distances
    $totalEditDistance+ = min(eds)$ ▷ most similar two sentences
  $p = max(0, 1 - \frac{totalEditDistance}{n})$ ▷ precision
  $r = max(0, 1 - \frac{totalEditDistance}{m})$ ▷ recall
  $f = \frac{2 \times p \times r}{p+r}$ ▷ f-1 score
  return p, r, f ▷ The gcd is b

which means minumum edit distance. And sums all those minumum edit distances. Note that the edit distance here means word edit distance not character edit distance. After finding difference(edit distance) we used a formula like q-gram similarity to compute precision and recall.

4. RESULTS

We are actually trying to improve the ROUGE which is state-of-the-art. Current results of ours show that correlations with human scoring may sometimes have bad results. This was the main reasons for us to start this project. We compared results of ROUGE as humans, and concluded that rouge metrics can be improved with some much more advanced semantic techniques with using the state-of-the-art. But we tried to combine all the good sides of the metrics and overcome disadvantages while making metric as simple as possible like rouge to make it more robust and basic. With this approach, we come up with edit distance idea like we mentioned in the methods part. The scores which are generated by our edit distance algorithm and other rouge metrics without stemming can be seen in the following correlation heat map. We named our

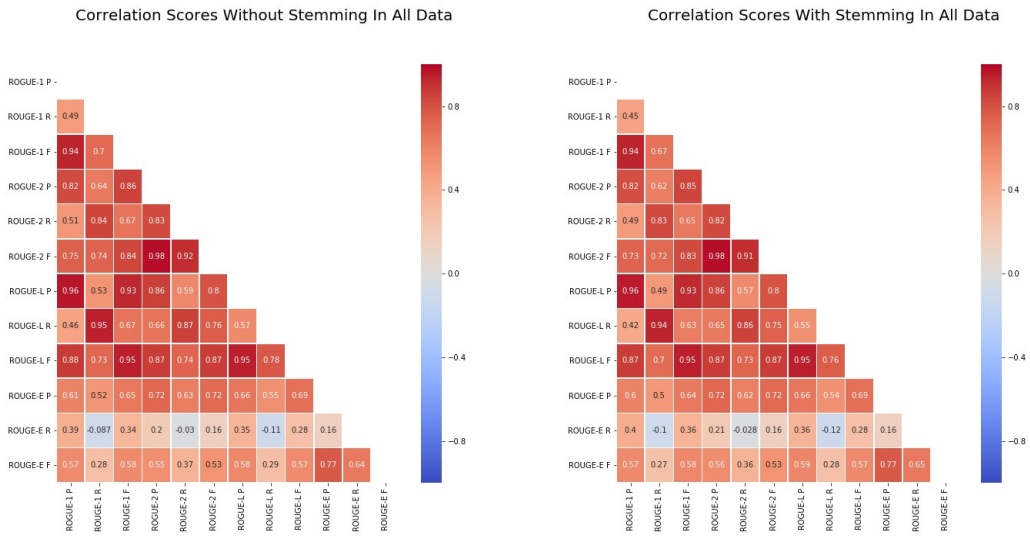

Figure 4.1. Correlation Scores with other rouge metrics.

metric as Rouge-E as you can see in the correlation heat maps. The first thing that takes our interest with those scores was our recall correlation with other rouge recalls. We discussed about it and concluded it is mostly because of the formulation of our edit distance metric. To make it more observable and comparable with rouge metrics we decided to score our metric as precision, recall and f score, but this decision came with a one flaw. The maximum edit distance of two separate sentences is the maximum count

of words of two sentences. If the calculated edit distance is bigger than denominator, this causes to our recall or precision to be negative, we handled it with setting minimum score of zero. We should not see this error if we used similar length summaries with reference summaries, but we wanted to observe exactly this type of behaviour with multiple scoring(precision, recall, f-1). The formula can be improved by making it prune to this flaw.

After observing those we also calculate correlations between human scores and all rouge scores including our new Rouge-E score. To make those calculations we scored 200 sample summaries(100 of pyteaser, 100 of gensim). You can see the correlations with human scores in the following heat map.

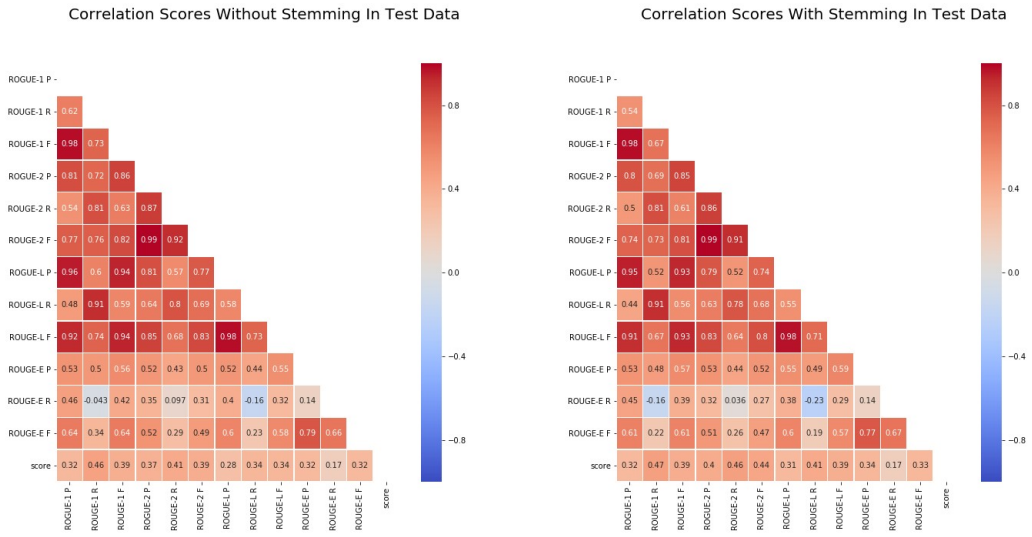

Figure 4.2. Correlation Scores with other rouge and human scores.

The main improvement for edit distance was for improving Rouge-L. We can see that the correlation score of precision is bigger than Rouge-L's score. The recall shortcoming shows its results here, too. In general we can say it performed fine compared other scores.

5. CONCLUSION AND DISCUSSION

Summarization systems are used in various places in today's world. Currently we are interpretting our summaries according to original text and we see the metrics used in ROUGE are not perfect in the sense of context. They are useful for deciding which summary is better than other one but we see that they don't show very good results with their scores according to original texts.

The most observable and maybe the most important usage of summarization systems is the world of internet. If we can improve the evaluation of summary successes, we can improve the ranking systems and thanks to that we can even improve recommendation systems which is used in very wide area.

Test results show that there's a correlation between ROUGE-E and other metrics as well as human score. Stemming has better scores but didn't affect correlations by much. The correlation with the existing ROUGE metrics are due to our implementation of precision and recall scores to ROUGE-E which has flaws in calculating since the reference and system summaries aren't of the same length. Considering our test data reference summaries are shorter than the system summaries that we're measuring. The created some issues especially in there call calculation. That's why recall correlation is weaker than the precision's.

In this project we created a new metric for the ROUGE package and correlation shows the success of it. It has been a great effort but it needs work to do to be accepted by world wide NLP communities. More people and greater effort creating a systematic scoring in human scores would make correlations more meaningful with the mentioned scoring problems solved.

6. FUTURE WORK

We need to address shortcomings of the current work, starting with the scoring system. More solid calculation of precision and recall would be needed to show it's correlation with ROUGE-1, ROUGE-2, ROUGE-L etc.

New metrics also can be thought and developed. The main thing we kept in mind is that we're trying to extend ROUGE, not creating completely new methodology for document summary evaluations. So it has to keep things simple and should be comparable with other metrics and show a correlation. Bigger and more systematic human score done by professionals would also help it to gain international recognition and thus it would produce more valuable outputs.

We could also add semantic weight with WordNet. Multiple reference evaluation implementation would improve the results. As to edit distance formula it could be improved by considering not only best matched sentences but all or part of the sentences if we're to keep working on perfecting the formula.

□

REFERENCES

1. Lin, C.-Y., "ROUGE: A Package for Automatic Evaluation of summaries", p. 10, 01 2004.

2. Ganesan, K., "ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks", , 2015.

3. Ganesan, K., C. Zhai and J. Han, "Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions", *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 340–348, Association for Computational Linguistics, 2010.

4. Barrios, F., F. López, L. Argerich and R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization", *CoRR*, Vol. abs/1602.03606, 2016, `http://arxiv.org/abs/1602.03606`.

5. Řehůřek, R. and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, ELRA, Valletta, Malta, May 2010, `http://is.muni.cz/publication/884893/en`.

APPENDIX A: DATA AVAILABILITY STATEMENT

https://www.dropbox.com/s/quios842qv8g1qb/DUC.rar?dl=0