The ROUGE-AR: A Proposed Extension to the ROUGE Evaluation Metric for Abstractive Text Summarization

Sydney Maples

Symbolic Systems Department, Stanford University

smaples@stanford.edu

Abstract

Abstractive text summarization refers to summary generation that is based on semantic understanding, and is thus not strictly limited to the words found in the source. Despite its success in deep learning, however, the task of text summarization has no reliably effective metric for evaluating performance. In this paper, we describe the standard evaluative measure for abstractive text summarization, the ROUGE metric. We then propose our extension to the standard ROUGE measure, the ROUGE-AR. Drawing from methodologies pertaining to latent semantic analysis (LSA) and part-of-speech tagging, the ROUGE-AR metric reweights the final ROUGE output by incorporating both anaphor resolution and other intrinsic methods that are largely absent from non-human text summary evaluation.

1. Introduction

Broadly defined, the goal of text summarization is to generate a short text summary from a source document, with the ideal summary consisting of none other than the most informative and relevant ideas from the source document. In the field of natural language processing (NLP), the majority of text summarization tasks can be categorized as either extractive and abstractive. *Extractive summarization* refers to summary generation that only uses the words fed from the input text, whereas *abstractive summarization* refers to summary generation based on semantic understanding, and thus is not limited to the words received as input. Other methods include deletion, in which words from the original text are deleted (thus preserving word order), and text extraction using features such as word or phrase frequency, key phrases, and word position in the text. [1]

One notable issue in the task of text summarization is the near-absence of a universal strategy to evaluate summarization systems. Additionally, there is less research in the area of abstractive summary evaluation;

most of summarization research thus far has focused on extractive tasks, as extractive measures can readily make use of pre-existing machine learning tools (such as Naïve Bayes classification and hidden Markov models). Thus, when it comes to abstractive summarization, the options for evaluation are even less idyllic. The most commonly-cited metric appears to be ROUGE [2], a package of metrics developed by the Document Understanding Conference (DUC). Hence, within the realm of abstractive summarization (and for the purposes of this project), ROUGE is considered the standard ("baseline") evaluation metric. We will go into more detail regarding ROUGE later in this section.

Given that humans tend to construct their own summaries 'in their own words' (i.e. not necessarily using words from the source text), human-curated summaries tend to be more abstractive in nature. In approaching this issue, we consider the recent success and popularity of deep learning techniques. In August 2016, Peter Liu and Xin Pan of Google Brain published a blog post called "Text summarization with Tensorflow". The post included a link to a repository containing their text summarization model, a recurrent neural network (RNN) encoder-decoder with attention mechanism. Liu and Xin trained this model to produce headlines for news articles, and then used the ROUGE-L metric for evaluation. [3]

In this paper, we will introduce an extension to the standard ROUGE metric called "ROUGE-AR". The "AR" refers to "Anaphora Resolution", an intrinsic measure the baseline metric does not account for. For our model, we implemented the RNN encoder-decoder provided in [3]. However, while testing our evaluation metric, we found that fine-tuning the parameters of the metric could be done more effectively when we could make use of example summaries that directly challenged anaphora resolution and text-quality performance. Before discussing our approach, we will first provide an overview of the more commonly-used evaluation metrics for modern-day text summarization.

1.1 Overview of Existing Evaluation Metrics

Summary evaluation measures can be divided into two categories: *intrinsic* and *extrinsic*. Intrinsic measures seek to assess *text quality* and *content evaluation*. Text quality measures the readability of a summary, and is typically evaluated manually (i.e. by using human judges). Content evaluation is measured differently depending on what is being summarized: sentence extracts are often evaluated using *co-selection*, while human abstracts often achieve better results with *content-based measures.*

In contrast to intrinsic measures, extrinsic measures essentially use *task-based methods*, such as question answering and information retrieval, in order to measure the performance of a summary for use during a specific task. For the purposes of this project as it pertains to abstractive text summarization, we are mostly interested in intrinsic measures.

1.1.1 Co-Selection

In co-selection, the main evaluation metrics are precision, recall, and F1-scores. For summary evaluation, we define *precision* (p) as the total sentence count in both system and ideal summaries divided by the total sentence count in system summaries. We define *recall* (r) as the total sentence count in both system and ideal summaries divided by the total sentence count in ideal summaries. The *F1-score* is computed as the harmonic mean of precision and recall.

$$F_1 = \frac{2pr}{p+r}. \qquad (1)$$

An evaluation based solely on a system's F1-score runs into the problem of determining which words are the most important in a source document. As an example, suppose that a manual summary for a source contains only words from set X. Suppose we have two systems: system A and system B. If system A writes a summary using only words from set X, and system B writes an equally-good (based on human evaluation) summary using words from set Y, the F1-score would be much higher for system A than system B because the summary produced by the former most closely matches the manual summary. To combat this issue, a *relative utility (RU)* measure may be incorporated into the metric. [4] The RU measure represents all sentences in a summary with confidence values, or sentence *utility,* that indicate the degree to which the given sentence should be a part of the summary. Formally, RU is defined as follows:

$$RU = \frac{\sum_{j=1}^{n} \delta_j \sum_{i=1}^{N} u_{ij}}{\sum_{j=1}^{n} \epsilon_j \sum_{i=1}^{N} u_{ij}}, \qquad (2)$$

With the addition of the RU measure, both the summaries would be regarded as equal and optimal, as no other combination of two sentences contains a higher utility.

1.1.2 Content-Based Measures

Co-selection measures are limited in that they can only match the exact same sentences; thus, these measures do not account for the fact that two summaries may contain the same information while using different words. For example, consider the following two summaries to describe a situation in which a customer must select between chocolate or vanilla ice cream:

(1) The customer selected chocolate.

(2) She didn't choose vanilla.

These two summaries are equally informative, yet they are worded completely differently. F1-scores would fail to capture this semantic difference. Thus, a metric for evaluating semantic similarity between summary vectors is warranted. In general, such a metric may be described as a comparison between term frequency (tf) vectors of source extracts to tf vectors of the full source text. Thus, summaries which disagree due to synonymy (such as the two example sentences above) may still contain similarly-distributed tfs, thus reflecting their semantic equivalence. Furthermore, summaries that contain synonyms that are infrequently distributed throughout the text will not be penalized in a comparison between an extract and the full source document.

The most basic of these measures is based on the vector space model [5]. It uses the inner product of document vectors to measure content similarity geometrically, using the cosine of the angle between two document vectors X and Y.

$$\cos(X,Y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i (x_i)^2} \cdot \sqrt{\sum_i (y_i)^2}}, \quad (3)$$

Since cos(0) = 1, document vectors with high cosine-similarity are considered similar. This measure can be enhanced further by term weighing using log-entropy or increasing the number of documents that the tf is calculated over.

Another content-based measure is the Longest Common Subsequence (LCS), in which X and Y represent sequences of words, and lcs(X,Y) is the longest common subsequence between X and Y. It is calculated as follows

$$lcs(X,Y) = \frac{length(X) + length(Y) - edit_{di}(X,Y)}{2},$$

where length(X) is length of string X, and edit(X,Y) is the edit distance between X and Y.

The third and most commonly used metric for summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which was introduced by DUC in 2003. [2] It is a package of metrics that is based on the similarity of *n*-grams that both the reference summary and the test summary have in common. ROUGE-N divides the number of matching *n*-grams by the number of total *n*-grams in the reference summary. Formally, N-ROUGE is defined as:

$$\frac{\sum_{S \in ReferenceSummary} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in ReferenceSummary} \sum_{n-gram \in S} Count(n-gram)}$$

Notably, the ROUGE metric was used to evaluate the model that we will be using [2], and will thus be incorporated as a baseline measure to determine the effectiveness of our own metric.

1.2 Proposal: ROUGE-AR

As ROUGE is a content-based measure, it does not directly measure a summary's readability. A text's *readability* is typically defined by its grammaticality, non-redundancy, referential clarity, and coherence of a summary. Furthermore, it fails to capture semantic context and a host of syntactic information that is latent in the text, such as anaphora (pronoun) resolution.[1]

With this in mind, we propose an evaluation metric, ROUGE-AR, that extends ROUGE to incorporate both anaphor resolution and measures for summary readability. In particular, we can use ROUGE-AR to tackle the following issues that standard ROUGE does not consider:

- ***Text redundancy and coherence.*** For this, we drew inspiration from *latent semantic analysis* (LSA), a method used to capture the main ideas from a text. Using LSA, we extracted the text from each input document and converted it to a sparse sentence-term matrix, which was then processed through singular value decomposition (SVD) as a dimensionality reduction method.[2] After separating each of the summaries into individual sentences, the vector for each sentence was computed (as the weighted sum of its weighted terms) and then was compared to the vector for the next sentence in the text. In determining the vectors, the English stop words indicated by the NLTK[3] library were omitted from the analyses. The cosine between these two vectors indicated their semantic relatedness or coherence. An overall coherence measure was then calculated for each text by averaging the cosines between the vectors for all pairs of adjoining sentences. As this measure only captures redundancy and coherence between adjacent sentences and not within each individual sentence, a separate measure was used to count other measures of redundancy, such as repeated words.

- ***Referential clarity***. Referential clarity was primarily assessed by the summaries' compliance with Chomsky's binding theory (Chomsky 1981).[4] Though we could not find a standard score for measuring the referential clarity of a summary, we ultimately used a F1-like scoring metric called a *reference score*, in addition to the baseline F1-score. We define the reference score (RS) as the number of correctly-resolved referential pronouns (RC) divided by the sum of the number of

[1] Though difficult to define (and defined variously throughout the linguistic literature), we formally recognize 'anaphora resolution' as a form of pronoun resolution in which the antecedents are the primary identities of interest.
[2] The use of SVD can reduce noise, model the relationship between words, and find a relationship (if any) between words and sentences.
[3] Further documentation for nltk can be found at 'www.nltk.org'.
[4] "Binding theory" is a theory of syntax that is concerned with how the interpretation of noun phrases (NPs) is constrained by syntactic considerations ("binding").

pronouns in the document (A) and the number of unresolved referents (U). Similar to standard F1, the reference score was weighted by an alpha value.[5]

- ***Anaphora resolution.*** This measure is very similar (and perhaps equinamous) to referential clarity. We calculated the number of anaphora candidates by tagging each pronoun (tagged as '2') and noun phrase (tagged as '1') We then built a vector space to represent each word in the document, populated with either 0 (not anaphora), 1 (noun phrase), or 2 (pronoun). Using recursive backtracking, for each pronoun in the anaphora matrix, we identified the noun phrases which preceded the anaphor within a distance of two sentences (or until we hit another pronoun). We then cross-checked to ensure that noun phrases and the pronoun agreed on gender, person (first, second, third), and number (singular or plural).[6]

1.2.1 Related Work and Other Evaluation Metrics

LSA is used in both evaluation and summarization tasks, and has enjoyed recent success in both. [7] When used as a summarization utility, LSA is quite good at performing anaphora resolution. However, LSA as an evaluation metric underperforms when compared to ROUGE on abstractive evaluation, as it was designed to compare summaries within the full source document.

Additionally, SERA (Summarization Evaluation by Relevance Analysis) has proven to be an effective evaluation metric for scientific article-based summaries. [8] SERA focuses primarily on the content relevance between an automatically-generated summary and the corresponding human written summaries. In this way, SERA relies less on word choice and lexical overlapping than does ROUGE.

Alongside ROUGE, Microsoft has also introduced ParaEval, which evaluates semantic equivalences more sensibly than does the former metric. [9] Thus, paraphrased sentences are punished less harshly under ParaEval than those evaluated under ROUGE.

2. Experiments and Results

2.1 Data Extraction and Preprocessing

We used the English Gigaword Fifth Edition corpus. The corpus was provided courtesy of Stanford University and was obtained through AFS. More information about the corpus and its contents can be found on the Linguistic Data Consortium catalog. [10] As the original built by Liu and Pan was built around the annotated edition of the corpus, a separate script was written to reformat the data to support the provided training scripts. From here, the vocabulary list was cut to include only the 200,000 most

[5] When $\alpha \approx 0$, the score favors recall; when $\alpha \approx 1$, the score favors precision.
[6] This method was originally proposed by Mitkov (1998). [6]

frequently-used words.

2.2 Training the Model

During training, the model uses the first sentence from the article as input in order to generate the headline. Articles with headlines or summaries containing less than two words or more than 120 words were dropped during training. We used a minibatch size of 64. When decoding, we used a beam search algorithm (with a beam size of 4) to find the best headline from the candidate headlines generated by the model.

Within the model itself, we used 256 LSTM hidden units with four bidirectional encoding layers, as these were the parameters that had reportedly achieved the best performance when evaluated by ROUGE.

As the training took a fair bit of time (often around 10 hours on a single GPU), we reduced the maximum number of steps down from 100,000 to 2,000 steps. With this change, the model would finish training at around ~10 hours, but the training loss would never converge (the lowest loss we received was 2, and this was via intentional overfitting). We adjusted and readjusted the parameters accordingly (batch size [ranged from 4 to 64], number of epochs [ranged from 50, to 240, to ~5000] and learning rate [anywhere from 0.001 to 0.2]) and even used a variety of datasets and vocabulary files, yet the loss function was unable to converge. As the model was not of our own creation and we could not come into contact with the original authors, the RNN was inevitably abandoned in the final implementation of our evaluation metric. This ended up being fine, as the model was not essential to the evaluation metric itself.

2.3 Testing Metric Functionality

To ensure that our metric was working effectively, we ran several baseline tests using datasets that sought to exploit specific properties. Notably, we built a data set that violated syntactic binding constraints. Sentences in this set were formatted like this:

(1) John lets Mary wash himself.
(2) Jane thinks that Joe owed her the money back.
(3) The boy rode her bike.

The first sentence is grammaticality incorrect (in terms of the antecedent relative to the reflexive pronoun). The second sentence is a case of improper tense relative to other verb tenses, and the third sentence is a case of conflicting gender pronouns (from a purely syntactic standpoint).

The anaphora resolution extension was successfully able to pick up on sentences of the form (1) and (3),

but it did not pick up on (2). Upon further observation, it seems that this is because the resolution extension interprets grammaticality per-pronoun and within-noun-phrases; as such, verb-phrase tenses are largely ignored by the anaphora resolution algorithm.

Aside from these and similar caveats, the anaphora resolution metric operated as expected. It is worth noting that the evaluation metric is not designed to deal with subsequent phora (namely, cataphora).

2.4 Performance, Successes, and Limitations

DUC, the conference that organizes ROUGE, held a document-summarization task in 2002 in which several systems and human judges produced summaries from a corpus of 567 documents. The DUC conference also provided 100-word baseline summaries, also assessed by humans. To assess the effectiveness of stand-alone LSA, we included the LSA class (incorporated into the Pythonic ROUGE class) as a separate evaluation metric while training on these summaries. We had to eliminate two of these systems so as not to bias the data (as they only produced headlines and not full summaries):

Table 1. Evaluation on DUC Summaries (2002)

ROUGE Variant	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L
ROUGE	0.116	0.119	0.122	0.116
ROUGE+LSA	0.088	0.089	0.091	0.85
ROUGE-AR	0.108	0.106	0.1	0.07

Though it is difficult to judge the performance of a metric based solely on these numbers, the previous confirmation that the evaluation metric was working as it should would lead us to believe that the metric is penalizing the features, and reweighing the scores as such. The large size of the document data makes it difficult to calculate individual scores.

Next, we tested the metrics on a small set of Apple product reviews (generated using abstractive text summarization from the NAMAS repository on github). This was to see how well the metrics performed on small data sets:

Table 2. Evaluation on Apple Product Reviews

ROUGE Variant	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L
ROUGE	0.399	0.399	0.399	0.241
ROUGE+LSA	0.220	0.220	0.220	0.19
ROUGE-AR	0.39	0.399	0.399	0.241

Here. It wasn't until later that we realized that it wasn't a fault with the implementation itself, but rather how we implemented the standard ROUGE metric into Python. The original ROUGE metric is implemented in a Perl script and requires all inputs to be passed through files on a disk. As this is clearly a cumbersome move for Python, we instead opted to separate each summary on a per-line basis, though this can cause a host of preprocessing and string tokenization issues, which greatly affects the sentence tagging implementation (and, thus, the anaphora resolution system). With these Apple Product reviews, all punctuation had been stripped, and thus the implementation of ROUGE we had built did not know where to locate the ends of each sentence when tokenizing the summary sentences for evaluation. To resolve this issue, it would be best to extend the ROUGE model to accommodate a variety of different file-types (i.e. summaries separated by lines, one-sentence summaries, summaries separated by documents, etc.).

In terms of successes, we think that ROUGE paired nicely with its referential score and anaphora resolution. Though a sophisticated model for anaphora resolution would be very complex (and far outside the scope of the two-week period provided by the class), this would likely be an interesting side-project to work on in the future.

3. Conclusion

3.1 Future Trajectories

There are a few things we'd like to implement into a future model. Instead of simply taking off points for repeated words on a per-anaphora-chain basis, we think it improve the metric even more to include a *brevity penalty*, which instead penalizes system summaries that are shorter than the average length of a reference summary. This has already been added to certain measures (such as BLEU, a metric similar to ROUGE, though it measures precision rather than recall).

Furthermore, we believe that there is room to reconstruct the standard ROUGE metric such that its file

input system is python-friendly and efficient. Aside from implementing a more complex pronoun resolution framework, we believe that reformatting the standard ROUGE metric for would be the first step in a very worthwhile path – particularly when considering the possibilities for extension that exist for the premier ROUGE metric.

Acknowledgements

Special thanks to Abi See for her wisdom and excellent mentorship.

References

[1] Das, D., and Martins, A. 2007. *A survey on automatic text summarization*.

[2] Lin, C. (2004) ROUGE: A Package for Automatic Evaluation of Summaries.Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain.

[3] https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html

[4] Radev, D. Jing, H. Budzikowska, M.: *Centroid-Based Summarization of Multiple Documents*. In ANLP/NAACL Workshop on Automatic Summarization, Seattle, USA, 2000.

[5] G. Salton, A. Wong, and C. S. Yang (1975), *A Vector Space Model for Automatic Indexing*. Communications of the ACM, vol. 18, nr. 11, pages 613–620.

[6] R. Mitkov. 1998. *Robust pronoun resolution with limited knowledge*. In Proceedings of COLING. Montreal.

[7] Steinberger, J., Jezek, K.: *Evaluation measures for text summarization*. Computing and Informatics 28(2), 251–275 (2009)

[8] A. Cohan and N. Goharian "Revisiting Summarization Evaluation for Scientific Papers", In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), May 2016. (SERA)

[9] Liang Zhou, Chin-Yew Lin, Dragos S. Munteanu, and Eduard Hovy. *ParaEval: Using Phrases to Evaluate Summaries Automatically*. In *Proceedings of 2006 Human Language Technology Conference (HLT-NAACL 2006)*, New York, USA, June 5 – 7, 2006.

[10] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, June. LDC2011T07.
