What is ROUGE and how it works for evaluation of summarization tasks?

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** (typically human-produced).Let us say, we have the following system and reference summaries:

System Summary (what the machine produced):

```
the cat was found under the bed
```

Reference Summary (gold standard – usually by humans) :

```
the cat was under the bed
```

If we consider just the individual words, the number of overlapping words between the system summary and reference summary is 6. This however, does not tell you much as a metric. To get a good quantitative value, we can actually compute the **precision** and **recall** using the overlap.

Precision and Recall in the Context of ROUGE

Simply put, Recall in the context of ROUGE means how much of the *reference summary* is the *system summary* recovering or capturing? If we are just considering the individual words, it can be computed as:

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

In this example, the Recall would thus be:

$$Recall = \frac{6}{6} = 1.0$$

This means that all the words in the **reference summary** has been captured by the **system summary**, which indeed is the case for this example.

Whoala! this looks really good for a text summarization system. However, it does not tell you the other side of the story. A machine generated summary (system summary) can be extremely long, capturing all words in the reference summary. But, much of the words in the system summary may be useless, making the summary unnecessarily verbose. This is where precision comes into play. In terms of precision, what you are essentially measuring is, *how much of the system summary was in fact relevant or needed*? Precision is measured as:

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

In this example, the Precision would thus be:

$$Precision = \frac{6}{7} = 0.86$$

This simply means that 6 out of the 7 words in the system summary were in fact relevant or needed. If we had the following system summary, as opposed to the example above:

System Summary 2:

```
the tiny little cat was found under the big funny bed
```

The Precision now becomes:

$$Precision = \frac{6}{11} = 0.55$$

Now, this doesn't look so good, does it? That is because we have quite a few unnecessary words in the summary. The **precision** aspect becomes really crucial when you are trying to generate summaries that are concise in nature. Therefore, it is always best to compute both the **Precision** and **Recall** and then report the **F-Measure**. If your summaries are in some way forced to be concise through some constraints, then you could consider using just the **Recall** since precision is of less concern in this scenario.

So What is ROUGE-N, ROUGE-S & ROUGE-L ?

ROUGE-N, ROUGE-S and ROUGE-L can be thought of as the granularity of texts being compared between the system summaries and reference summaries. For example, **ROUGE-1** refers to overlap of ***unigrams*** between the system summary and reference summary. **ROUGE-2** refers to the overlap of ***bigrams*** between the system and reference summaries. Let's take the example from above. Let us say we want to compute the **ROUGE-2 precision and recall** scores.

System Summary :

```
the cat was found under the bed
```

Reference Summary :

```
the cat was under the bed
```

System Summary Bigrams:

```
the cat,
cat was,
was found,
found under,
under the,
the bed

```

Reference Summary Bigrams:

```

the cat,
cat was,
was under,
under the,
the bed

```

Based on the bigrams above, the ROUGE-2 recall is as follows:

$$ROUGE2_{Recall} = \frac{4}{5} = 0.8$$

Essentially, the system summary has recovered 4 bigrams out of 5 bigrams from the reference summary which is pretty good! Now the ROUGE-2 precision is as follows:

$$ROUGE2_{Precision} = \frac{4}{6} = 0.67$$

The precision here tells us that out of all the system summary bigrams, there is a 67% overlap with the reference summary. This is not too bad either. Note that as the summaries (both system and reference summaries) get longer and longer, there will be fewer overlapping bigrams especially in the case of abstractive summarization where you are not directly re-using sentences for summarization.

The reason one would use ROUGE-1 over or in conjunction with ROUGE-2 (or other finer granularity ROUGE measures), is to also show the fluency of the summaries or translation. The intuition is that if you more closely follow the word orderings of the reference summary, then your summary is actually more fluent.

Short Explanation of a few Different ROUGE measures

- **ROUGE-N** – measures unigram, bigram, trigram and higher order n-gram overlap
- **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.
- **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"**.

For more in-depth information about these evaluation metrics you can refer to Lin's paper. Which measure to use depends on the specific task that you are trying to evaluate. If you are working on extractive summarization with fairly verbose system and reference summaries, then it may make sense to use ROUGE-1 and ROUGE-L. For very concise summaries, ROUGE-1 alone may suffice especially if you are also applying stemming and stop word removal.

ROUGE Evaluation Packages

- Perl implementation of ROUGE – original implementation of ROUGE
- ROUGE 2.0 – Simplification of the Perl version implemented in Java (Evaluate ROUGE-[L,S,SU,N] & support for unicode texts)
- Javascript implementation of ROUGE

Papers to Read

- ROUGE: A Package for Automatic Evaluation of Summaries
- ROUGE 2.0 report