Natural Language Processing 365

Data Science Natural Language Processing NLP Papers Summary

Day 109: NLP Papers Summary – Studying Summarization Evaluation Metrics In The Appropriate Scoring Range

By Ryan 18th April 2020 No Comments

Objective and Contribution

The role of evaluation metrics is extremely important as they heavily guide the research progress of a particular field. The goal of automatic evaluation metrics is to accurately evaluate generated summaries that's close to human judgements. The paper shows that there is a strong disagreement between evaluation metrics that behave similarly and the ones in higher-scoring range. This disagreement means we don't know which metrics to trust when ıting our generated summaries. The contributions of this paper is as follows:

1. Introduce a methodology to study the evaluation metrics in high-scoring range and found that there are low / negative correlations between metrics. This work hopes to encourage researchers to collect more human annotations in the appropriate scoring range

There aren't many manually annotated datasets and existing ones are created during shared tasks in 2008 and so annotated summaries are average compared to current level. This is illustrated in the figure below. As you can see, the score distribution of the ground-truth summaries (blue) differ from the score distribution of generated summaries by modern summarisation systems (red). There is no guarantee that evaluation metrics behave similar to human evaluation in the red distribution (high-scoring range) and the objective of this paper is to evaluate evaluation metrics in the high-scoring range, to assess if their ability to evaluate are consistent and correlates with human evaluation. The paper computes correlation between pairs of metrics in different scoring ranges that doesn't have human evaluation.

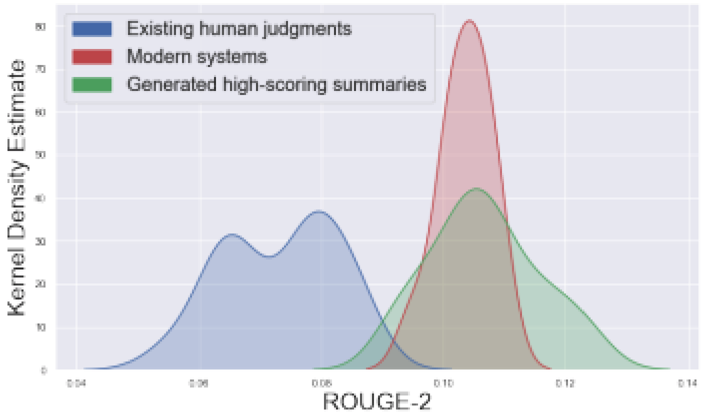


Figure 1: The blue distribution represents the score distribution of summaries available in the human judgment datasets of TAC-2008 and TAC-2009. The red distribution is the score distribution of summaries generated by mordern systems. The green distribution corresponds to the score distribution of summaries we generated in this work as described in section 3.

Experimental Setup – Data Generation

The paper studies the following metrics:

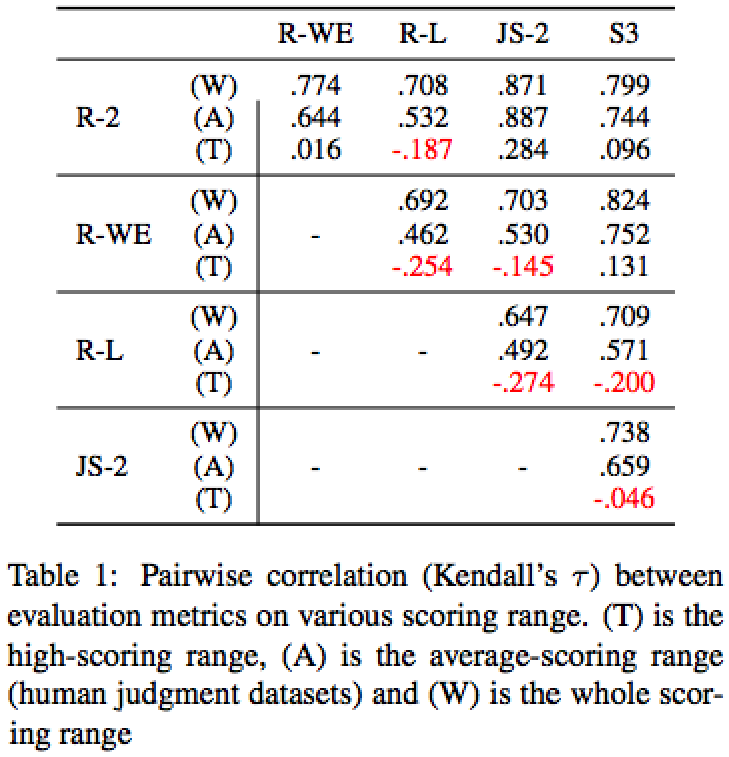
1. *ROUGE-2 (R-2)*. Bigram overlap between generated summaries and ground-truth

2. *ROUGE-L (R-L)*. Size of longest common subsequence between generated summarie nd ground-truth
3. *ROUGE-WE (R-WE)*. Soft matching based on cosine similarity and word embeddings
4. *JS divergence (JS-2)*. Uses Jensen-Shannon divergence to measure difference between bigram distributions
5. *S3*. a metric that maximise its correlation with manual Pyramid annotations

The authors use genetic algorithm for summarisation to generate summaries that optimise each metrics. The generated dataset (denoted W) consists of 160,523 summaries, with around 1763 summaries per metric. In order to focus on high-scoring summaries, we use LexRank to filter out summaries that underperform the benchmark. This led to the final dataset (T) of around 102 summaries per topic after removing duplicates and filtering. Human judgment summaries are denoted as A.

Correlation Analysis

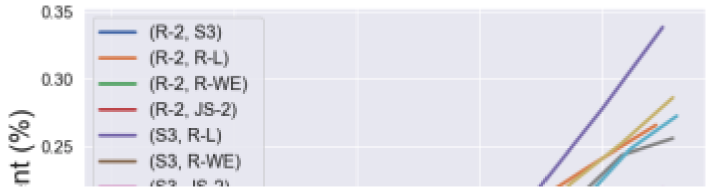
SIMPSON PARADOX



		R-WE	R-L	JS-2	S3
R-2	(W)	.774	.708	.871	.799
	(A)	.644	.532	.887	.744
	(T)	.016	-.187	.284	.096
R-WE	(W)	-	.692	.703	.824
	(A)		.462	.530	.752
	(T)		-.254	-.145	.131
R-L	(W)	-	-	.647	.709
	(A)			.492	.571
	(T)			-.274	-.200
JS-2	(W)	-	-	-	.738
	(A)				.659
	(T)				-.046

Table 1: Pairwise correlation (Kendall's τ) between evaluation metrics on various scoring range. (T) is the high-scoring range, (A) is the average-scoring range (human judgment datasets) and (W) is the whole scoring range

From the figure above, we can see that for dataset A and W, there's a high correlation between evaluation metrics, with R-2 and JS-2 having the strongest correlation. This can be n by the fact that they are both based on bigrams. R-L has the least correlation with the other metrics. However, in the high-scoring summaries (T), correlations between metrics are low and some are even negative. There's no global agreement between metrics to measure improvements when we examine summaries that are better than LexRank. In fact, the results show that this disagreement increases with higher-scoring summaries as shown in the figure below.



This is known as the Simpson paradox whereby different conclusions are reached depending on which sub-populations you drawn from. The results tell us that our current evaluation metrics are good at distinguishing very bad summaries from very good summaries but aren't able to distinguish between high-scoring summaries.

MEASURING CONSISTENT IMPROVEMENTS ACROSS METRICS

Given a set of evaluation metrics, to measure consistent improvements across metrics (metrics agreeing with each other), we compute the following F/N ratio:

1. Select a summary s
2. Among the summaries, which are better than s for one metric (N)
3. Among the summaries, which are better than s for all metrics (F)
4. F divide by N to obtain the ratio

The figure below shows this process repeated for 5000 random sample summaries. The r lts show that the proportion of consistent improvement (F/N ratio) decrease rapidly as the average score of summaries increases. By using multiple evaluation metrics that disagree with each other, it is very difficult to identify high-scoring summaries with high confidence.

Conclusion and Future Work

The disagreement between evaluation metrics in the high-scoring summaries means that it would be difficult to evaluate high-scoring summaries and researchers have a high risk of facing the Simpson paradox. The analysis was performed on TAC-2008 and TAC-2009 datasets as they are the standard dataset for evaluating evaluation metrics. Future work could extend this analysis on other datasets and/or other NLP tasks such as machine transl

Source: aclweb.org/anthology/P19-1502.pdf

Ryan

Data Scientist

Previous Post

Day 108: NLP Papers Summary - Simple BERT Models for Relation Extraction and Semantic Role Labelling

Next Post

Day 110: NLP Papers Summary - Double Embeddings and CNN-based Sequence Labelling for Aspect Extraction

Data Science Natural Language Processing NLP Papers Summary

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan
30th December 2020

Data Science Implementation Natural Language Processing

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan
29th December 2020

Data Science Ryan's PhD Journey

Day 363: Ryan’s PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

Ryan
28th December 2020