Abstract:

The quality of a summarization evaluation metric is quantified by calculating the correlation between its scores and human annotations across a large number of summaries. Currently, it is not clear how precise these correlation estimates are, nor whether differences between two metrics\textsuperscript{a} \& \textsuperscript{b}; correlations reflect a true difference or if it is due to random chance. In this talk, I will address these two problems by proposing methods for calculating confidence intervals and running hypothesis tests for correlations measured in summarization. After evaluating which of the proposed methods is most appropriate for summarization through two simulation experiments, I will analyze the results of applying these methods to several different automatic evaluation metrics across three sets of human annotations. In this research, we find that the confidence intervals are rather wide, demonstrating high uncertainty in how reliable automatic metrics truly are. Further, although many metrics fail to show statistical improvements over ROUGE, two recent works, QAEval and BERTScore, do in some evaluation settings. This work is published at TACL 2021: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00417/107833/A-Statistical-Analysis-of-Summarization-Evaluation.

In the second part of this talk, I will present an ongoing study that identifies two ways in which the definition of the system-level correlation is inconsistent with how metrics are used to evaluate summarization systems in practice and propose changes to rectify this disconnect. The results from these analyses point to the need for future research to focus on developing more consistent and reliable human evaluations of summaries.

This research was done in collaboration with Daniel Deutsch, a Ph.D. student from the Cognitive Computation Group at the Department of Computer and Information Science, University of Pennsylvania.