A "Diamond-Standard" for Evaluating Extractive Summaries and Keyphrases using Abstractive References

ABSTRACT

Automatic Summarization and Keyphrase Extraction are two widely used areas in Natural Language Processing and Information Retrieval. A reliable dataset is extremely helpful to the systems performing these tasks. To develop "gold-standards" manually on a large scale dataset is time-consuming and expensive, as well as subjective - differences between summaries written by different people can be significant.

Our contribution, in this work, is two-fold: firstly, we provide a dataset with a highly trustable reference set for summaries and keywords, and secondly, we provide an evaluation mechanism for systems that are mostly extractive whereas the references are abstractive by nature.

In case of scientific articles, author assigned abstracts and key-phrases can provide a high quality reference for summarization and key-phrase extraction tasks. We call author generated summary (i.e., abstract) and key-phrases as our *"diamond standard"*. We collect a set of scientific articles, along with author assigned abstracts and keywords, and provide the major sections of the article (excluding abstract, title, and reference list) as a dataset. In order to test the usability of the dataset and the diamond-standard, we use TextRank as a candidate system. As author provided samples can be abstractive by nature, and can vary over length, we compute a relativized score using *i-measure* between the TextRank generated output and the diamond-standard. Systems using our dataset can either use the state-of-the-art evaluation methodologies (precision, recall, f-measure) or the relativized scale (i-measure) to evaluate their performances.

Keywords

extractive approach, defence, scientific article, author, summary, keyphrase

ACM ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235