Abstract

The increasing availability of automatically generated summaries has prompted intensive research in the area of automatic text summarization evaluation within the Natural Language Processing community. The development of automatic summarization systems has mainly focused on improving content selection, which has led to a focus on automatic methods for evaluation of the content of automatically generated summaries. The problem of evaluating the linguistic quality of summarization system outputs has been addressed only in recent years.

Linguistic quality of a summary is associated with its readability, which includes various linguistic factors that determine the overall quality of a summary. This project investigates and defines a set of potentially important factors of linguistic quality for further manual annotation, with the purpose of revealing linguistic quality violations that occur frequently in machine-produced summaries. The results of this analysis motivate a scheme and process for annotation of such violations in automatically produced summaries. The annotation scheme that has been developed for this task and further employed in the process of annotation includes five annotation tags: entity mention, clause, relation between entities, relation between clauses and misleading discourse connective.

This project's outcome is a corpus of summaries annotated for the identified factors of linguistic quality. The corpus contains 1935 annotated extractive summaries from the TAC dataset and 50 annotated extractive summaries from the G-Flow dataset. Statistical analysis of both annotated datasets has been carried out and evaluated. The results actually show that the G-Flow system produces summaries with fewer violations on average than the TAC systems. This is the most we can say, especially given the relatively small G-Flow dataset.

For producing the corpus a detailed set of annotation guidelines containing the explanation of annotation tags, the classification into attribute types followed by examples, has been developed. Inter-annotator agreement has been measured to verify the reliability of the guidelines and returns high agreement rates for most annotation tags.