A Truth-Preserving Compression Scheme for Scientific Documents

Philippe Suter, EPFL, Switzerland

ABSTRACT

In the internet age, research continues to be hampered by constraints stemming from the limitations of printed media, as conferences and journals typically require submitted articles to fit within a strict page limit. As a result, researchers regularly face the challenge of shortening their work while preserving its substance. To alleviate this task, we propose an algorithm that automatically compresses scientific articles, while completely preserving their scientific truth. The algorithm builds on state-of-the-art techniques for automated summarization and incorporates a range of increasingly complex heuristics that were crafted to incorporate domain-specific knowledge about scientific literature. The simplest are keyword-driven and will, e.g., suppress any sentence containing the noun or verb “conjecture”. More advanced heuristics rely on external knowledge bases, such as social networks. For instance we found it particularly effective to remove paragraphs mentioning works by friends of the authors. In the final and most computationally intensive phase, each sentence is dispatched to the IBM Watson supercomputer for an in-depth semantic analysis, where statements that cannot be proved uncontroversial using a standard axiomatization of set theory (potentially assuming the existence of strongly inaccessible cardinals) are discarded. We evaluated our algorithm on papers published in the last five years in forty-two first-, second- and third-tier conferences. Among our findings is a strong correlation between the achievable compression rate and the conference acceptance rate. We also applied our algorithm to the body of this paper.

BODY

Truth-preserving text shortening is an important task. We claim it is feasible. In reality, it is not. We thank our colleagues and friends.

Volume X of Tiny Transactions on Computer Science