Using Context-Free Grammar to Generate Synthetic Technical Short Texts

Conference Publishing

Tyler Bikaun

Authors: Tyler Bikaun, Michael Stewart, Melinda Hodkiewicz
2022-12-03

Publication

AI 2022: AI 2022: Advances in Artificial Intelligence pp 325-338

volume 13728

Part of the Lecture Notes in Computer Science book series (LNAI,volume 13728)

Quality Indicators

Peer Reviewed

Relevance to the Centre

VValuable technical information are buried in the under-utilised, user-generated technical texts in engineering domains, such as manufacturing, logistics and maintenance. For maintenance and reliability personnel, the unstructured technical text in maintenance work orders (MWO) hold crucial information about failures and work performed on physical assets. However, the domain-specific language used and scarcity of shared labelled data sets in these contexts present formidable challenges to contemporary natural language processing (NLP) techniques, resulting in inability to achieve performance similar to those in non-engineering domains. In this work, we explore the structure of language in technical short texts by learning a context-free grammar (CFG) through unsupervised grammar induction on industrial MWO texts. We exploit the grammar's generative properties for novel sentence generation and corpus construction and assess its viability for developing synthetic MWO data sets. The results demonstrate a) there exists a grammar in the MWOs, b) the grammar was able to model aspects of the maintenance technical language to produce 12k of synthetic MWO texts 93% as natural and 87% as correct as real texts, and c) the domain-specific language used in technical short text remains challenging to parse due to low data quality and sparsity. Contributions of this work include baseline results for a grammar-based synthetic technical text generation and an appreciation for challenges in assessing the engineering correctness and naturalness of the new synthetic texts.

DOI: 10.1007/978-3-031-22695-3_23