Struct2Unstruct: Creating Tender NER Datasets from Structured Procurement Records using Large Language Models

Asim Abbas, Mark Lee, Niloofer Shanavas, Venelin Kovatchev, Mubashir Ali

UNIVERSITY OF BIRMINGHAM

Introduction

Tender is a process by which an organization who needs goods or services invites other parties to submit a proposal or bid to provide these goods or services for a specific project, at specified cost and time.

Models trained on general corpora - perform poorly on Tender

Problem Statement

- ❑ Tender data scarcity due to domain sensitivity
- ❑ Data hallucination using LLMs
- ❑ Data annotation is tedious for NER task

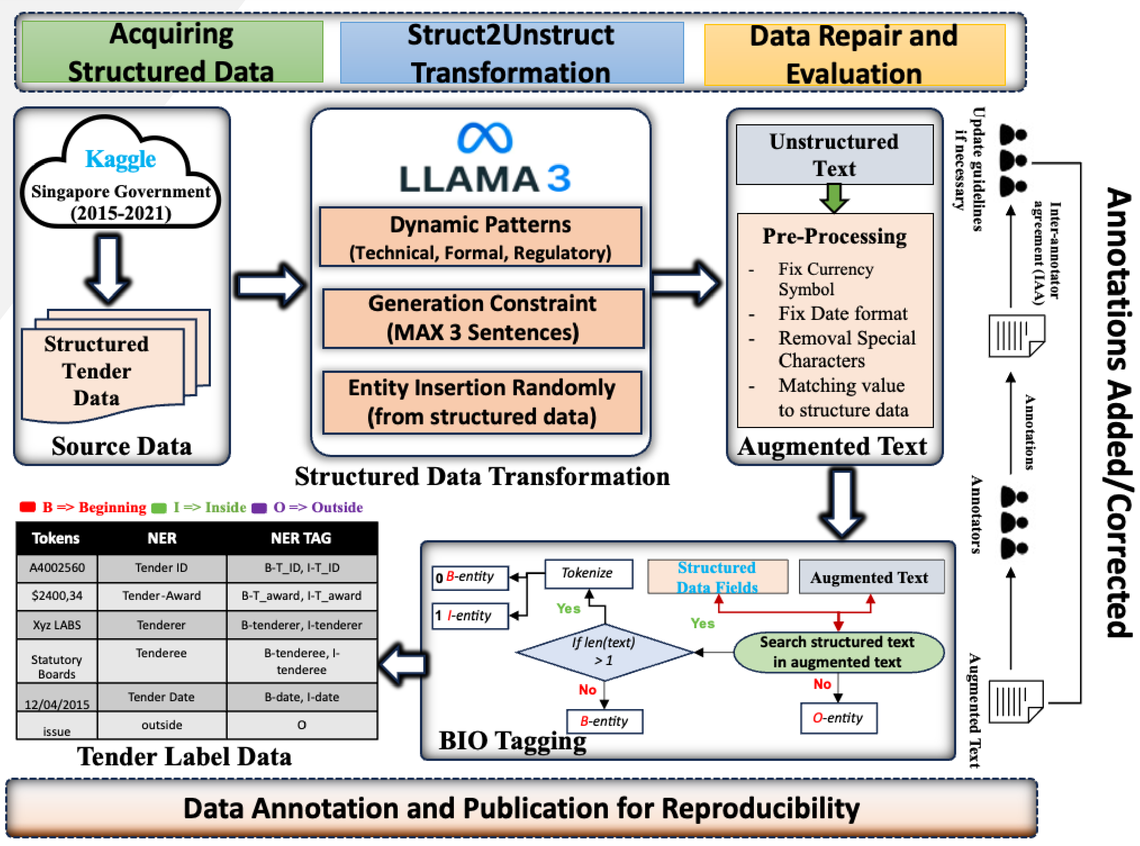
Research Question

To what extent can LLM-generated synthetic tender data mitigate data scarcity and domain sensitivity challenges in NER tasks?

Research Gap

Despite advances in LLM-based synthetic data generation, there is a lack of rigorous evaluation of data to ensure the reliability of such data for downstream tasks such as NER in sensitive domains like tender and procurement.

Proposed Approach



Experiments and Results

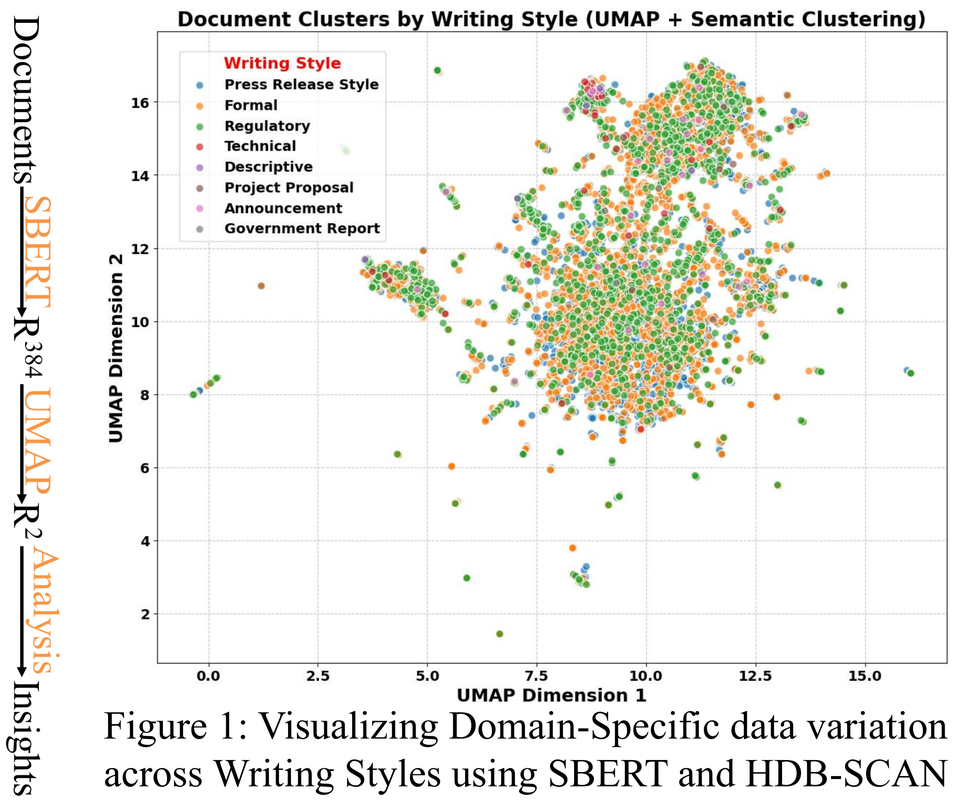


Figure 1: Visualizing Domain-Specific data variation across Writing Styles using SBERT and HDB-SCAN

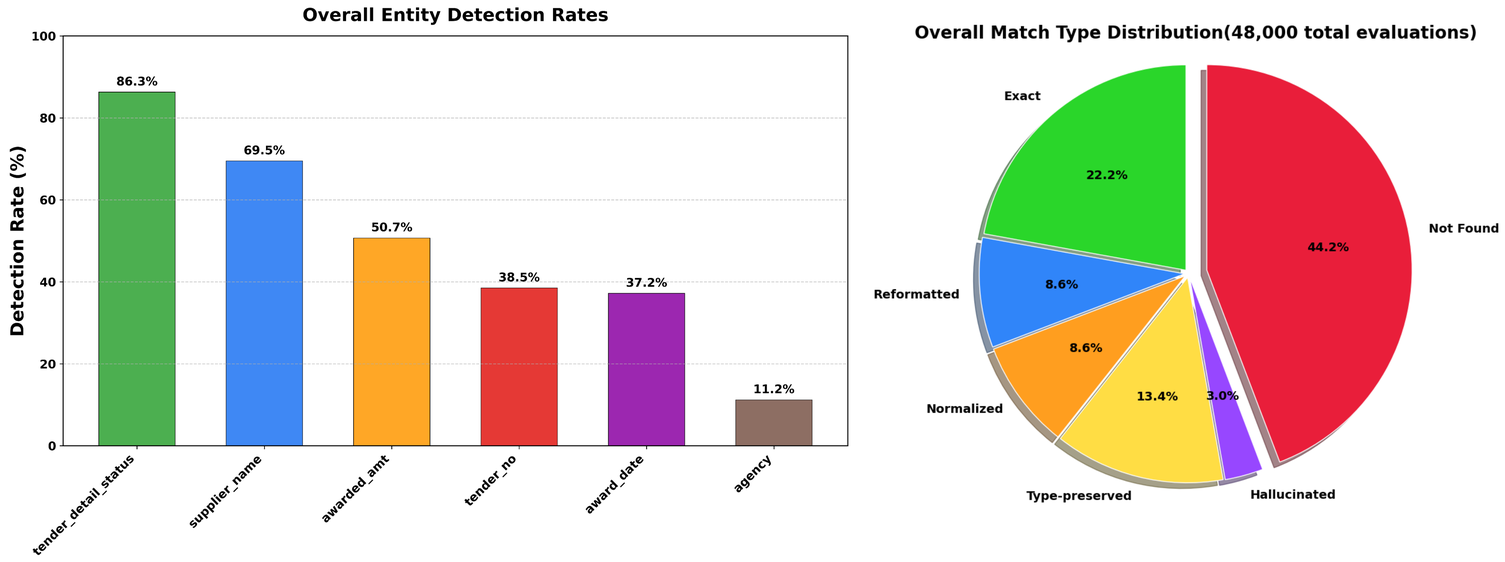


Figure 2: Overall match type distribution and entity level detection rates beyond match type

Expert A \ Expert B	Entity	Not Entity	Total
Entity	33,227	3,660	36,887
Not Entity	2,031	349,338	351,369
Total	35,258	352,998	388,256

Metric	Value
Observed Agreement	98.53%
Expected Agreement (by chance)	83.14%
Cohen's Kappa	0.913 **(3)**
Standard Error	0.001
95% Confidence Interval	0.911 – 0.915

Table 1: Tender Entities annotation agreement confusion matrix between two annotators with observed and expected agreement, Cohen's κ, and its standard error.

Conclusion and Future Work

- We propose a pipeline for generating and validating synthetic tender NER data, addressing data scarcity and enabling reliable domain-specific dataset creation.
- Future work will focus on fine-tuning transformer models and improving dataset generalization across diverse procurement domains

School of Computer Science, University of Birmingham,
Edgbaston , B15 2TT, England.
axa2233@student.bham.ac.uk

UNIVERSITY OF BIRMINGHAM