A FRAMEWORK FOR SEMANTIC DATA INTEGRATION THROUGH NON-TAXONOMIC AND TERNARY RELATIONS: A CASE STUDY OF PATIENT-GENERATED TEXTS

ABSTRACT

Patient-generated texts contain useful information, but lack a definite internal structure and often contain informal vocabularies, rendering them unsuitable for structured databases. This creates a lexico-semantic gap between patient expressions and the formal, standardized domain vocabularies found in formal domain ontologies. Non-taxonomic, ternary and structural relations are necessary for comprehensive semantic understanding. However, they remain underexplored, as current research has focused on identifying taxonomic relations. This thesis addresses this through semantic alignment based on non-taxonomic and ternary relational components. By adopting a Design Science Research (DSR) approach, with a pragmatic research philosophy, the study develops and evaluates a knowledge-infused neural framework for cross-domain ontology integration that supports capturing and representing non-taxonomic and ternary relationships beyond general hierarchical relations. The framework adopts a four-layered architecture that includes: Data Acquisition and Ontology Pre-processing, Semantic Processing and Reasoning, Ontology Integration and Link Analysis, and Evaluation and Quality Assurance. A key contribution is the implementation of the delayed fusion strategy that balances the need for contextual neural learning and interpretable rule-based dictionary knowledge, and incorporating BioBERT as a plausibility filter to ensure domain factual grounding in the integration. The framework was evaluated on 38,115 patient-generated sentences from anxiety and depression datasets, from which 27,183 were key phrases forming the relevant set. The concepts were extracted and classified into disease, symptom and treatment domains. The hybrid per-class adaptive strategy extracted 113 unique concepts, prioritizing a more conservative prediction. The hybrid union strategy extracted 222 unique concepts, prioritizing wider coverage of domain concepts for the construction of the knowledge graph. The framework achieved a Tier-1 concept-level F1 score of 77.6%, a 9.9 percentage-point improvement over the standalone BiLSTM (67.7%), with an independent Tier-2 micro F1 of 0.31–0.35.

The framework also retained 384 semantic relations at a 92.7% filter retention rate; of these, a stratified sample of 126 relations was independently judged by a domain expert and achieved a precision of 0.50. Of the semantic relations validated, 144 were non-taxonomic relations organized into different semantic categories including associative, functional, causal, risk factors, and statistical, and 240 were ternary relations that captured multiple contexts of interaction between the concepts. The resulting knowledge graph, built with a support rule of at least two posts, had 75 nodes and 365 edges, organized into three loose but clinically coherent communities (modularity Q = 0.161). The framework transforms unstructured patient-generated texts into structured interoperable knowledge while preserving different clinical contexts. It mapped 58 of 75 graph concepts (77%) to existing biomedical standards including SNOMED CT and MeSH via BioPortal, with 40 exact matches, 17 synonym matches and 1 fuzzy match. This helped advance semantic interoperability of health data that improves clinical decision making and biomedical knowledge reuse. The study also contributes to design knowledge comprising guidelines and principles that respond to the development of semantic interoperability solutions, particularly in specific NLP environments where computational resources are limited.

Signature: 

Jael Adhiambo Gudu (2014/HD05/18686K)



Signature: ……..

Dr. Joseph Balikuddembe

Signature: ……

Dr. Johnson Mwebaze