Enhancing Knowledge Management in Healthcare: An Embedding Fusion Approach to Business Rule Representation

Brady Berg, Mukundan Agaram, Shikha Mohindra, Abhinav Nadh Thirupathi
Knowledge and Data Science, Roosevelt Innovations
Okemos, MI 48864
{bberg, magaram, smohindra, athirupathi}@rooseveltinnovations.com

Abstract

Enterprise Decision Management systems are vital for delivering efficient healthcare services. However, the ever-changing clinical terminology creates complex business rules, making healthcare IT systems difficult to maintain. In this study, we present an embedding fusion technique using unsupervised Natural Language Processing (NLP) to represent business rules as semantic vectors by incorporating multiple text data sources for each rule. We apply this method to a dental insurance administration case study and find that our approach is over 200 times more likely to identify redundant rule pairs compared to random pairs. This case study suggests that an embedding-based technique can significantly improve knowledge management efficiency in healthcare IT systems.

Introduction

Businesses can potentially benefit from implementing Business Rules Management Systems (BRMSs), as these systems offer a way to automate and manage complex decision-making processes, improving efficiency and effectiveness in various operations. In the healthcare sector, the high complexity of clinical terminologies, regulatory frameworks, and patient treatment histories, necessitate a comprehensive and efficient approach to managing business rules. Despite the importance of this problem, it has not been extensively studied in the research literature, leaving room for the development of novel techniques to enhance the understanding and management of complex business rules in healthcare.

These rules can take various forms, including natural language, flowcharts, decision tables, and decision trees. At Delta Dental of Michigan, the Business Rules Integrated Development Environment (BRIDE) system is utilized to manage rules, which are mainly expressed in a form of structured English (Agaram and Laird 2010). Fig. 1 illustrates an example rule from BRIDE.

Given the inherent complexity of healthcare, Delta Dental's rule corpus has grown to a substantial size of about 35,000 rules used in millions of combinations, posing challenges in maintaining and managing these rules effectively. Traditional analysis methods, such as bag-of-words algorithms and hard-coded logic, have been helpful to some extent in capturing the semantics within and between rules. However, due to their reliance on strict definitions, they may not fully comprehend more complex relationships (Agaram 2018) (Agaram 2019).

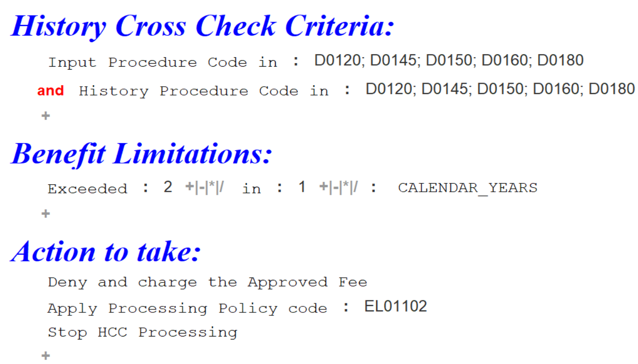

Figure 1: An example of a claim adjudication rule within the BRIDE system. The rule contains several elements that define its criteria and action. These include the History Cross Check Criteria, which defines the historical and current dental procedures on a claim line using CDT codes, the Benefit Limitations, which outline contractual limitations to benefits, and the Action to take, which includes a processing policy code, EL01102.

Moreover, healthcare business rules often encompass various types of codified language, each contributing distinct semantics and accompanied by a natural language descriptor. To tackle these challenges, we propose a method that employs unsupervised Natural Language Processing (NLP) to generate semantic vectors for business rules by incorporating multiple text data sources for each rule. By doing so, this approach aims to enhance knowledge management efficiency in healthcare IT systems through a deeper understanding and improved maintenance of complex business rules.

Approach

As depicted in Fig. 2, our approach focuses on utilizing the textual descriptions of codified CDT codes and processing policies within the rules, alongside the rule's text, to create

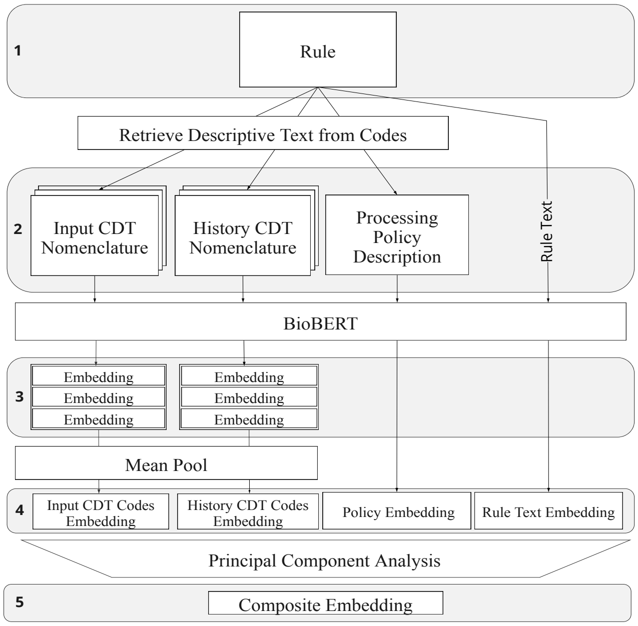

Figure 2: Overview of the embedding method. A rule and its natural language descriptors are embedded with BioBERT and composited to form a single vector.

meaningful representations. CDT codes, representing dental procedures, are obtained from the American Dental Association (ADA) (Association 2021). In contrast, processing policies define the application of benefits using proprietary company data. Both CDT codes and processing policies are essential components in understanding a rule's semantics. By integrating the complete text of the rule, we can encompass the rule's syntax and vocabulary within the representation, allowing for a more in-depth comprehension of the rule's structure and function. The first two numbered steps in the figure illustrate the extraction of these codified elements and the retrieval of their corresponding natural language descriptions.

BioBERT (Lee et al. 2019), a medically trained LLM, is used to embed these textual data. We use BioBERT in this work as the dental procedure nomenclatures contain medically-dominant vocabulary e.g.

> Immediate maxillary partial denture resin base including any conventional clasps, rests and teeth.

Similarly, Delta Dental's processing policies commonly reference dental procedures in the context of how they apply to business policies e.g.

> The opening and drainage of a tooth or palliative treatment when done by the same dentist or dental office on the same day as the root canal treatment is considered part of the root canal treatment.

To generate a holistic semantic vector for the rule, the embeddings undergo aggregation and fusion. First, the CDT codes are mean-pooled to maintain a fixed-length vector for each set of codes, as depicted in the transition between steps 3 and 4. Next, the four representative vectors—mean-pooled input CDT code nomenclatures, mean-pooled history CDT code nomenclatures, processing policy description, and rule text—are concatenated to form a 3,072-dimensional vector, as shown in step 4 of the figure. Lastly, Principal Component Analysis (PCA) is applied to reduce the dimensionality, as illustrated by the transformation from steps 4 to 5 in fig. 2 (F.R.S. 1901). We chose to preserve 98% of the original variance in this application.

Analysis

This study assessed a framework designed to identify redundant rules by employing affinity analysis to compare machine-generated rules with those created by humans. The goal was to enhance data maintenance, with an emphasis on rules governing the processing of multiple dental procedures performed within a single day for Fraud, Waste, and Abuse (FWA) policies.

Domain experts manually create human-authored rules, while machine-generated rules are produced through automated processes that utilize domain knowledge and algorithms. The methodology involved obtaining rule embeddings and generating an affinity matrix using cosine similarity. To find matches, the most similar machine-generated rule to each human-generated rule was identified from the affinity matrix. Subject Matter Experts (SMEs) then labeled the resulting matches to assess their accuracy.

Out of 360 human-authored and 399 machine-generated rules, the algorithm produced 156 matches involving 52 human-authored rules. SMEs found 62% (97/156) of these matches accurate, improving the redundant rule discovery rate by at least 221 times compared to random matches.

The framework was also evaluated for uncovering hidden relationships by comparing similarity measures derived from the proposed method with a measure derived from explicit attributes of the knowledge artifacts, such as the Jaccard Index for input and history CDT codes, and the Identity Similarity for processing policies. A Pearson correlation coefficient of 0.65 indicated a moderate positive correlation between explicit and implicit measures, indicating that the implicit measure is fairly based on explicit features. Additional research is needed to better understand the factors causing the remaining unexplained variance.

Conclusion and Future Work

This study presents an unsupervised NLP-based embedding fusion technique using a domain-specific language model BioBERT to encode healthcare business rules as semantic vectors. The method improves rule management, increasing the likelihood of identifying redundant rule pairs, which has implications for knowledge management efficiency, cost reduction, enhanced service quality, and better data maintenance in healthcare IT systems.

Future work should examine the generalizability and scalability of the proposed method across other use cases and domains, explore the implications of the framework for developing explainability for latent patterns, and consider integrating more interpretable and interactive tools for experts.

References

Agaram, M. K., and Laird, B. 2010. A componentized architecture for externalized business rules. In *2010 14th IEEE International Enterprise Distributed Object Computing Conference*, 175–183.

Agaram, M. 2018. Intelligent discovery features for EDM and MDM systems. In *2018 IEEE 22nd International Enterprise Distributed Object Computing Workshop (EDOCW)*, 135–144.

Agaram, M. 2019. Intelligent foundations for knowledge based systems. *Advances in Science, Technology and Engineering Systems Journal* 4.

Association, A. 2021. *CDT 2022: Current Dental Terminology*. American Dental Association.

F.R.S., K. P. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11):559–572.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR* abs/1901.08746.