A Scalable Framework for Synonymy Prediction: Inserting New Biomedical Terms into the UMLS

Vishesh Javangula
Lister Hill National Center for Biomedical Communications,
U.S. National Library of Medicine, National Institute of Health, Bethesda, Maryland Mentor: Dr. Olivier Bodenreider

1. Introduction

The National Library of Medicine developed the Unified Medical Language System (UMLS) as a biomedical vocabulary database that integrates terms and concepts from a variety of biomedical sources, e.g, electronic health records, scientific literature, and public health data. The purpose of the UMLS is to enable computer systems to "understand" the meaning behind this data to interconnect previously disparate vocabularies. By appropriately unifying these information systems, the UMLS can facilitate coordination of patient care across physicians, medical departments, pharmacies, and insurance. Furthermore, it enables research into terminologies, the development of information retrieval systems, and the extraction of relevant information from text data[1]. Three knowledge sources comprise the UMLS: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon.

The Metathesaurus is a large, multi-purpose, multi-lingual vocabulary database that houses information about biomedical and health-related data and organizes them into concepts, their synonyms, and the relationships between them. Each term present in the Metathesaurus is represented by a unique atom identifier (AUI), each of which belongs to a unique concept identifier (CUI). To facilitate consistent categorization of all the concepts, the Semantic Network assigns at least one semantic type to each concept. Furthermore, the Semantic Network provides users with information about each semantic type as well as the relationship between them.
To address the high degree of variability in natural language, the SPECIALIST Lexicon abstracts away inflectional, alphabetical, and other types of variants. This provides different kinds of lexical information for natural language systems to process.

2. Construction of Metathesaurus

Currently, the Metathesaurus construction process relies on the assumption that a combination of semantic preprocessing, lexical similarity models and human annotators will yield high accuracy when inserting new AUIs. However, recent literature suggests the process remains costly, demanding, and error-prone [2]. To reduce the burden on the human annotators as well as improve accuracy deep learning Siamese models have been developed with promising results

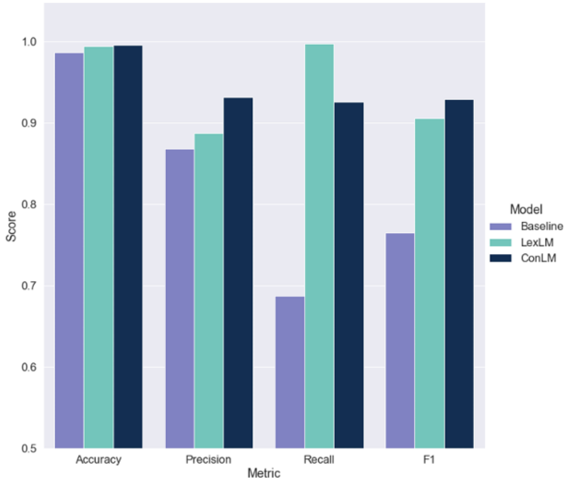

Figure 1. baseline and model metrics

3. Prior Works

Two high-performing models, LexLM and ConLM, were trained and evaluated using the active subset of the 2020AA UMLS version with test sets comprising of $1.7 * 10^8$ AUI pairs. As seen in figure 1, the models were compared to the current baseline via accuracy, precision, recall, and F1-score. LexLM relies on lexical similarity and showed improvements in every metric, namely F1 and recall. ConLM then incorporated contextual information and was able to exceed LexLM's results for every metric except recall [3,4].

4. Goals

While these experiments suggest a significant improvement over the baseline, the pairs in both training and testing datasets relied on the same set of AUIs from 2020AA. Therefore, it remains unclear how these models will perform in production, where they will classify on new terms from the next release of the UMLS. This summer, our goal was to create a framework to apply prior models on a use case: inserting new biomedical terms into an existing version of the UMLS Metathesaurus. To improve the feasibility of adoption in the UMLS construction process, we also aimed to minimize the running time when predicting synonymy.

5. Challenges and Contributions

With a test set size of $2 * 10^{12}$, it would take the models 5 months to insert the new AUIs and 17 years if we had to completely reconstruct the UMLS Metathesaurus. Scalability was therefore the primary challenge. To address this, we developed a model splitting approach that can efficiently calculate prediction scores.

6. Dataset

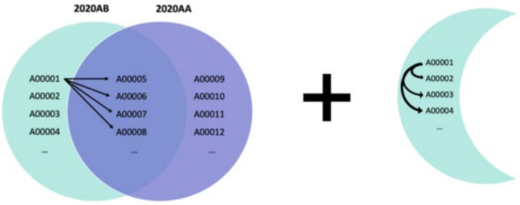

Figure 2. Dataset Contents

To simulate the insertion of new terms into the UMLS, we extracted unique AUIs present in 2020AB and formed pairs with every unsuppressed AUI present in 2020AA. As depicted in figure 2, this meant taking the cross-product between the 2020AB specific terms and the intersection of AUIs in 2020AA and 2020AB. Additionally, we generated every combination of pairs within the set of 2020AB-specific terms.

7. Approach

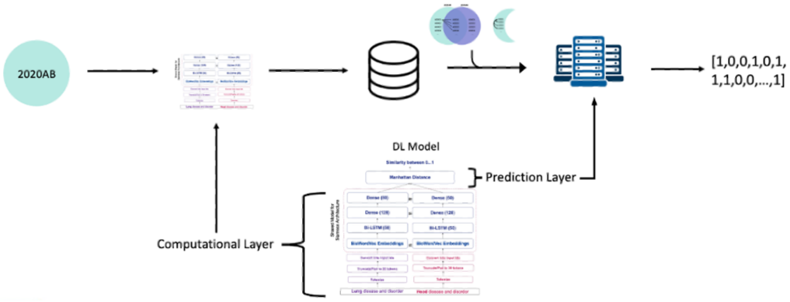

Figure 3. Model Splitting

Traditionally, pairs from the test set would be fed into a supervised learning model with the final output being the Manhattan score between two vectors corresponding to each AUI. Rather than going through this computationally expensive process, we precomputed each AUI within our test set and stored the vectorized representation in a database. This allowed us to fetch the vectorized representation of each AUI in a pair and calculate the Manhattan distance. To accelerate this process further, we parallelized test set generation and prediction across 500 nodes in the Biowulf cluster. Each node was given an equal subset of AUIs from the 2020AB-specific and each core was tasked with fetching these AUI to generate all its relevant pairs for an AUI. The pairs, along with the stored vectors, were directly into the model for prediction. Finally, the total true positives, true negatives, false positives, and false negatives for a given AUI were written into a file for analysis.

6. Experiment Validation

To ensure accurate results of our approach, multiple additional experiments were run. The first experiment was to ensure that the model splitting process yielded identical scores to the traditional approach. To do this we generated a mini-test set and compared the aforementioned metrics from both approaches. Following validation, we tested whether the parallelized process of mini-test set generation and prediction yielded identical results to predicting on a single node. The remaining validation component is to compare the parallelized prediction process on an older test set and compare it with the original results of prior experiments.

7. Results

While the model scores are still being evaluated, our model splitting approach was able to generate and predict $2 * 10^{12}$ pairs of AUIs in approximately 3 hours.

8. Acknowledgements

I would like to thank Dr. Bodenreider for his guidance, patience, and time throughout my summer internship. Additionally I would like to thank Dr.Vinh Nguyen for her mentorship, assistance, and technical guidance throughout the implementation process.

References

1. Bodenreider, O. (2004). The unified medical language SYSTEM (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(90001). https://doi.org/10.1093/nar/gkh061

2. Cimino, J. J. (1998). Auditing the unified medical language system with semantic methods. *Journal of the American Medical Informatics Association*, *5*(1), 41–51. https://doi.org/10.1136/jamia.1998.0050041

3. Nguyen, V., Yip, H. Y., & Bodenreider, O. (2021). Biomedical vocabulary alignment at scale in the umls metathesaurus. *Proceedings of the Web Conference 2021*. https://doi.org/10.1145/3442381.3450128

4. Nguyen, V. (n.d.). Context-Enriched Learning Models for Aligning Biomedical Vocabularies in the UMLS Metathesaurus. *Semantic Scholar*.