Internal Report
Translational UMLS Vocabulary Alignment

Bernal Jiménez Gutiérrez
Mentor: Olivier Bodenreider

Summer 2022

1 Introduction

The Unified Medical Language System (UMLS) [Bodenreider, 2004] is a large-scale biomedical knowledge base which combines over 200 medical terminologies, making it a vital biomedical interoperability resource and one of the most comprehensive biomedical knowledge bases available. Currently, the UMLS contains around 4 million unique concepts and approximately 16 million atoms (term strings associated with specific sources). Most of this large-scale knowledge base was created manually by combining concepts from different medical vocabularies into one normalized knowledge base, a process known as the vocabulary or ontology alignment which has yet to be successfully automated, especially in the biomedical domain. Therefore, UMLS is currently maintained by a team of expert editors who use lexical similarity, synonymy information obtained from the sources themselves as well as their own domain knowledge to perform vocabulary alignment. Unfortunately, the ever growing scale of UMLS as well as the challenging and time-consuming nature of this task makes it very challenging for even a well equipped team to do comprehensively.

Given the success of deep learning representation methods such as pretrained language models (PLMs) in a wide range of NLP tasks both in the biomedical domain as well as the general domain, we seek to study to leverage these methods for the task of biomedical vocabulary alignment. Existing work on vocabulary alignment on the UMLS task [Nguyen et al., 2021, Nguyen et al., 2022] formulates the problem as a synonymy classification task, where each pair of terms or atoms from different sources are compared one by one. Given that the number of lexical atoms in UMLSss are on the order of 10^7, the total number of possible comparisons is in the order of 10^{14}. The sheer number of comparisons makes this approach infeasible in practice since even basic modern deep learning techniques can take around 1ms to run, leading to total times in the order of 10^3 years. Additionally, given that there are only on the order of 10^7 positive synonym pairs, the ratio of positive to negative edges can be as low as 10^{-7}, an imbalance which is untenable for modern machine learning systems.

In this paper, we introduce a dataset which faithfully represents the real-world problem of UMLS vocabulary alignment. We propose a strong baseline to tackle this new dataset using a two-step approach which manages to avoid both the high computational cost as well as the extremely low prevalence by first narrowing down the number of candidate synonyms without losing many potential candidates and then performing standard synonymy classification using modern methods. Our task formulation, dataset and proposed system provide a solid foundation from which to perform realistic evaluation of future systems which tackle this challenging and important real-world problem.

2 Related Work

2.1 UMLS Vocabulary Alignment

Previous work on UMLS vocabulary alignment has demonstrated that deep learning methods based on both lexical and structural signal can be quite effective in predicting synonymy for biomedical terms, obtaining and F1 above 90% in their test set using simple word embedding methods [Nguyen et al., 2021, Nguyen et al., 2022, Wijesiriwardene et al., 2022]. However, the strong performance reported in these studies is obtained on a test set which was sampled from all UMLS pairs in a stratified way based on lexical similarity scores yielding a ratio of synonymous to non-synonymous examples which is approximately 10^5 larger than the actual ratio. This approach to constructing a dataset with a larger percentage of synonymy pairs is not deterministic or reproducible and thus limits the applicability of these results to the real-world UMLS vocabulary alignment task.

2.2 Biomedical Entity Linking

Our approach was inspired by strong parallels between UMLS vocabulary alignment and the biomedical entity linking task. In the task of entity linking, terms mentioned within text must be linked to existing concepts in a knowledge base. In biomedical entity linking, the knowledge base used is often UMLS, making it particularly useful for our purposes. Current SOTA methods for biomedical entity linking use a two-step method: candidate generation and candidate re-ranking. For candidate generation, candidate concepts from the knowledge base and the mention found in text are first transformed into a dense representation using an encoder such as a pretrained language model. Second, the knowledge base concepts vectors are ranked by their distance to the mention's vector representation. This is often done using a k-nearest neighbors algorithm that allows for quick filtering of unrelated candidates. The most effective biomedical entity linking models use specialized encoders such as [Liu et al., 2021, Zhang et al., 2021, Yuan et al., 2022], which were explicitly trained to align euclidean distance with synonymy semantics using a contrastive learning training scheme. We borrow this candidate generation approach directly in our work. For candidate re-ranking, most modern systems fine-tune a pretrained language model to take a concatenated string with mention and candidate concept as input and output a score used to re-rank the current candidate list. However, since some terms in UMLS have no synonymous terms, we replace the candidate re-ranking module with a candidate synonymy classification module which outputs a label rather than a relative score.

3 Task & Datasets

In order to represent the real-world vocabulary alignment task as closely as possible, we focus on the important sub-task of integrating new atoms from different sources into the UMLS. This task requires UMLS editors to link around half a million new atoms into UMLS twice a year in order to keep the knowledge base up-to-date. To create a faithful dataset for this task, we compare the first and second versions of UMLS in 2020, 2020AA and 2020AB respectively. We then separate all atoms which are present in the 2020AB version but not in 2020AA, meaning that they were added in that update. We define this subset of the dataset as the query data Q and the rest as the database D. The goal for this task is therefore to find all synonyms in D for each term in Q.

4 Approach

Our approach consists of two steps. First, for each new query term, we retrieve a small set of plausible candidate terms (candidate generation) to limit computational costs and increase the percentage of synonym pairs present in the second stage. Secondly, we use a synonymy prediction models to label each query-candidate pair as synonymous or not. This system ultimately produces a set of predicted synonym terms from D for each query term from Q which can be evaluated against the true synonyms chosen by UMLS editors.

4.1 Candidate Generation

Modeled after biomedical entity linking, we first encode both all terms in D and Q using some textual encoder model. Then, for every encoded query q in Q, we use the Faiss GPU implementation of the k-nearest neighbor algorithm [Johnson et al., 2019] to extract the k most *similar* terms to q in D. We chose 2000 as a maximum number of candidates to extract in our experiments. We chose a variety of models as textual encoders for candidate generation including previous UVA models such as LexLM, ConLM and UBERT as well as biomedical PLMs such as PubMedBERT and biomedical PLMs infused with UMLS signal such as SAPBERT and KRISSBERT. Apart from using the k-NN algorithm to extract plausible candidates for each query term, we also leverage source synonymy information (synonymy labels which come directly from the source vocabularies) to augment the list of candidates.

In order to measure the performance of the candidate generation step, we use the average recall at k metric which is a sample based average of the number of true synonyms within the top k candidates over the number of total synonyms for each query q.

4.2 Synonymy Classification

After restricting the number of possible candidates to a reasonable number, we use previous UVA methods for synonymy prediction such as LexLM, ConLM and UBERT as well as standard pre-trained language model fine-tuning. For PLM fine-tuning, the task is formulated as a binary classification task where each query-candidate pair (q,c) is concatenated as follows: `"[CLS]` q `[SEP]` c`"`. The [CLS] token was passed through a linear classification layer and all models were trained using cross-entropy loss.

To evaluate synonymy classification, we split the set of query terms Q into train, dev and test datasets. To create this split, we separate 1000 and 2000 concepts for dev and test sets respectively, making sure that the distribution of concepts with and without synonyms is the same in all three splits. We then add the top 100 candidates from the best candidate generation model, SAPBERT, creating a dev and test set of 100,000 and 200,000 examples respectively with a positive:negative imbalance of around 1:20. The rest of the term pairs are added to the training dataset. Table **??** shows the statistics for this dataset. In order to keep the errors from the first step to influence our results in the second step, we add the missing synonym terms into the train, dev and test sets. Performance in this second step is measured using the standard f1, precision and recall on the positive synonymy class.

5 Results & Discussion

5.1 Candidate Generation

Drilling into the candidate generation results in Tables 1 and 2, we notice that the SAPBERT model is by far the most effective candidate generation system. Considering that SAPBERT was optimized using contrastive learning to reduce the distance between synonymous phrases and increase the distance between the rest, it is not surprising that it would perform exceptionally in synonymy retrieval. KRISSBERT adopts a similar training scheme to SAPBERT but encodes concepts using textual excerpts rather than phrases, meaning that its training setting does not coincide perfectly with our setting. We also note that PubMedBERT dramatically under-performs both of these ontology infused PLMs as well as LexLM (a word embedding based model), suggesting that, at least in the biomedical domain, the masked language modeling objective does not enforce a strong correlation between euclidean distance and semantic similarity but rather that this bias must be introduced explicitly. Furthermore, we see that even though only 35% of all query atoms can be linked using source synonymy, including this signal improves recall significantly (by 10 points at 100 candidates). This boost in recall yields scores in the high 80's for even relatively small k values such as 100, making it a feasible first step for a practical vocabulary alignment system.

	R@1	**R@5**	**R@10**	**R@50**	**R@100**	**R@200**	**R@500**	**R@1000**	**R@2000**
LexLM	10%	22%	28%	42%	47%	51%	56%	59%	62%
PubMedBERT	9%	16%	18%	23%	25%	28%	31%	34%	37%
KRISSBERT	13%	25%	30%	43%	48%	53%	59%	64%	68%
SAPBERT	20%	44%	53%	71%	76%	81%	86%	88%	89%

Table 1: Candidate generation results with only k-NN candidates.

	R@1	**R@5**	**R@10**	**R@50**	**R@100**	**R@200**	**R@500**	**R@1000**	**R@2000**
Source Synonymy	35%	35%	35%	35%	35%	35%	35%	35%	35%
LexLM	42%	51%	56%	66%	70%	73%	77%	79%	81%
PubMedBERT	41%	46%	48%	52%	53%	55%	57%	59%	61%
KRISSBERT	43%	51%	55%	64%	68%	71%	76%	79%	82%
SAPBERT	46%	63%	70%	83%	86%	90%	93%	95%	95%

Table 2: Candidate generation results with k-NN candidates as well as source synonymy information.

5.2 Synonymy Classification

Our preliminary experiments show that both versions of UBERT [Wijesiriwardene et al., 2022], the best performing model in previous UVA work, dramatically underperforms PubMedBERT when fine-tuned on a dataset created from the candidate generation process described in 4.2 on the order of 10^5. In contrast, UBERT was trained on a dataset with over 170 million term pairs sampled in a stratified manner based on lexical similarity scores [Nguyen et al., 2021] to obtain a more balanced ratio of negative to positives. Therefore, the training dataset distribution for UBERT is very different than the datasets used for evaluation, degrading its performance significantly. We note that all three models achieve very high recall, a desirable feature for synonymy detection.

However, precision is only at 30-40%, meaning that approximately two-thirds of all predicted synonym pairs are incorrect. These results demonstrate that, although this two-stage approach is unable to provide a very robust solution to the vocabulary alignment task, it does provide a powerful tool for suggesting a wide net of synonyms which UMLS editors can reject if necessary.

	F1	**Precision**	**Recall**
UBERT (SAPBERT)	27.4	16.3	87.4
UBERT	35.0	22.0	85.0
PubMedBERT	52.5	37.1	90.0

Table 3: Results for synonymy prediction module for baseline models as well as the best performing model fine-tuned on the training data created from Q.

6 Conclusion

This work provides the first investigation into the real-world capabilities of modern NLP technologies for the task of UMLS vocabulary alignment. First, we provide the first task formulation and dataset which can be used directly to evaluate the practicality of new systems for integrating new concepts into UMLS. We provide a strong baseline for this dataset which uses a two-step approach: we first use dense textual representations and a fast k-nearest neighbor system to retrieve the most likely synonyms for each query term and then classify each query-candidate pair as synonymous or not using standard supervised classification models. We obtain high recall on our methods first step when using SAPBERT, a knowledge-infused PLM, demonstrating that our two-step method can successfully narrow the number of candidates to a manageable set. Finally, even though our supervised synonymy classification results are somewhat low for a practical system, we find that standard fine-tuning outperforms previous methods in this more realistic setting. We hope that this study provides a solid foundation for future work on the important and under-studied task of translational UMLS vocabulary alignment.

7 Future Work

There are several important avenues for future work which were not explored in this report. In terms of the synonymy classification task, it is important to address the more moderate but still important class imbalance problem which exists in this dataset. Techniques such as distributionally robust optimization [Lévy et al., 2020] which help models properly weigh the less prevalent but more important classes could be valuable to explore. In more data specific terms, future work should attempt to introduce more broad information about a term in a source vocabulary (such as its source synonyms, its semantic categories, source hierarchical structure, etc.) instead of focusing only on their lexical features. Additionally, it is important to note that the large degree of ambiguity which exists in the synonymy prediction task. Since the concept of synonymy is a very narrow subset of the general concept of semantic relatedness, it is plausible that including more detailed signals concerning different types of pairwise relations such as ”part of”, ”is a”, ”sibling of”, ”related to”, etc. could yield significant improvements in synonymy prediction.

References Cited

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

[Johnson et al., 2019] Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

[Lévy et al., 2020] Lévy, D., Carmon, Y., Duchi, J. C., and Sidford, A. (2020). Large-scale methods for distributionally robust optimization. *ArXiv*, abs/2010.05893.

[Liu et al., 2021] Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021). Self-alignment pretraining for biomedical entity representations. In *NAACL*.

[Nguyen et al., 2022] Nguyen, V. P., Yip, H. Y., Bajaj, G., Wijesiriwardene, T., Javangula, V., Parthasarathy, S., Sheth, A. P., and Bodenreider, O. (2022). Context-enriched learning models for aligning biomedical vocabularies at scale in the umls metathesaurus. *Proceedings of the ACM Web Conference 2022*.

[Nguyen et al., 2021] Nguyen, V. P., Yip, H. Y., and Bodenreider, O. (2021). Biomedical vocabulary alignment at scale in the umls metathesaurus. *Proceedings of the ... International World-Wide Web Conference. International WWW Conference*, 2021:2672 – 2683.

[Wijesiriwardene et al., 2022] Wijesiriwardene, T., Nguyen, V. P., Bajaj, G., Yip, H. Y., Javangula, V., Mao, Y., Fung, K. W., Parthasarathy, S., Sheth, A. P., and Bodenreider, O. (2022). Ubert: A novel language model for synonymy prediction at scale in the umls metathesaurus. *ArXiv*, abs/2204.12716.

[Yuan et al., 2022] Yuan, Z., Zhao, Z., and Yu, S. (2022). Coder: Knowledge infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, page 103983.

[Zhang et al., 2021] Zhang, S., Cheng, H., Vashishth, S., Wong, C., Xiao, J., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Knowledge-rich self-supervision for biomedical entity linking.