A GCN-based approach to uncover misaligned synonymous terms in the UMLS Metathesaurus

Xubing Hao[1], Rashmie Abeysinghe, PhD[2], Jay Shi, MD[3], Licong Cui, PhD[1,⋆]
[1]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
[2]Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX
[3]Intermountain Healthcare, Denver, CO

Abstract

The Unified Medical Language System (UMLS), a large repository of biomedical vocabularies, has been used for supporting various biomedical applications. Ensuring the quality of the UMLS is critical to maintain both the accuracy of its content and the reliability of downstream applications. In this work, we present a Graph Convolutional Network (GCN)-based approach to identify misaligned synonymous terms organized under different UMLS concepts. We used synonymous terms grouped under the same concept as positive samples and top lexically similar terms as negative samples to train the GCN model. We applied the model to a test set and suggested those negative samples predicted to be synonymous as potentially misaligned synonymous terms. A total of 147,625 suggestions were made. A human expert evaluated 100 randomly selected suggestions and agreed with 60 of them. The results indicate that our GCN-based approach shows promise to help improve the synonymy grouping in the UMLS.

1 Introduction

The Unified Medical Language System (UMLS), developed and maintained by the US National Library of Medicine, integrates more than 16 million terms from over 180 biomedical vocabularies and coding systems including SNOMED CT, LOINC, RxNorm, MeSH, and ICD-10-CM[1]. Synonymous terms from these source vocabularies are mapped and grouped into concepts in the UMLS Metathesaurus to facilitate interoperability and data sharing across computer systems. The UMLS has been used in a wide variety of biomedical applications, including entity recognition and relation extraction from biomedical text[2,3], data integration from different electronic health record (EHR) systems[4], clinical decision support[5], information retrieval[6], and various biomedical research[7–9].

As can be imagined, mapping and integrating terms from over 180 source vocabularies is a challenging task[10]. The current construction and maintenance process of the UMLS Metathesaurus leverages lexical and semantic techniques to suggest candidates for synonymous terms, and relies on human editors to review the suggestions and make final decisions[11], which is time-consuming and labor-intensive. It is inevitable that inconsistencies or errors may exist due to the large size of terms involved and the constant addition of new terms in source vocabularies. Researchers have developed automated methods for auditing or quality assurance of the UMLS Metathesaurus regarding different characteristics of concepts including names, synonyms, semantic type assignments, and hierarchical (IS-A) relationships[12].

In this paper, we focus on a particular aspect regarding misaligned (or missed) synonymous terms, that is, synonym pairs grouped under different UMLS concepts. Most previous works on such synonym detection for auditing the UMLS Metathesaurus are rule-based lexical or heuristics approaches[13–17]. In this work, we develop a learning-based approach leveraging Graph Convolutional Network (GCN) to detect synonymous terms mapped to different UMLS concepts. More specifically, we train a GCN-based model to predict whether a pair of terms from different source vocabularies are synonymous. The positive samples include synonymous pairs grouped under the same concept, while the negative samples are formed by selecting top lexically similar pairs of terms. The trained model is applied to a test set, where the negative samples in the test set which are predicted as synonymous by the model are suggested to be potentially misaligned synonymous terms. A randomly selected collection of such suggestions is manually evaluated by a human expert. In addition, a newer version of the UMLS Metathesaurus is leveraged to validate the suggestions.

⋆Corresponding author. Email: licong.cui@uth.tmc.edu

2 Background

2.1 UMLS Metathesaurus

The UMLS Metathesaurus, which is an integration system of biomedical vocabularies created by the US National Library of Medicine, groups synonymous terms into concepts to facilitate the integration and alignment of biomedical terminologies[11]. The UMLS Metathesaurus integrates hundreds of biomedical terminologies such as SNOMED CT, National Cancer Institute Thesaurus (NCIt) and Gene Ontology[1]. The most recent 2022AB release of the UMLS Metathesaurus contains over 4 million concepts and 16 million names from 182 source vocabularies[18]. The key to the UMLS Metathesaurus are the notions of atom and concept. An atom is a term from a specific source vocabulary while a concept is a grouping of synonymous atoms[19]. In the UMLS Metathesaurus, a concept denotes a particular meaning aggregating all the atoms from any vocabularies that convey this particular meaning in any form. Each concept is designated with a unique identifier known as a Concept Unique Identifier (CUI) to distinguish that particular meaning. Atoms are allocated a distinct Atom Unique Identifier (AUI). All of the atoms within a concept are synonymous and every concept is linked to at least one atom[20]. Table 1 shows five examples of atoms from five different source vocabularies and their corresponding AUIs. These five atoms are grouped under the UMLS concept representing "Adrenal Gland Neoplasms" with a CUI of C0001624.

Table 1: Five atoms grouped under the UMLS concept "Adrenal Gland Neoplasms" with a CUI of C0001624.

Atom	AUI	Vocabulary
Neoplasm of adrenal gland	A3577517	SNOMEDCT_US
Neoplasm of the adrenal gland	A24683942	HPO
Adrenal Gland Neoplasm	A7568581	NCIt
neoplasm of adrenal gland	A14015726	MEDCIN
Adrenal Gland Neoplasms	A0020274	MSH

2.2 Related work

A number of approaches have been explored to audit various aspects of the UMLS including concepts, concept names, and synonymy; semantic type assignments; hierarchical relationships; lateral relationships; ontology enrichment; and ontology alignment[12]. Synonym detection serves as an important quality check since not identifying synonyms would result in the creation of redundant UMLS concepts. In one of the earlier works, Cimino et al. have investigated [13,14] synonymous (hence redundant) UMLS concepts using a lexical approach by looking for concepts containing same words in a different order or that contains different punctuation. Hole et al. have introduced techniques such as lexical tweaks like trimming space or punctuation, swapping synonymous words to enhance the identification of synonyms[15]. Huang et al. have investigated piecewise synonym identification method where multi-word source terms are broken down to their components words and these words are replaced by their synonyms to generate synonymous terms for the original term. If a generated candidate already existed in the UMLS, then a synonym was considered to be found[16]. Huang et al. have also further investigated a similar synonym replacement method but leveraging WordNet as the synonym source. They have experimented with tuning the maximum number of allowed synonym substitutions per term and maximum term length[17].

3 Methods

In this work, we train a deep learning model to predict whether two atoms (or terms) are synonymous in the 2022AA full version of the UMLS Metathesaurus. We leverage lexical features of atom names and hierarchical features extracted by Graph Convolutional Networks (GCNs) to train our model.

Our approach contains five major steps: (1) data preprocessing; (2) sample selection; (3) model training; (4) misaligned synonymous terms identification; and (5) evaluation.

3.1 Data pre-processing

We only consider a UMLS atom if it satisfies all the following conditions: (1) the atom is in English; (2) the atom is not obsolete; and (3) the atom contains at least one letter or number.

The atoms are pre-processed by normalizing their lexical features in a similar manner to the approach in our prior work for normalizing the lexical features of SNOMED CT concepts[21]. Our pre-processing steps include converting the atom to lowercase; removing punctuations such as !, ", #, $, %; eliminating unnecessary white spaces; and lemmatizing the lexical feature using the WordNet lemmatizer in NLTK[22].

3.2 Sample selection

We frame the problem of synonym prediction in the UMLS Metathesaurus as a binary classification task where we have two classes: (1) positive class with synonymous atom-pairs, and (2) negative class with non-synonymous atom-pairs. To train our classifier, we construct a dataset with positive and negative samples as follows.

3.2.1 Positive sample generation

Given two atoms A and B that are grouped under the same UMLS concept, if A and B belong to different source vocabularies, they will form a positive sample (A, B). For example, in Table 1, "Adrenal Gland Neoplasm" from the NCIt with an AUI of A7568581 and "Neoplasm of the adrenal gland" from Human Phenotype Ontology (HPO) with an AUI of A24683942 will form a positive sample.

3.2.2 Negative sample generation

We use a positive sample corruption strategy to generate a negative sample for each positive sample. For a positive sample (A, B), we replace B with A's most lexically similar atom X that is not from A's source vocabulary to form a negative sample (A, X). The lexical similarity between two atoms is calculated using cosine similarity score based on bag of words. Note that in some situations, multiple positive samples could contain the concept A. In such cases, we will pick the next top lexically similar atoms to generate negative samples. For example, if there are positive samples (A, B) and (A, C), we will generate corresponding two negative samples (A, X) and (A, Y) where X is an atom that is the most lexically similar to A and Y is the atom that is the second most lexically similar to A. Note that if there does not exist an atom that is lexically similar to A (i.e., cosine similarity between A and all other atoms are 0), there will be no negative sample generated for the positive sample. This will result in the positive to negative sample ratio of our dataset to be approximately 1 to 1.

After generating all the positive and negative samples, we group the positive samples and the corresponding negative samples based on which UMLS concepts the positive samples were generated from. Then we randomly split these groups to training, validation, and testing sets based on the ratio of 8:1:1. This was done to ensure that two synonymous atom-pairs do not spread across training/validation/testing sets. For instance, if atoms A, B, C, D are grouped under the same UMLS concept, they may generate positive samples (A, B) and (C, D). A and C can be from the same source terminology while B and D can also be from the same source terminology. In such a case, if (A, B) is in the training set while (C, D) is in the validation set, this would be akin to a situation where we are predicting on the training samples. Our sampling strategy avoids such scenarios.

3.3 Sample preparation

For each training sample consisting of an atom-pair, we prepare two types of inputs to the model: (1) atom name embeddings for each atom, and (2) ancestor subgraphs of the two atoms.

3.3.1 Atom name embeddings

We represent each word in an atom's name leveraging BioWordVec which contains biomedical word embeddings pre-trained using PubMed and the clinical notes from MIMIC-III Clinical Database[23]. If a word can not be found in the

pre-trained BioWordVec embeddings, we randomly generated a 200-dimension word vector by Gaussian distribution using NumPy package[24]. Then, we average the BioWordVec embeddings of all words in an atom to obtain atom name embeddings.

3.3.2 Ancestor subgraphs

For each atom in a sample, we generate an ancestor subgraph containing the atom itself, its parents, its grandparents, and the hierarchical (is-a) relations among those concepts within their respective terminologies. Figure 1 shows the two ancestor subgraphs obtained for the positive sample atoms "Adrenal Gland Neoplasm " from NCIt and "Neoplasm of the adrenal gland" from HPO. These subgraphs are generated to learn hierarchical relations with Graph Convolutional Networks as discussed later.

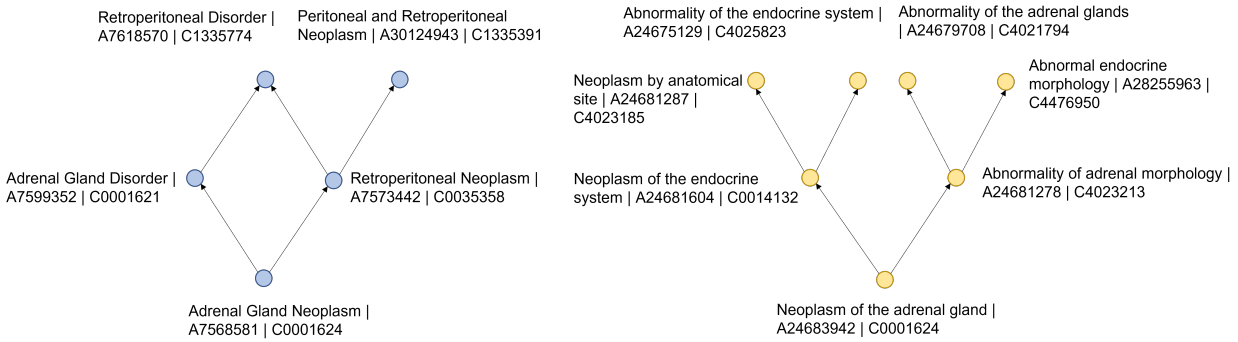

Figure 1: Ancestor subgraphs obtained for two atoms: "Adrenal Gland Neoplasm" from NCIt (left) and "Neoplasm of the adrenal gland" from HPO (right).

3.4 Model training

Figure 2 demonstrates the overall architecture of our model. In this study, we leveraged GCNs to automatically learn the hierarchical features of each atom[25]. GCNs are a type of neural network that directly operates on graph data. GCNs can encode information about the neighborhood of a node as a real-valued vector. The inputs of a GCN are feature vectors of nodes and the structure of the graph. The output of a GCN is representations of nodes aggregated with neighborhood information[26]. The input to the l-th layer of the GCN model is a vertex feature matrix, $H^{(l)} \in \mathbb{R}^{n \times d^{(l)}}$, where n is the number of vertices and $d^{(l)}$ is the number of features in the l-th layer. The output of the l-th layer is a new feature matrix $H^{(l+1)}$ by the following convolutional computation:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (1)$$

where σ is an activation function such as ReLU, $\tilde{A}$ is the adjacency matrix that stores the hierarchical information of the graph with added self-connections, $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$, and $W^{(l)}$ is the layer-specific trainable weight matrix[25].

In our approach, we leverage two GCNs with each consisting of two layers that individually learns hierarchical features of each atom in a sample. The GCN layers are provided with the ancestor subgraphs of each atom in a sample. Atom name embeddings serve as the initial features of each node in the ancestor subgraph (A_1 and B_1 for the two atoms in the sample as in Figure 2. The parents and grandparents will also be assigned with their own atom name embeddings). After the graph convolutions, the feature vectors of the nodes corresponding to the two atoms in the sample (A_2 and B_2 in Figure 2) would embed the hierarchical features based on their parents and grandparents.

Then the atom name embeddings of the two atoms are multiplied with each other ($A_1 * B_1$ in Figure 2). Similarly, the hierarchical feature embeddings of the two atoms are also multiplied with each other ($A_2 * B_2$ in Figure 2). We also obtain the difference between the atom name embeddings and hierarchical feature embeddings ($A_1 - B_1$ and $A_2 - B_2$ in Figure 2). Then the resulting vectors of these multiplications and differences are concatenated together

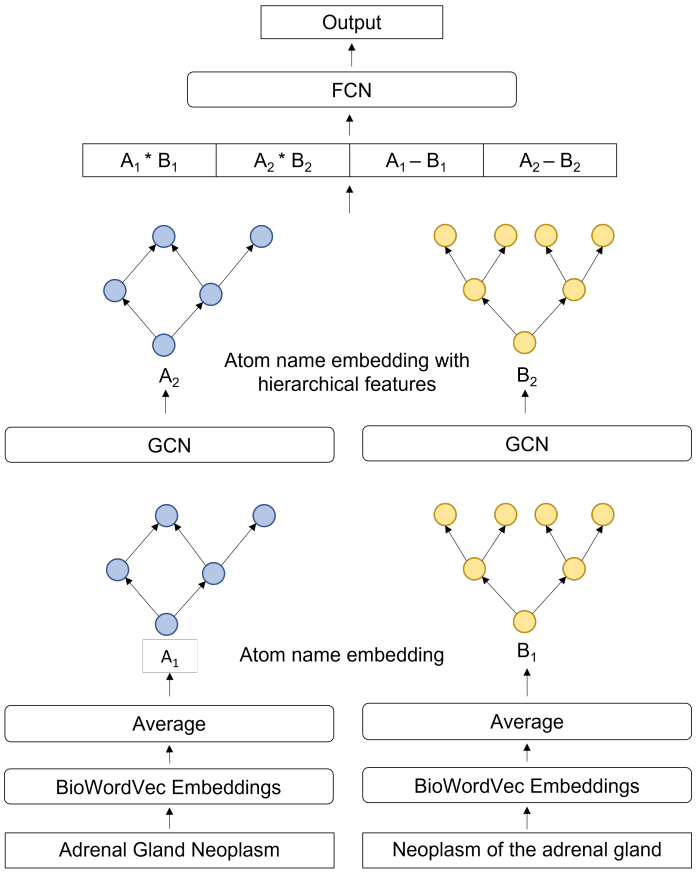

Figure 2: Overall architecture of the model.

and forwarded through two fully connected layers to perform the classification.

3.5 Misaligned synonymous atoms identification

The goal of this study is to identify the potentially misaligned synonymous atoms in the UMLS Metathesaurus. To achieve this, after training of the model is complete, we apply the trained model on the testing set. If the model predicts the atoms in a negative sample to be synonymous, we consider the two atoms in the negative sample as potential misaligned synonymous atoms in the UMLS. This is because negative samples in the test set serve as our candidate pairs where we look for misaligned synonymous atoms.

3.6 Performance evaluation

The performance of our model is evaluated in terms of precision, recall, and F1 score on the validation set. Furthermore, to evaluate our model's performance in suggesting actual misaligned synonymous atoms, we conduct: (1) an automated evaluation leveraging the newer version of the UMLS Metathesaurus, and (2) a manual evaluation by a domain expert (author JS).

For the automated evaluation, we leverage the newer 2022AB full version of the UMLS Metathesaurus released on November 7, 2022. If a pair of suggested misaligned synonymous atoms are grouped together under the same UMLS concept in the newer version, then this suggestion is considered to be valid.

For the manual evaluation, we randomly select a small subset of the suggested potential misaligned synonymous atoms for manual review by the domain expert who has experience in evaluating clinical terminologies to assess their validity.

We report this performance in terms of precision only as there is no gold standard for computing the actual recall and creating a gold standard could be very time-consuming and labor-intensive.

4 Results

4.1 Sample selection

After filtering the atoms that were not in English, were obsolete, and did not contain any letter or number, there existed a total of 9,351,912 atoms grouped under 4,324,051 UMLS concepts from 123 source terminologies in the 2022AA full version of the UMLS Metathesaurus. After applying our sample selection strategy, a total of 17,710,981 positive samples and 17,162,449 negative samples were selected as our dataset. After splitting the dataset, there existed a total of 27,962,212 samples in the training set, 3,414,455 samples in the validation set, and 3,496,793 in the testing set.

4.2 GCN-based model

We implemented our GCN-based classifier using TensorFlow 1.15.0. Table 2 summarizes the hyperparameters used in our model. The model was trained with an NVIDIA A100-SXM4-80GB graphics card with CUDA version 11.6 on a server running CentOS Linux (release 7.9.2009). It took 73 hours in total to train, validate, and test the model.

Table 2: Model hyperparameters.

Source	Value
GCN layer size	200
Fully connected layer 1 (FC1) size	512
Fully connected layer 2 (FC2) size	2
Activation for GCN	ReLU
Activation for FC1	ReLU
Activation for FC2	Softmax
Dropouts for FC1	0.5
Optimizer	Adam optimizer
Learning rate	0.005
Loss function	Softmax cross entropy with logits
Batch size	8192
Epochs	10

The model achieved a precision of 0.9152, a recall of 0.8338, and an F1 score of 0.8726 on the validation set. In the test set, there were a total of 147,625 atom pairs that originally had a negative label but were predicted to be synonymous by our model. These 147,625 pairs were considered as potential misaligned synonymous atoms for evaluation.

4.3 Evaluation

For the automated evaluation, we found that 239 of the misaligned synonymous atoms identified by our approach were denoted as synonymous (under the same UMLS concept) in the newer release of the UMLS. Table 3 demonstrates ten examples of misaligned synonymous atoms that were validated this way. For instance, atom "Spastic paraplegia 2, X-linked" from MSH with an AUI of A18470932 and atom "X-linked spastic paraplegia type 2" from SNOMEDCT_US with an AUI of A28441616 were grouped under different CUIs C1839264 and C0751604 respectively in the 2022AA version. However, in the newest 2022AB version, they are grouped under the same CUI C1839264 which validates our identification that they were misaligned synonyms in the 2022AA version.

For the manual evaluation, we randomly selected 100 samples from the total 147,625 identified potential misaligned synonymous atoms for the domain expert's manual review. The domain expert confirmed that 60 out of 100 are valid cases (a precision of 60%). Table 4 shows ten examples of misaligned synonymous atoms validated by the domain expert. For example, the domain expert confirmed that the atom "Operation on heart" from SNOMEDCT_US with an AUI of A3600156 under CUI C0018821 and the atom "Heart surgery operation" from LNC with an AUI of A20077783

Table 3: Ten misaligned synonymous atoms validated by the newer version of the UMLS Metathesaurus. The CUI of the UMLS concept that they are grouped in the newer version of the UMLS is also given.

Atom name-1	AUI-1|Vocabulary-1	Atom name-2	AUI-2|Vocabulary-2	CUI
Otopalatodigital Spectrum Disorder	A20982774|MSH	OPD(otopalatodigital) spectrum disorder	A31050408|SNOMED CT_US	C2748918
Magnesium loss, isolated renal	A18464844|MSH	Isolated renal magnesium wasting	A28434182|SNOMED CT_US	C1835171
Spastic paraplegia 2, X-linked	A18470932|MSH	X-linked spastic paraplegia type 2	A28441616|SNOMED CT_US	C1839264
Gluconate, Copper	A0063739|MSH	copper (as gluconate)	A13280242|MMSL	C0009975
butyl alcohol	A29797455|RXNORM	Alcohol Butyl	A9444964|MMSL	C0089147
Peripheral Arterial Diseases	A26621765|MSH	Peripheral arterial disease (disorder)	A32325946|SNOMED CT_US	C1704436
Miscarriage	A0087002|ICPC2P	Miscarriage of pregnancy	A25702967|MDR	C4552766
Juvenile polyposis syndrome	A2970824|SNOMED CT_US	Juvenile GI polyposis	A30926680|HPO	C0345893
peripheral arterial disease (diagnosis)	A16890157|MEDCIN	Peripheral arterial disease	A32317850|SNOMED CT_US	C1704436
Dissection of aorta, thoracic	A20881316|ICD9CM	dissection of thoracic aorta	A18581774|CHV	C0729233

under CUI C3261232 are synonymous atoms.

Table 4: Ten misaligned synonymous atoms validated by the domain expert.

Atom name-1	AUI-1|Vocabulary-1	Atom name-2	AUI-2|Vocabulary-2
Hypophosphatemia	A12029715|OMIM	Hypophosphataemia	A24675387|HPO
Cor Pulmonale	A26679742|MSH	Right ventricular dysfunction (cor pulmonale)	A29168389|NCI_CTCAE_3
Operation on heart	A3600156|SNOMED CT_US	Heart surgery operation	A20077783|LNC
Drug withdrawal syndrome in newborn	A25692959|MDR	drug; reaction, withdrawal, newborn)	A4408758|ICPC2ICD10ENG
Blood alcohol level measurement	A150610|MTH	Blood Alcohol Level	A26662450|MSH
Eptacog alfa (substance)	A30198910|SNOMED CT_US	eptacog alfa	A20682663|HGNC
DHT	A23909510|NCI_NCI-GLOSS	Dihydrotestosterone (DHT)	A29929970|CPT
neuropathy optic	A18667447|CHV	Optic neuropathy	A12021502|OMIM
Hospital Services, Emergency	A0558085|MSH	Emergency medical service	A4367912|MTH
Area 20 of Brodmann of guenon	A24156045|MTH	Area 20 of Brodmann	A15456400|FMA

5 Discussion

In this study, we developed a Graph Convolutional Neural Network-based approach to identify potentially misaligned synonymous atoms in the UMLS Metathesaurus. We generated positive samples by synonymous atoms grouped under the same UMLS concept. Negative samples were generated by non-synonymous but top lexically similar atom pairs. We used BioWordVec embeddings to represent lexical features of atoms and generated hierarchical features by feeding a Graph Convolutional Network with the ancestor subgraphs of samples. From the evaluation by a domain expert, it can be seen that the performance of the model is promising and hence the misaligned synonymous terms identified could be valuable in the quality improvement process of the UMLS.

5.1 Comparison with related work

Recently, there have been a number of investigations leveraging deep learning techniques for vocabulary alignment in the UMLS. For instance, Yip et al. have developed a deep learning approach leveraging a Siamese network with Long Short Term Memory (LSTM) and Convolutional Neural Network models to identify synonymy and non-synonym among atoms so that it could emulate the rule-based UMLS Metathesaurus building process[27]. Tran et. al further improved upon this work by introducing semantic features extracted from knowledge graph[28]. Nguyen et al. has also improved the original model by experimenting with adding an attention layer on top of the LSTM layer. In terms of the model architecture, our approach differs with the above approaches as we have used GCNs to automatically obtain hierarchical information from ancestors. The above approaches are more targeted towards UMLS construction with vocabulary alignment while the aim of our approach is to audit the existing synonymy in the UMLS.

5.2 False positives

The review by the domain expert revealed that 40 out of 100 of our identified suggested misaligned synonymous atoms are not valid synonyms. Table 5 demonstrates five such cases pointed out by the domain expert. For example, our model suggested that atom "Malignant neuroleptic syndrome" from SNOMEDCT_US with an AUI of A3041909 under CUI C0027849 is synonymous with atom "Neuroleptic-Malignant Syndrome, Neuroleptic Induced" from MSH with an AUI of A26606809 under C0751376. However, the domain expert pointed out that these two atoms are not synonymous since this syndrome can be both neuroleptic induced and antipsychotic induced.

Table 5: Five invalid misaligned synonymous atom predictions as pointed out by the domain expert.

Name	AUI	Vocabulary	CUI	Domain expert's comment
tartaric acid L-tartaric acid	A10337857 A10980428	RXNORM MTH	C0075821 C1289966	"L-" is a specific type of enantiomer, the other type of tartaric acid is "D-"
Malignant neuroleptic syndrome Neuroleptic-Malignant Syndrome, Neuroleptic Induced	A3041909 A26606809	SNOMEDCT_US MSH	C0027849 C0751376	This syndrome can be both neuroleptic induced and antipsychotic induced
Myeloproliferative Leukemia Protein myeloproliferative leukemia K protein, human	A24387015 A3831963	NCI MSH	C0218227 C0652198	k protein is specific type of protein
b complex deficiencies vitamin Vitamin B complex deficiency symptom	A18669031 A25688112	CHV MDR	C0042850 C0920232	the deficiency of the vitamin and the symptoms of that deficiency are different concepts
Administration of prophylactic antimalarial Administration of prophylactic treatment	A33687081 A33558496	SNOMEDCT_US ICNP	C0199244 C4039267	prophylactic means preventive, but preventive malarial treatment is very specific to malaria

5.3 Limitations and future work

In this work, we selected a specific test set to apply the trained model and identify misaligned synonymous atoms. However, to uncover misaligned synonymous atoms in the entire UMLS, in the future we will leverage a cross-validation approach that we introduced in a previous work[29]. With this cross-validation approach, in different runs, different splits will be used for training and identification of synonyms so that potential misaligned synonymous atoms can be identified from the entire UMLS Metathesaurus.

Also, in the current work we only predicted misaligned synonymous atoms among the negative samples in the test set. In the future, we will explore whether we could predict incorrectly aligned synonymous atoms among the positive samples in the test set. In addition, the focus of the current work was detecting synonymy at UMLS atom level. Another interesting future direction is to develop an approach that can identify synonymy at UMLS concept level.

In this work, we averaged the embeddings of each word in an atom to obtain atom name embeddings. However, this means that we lose important positional information of the words in an atom. In the future, we plan to explore approaches that take this important aspect into account.

Since our approach relies on ancestor subgraphs that only contain parents and grandparents, our model might lack information regarding the broader categories that the atoms belong to in their respective source terminologies. In the future, we would like to address this issue by investigating a mechanism where this information can be infused into

the model. In addition, we only leveraged the concepts' ancestor information when aggregating hierarchical features with the GCNs. In the future, we plan to also incorporate the concepts' descendants information to explore whether the performance could be improved. Furthermore, we plan to train knowledge graph (KG) embeddings using KG embedding techniques such as TransE and TransR to incorporate different relations between atoms in addition to is-a relations[30,31]. We will also explore whether incorporating additional information such as semantic groups of concepts will improve the model performance[32].

Importantly, in this work, we only predict whether two atoms currently deemed to be non-synonymous by the UMLS, are actually synonyms. Since the predicted synonymous atoms are currently aggregated in different UMLS concepts, we further need to investigate how these UMLS concepts could be merged. For example, in the 2022AA UMLS release, the atom "Amitriptyline-Chlordiazepoxide" with AUI A1529804 is grouped under the UMLS concept with CUI C0717408 and the atom "amitriptyline / chlordiazepoxide" with AUI A31645297 is grouped under the UMLS concept with CUI C2742631 (hence these were not synonyms in the 2022AA release). However, in the newer 2022AB release of the UMLS, atom A1529804 has been regrouped under C2742631 (C0717408 has been merged into C2742631). While our method is able to identify these two atoms as synonymous, it cannot determine how the atoms need to be reassigned to UMLS concepts.

6 Conclusion

In this study, we developed a Graph Convolutional Neural Network based approach to identify potentially misaligned synonymous atoms in the UMLS Metathesaurus. We trained our model with synonymous atom-pairs as recorded by the UMLS as positive samples and top lexically similar non-synonymous atom-pairs as negative samples. The model leveraged atom name embeddings as lexical features and hierarchical features generated from ancestor subgraphs. Based on the validation set, the model achieved a precision, recall, and F-1 score of 0.9152, 0.8338, 0.8726 respectively. Applying the trained model on the test set, we identified 147,625 potential misaligned synonymous atoms. Out of these, 239 cases were found to be synonymous in the newer release of the UMLS. Evaluation by a domain expert on a random sample of 100 cases revealed that 60 are valid. This indicates that the approach has the potential to identify valid misaligned synonymous atoms contributing to the important quality improvement process of the UMLS.

Acknowledgment

This work was supported by the National Science Foundation (NSF) through grant 2047001, and National Institutes of Health (NIH) through grants R01LM013335 and R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

References

1. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl_1):D267–D270.
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. p. 17.
3. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229–236.
4. Reimer AP, Milinovich A. Using UMLS for electronic health data standardization and database design. Journal of the American Medical Informatics Association. 2020;27(10):1520–1528.
5. Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. Journal of the American Medical Informatics Association. 2001;8(4):351–360.
6. Eichmann D, Ruiz ME, Srinivasan P. Cross-language information retrieval with the UMLS metathesaurus. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval; 1998. p. 72–80.
7. Amos L, Anderson D, Brody S, Ripple A, Humphreys BL. UMLS users and uses: a current overview. Journal of the American Medical Informatics Association. 2020;27(10):1606–1611.
8. Brody S. The Unifed Medical Language System: A Scoping Review of its Use in Research. 2020;.
9. Humphreys BL, Del Fiol G, Xu H. The UMLS knowledge sources at 30: indispensable to current research and applications in biomedical informatics. Journal of the American Medical Informatics Association. 2020;27(10):1499–1501.

10. Bodenreider O, Burgun A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. Studies in health technology and informatics. 2004;107(0 1):327.
11. Nguyen V, Bodenreider O. UVA Resources for the Biomedical Vocabulary Alignment at Scale in the UMLS Metathesaurus. arXiv preprint arXiv:220510575. 2022;.
12. Zheng L, He Z, Wei D, Keloth V, Fan JW, Lindemann L, et al. A review of auditing techniques for the Unified Medical Language System. Journal of the American Medical Informatics Association. 2020;27(10):1625–1638.
13. Cimino JJ. Auditing the unified medical language system with semantic methods. Journal of the American Medical Informatics Association. 1998;5(1):41–51.
14. Cimino JJ. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. p. 120.
15. Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented Metathesaurus. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000. p. 354.
16. Huang Kc, Geller J, Halper M, Cimino JJ. Piecewise synonyms for enhanced UMLS source terminology integration. In: AMIA Annual Symposium Proceedings. vol. 2007. American Medical Informatics Association; 2007. p. 339.
17. Huang KC, Geller J, Halper M, Perl Y, Xu J. Using WordNet synonym substitution to enhance UMLS source integration. Artificial intelligence in medicine. 2009;46(2):97–109.
18. UMLS Metathesaurus Browser;. (Online; accessed March, 2023). https://uts.nlm.nih.gov/uts/umls/home.
19. Nguyen V, Yip HY, Bodenreider O. Biomedical vocabulary alignment at scale in the UMLS metathesaurus. In: Proceedings of the Web Conference 2021; 2021. p. 2672–2683.
20. Mohtashamian M, Abeysinghe R, Hao X, Cui L. Identifying Missing IS-A Relations in Orphanet Rare Disease Ontology. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2022. p. 3274–3279.
21. Hao X, Abeysinghe R, Zheng F, Cui L. Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2021. p. 1805–1812.
22. Text preprocessing in Python: Steps, Tools, and Examples;. (Online; accessed August, 2021). https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908.
23. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data. 2019;6(1):52.
24. numpy.random.normal;. (Online; accessed March, 2023). https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html.
25. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016;.
26. Wang Z, Lv Q, Lan X, Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018. p. 349–357.
27. Yip HY, Nguyen V, Bodenreider O. Construction of UMLS Metathesaurus with Knowledge-Infused Deep Learning. In: BlockSW/CKG@ ISWC; 2019. .
28. Tran TT, Nghiem SV, Le VT, Quan TT, Nguyen V, Yip HY, et al. Siamese KG-LSTM: A deep learning model for enriching UMLS Metathesaurus synonymy. In: 2020 12th International Conference on Knowledge and Systems Engineering (KSE). IEEE; 2020. p. 281–286.
29. Abeysinghe R, Zheng F, Bernstam EV, Shi J, Bodenreider O, Cui L. A deep learning approach to identify missing is-a relations in SNOMED CT. Journal of the American Medical Informatics Association. 2023;30(3):475–484.
30. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems. 2013;26.
31. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29; 2015. .
32. Nguyen V, Yip HY, Bajaj G, Wijesiriwardene T, Javangula V, Parthasarathy S, et al. Context-Enriched Learning Models for Aligning Biomedical Vocabularies at Scale in the UMLS Metathesaurus. In: Proceedings of the ACM Web Conference 2022; 2022. p. 1037–1046.