Creation of Unambiguous Centralized Knowledge Base from UMLS Metathesaurus

1st Vatsal Shah
Research Lab
ezDI Inc.
Louisville, USA
vatsal.s@ezdi.com

2nd Dr. Binni Shah
Research Lab
ezDI Inc.
Louisville, USA
binni.s@ezdi.com

3rd Raxit Goswami
Research Lab
ezDI Inc.
Louisville, USA
raxit.g@ezdi.com

4th Saket Kumar
Research Lab
ezDI Inc.
Louisville, USA
saket.k@ezdi.com

5th Chetan Moradiya
Research Lab
ezDI Inc.
Louisville, USA
chetan.m@ezdi.com

***Abstract*—Efficient clinical information retrieval tasks like entity recognition, relation suggestions, summarization etc, require a comprehensive, extensive, unambiguous and well structured medical knowledge base. One of the largest Metathesaurus, UMLS (Unified Medical Language System), is a repository of biomedical dictionaries developed by the US National Library of Medicine (NLM) and widely used in medical domain. UMLS Metathesaurus includes the dictionaries like SNOMED-CT, National Center for Biotechnology Information (NCBI) taxonomy, Medical Subject Headings (MeSH), Online Mendelian Inheritance in Man (OMIM) etc. As it has integrated data from different sources, it contains different kinds of ambiguity, which is problematic for all clinical information retrieval tasks that use it. In this paper, we describe our methodology of curating the UMLS metathesaurus to create a centralized knowledge base that can be used as a knowledge base for a variety of clinical NLP systems. We have also developed a process of updating the curated centralized knowledge base with a newer version of UMLS such that there is no need to repeat the whole process. We have also presented the comparative results of a Clinical Entity Recognition (CER) using our curated centralized knowledge base and original UMLS database.**

***Index Terms*—UMLS, Unified Medical Language System, Clinical information retrieval, Medical knowledge base, NLP (Natural Language Processing), Ambiguity, Biomedical, Metathesaurus**

I. Introduction

The application of natural language processing (NLP) is increasing in the clinical industry for variety of tasks like medical transcription, documentation improvement, information extraction, document indexing, medical coding etc. Structured data is inevitable for any of NLP applications regardless of whether it uses a rule based or statistical approach. Precision in the data is an important aspect of it. Since any such data come from human language, they inherit the ambiguity from the source language. UMLS metathesaurus is the largest source of structured information for the medical domain. It is a huge collection of biomedical vocabularies which are collected from various sources [1] like SNOMED-CT, NCBI taxonomy, MeSH etc. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts [2]. This metathesaurus is widely used for different kinds of applications of medical domain like information extraction [3], medical coding [4], [5], document indexing [6], word sense disambiguation [7] etc.

In the UMLS there are dictionaries for different purposes like diagnosis, procedure, genetics, species and organisms, drugs, medical coding, HL7 standards, medical devices, dental terminology and procedures, nursing terminology, medical products and veterinary etc. as there are multiple audiences who use these dictionaries for their respective fields. The users may be physicians, nurses, researchers, health plan and policy makers, administrators, educators, pharmaceutical companies, therapists, hospitals, students, dental professionals, etc.

A string of characters can represent different concepts or meanings. There are the cases where the same concept is given different concept IDs in different dictionaries, and sometimes in the same dictionary too. We cannot rely on a single best dictionary because no dictionary in the UMLS provides complete coverage of the clinical domain. Some dictionaries are such that all the concepts and their corresponding codes in them are contained in some larger ones. These dictionaries become redundant and they just cost efficiency. Some dictionaries must be excluded which are not of the domain of our relevance (e.g., veterinary dictionary is irrelevant for clinical domain). Including them may result in out-of-domain concept mapping which is strictly undesirable. Therefore dictionary selection is a necessary process for correct information extraction. This drives a conflict of concept codes within the selected subset of the UMLS, which further requires resolution of these conflicts or ambiguity.

There have been good research work on analyzing the ambiguity in the UMLS [4], [8], [9], and then there has been work on resolving this ambiguity using different approaches [10]. They first perform mapping of free text (of medical domain) to the metathesaurus (UMLS), then resolve the ambiguity using various probabilistic rules. Our work differs from this approach in that we resolve the ambiguity in the metathesaurus itself. We implemented a process to exclude the least effective dictionaries followed by filtering and curating the resulting database to reduce the ambiguity within it.

There are also various tools available for the dictionary filtration and generation of new terms, multilingual and language specific as well. Some of them are JuFiT [11], MetaMap [12], Casper [13] etc. But these tools perform filtration by

correcting syntactic inversions or generating variants during text processing. These tools mainly focus on UMLS term strings only. However we use the concept and semantic type of term, and relevance of dictionary for the dictionary selection and filtration.

In this paper, we first give an overview of the UMLS metathesaurus, followed by a description of ambiguity present in it. Then we describe our process of creating a centralized knowledge base from selected UMLS dictionaries, and curating it for resolving the ambiguity through an automated process[1]. Then we present statistical details of the knowledge base and performance of a CER using the centralized knowledge base.

II. Overview of the UMLS

The UMLS Metathesaurus is a very large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. Designed for use by system developers, the Metathesaurus is built from the electronic versions of many different thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and basic, clinical, and health services research. These are referred to as the 'source vocabularies' of the Metathesaurus. The term Metathesaurus draws on Webster's Dictionary third definition for the prefix 'Meta', i.e., 'more comprehensive, transcending'. In a sense, the Metathesaurus transcends the specific thesauri, vocabularies, and classifications it encompasses. The Metathesaurus is organized by concept. In essence, its purpose is to link alternative names and views of the same concept together and to identify useful relationships between different concepts. Metathesaurus users may select from two relational formats: the Rich Release Format (RRF), introduced in 2004, and the Original Release Format (ORF) [1]. In addition to retaining all identifiers that are present in the source vocabularies, the Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains.

The Metathesaurus **'concept structure'** includes concept names, their identifiers, and key characteristics of these concept names (e.g., language, vocabulary source, name type). The entire concept structure appears in a single file in the Rich Release Format (MRCONSO.RRF). There are other RRF/ORF files in the UMLS database that specify the Attribute, Relationship, indexes and Data about the Metathesaurus. We have mainly focussed on the following terms and identifiers [14] for ambiguity reduction process:

1) **Concepts and Concept Identifiers (CUI):** A concept is an abstract idea represented by a string. A concept can have different names. Each concept in the Metathesaurus has a unique concept identifier (CUI). The CUI has no intrinsic meaning. In other words, you cannot infer anything about a concept just by looking at its CUI.
2) **Concept Names and String Identifiers (SUI):** Each unique concept name or string in each language in the Metathesaurus has a unique and permanent string identifier (SUI). Any variation in character set, upper-lower case, or punctuation, is a separate string, with a separate SUI. If the same string, e.g., 'Cold', has more than one meaning, the string identifier will be linked to more than one concept identifier (CUI) but it will have only one SUI.
3) **Atoms and Atom Identifiers (AUI):** The basic building blocks or 'atoms' from which the Metathesaurus is constructed are the concept names or strings from each of the source vocabularies. Each and every occurrence of a string in each source vocabulary is assigned a unique atom identifier (AUI). If exactly the same string appears more than once in the same vocabulary, a unique AUI is assigned for each occurrence. When the same string appears in multiple source vocabularies, it has different AUIs for every time it appears as a concept name. All of these AUIs are linked to a single string identifier (SUI), since they represent occurrences of the same string.
4) **Terms and Lexical Identifiers (LUI):** For English language entries in the Metathesaurus, each string is linked to all of its lexical variants or minor variations by means of a common term identifier (LUI). Like a string identifier, the LUI for an English string may be linked to more than one concept. This occurs when strings that are lexical variants of each other have different meanings. In contrast, each string identifier and each atom identifier can only be linked to a single LUI.
5) **Terms and Semantic Type Identifiers (TUI):** Each Metathesaurus concept is assigned at least one semantic type along with its identifiers (TUI), that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. In all cases, the most specific semantic type available in the hierarchy is assigned to the concept. For example, the concept "Colonoscop" is categorized in 'Diagnostic Procedure' semantic type and assigned 'T060' identifier (TUI). There are overall 133 semantic types present in the UMLS.
6) **Semantic Group (TUI group):** The UMLS semantic network reduces the complexity of the Metathesaurus by grouping concepts according to the semantic types that have been assigned to them [15]. For certain purposes, however, a set of semantic type groupings may be desirable. The following principles were used to design the groupings: semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. For example, semantic types like 'Diagnostic Procedure','Laboratory Procedure' and 'Therapeutic or Preventive Procedure' are grouped into 'Procedure'.
7) **Rank:** Its is defined as precedence of vocabulary source or dictionary and term types that is used to compute the default preferred concept name for each concept in the Metathesaurus. For example, the concept "Abdomen" has two term types 'Anatomical structure' and 'Finding'

[1]The system is not available for public at this time. If someone wants to use it, one can contact the company for it.

from the source vocabularies 'SNOMEDCT_US' and 'CCPSS' with RANK numbers '467' and '33' respectively. Here the higher rank is '467'. The preference of the source dictionary can be changed by using UMLS MetamorphoSys tool [16].

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only	Term Type (TUI)
C0020538 High blood pressure high blood pressure Hypertension Hypertension, NOS	**L0005827** High blood pressure high blood pressure	**S0004538** High blood pressure	**A0011516** High blood pressure (from ICPC2P) **A0011517** High blood pressure (from RCD)	**T047** Disease or Syndrome
		S1344447 high blood pressure	**A1303965** high blood pressure (from AOD)	
	L0020538 (synonym) Hypertension Hypertension, NOS	**S0050509** Hypertension	**A0070973** Hypertension (from CCS)	
		S0490041 Hypertension, NOS	**A0558763** Hypertension, NOS (from SNMI) **A22871327** Hypertension, NOS (from SNOMEDCT_US)	

Fig. 1. An example of Concept, String, Atom, and Term identifiers in the UMLS.

As shown in Fig. 1, 'High blood pressure' appears as an atom in more than one source vocabulary and has a distinct AUI for each occurrence. Since each of these atoms has an identical string or concept name, they are linked to a single SUI. But, 'high blood pressure', the lowercase of 'High blood pressure' has a different string identifier. Since both are lexical variants of each other, they are linked to the same LUI. There is a different LUI and different SUIs and AUIs for 'Hypertension' and 'Hypertension, NOS'. Since both have been judged to have the same meaning (synonyms), they are linked to the same CUI.

III. Ambiguity in the UMLS

There are different types of ambiguities commonly occurring in the Metathesaurus. Reference [8] have described four major classes of ambiguity which are explained below:

A. Classes of ambiguity

1) **Contextual ambiguity:** They call it false ambiguity because its source is not the actual meaning of the string itself. This class of ambiguity arises from terms which require context within their vocabulary in order to be properly understood. For example, the word 'prostate' does not mean a disease, but it means 'prostatic diseases' when it is used in the hierarchy of 'disease'. Contextual ambiguities can be further classified according to their participants:
   - **Body part/disease** ambiguity as in 'Heart' and 'Heart failure'
   - **Body part/procedure** ambiguity as in 'Graft' and 'Graft Procedures'
   - **Pathology/procedure** ambiguity as in 'Pathology' and 'Pathology procedure'
   - **Medical device/procedure** ambiguity as in 'Prosthesis' and 'Prosthesis Implantation'
2) **Generalization ambiguity:** This is also a false ambiguity caused by grouping several concepts together using a more general term. For example, 23 concepts including 'Protocols: Activities' and 'Protocols: Pre- or Intra- or Post-Procedure' are generalized to 'Protocols' which does seem to be a legitimate synonym of the concept 'Protocols documentation'.
3) **Meta ambiguity:** This class of ambiguity, represented by strings such as 'Stress fracture, NEC in ICD10_1998', contains meta information. In this case, 'ICD10_1998', is the name of the vocabulary. These strings disclose excessive information. The meaning of a string containing 'NEC', 'not elsewhere classified' or similar phrase depends upon its vocabulary, but such information is already available in the MSRO file (where it belongs). However, for practical purposes, most users do not want or need to resolve this ambiguity.
4) **Abbreviation ambiguity:** This is another, large class of ambiguity caused by distinct concepts having the same acronyms or abbreviation. An example of this type of ambiguity is that the phrases -'Mitral Valve Stenosis', 'Multiple Sclerosis', 'Morphine Sulfate' and 'millisecond' all have the same abbreviation 'MS' or 'ms'.

B. Conceptual Ambiguity in the UMLS

The definition of ambiguity we deal with here is very specific which is related to 'concept' in the UMLS. A term (label with all its string variants) can have multiple CUIs. *When more than one CUI of a term belongs to same TUI group, the term is ambiguous.* Thus after ambiguity resolution, there is only one CUI in a TUI group corresponding to one term.

IV. Centralized Knowledge Base Creation Methodology

The ambiguity is one of the major cause of poor performance of the most NLP systems. The ambiguity in the data on which a system operates creates limitations of the system performance. Thus, reducing ambiguity in the source data only can improve the systems significantly. The methodology used here for creation of unambiguous centralized knowledge base of clinical domain involves the following major steps:

1) Dictionary Selection
2) Ambiguity Reduction

A. Dictionary Selection

Dictionary selection was done to preserve the CUI and text and to remove the irrelevant content. The UMLS metathesaurus constitutes one of the most important biomedical term repositories, integrating 179 source vocabularies in 21 natural languages [1]. The UMLS version - 2016AB contains 128

English language dictionaries which we used for the central knowledge base creation. There are two major steps in this process. First, we run redundancy check process on all 128 English dictionaries and exclude the redundant dictionaries from the list. Then we calculate relevance score for all the remaining dictionaries in the list. The dictionaries below a threshold of relevance score are removed.

1) Redundant dictionary exclusion: We create the overlap matrix for all dictionaries against all other dictionaries. We define a 'concept' as "a unique pair of case insensitive concept name and concept type (TUI)". The overlap score of dictionary X in dictionary Y is calculated as:

$$\frac{\text{Number of concepts } (X \cap Y)}{\text{Number of concepts } (X)} \times 100 \tag{1}$$

Each cell in the overlap matrix represents the overlap score of 'the dictionary in the row' over 'the dictionary in the column'.

The dictionaries in the row, which have 100% overlap score against some other dictionary in a column, are considered a proper subset of the other ones. All such 'proper subset' dictionaries are considered redundant and excluded from further process.

2) Relevance score calculation: After the removal of redundant dictionaries, we calculate the relevance score for all of the remaining dictionaries. There are two prerequisites for relevance calculation: a corpus annotated with named entities and entity types, and mapping of the UMLS concept types to these entity types if these are different.

We have an entity annotated corpus of clinical documents comprising of 5,160 documents with 398,568 sentences. These documents are physicians notes from hospitals and clinics of US. The documents have large part as unstructured text and also some tables and template sentences. The corpus contains annotated 443,328 clinical entities. Each annotated entity is associated with one of eleven entity types: problem, finding, procedure, anatomical structure, body function, laboratory data, medical device, medicine, body measurement, measurement value and modifier. We extract case insensitive list of unique entities with entity types from the corpus. Our list of entity types in the corpus is same as the concept types in the UMLS. In case it is different, one needs to convert the entity types of the unique entity list into the corresponding concept types of the UMLS.

For each dictionary, we calculate two types of coverage scores - corpus coverage in dictionary (CCID) score and dictionary coverage in corpus (DCIC) score.

Corpus coverage in dictionary (CCID) score: each entity from the unique entity list with its entity type is searched against the concept name and concept type in the dictionary. If it is found, the searched entity is considered present in the dictionary. Then the CCID score is calculated as proportion of common entities in the corpus and dictionary to the total number of entities in the unique entity list:

$$\frac{\text{Unique entity list } \cap \text{ dictionary}}{\text{Total entities in the unique entity list}} \times 100 \tag{2}$$

Dictionary coverage in corpus (DCIC) score: each concept name from the dictionary with its concept type is searched against the entity and entity type in the unique entity list. If it is found, the searched concept is considered present in the unique entity list. TThen the DCIC score is calculated as proportion of common entities in the corpus and dictionary to the total number of concepts in the dictionary:

$$\frac{\text{Unique entity list } \cap \text{ dictionary}}{\text{Total concepts in the dictionary}} \times 100 \tag{3}$$

To calculate the relevance score, we borrow the formula from F1 score. The relevance score is the harmonic mean of the CCID and DCIC scores.

$$\frac{2 \times CCID \times DCIC}{CCID + DCIC} \tag{4}$$

After calculating the relevance score of all the dictionaries, we sorted them in decreasing order of the score. The dictionaries which had the score below our threshold[2] (0.12) were removed. In this process, we removed 16 redundant and 26 irrelevant, and a total of 42 dictionaries out of 128 and we came up with 86 selected dictionaries. Following are the examples of dictionaries which were excluded as redundant or irrelevant.

Example 1) NCI (National Cancer Institute) Thesaurus - It covers vocabulary for cancer-related clinical care, translational and basic research, and administrative activities. It also has 23 sub-dictionaries like NCI_CareLex, NCI_CDC, used for different purposes. We removed those sub-dictionaries as all the concept in those are covered in NCI itself.

Example 2) All the concepts in Healthcare Current Dental Terminology (HCDT) are included in Healthcare Common Procedure Coding System (HCPCS) which is a collection of standardized codes that represent medical procedures, supplies, products and services. So this dictionary was removed as redundant.

Example 3) All the contents of National Drug Data File (NDDF) are contained in RXNORM dictionary which covers all prescription medications approved for human use in the United States. So NDDF was removed as redundant dictionary.

Example 4) Dictionaries like GO, HGNC, OMIM, NCBI are gene based dictionaries and these were removed as irrelevant dictionaries.

Example 5) There are other source dictionaries like SOP (Source of Payment Typology), SRC (Source Terminology Names) containing abstract information like metadata etc. These are removed as irrelevant dictionaries.

After the dictionary selection process we obtained the list of dictionaries sorted by relevance in decreasing order. We

[2]To select the threshold, we followed an iterative process that included (a) removing a dictionary from the bottom (least relevance score), (b) building the PKB (c) curation of PKB as version 4 (described in Ambiguity Reduction section), and (d) checking the performance of CRF based CER. When removing a dictionary decreased the performance of CER significantly, we chose the relevance score of the last removed dictionary as the threshold.

pass on this list to the **MetamorphoSys** tool (the UMLS installation and customization program) [16]. This tool gives ranking to the selected dictionaries and helps creating **Primary Knowledge Base** (PKB) which is used for the next step. This PKB is designed as a Relational Database and has the data structure similar to UMLS, but differs in number of dictionaries and number of concepts in the dictionaries.

B. Ambiguity Reduction

We started the ambiguity reduction process in an incremental way, increasing the complexity at each level to obtain better results. We created 4 versions of the knowledge base from the PKB. Out of all these versions, the results of the CER system are the best with the version 4 which is our final centralized knowledge base.

Version 1: In this version we take each row of PKB one by one. From each row we take the label (Text) and the CUI. Now for the same label we find all the highest ranked CUIs from PKB. If the current row's CUI is among the highest ranked CUIs then it is placed in centralized knowledge base, otherwise ignored (Fig. 2). However, for a single label, we might get different CUIs that have the same rank. We solved these conflicts manually because the number of such conflicts was small and human validation was needed for selecting correct CUI for these labels. This manual process is applied to all the versions.

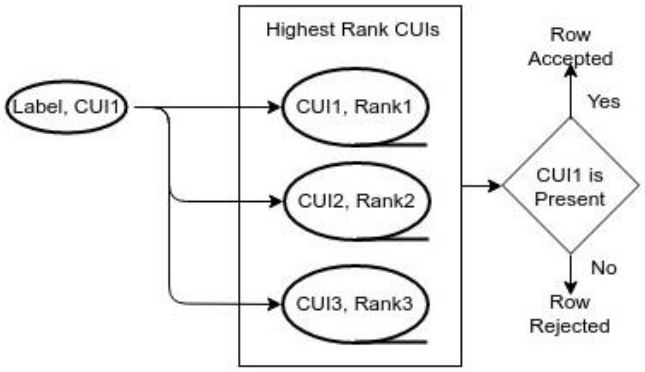

Fig. 2. Ambiguity reduction process for version 1.

Version 2: In previous version we were taking decision only based on CUI and Rank, but not considering the term type (TUI). So, in this version, we add one more level of filter using TUI. There are multiple TUIs for each label, and for each TUI, there are multiple CUI and RANK pairs. We retain the highest ranked CUIs for each TUI of a label. If the CUI of the label is among these highest ranked CUIs, then it is placed in centralized knowledge base, otherwise ignored (Fig. 3). The same process describe in version 1 is used when there is conflict of CUIs having same Rank.

Version 3: In this version we have also considered semantic group (TUI group) along with TUIs. Now for a label, we can have more than one TUI group. We select the highest rank CUIs in every TUI group for a label. If the current row's CUI is among these highest ranked CUIs, then it is placed in centralized knowledge base, otherwise ignored (Fig. 4).

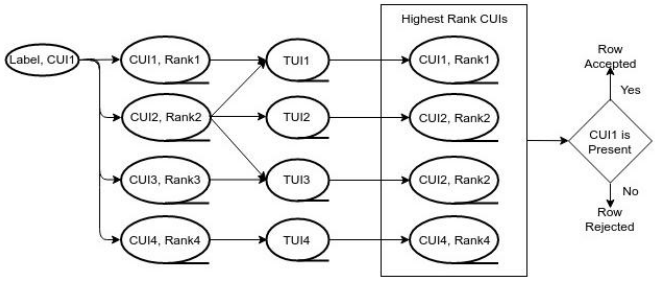

Fig. 3. Ambiguity reduction process for version 2.

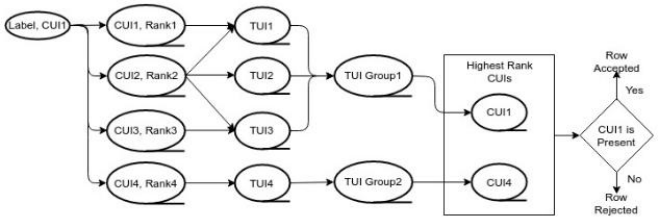

Fig. 4. Ambiguity reduction process for version 3.

Version 4: In previous versions of curation methods, we were performing text matching on the labels, however, here we perform term matching. For example, 'High Blood Pressure' and 'Blood Pressure, High' are two different texts but as a term both are the same. For this we created two more tables using PKB.

1) **Word_WordId_Map:** This table contains a mapping for word to wordId (unique). These words are obtained by tokenizing the labels in PKB.
2) **SUI_WordId_Map:** In this table, the rows contain SUI and corresponding WordId Combination (from Word_WordId_Map), where the WordIds are stored in ascending order.

Following are the steps involved in the this approach:

1) One row was taken from PKB.
2) We took the the CUI (for example C1) and SUI (for example S1) from that row.
3) Then found the corresponding row in SUI_WordId_Map table for the S1 to obtain a combination of WordIds.
4) With this wordId combination, we searched all the rows in SUI_WordId_Map that had this wordId combination, to obtain a set of SUIs.
5) Found all the CUIs (for example C1, C2, C3, C4) for the corresponding SUIs and filtered the highest ranked CUI for each TUI as done in the previous (third) version.
6) If the selected C1 is among the highest ranked CUIs, then the row and its corresponding CUI is placed in centralized knowledge base, otherwise ignored (Fig. 5).

The knowledge base created from version 4 is our final centralized knowledge base which is used by our clinical modules.

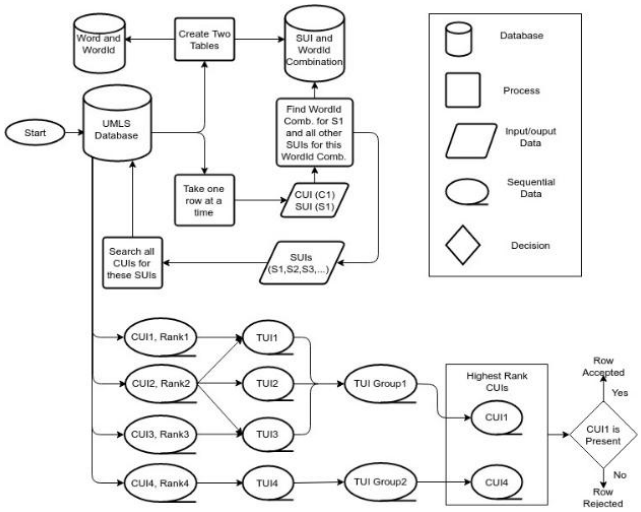

Fig. 5. Ambiguity reduction process for version 4.

V. Updation of Centralized Knowledge Base with new UMLS Version

Each version of the Metathesaurus contains a set of files that summarize the changes from the previous version. With each UMLS release, we have to update the knowledge base so that the systems can utilize the latest data. Though we can directly commit all the changes that are coming from UMLS, we prefer validating changes for our requirement.

We find the new dictionaries which were added new in the UMLS release, and apply dictionary selection process. The new dictionaries go through redundancy check against existing dictionaries in PKB. If it is not redundant, it is checked for relevance against our corpus. If the relevance score is above the threshold, then the list of dictionaries in PKB is updated. Now using Metamorphosys tool, we create a new version of primary knowledge base (PKB_NEW) from the updated list. Then, we find the differences in PKB_NEW with respect to PKB using SUI, label, dictionary name and TUI. We commit new changes in centralized knowledge base following the curation process of version 4. The new changes are recorded and can be validated by human expert if required.

VI. Statistics of the Knowledge Base

In this section we present different statistical data about our knowledge base curated in four steps.

Table I lists different types of contents present in the original UMLS database and our centralized knowledge base, as well as the reduction percentage of it in size as compared to UMLS.

Table II presents the comparison of number of entities in the two databases, belonging to top ten dictionaries out of 86.

Table III presents the comparison of the two databases in terms of number of entries categorized as top ten concept type.

Finally, the Table IV presents the number of ambiguous entries in the database that are resolved with this process.

TABLE I
Content comparison of original UMLS and centralized knowledge base.

	UMLS	Centralized knowledge base	Reduction
Dictionary	128	86	32.81%
Concepts (CUI)	3,436,328	2,187,825	36.33%
Label(Text)	9,417,451	6,914,311	26.57%
String (SUI)	7,772,712	5,607,818	27.85%

TABLE II
Dictionary-wise content comparison (Top 10)

	UMLS	Centralized knowledge base	Reduction
SNOMEDCT_US	1,304,951	1,290,875	1.07%
MEDCIN	870,686	868,885	2.06%
MSH	843,848	840,639	0.38%
LNC	436,981	433,607	0.77%
RCD	347,568	344,466	0.89%
ICD10PCS	337,153	337,149	0.0012%
RXNORM	308,962	308,043	0.29%
NCI	287,420	268,244	6.67%
MTH	185,389	184,051	0.72%
ICD10CM	176,633	176,053	0.32%

TABLE III
Semantic group-wise content comparison (Top 10)

	UMLS	Centralized knowledge base	Reduction
Chemicals and Drugs	3,049,456	2,691,787	11.72%
Disorders	2,387,807	2,380,093	0.32%
Procedures	1,143,924	1,125,121	1.64%
Anatomy	496,336	451,663	9.00%
Physiology	542,468	385,309	28.97%
Living Beings	1,937,812	249,680	87.11%
Genes and Molecular Sequences	352,895	176,245	50.05%
Concepts & Ideas	192,912	162,128	15.95%
Devices	166,413	151,442	8.99%
Objects	65,733	47,435	27.83%

TABLE IV
Overall ambiguity in UMLS and number of concepts deleted by Version 4 to resolve the ambiguity.

Total concepts in UMLS	3,436,328
Total ambiguous concepts in UMLS	111,909
Removed ambiguous concepts to resolve ambiguity	52,486 (46.90%)

VII. Result Analysis

To calculate the impact of curation exercise, we compared the performance of a CER using both the databases. We processed 12,364 clinical documents using original UMLS database, and using our centralized knowledge base (version 4). The Table V presents the results of the CER system using both the databases.

TABLE V
PERFORMANCE OF A CER USING BOTH THE DATABASES.

	UMLS	**Centralized Knowledge base**	**% increase in CER**
Entity detected with lookup	6,465,439	6,689,570	3.35%
Entity detected with CRF	1,195,476	1,233,765	3.10%

As we can see in Table V that the CER system shows significant improvement when using our curated knowledge base as compared to when using the original UMLS.

TABLE VI
CUI DETECTION RESULTS ON SEMEVAL 2015 DATA.

	UMLS	**Centralized knowledge base**
Precision	47.19%	63.62%
Recall	77.14%	71.01%
F1-Score	58.55%	67.12%

Table VI shows results of CUI detection on dataset of SemEval-2015 Task 14: Analysis of Clinical Text [17]. We got more than 8% improvement in F1-score when we used curated centralized knowledge base in place of original UMLS.

VIII. Conclusion

UMLS is the largest metathesaurus for biomedical and health related concepts. We only need a relevant part of it. So we used an automated process to select the most relevant dictionaries and remove ambiguities. As shown in Table IV, we have removed 47% entities out of all entities of ambiguous nature to resolve the conflicts. This resulted in an immediate increase in CER performance by $>3\%$. Moreover, the time complexity of the system was reduced and we obtained more accurate outputs. The larger part of the process is automated and thus does not require deep knowledge of the UMLS for implementation. We have applied the dictionary selection and curation process on English dictionaries from the UMLS. The process is language neutral and can be applied for creating such knowledge base for any language. The process is also neutral to use case. For example, if someone wants to create knowledge base for a use case that frequently addresses gene related issues, the same process can be used and it will select a different set of dictionaries which are more relevant for that purpose. In future, We can improve our relevance score calculation such that it can find relevant dictionaries even with smaller entity corpus.

References

[1] National Library of Medicine (US), "Umls manual," 2009, https://www.ncbi.nlm.nih.gov/books/NBK9675/, Last accessed on 2018-03-08.

[2] Wikipedia article on UMLS, "Unified medical language system," 2017, https://en.wikipedia.org/wiki/Unified,Last accessed on 2018-02-08.

[3] A. R. Aronson, T. C. Rindflesch, and A. C. Browne, "Exploiting a large thesaurus for information retrieval," in *Intelligent Multimedia Information Retrieval Systems and Management-Volume 1*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1994, pp. 197–216.

[4] C. Chute, M. Tuttle, Y. Yang, D. Sherertz, N. Olson, and M. Erlbaum, "A preliminary evaluation of the umls metathesaurus for patient record classification." in *Proceedings. Symposium on Computer Applications in Medical Care*. American Medical Informatics Association, 1990, pp. 161–165.

[5] N. Shah, A. Sheth, S. Bhatt, R. Goswami, V. Shah, R. Kanani, A. Patel, and P. Pathak, "Data processing system and method for computer-assisted coding of natural language medical text," May 12 2016, uS Patent App. 14/918,881.

[6] W. R. Hersh, D. H. Hickam, and T. Leone, "Words, concepts, or both: optimal indexing units for automated information retrieval." in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1992, p. 644.

[7] M. Stevenson and Y. Guo, "Disambiguation of ambiguous biomedical terms using examples generated from the umls metathesaurus," *Journal of biomedical informatics*, vol. 43, no. 5, pp. 762–773, 2010.

[8] F.-M. Lang, S. E. Shooshan, J. G. Mork, and A. R. Aronson, "Ambiguity in the umls metathesaurus 2010 edition," 2009, https://ii.nlm.nih.gov/Publications/Papers/ambiguity10.pdf,Last accessed on 2018-03-05.

[9] X. Zhu, J.-W. Fan, D. M. Baorto, C. Weng, and J. J. Cimino, "A review of auditing methods applied to the content of controlled biomedical terminologies," *Journal of biomedical informatics*, vol. 42, no. 3, pp. 413–425, 2009.

[10] T. C. Rindflesch and A. R. Aronson, "Ambiguity resolution while mapping free text to the umls metathesaurus." in *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 1994, p. 240.

[11] J. Hellrich, S. Schulz, S. Buechel, and U. Hahn, "Jufit: A configurable rule engine for filtering and generating new multilingual umls terms," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 604.

[12] Alan Aronson, "Metamap - a tool for recognizing umls concepts in text," 2000, last accessed on 2000-11-04.

[13] K. M. Hettne, E. M. van Mulligen, M. J. Schuemie, B. J. Schijvenaars, and J. A. Kors, "Rewriting and suppressing umls terms for improved biomedical term identification," *Journal of biomedical semantics*, vol. 1, no. 1, p. 5, 2010.

[14] National Library of Medicine (US), "Umls 2006aa documentation," 2006, https://www.nlm.nih.gov/research/umls/2006AA_umls_documentation.pdf, Last accessed on 2018-01-10.

[15] National Library of Medicine (NLM) (US), "Semantic types and groups," 2009, https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml.,Last accessed on 2018-03-10.

[16] W. T. Hole, B. L. Humphreys, and M. S. Srinivasan, "Customizing the umls metathesaurus for your applications," in *Proceedings of the AMIA Symposium*, 2000.

[17] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, and G. Savova, "Semeval-2015 task 14: Analysis of clinical text," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 303–310.