Towards a semantic lexicon for clinical natural language processing

Hongfang Liu[1], Stephen T. Wu[1], Dingcheng Li[1], Siddhartha Jonnalagadda[1], Sunghwan Sohn[1], Kavishwar Wagholikar[1], Peter J. Haug[2], Stanley M. Huff[2], Christopher G Chute[1]

[1]Mayo Clinic College of Medicine, Rochester, MN [2]Intermoutain HealthCare, Murray, UT

Abstract

A semantic lexicon which associates words and phrases in text to concepts is critical for extracting and encoding clinical information in free text and therefore achieving semantic interoperability between structured and unstructured data in Electronic Health Records (EHRs). Directly using existing standard terminologies may have limited coverage with respect to concepts and their corresponding mentions in text. In this paper, we analyze how tokens and phrases in a large corpus distribute and how well the UMLS captures the semantics. A corpus-driven semantic lexicon, MedLex, has been constructed where the semantics is based on the UMLS assisted with variants mined and usage information gathered from clinical text. The detailed corpus analysis of tokens, chunks, and concept mentions shows the UMLS is an invaluable source for natural language processing. Increasing the semantic coverage of tokens provides a good foundation in capturing clinical information comprehensively. The study also yields some insights in developing practical NLP systems.

Introduction

A critical component in achieving semantic interoperability for secondary use of EHRs is to normalize clinical information stored in EHRs to standard structure and value. However, much of the data in EHRs is in free text form since compared to structured data, the free text data are a more efficient way in the health care environment to express concepts and events. The free text data is often a result of dictation transcription, clinician direct entry, or deployment of speech recognition applications. However, free text is very challenging for searching, summarization, decision-support, or statistical analysis. Natural language processing (NLP) has the promise to automatically extract and encode clinical information in clinical narratives. A semantic lexicon which associates words and phrases in text to concepts is needed. The construction of a semantic lexicon has been acknowledged as one of the most challenging NLP tasks (1-3). Existing standard terminologies such as SNOMED-CT, CPT, or LOINC may have limited coverage with respect to concepts (i.e., concrete meanings; e.g., clinical finding “chest pain”) and concept mentions (i.e., there are multiple different ways to mention a concept; e.g., “chest pain”, “pain in chest”, “pain in the chest”, and “pain in her chest”) in clinical narratives. In general, a semantic lexicon for clinical NLP should describe observed usage of words in clinical narratives, be comprehensive, and include a wide range of variants of a concept appearing in the text.

In this paper, we present our investigation towards building a semantic lexicon, MedLex, for processing clinical text for the purpose of extracting information from clinical narratives. Tokens, phrases, and concept mentions observed in a large corpus are analyzed for assessing the UMLS with respect to morphosyntactic and semantic coverage. The resulting lexicon has been utilized by Mayo Clinic’s production NLP pipeline to process millions of clinical documents. We will release MedLex as an open source resource through open health natural language processing (OHNLP).

Background and Related work

Major sources of clinical domain semantics are terms in ontologies and controlled terminologies gathered by the Unified Medical Language System (UMLS), which is developed and maintained by the National Library of Medicine (NLM). The main purpose of the UMLS is to integrate multiple different electronic clinical and biomedical ontologies and terminologies, where most of those sources were created by a group of domain experts. It captures domain semantics well but since it is not data-driven, it does not reflect observed usage of words in clinical narrative: many terms in the UMLS may never occur in text and verse visa. Filtering of terms that never occur in text has been attempted using statistical models and rule-based systems. Statistical models typically identify a number of properties that allow prediction of the likelihood that a given term is found or not found in a particular corpus (4). Recent rule-based efforts by Hettne et al.(5) recommend applying five rewrite rules and seven suppression rules to the UMLS before it is used for biomedical term identification in MEDLINE. McCray et al. evaluated the nature of terms in the UMLS with respect to their likelihood of appearing in a natural language corpus (6). They found that only 10% of the terms in the UMLS occurred in their MEDLINE corpus (representing 1 year of abstracts). Moreover, they identified term properties, which could be used to filter out terms that are unlikely to occur in a corpus.

There are several studies exploring the use of the UMLS for building a semantic lexicon for clinical NLP. Johnson et al utilized a corpus of discharge summaries from hospital visits for the construction of a lexical resource from the UMLS in support of processing of medical narrative, specifically(3). A semantic lexicon with 75,000 entries was automatically generated from the UMLS. By evaluating the coverage of the lexicon, they estimated that 38% of the terms in the corpus had both morphosyntactic and semantic information derived from the UMLS. Previously, Friedman et al quantitatively compared a lexicon developed manually for the MedLEE system with a lexicon derived automatically from the UMLS, with respect to the task of processing clinical information in patient reports (7). The UMLS-derived lexicon led to degraded performance relative to MedLEE's own lexicon. The results do not, however, invalidate the UMLS as an important source of lexical information, as they may simply be a reflection of the completeness of the existing MedLEE lexicon for the task evaluated. Liu et al applied a corpus-based method for facilitating vocabulary development that displays term frequency, relations of terms to other terms, the compositional components of terms, and mappings to controlled terminologies (8, 9). Our previous corpus analysis work includes comparisons between concepts in the clinical and biomedical domains (10) and comparisons of lexicon filtering based on sources, rules, and corpus statistics (11). We found that features intrinsic to UMLS terms (well-formedness, length, and language) generalize easily across clinical institutions, but term frequencies should be adapted with caution.

Built upon related and prior work, in this paper, we analyze how tokens and phrases in a large corpus distribute and how well the UMLS captures the morphosyntactic and semantic information. A raw corpus-driven semantic lexicon, MedLex, was acquired where the semantics is based on the UMLS assisted with terminology mining from EHRs and usage information gathered from clinical text. The main hypothesis driven our research in clinical NLP is that "clinical documents intend to capture information critical for clinical care and tokens/phrases in those documents should convey or provide assistance in conveying some kind of clinical information". However, concepts can appear in documents in different forms. Discovery, mapping and integration of new terms and their variants into existing terminologies have constituted a major challenge for practical NLP systems. Some common types of variant forms include orthographical (tumour vs tumor), punctuations (B.I.D and BID), morphological (foot cramps vs foot cramping), structural (breast cancer vs cancer of breast), short forms (SOB vs short of breath) and semantics (dyspnea vs short of breath). Identification of these variant forms from text is critical in information extraction and encoding. Additionally, there are many tokens and phrases pertaining to specific care event (e.g., names of patients or physicians or address information) and institutions (e.g., names of rooms or buildings). A list of terms pertaining to local institutions is also needed when building practical NLP systems.

Methods

Our study is based on a large collection of clinical documents collected at Mayo Clinic in our clinical data warehouse and the UMLS (2011AA version) distributed by NLM. In the following, we describe the resources used in the study. Detail methods in acquiring a preliminary version of the MedLex are presented next.

Resources

NLP-annotated Enterprise Data Trust (EDT) Corpus - Mayo Clinic has established the EDT data warehouse that enables access of all data generated from patient care, education, research, and administrative transaction systems. It is organized to support information retrieval, business intelligence, and high-level decision making (12). The EDT corpus used in the study includes 55M notes, all annotated with part-of-speech (POS) and chunk information from IBM shallow parser with our production NLP pipeline, Clinical Note Indexer (CNI) (13). Additionally, a broad range of note types at Mayo were represented, including Clinical Note, Hospital Summary, Post-procedure Note, Procedure Note, Progress Note, Tertiary Trauma, and Transfer Note.

The Unified Medical Language System (UMLS) – The UMLS contains three knowledge sources: the Metathesaurus (META), the Specialist Lexicon (LEX), and the Semantic Network (SN) (14). META provides a uniform, integrated distribution format for over 160 biomedical vocabularies, terminologies, and ontologies, and contains rich semantic relationships of terms. LEX contains syntactic information for many terms, component words, and English words, including verbs. SN contains information about the semantic categories (e.g., "Disease or Syndrome", "Virus") to which all META concepts have been assigned and the permissible relationships among these types (e.g., "Virus" causes "Disease or Syndrome"). There are 133 semantic types grouped into 15 groups following Bodenreider and

McCray's breakdown(15). The version of the UMLS used is 2011AA. Tables in the UMLS used in the study include i) MRCONSO and MRSTY from META and ii) LEXICON, LRAGR, LRABR, LRNOM from LEX.

MedTagger - MedTagger takes advantage of large corpus statistics, machine learning, knowledge bases, and syntactic parsing to detect concept mentions in clinical text (16). It was adapted from BioTagger-GM (17), a tool developed for literature mining domain. Our previous research has shown that both systems achieve the-state-of-the-art performance where the F-Measure of MedTagger at 2010 i2b2 NLP challenge on the concept mention task (18) was 84% (ranked 4th among 20 teams) and the F-Measure of BioTagger-GM at 2007 BioCreAtive II on the gene mention task (19) was 0.87 (also ranked 6th among 20 teams). In our current study, all notes were annotated by the mapping module in MedTagger to detect UMLS terms occurring in the text.

Short Forms in the clinical document - Medical writing favors brevity due to a large amount of information required to be conveyed in a short time and limited space. The problem with short forms (i.e., abbreviations and acronyms) is that their corresponding long forms may differ according to the context. For example, *BPD* can mean *broncho-pulmonary dysplasia* or *borderline personality disorder*, *Hb* can refer to *hemoglobin* or *hepatitis*, and *TOF* can refer to *tetralogy of Fallot* or *tracheo-oesophageal fistula*. The expansions of these short forms differ widely in their intended meanings. Moreover some medical terms are shortened in different forms. For example, *normal* can be abbreviated as *N*, *Nl*, or *NAD* or *hemoglobin* as *Hb*, *HGB*, or *HbA*. Previously, we have studied short forms in the UMLS as well as in the MEDLINE corpus and notice the meaning of short forms in EHRs can differ from ones defined in the knowledge base (20-22). From those studies, we have developed a set of short form extraction tools that we can deploy to gather short form definitions from the UMLS and the EDT corpus.

Data preparation

We gathered all tokens and chunks detected by the IBM Shallow Parser and all concept mentions detected by MedTagger in the EDT corpus. Observing a large amount of ambiguity of clinical terms in the UMLS is associated with short forms (i.e., abbreviations and acronyms), but the corresponding long forms (i.e., expansions or definitions) may never occur in the clinical text, we also gathered short form definitions from the corpus using systems built before (20-22) .

Data analysis and lexicon construction

Single-word lexeme - Since single-word lexemes are building blocks for the text and the meaning of the majority of the multi-word lexemes can be inferred from its constituent single-word lexemes, we pay special attention in our data analysis on single-word lexemes. We begin with tokens with occurrences at least 10 in the corpus. Token mentions are then grouped by ignoring cases and classified into four groups: I) occur as entries in the Specialist Lexicon variant table (LRAGR) and as tokens in META; II) appear as entries in LRAGR only; III) appear as tokens in META only; and IV) do not appear in META or LEX. For the last group, several subgroups are defined: IV.a) proper nouns (only occurring in text with some characters in capital form); IV.b) contain numbers and punctuation marks; IV.c) potential typos with high confidence corrections, and IV.d) potential typos with low confidence correction. The spelling check was conducted using Aspell, a free software spell checker, with default English dictionaries supplemented with tokens found in the first three groups. For tokens in group I, we include them in MedLex where morphosyntactic and semantic information can be acquired from the UMLS in different ways. In the remaining three groups, some manual curation is needed at least for those popular tokens.

If a token is a term in META, we include it in MedLex as a single word lexeme without any curation where morphosyntactic information is imported from the corresponding entry in LEX and semantic information is imported from META. The usage information of the lexeme and the corresponding concepts are represented by an integer from 0 to 10. Since we only consider tokens with occurrences at least 10 and the most popular term is about 4 power of 10, we define the usage statistics to be the integer of the logarithm of the number of occurrences minus 9 base 4. For example, "vasectomies" as a token occurred 143 times in our corpus, it corresponds to a lexicon entry "vasectomy" in LEX with a unique identifier "E0064093". The corresponding concept is "vasectomy" with unique identifier "C0042387" and semantic group "Procedure". The term usage statistics is 3 (i.e., *log*(143-9, 4) = 3.559 and the integer part of 3.559 is 3) and the concept usage statistics is 8 since the total number of occurrences for C0042387 in our corpus is 73927 and log(73927-9,4)=8.09. The most frequent concept target for vasectomies is "vasectomy" which has a usage statistics of 8.

We use a set of heuristic rules to infer lexical information for tokens that cannot be directly imported from the UMLS. For tokens failed to be assigned by rules, we leave the semantic assignment as "TBA".

Stemming rule: Tokens sharing a common root are semantically similar. We can infer the meaning of one from another. For example, “diencephalic” (LEX unique identifier as E0022543), occurred 139 times in the corpus, is an adjectival form for “diencephalon”, an anatomical site with concept identifier “C1281065”. The semantic information of “diencephalon” will be ported to “diencephalic”. The term usage information will be 3. The concept usage information will be updated to include the adjectival form. We obtain the mapping between theh adjectival forms and the corresponding noun through stemming. For an adjective form, we generate candidate stems by removing zero to four letters. For example, “spectroscopic” generates five candidates: “dienceph”, “diencepha”, “diencephal”, “diencephali”, and “diencephalic”. For all tokens with known semantics, we also generate five candidates by removing zero to four letters. The token with the most number of matches is considered to be the corresponding nouns. The one with the most number of matches for “diencephalic” is “diencephalon” with three matches “dienceph”, “diencepha”, “diencephal”. We then assign the semantics of “diencephalon” to “diencephalic”.

Single concept rule: For a token that appears in just one concept, the semantics of the corresponding concept will be associated with the token after manual review. For example, “humidification” is a token uniquely associated with Procedure “Humidification therapy” (C0418987). We consider the semantics of “humidification” the same as that of “Humidification therapy”.

Medical morpheme rule: The meaning of a medical term sometimes can be inferred from roots, suffixes, and prefixes used (23). If the root of a term is associated with a UMLS concept, we can assume the semantic type of the term is the same as its root. For example, “glenohumeral” ends with “humeral” (C0020164, an anatomical site). We can infer the semantic group for “glenohumeral” as anatomical site. The prefixes “hyper” and “hyp(o)” denote something as “beyond normal” or “below normal”. The corresponding semantic groups for tokens beginning with “hyper” or “hyp(o)” would be “Finding”. The suffixes “copy” and “graphy” refer to processes and the corresponding semantic groups for tokens ending with “copy” and “graphy” would be “Procedure”.

Short forms – We focus on short forms that are at most six characters with at least one letter capitalized. We filter out definitions that are never occurring in the EDT corpus from the list of definitions gathered from the UMLS. The definitions gathered from the corpus are curated after mapping the definitions to META to capture corpus-driven short form definitions.

Multi-word lexemes – We directly import concept mentions detected in our corpus by MedTagger as multi-word lexemes. To mine additional variants of those concepts, we begin with chunks obtained from shallow parsing with occurrences at least 10 in the corpus. We consider chunks that can be mapped to the UMLS after ignoring a list of stop words as potential variants for the corresponding UMLS concepts. A lot of them are prepositional phrases (e.g., “weakness in the legs” or “pain in the low-back” or “tightness in the chest”) that can be collected automatically as variants for the corresponding concepts.

Results

Statistics

Tokens - There were a total of 1.43G unique tokens with 3.277G number of occurrences in the corpus. 410,417 tokens occurred at least 10 times and accounted for 3.274G number of occurrences (99.9%). After transforming to lower case, there were a total of 314,866 unique tokens. Table 1 shows the detailed statistics: 56,828 (18.05%) of them can be found in both LEX and META; 24,000 (7.62%) can only be found in the Specialist Lexicon and 23,418 (7.44%) can only be found in META. The remaining 210,620 tokens (66.9%) could not be found in the UMLS. Figure 1 shows the histogram when grouping according to the logarithm of the number of occurrences base 4 minus 9 in the corpus where left Y-Axis is the percentage of tokens in a group and right Y-Axis is the total number of tokens for a group.

Table 1. Statistics of tokens.

Groups	Description	#Tokens	% Tokens	#Occurrences	% Occurrences
I	In both META and LEX	56,828	18.05%	3,098,921,686	94.65%
II	In LEX	24,000	7.62%	33,170,025	1.01%
III	In META	23,418	7.44%	32,499,233	0.99%
IV.a	Proper Nouns	94,684	30.07%	12,769,286	0.39%
IV.b	Contain Numbers and Punctuations	84,970	26.99%	83,602,469	2.55%
IV.c	Spelling error with high confidence	19,036	6.05%	2,841,759	0.09%
IV.d	Spelling error with low confidence	11,930	3.79%	10,174,054	0.31%

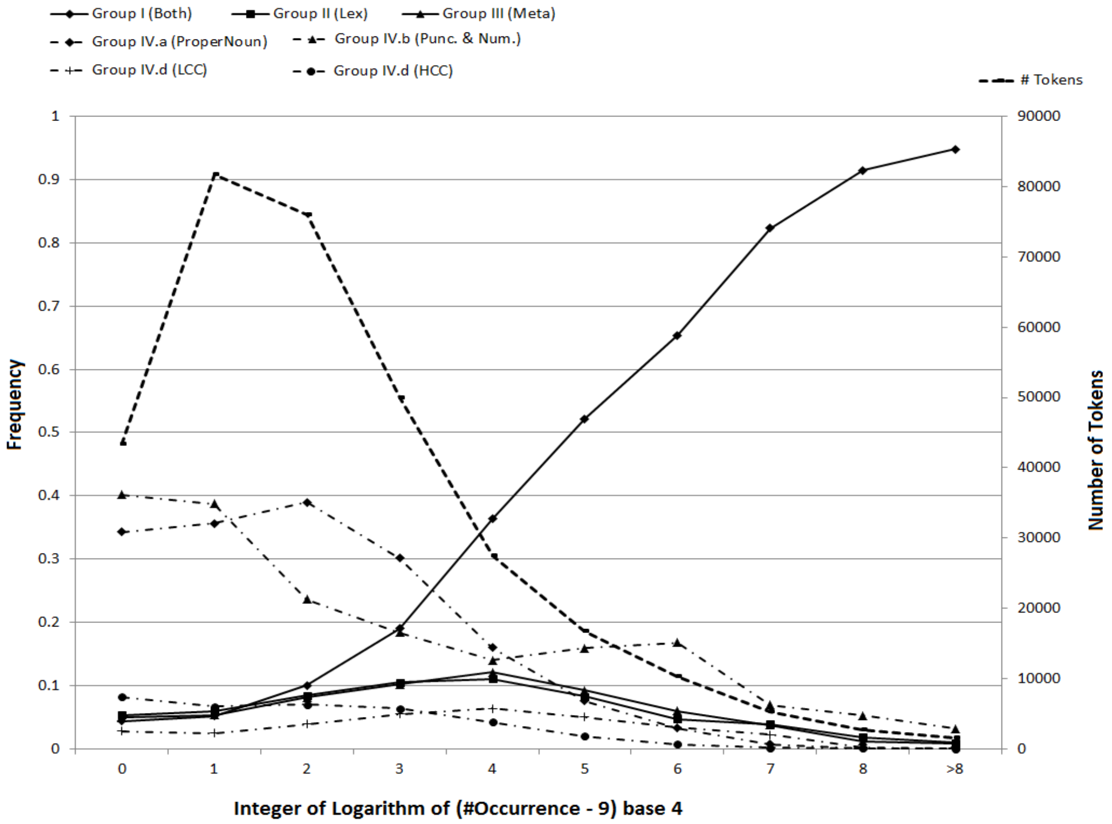

Figure 1. Histogram of different token groups.

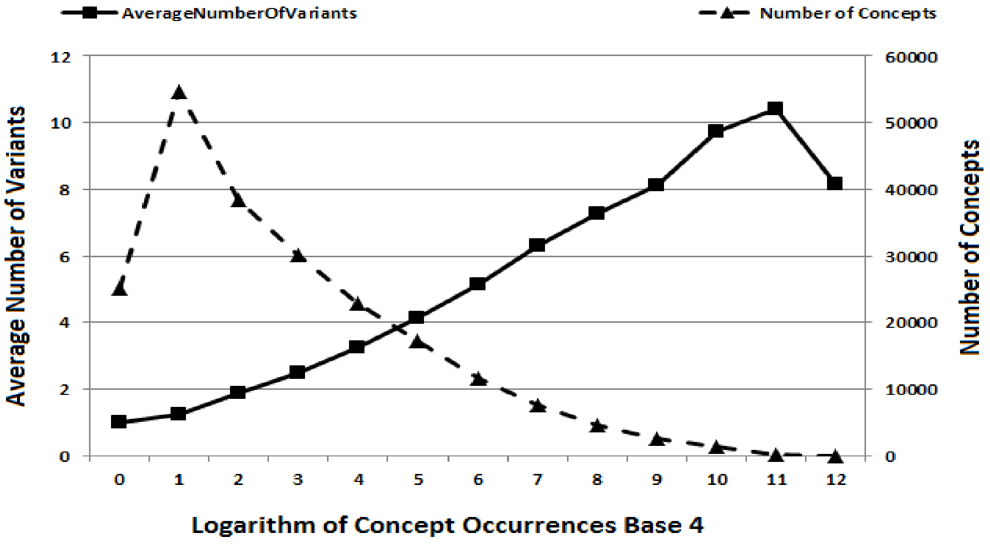

Figure 2. Histogram of different token groups.

Concept mentions and chunks - When using MedTagger to identify all possible concept mentions, 485,294 unique concept mentions were identified corresponding to 208,347 unique concepts and 2.37G total occurrences. Note that terms in the UMLS occur dominantly as noun phrases in the text (11). Figure 2 shows the relationship of the average number of variants per concept (Left Y-Axis) and the number of concepts (Right Y-Axis) when grouping the mentions according to the logarithm base 4 of the number of concept occurrences. There are a total of 107.5M unique chunks with 1.802G number of occurrences in the corpus. Among them, 4.777M (1.649G occurrences) chunks occurred at least 10 times. After transforming to lower case, we have a total of 4.118M unique chunks. After ignoring stop words, we have 0.957M (23.3%) chunks mapped to terms in META (0.908G occurrences; 50.4% of the total occurrences).

Short forms – We obtained 34,169 (short form, concept) pairs corresponding to 17,234 unique short forms from the UMLS where 5,440 pairs (corresponding to 2,803 unique short forms) have the corresponding definition(s) occurred in the corpus. The average ambiguity of short forms for the 2,803 short forms (i.e., the average number of concepts per short form) reduced from 3.65 to 1.94 when restricted to clinical concepts that occur in text. There were 32,268 unique (short form, long form) pairs corresponding to 9,897 short forms extracted from the clinical corpus. Only 2,440 of the short forms were also present in the short form set acquired from the UMLS, and the remaining 7,464 are short forms defined only in the corpus. Manual curation is required to group long forms that have same meanings. For example, ("AVLT", "Auditory Verbal Learning Test") is defined in the corpus but the pair cannot be found in the UMLS. Table 2 shows some of the interesting examples mined from the text.

Table 2. Examples of short forms.

24 full definitions of EGD observed from the text, they differ from each other on punctuation marks	esophagogastro-duodenoscopy; esophagogastro- duodenoscopy; esophago-gastroduodenoscopy;; …
64 full forms of PAME corresponding to the same clinical concept	Pre-surgical medical examination; pre-operative medical evaluation; pre-anesthetic medical exam; ….
BMD has been defined 18,055 times as "Bone Mineral Density" but there are two other definitions	Bone marrow density; Becker Muscular Destrophy

Table 3. Semantic groups defined (see http://semanticnetwork.nlm.nih.gov/SemGroups/SemGroups.txt for more information).

Anatomical Site (ANAT)	T017, T018, T021, T022, T023, T024, T025, T026, T029, T030, T031
Procedure (PROC)	T058, T059, T060, T061, T062, T063, T065
Finding (FIND)	T032,T033,T034,T039,T040,T041,T042,T043,T044,T045,T184,T051,T052,T053,T054,T055,T056,T057
Diseases (DISO)	T019,T020,T037,T046,T047,T048,T049,T050,T190,T191
Clinical Drug (DRUG)	T200,T121,T126,T103,T119,T195,T127,T129,T130
Chemical (CHEM)	T103,T104,T109,T110,T111,T114,T115,T116,T118,T119,T120,T122,T123,T124,T125,T126,T131,T192,T196,T197

MedLex

We adopted the semantic groups proposed by Bodenreider and McCray (15) but with some modification for clinical semantic groups. Specifically, we define six categories that are clinically relevant as shown in Table 3. Table 3 also shows the constitutional semantic types of each semantic group.

There were 35,994 tokens directly imported from the UMLS to MedLex since they occur as lexical entries in LEX and as terms in META. For those could not be imported directly, the semantics of 9,620 tokens can be automatically inferred from tokens sharing the same root. The majority of them are adjectives with the corresponding nouns defined in the META. There are 2,800 tokens associated uniquely to META concepts. The semantics of about 443 tokens can be inferred from medical suffixes, roots, and prefixes but the corresponding concept identifiers are missing. For example, 123 tokens begin with "hyper" and "hypo", we can associate them with "Finding". 15 tokens end with "pathy", we can associate them with "Disease". Table 4 shows some of the examples. The semantics of the remaining about 5000 tokens are currently assigned to "TBA". All inferred semantics require some manual verification and curation.

All 485,294 unique concept mentions identified corresponding to 208,347 unique concepts are also directly ported to MedLex as single- or multi-word lexeme entries where the dominant syntactic information for them is noun or noun phrases. Among the 23.3% chunks that can be mapped to META terms, most of them have already captured in MedLex. There are 97,050 chunks that are prepositional phrases and we added them as variants for the corresponding concepts. Table 5 shows the detail statistics of the conceptual space of MedLex related to each semantic group in our corpus. Clinical semantic groups such as disorders (DISO), medications (DRUG), findings (FIND), and procedures (PROC) dominate in the number of concepts and the number of concept occurrences. The semantic group CONC (concepts) has the most coverage since most of the terms there are modifiers.

Table 4. Statistics of tokens with semantics inferred using heuristics.

Rules	**# Tokens**	**Examples**		
Stemming	9620	"ic" <- "ia"	127	*heterochromic* <-Finding <- heterochromia (C0423318); *cachexic*<-Finding <- cachexia (C0006625)
		"al" <- "is"	53	*perichondral* <- Disease <- perichondritis (C0031053); *diaphysial* <- Anatomical Site <- diaphysis (C0242696)
Single concept	2800	*Evasiveness* <- Finding <- Marked evasiveness; behavior (C1398202) *Mesogastric* <- Anatomical Site <- Mesogastric region (C0230188)		
Medical suffixes, roots, and prefixes	3172	Vagectomy <- Procedure <- "tomy" is the suffix for procedures Ulnohumeral <- Anatomical Site <- Humeral (C0020164)		

Table 5. The detailed statistics of different semantics groups.

SemG	**#Con**	**#Occu**	**SemG**	**# Con**	**#Occu**	**SemG**	**#Con**	**#Occu**
ACTI	284	5,988k	DRUG	50,602	166,886k	OCCU	999	53,768k
ANAT	17,874	284,870k	FIND	44,032	772,496k	ORGA	1,261	58,945k
CHEM	11,206	109,895k	GENE	8,329	114,453k	PHEN	1,196	54,017k
CONC	14,501	1,345,707k	GEOR	865	32,124k	PHYS	2,530	12,8891k
DEVI	5,232	67,274k	LIVB	9,830	145,323k	PROC	27,083	416,056k
DISO	43,398	316,767k	OBJC	3,642	9,2727k			

Discussion and future work

The construction of a semantic lexicon is not only critical for clinical information extraction. It can also play an important role in defining value sets for secondary use of EHRs. Specifically, practical secondary use of EHRs requires detailed data models and value sets to capture the semantics of clinical information in EHRs. Mayo Strategic Healthcare IT Advanced Research Project (SHARPn)which aims to enable secondary use of EHRs for improving health care quality and facilitating clinical and translational research, has adopted Intermountain Healthcare's Clinical Element Models (CEMs, accessible at clinicalelement.com) as a way to store clinical information. Defining allowable clinical concepts when instantiating CEM Instances is a challenge task. Similar effort in national scale is currently carried out by the Office of National Coordinator (ONC) Standards on clinical element data dictionary (CEDD) which intends to capture data elements, their corresponding definitions and attributes, and value sets of the clinical information used in a clinical context.

Note that we did not use MetaMap (24), the system popularly used for UMLS concept mapping, in our study for a couple of reasons. First, MetaMap is a rule-based system where some of the morphology rules have already been deployed in the implementation. As pointing out by McCray et al, manual verification is needed when inferring the semantics of a lexeme from its corresponding morphemes (25). Using the simple UMLS mapping method deployed in MedTagger together with some heuristic rules, we know explicitly how the semantics of each entry is obtained. Secondly, MedTagger uses a very fast string mapping algorithm and it finished the annotation of our corpus in 15 days and the estimation of MetaMap on the same corpus would require 3 months (inferred from the annotation time required for the baseline MedLINE corpus).

Our analysis indicates that the UMLS is quite comprehensive in providing morphosyntactic and semantic information for tokens occurring in the text. Even over 60% of the tokens are not found in the UMLS, but majority of them are proper nouns or spelling errors. From Figure 1, we can see that it achieves over 90% for most frequent tokens. The majority of the tokens occur in text about a couple of hundred times. For less popular tokens, the coverage drops. However, the semantic coverage of atomic lexeme is limited in the UMLS especially for adjective forms of the clinical terms. In MedLex, semantics were assigned to adjective forms through inference followed by manual verification.

Our analysis has provided some insights on what clinical NLP systems need to pay attention to when processing tokens. We estimate there are about 0.1% spelling errors in medical documents. Besides some common spelling errors "reoprts" for "reports", some of the spelling errors happen to medical terms. For example, "elliptical" appears as "elliptical". An NLP system may need to have spelling correction ability. For extreme long tokens, they can be split into different fragments. For example, the first row in Table 2 shows the long form of EGD appears in text in 24 different forms. To capture them correctly in text, an NLP system may need to purposely split those long tokens into fragments. Additionally, most proper nouns in clinical narratives are addresses and person's names. However, medical concepts can be named after a person. An NLP system also needs to be able to disambiguate proper nouns that are also medical concepts. We also notice some of the lexemes are themselves compositional and some decomposition may be needed for data normalization. For example, "headache" can be decomposed to "pain" as a finding and "head" as an anatomical site.

Only a portion of the UMLS concepts are present in the clinical domain. The usage information of a lexeme and its corresponding concept can be used for ambiguity resolution. For an ambiguous lexeme, if there are multiple concepts associated with it, NLP systems with limited disambiguation capability can simply pick the most frequent concept as the normalization target. For example, if "bone mineral density" is the most frequent meaning for "BMD" in clinical text, a default inference of the meaning of "BMD" would be "bone mineral density" rather than "Becker Muscular Destrophy".

We have done preliminary verification of entries added to MedLex. We will continue automated as well semi-automated curation of MedLex and it will be distributed to the public domain as an open source resource.

Conclusion

In this study, we reported our effort in using the UMLS and a large clinical corpus to acquire a raw semantic lexicon for semantic parsing of clinical narratives. The detailed corpus analysis of tokens, chunks, and concept mentions shows the UMLS is an invaluable source for natural language processing. Increasing the semantic coverage of tokens provides a good foundation in capturing clinical information comprehensively. The study also yields some insights in developing practical NLP systems.

Acknowledgement

This study was made possible by National Science Foundation ABI:0845523 and National Institute of Health R01LM009959A1 as well as funding from the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology. The content of the manuscript is solely the responsibility of the authors.

References

1. Boguraev B, Pustejovsky J. Corpus processing for lexical acquisition: MIT press; 1996.
2. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp. 2001:189-93.
3. Johnson SB. A semantic lexicon for medical language processing. Journal of the American Medical Informatics Association. 1999;6(3):205-18.
4. Kalfa VC, Jia HP, Kunkle RA, McCray PB, Jr., Tack BF, Brogden KA. Congeners of SMAP29 kill ovine pathogens and induce ultrastructural damage in bacterial cells. Antimicrob Agents Chemother. 2001 Nov;45(11):3256-61.
5. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA. Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics. 2010;1(1):5.
6. McCray AT, Bodenreider O, Malley JD, Browne AC. Evaluating UMLS strings for natural language processing. Proc AMIA Symp. 2001:448-52.
7. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:189-93.
8. Liu H, Friedman C. A method for vocabulary development and visualization based on medical language processing and XML. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:502-6.
9. Friedman C, Liu H, Shagina L. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. Journal of biomedical informatics. 2003 Jun;36(3):189-201.
10. Wu S, Liu H. Characteristics of NLP-extracted Concepts in Clinical Notes vs. Biomedical Literature. Annual Symposium of American Medical Informatics Association; 2011; 2011.
11. Wu S, Liu H, Li D, et al. UMLS Term Occurrences in Clinical Notes: A Large-scale Corpus Analysis. Journal of American Medical Informatics Association. 2012.
12. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association. 2010;17(2):131-5.
13. Pakhomov S, Buntrock J, Duffy P. High throughput modularized NLP system for clinical text. 2005: Association for Computational Linguistics; 2005. p. 25-8.
14. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl 1):D267-D70.
15. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of biomedical informatics. 2003;36(6):414-32.
16. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011;18(5):580-7.
17. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc. 2009;16(2):247-55.
18. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011;18(5):552-6.
19. Smith L, Tanabe LK, Ando RJ, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;9(Suppl 2):S2.
20. Liu H, Lussier YA, Friedman C. A study of abbreviations in the UMLS. Proc AMIA Symp. 2001:393-7.
21. Liu H, Aronson AR, Friedman C. A study of abbreviations in MEDLINE abstracts. Proc AMIA Symp. 2002:464-8.
22. Torii M, Liu H. Enhancing acronym/abbreviation knowledge bases with semantic information. AMIA Annu Symp Proc. 2007:731-5.
23. Dennerll JT, Davis PE. Medical terminology: a programmed systems approach: Delmar Pub; 2010.
24. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229-36.
25. McCray AT, Browne AC, Moore DL. The semantic structure of neo-classical compounds. 1988: American Medical Informatics Association; 1988. p. 165.