Similarity-Based Recommendation of New Concepts to a Terminology

***Praveen Chandar, PhD, *Anil Yaman, MS, Julia Hoxha, PhD, Zhe He, PhD, Chunhua Weng, PhD**

Department of Biomedical Informatics, Columbia University, New York, NY USA

(*equal contribution first-authors)

Abstract

Terminologies can suffer from poor concept coverage due to delays in new concept insertion. This study tests a similarity-based approach to recommending concepts from a text corpus to a terminology. Our approach involves extraction of candidate concepts from a given text corpus, which are represented using a set of features. The model learns the important features to characterize a concept and recommends new concepts to a terminology. Further, we propose a cost-effect evaluation methodology to estimate the effectiveness of terminology enrichment methods. To test our methodology, we use the clinical trial eligibility criteria free-text as an example text corpus to recommend concepts for SNOMED CT. We computed precision at various rank intervals to measure the performance of the methods. Results indicate that our automated algorithm is an effective method for concept recommendation.

Introduction

Over years, researchers have stressed on the importance and usefulness of developing domain-specific terminologies in biomedical informatics. Well-curated terminologies provide cohesive and structured domain knowledge to clinical information systems, such as clinical decision support systems and Electronic Health Record systems, further facilitating the interoperability among them. Natural Language Processing (NLP) applications in biomedical informatics also largely rely on controlled terminologies and ontologies. Concept coverage is an important aspect of Cimino's desiderata for terminologies (1). Good coverage requires timely and continuous concepts updating with terminologies to keep up with the none-stop growth of domain knowledge and the evolution of professional vocabularies. However, developing and maintaining a high-coverage terminology is a complex and expensive task that often requires constant laborious curation. Such manual effort involves domain experts who are required to read as much domain-specific literature as possible in order to identify meaningful concepts to insert to a terminology.

Due to limited resources and subjectivity of domain experts in the manual curation process, terminologies vary in their coverage of domain concepts. As such, they are often criticized for insufficient coverage of concepts and concept relations such as synonyms (2). Recently, informatics tools are being used in order to ease the load on curators. BioPortal developed by Nation Center for Biomedical Ontology (NCBO) (3) and Collaborative Protégé by Stanford (4) are some of the recent developments that facilitate continuous development of terminologies from geographically distributed locations. These tools facilitate continuous interaction between users and developers with the intent to improve terminology development through constructive feedback.

To complement the top-down approach for terminology development driven by domain experts, symbolic and statistical ontology learning methods have been proposed for identifying important concepts from human-generated texts. Hearst introduced a symbolic method to discover hyponyms within Lexical-Syntactic Pattern (LSPs) extracted from text (5). Liu et al. (6) further applied Hearst's LSP method to identify clinically meaningful concepts and relationships from medical documents. They used regular expressions over Part-of-Speech (POS) tags to extract LSPs in medical documents and asked human annotators to identify meaningful concepts in these patterns. Church and Hanks proposed a statistical method for estimating word association norms to identify useful concepts (7). A symbolic-and-statistical hybrid approach was proposed to identify meaningful concepts by computing the similarity between unrecognized concepts and existing concepts in controlled terminologies (8). Recently, He et al. (9) proposed a structural method leveraging UMLS's native term mapping and hierarchical structure to identify potentially missing concepts for a source terminology.

Even though the aforementioned approaches have achieved promising results, they all require human experts to suggest and review the concepts. In this work, we present a purely computational method for recommending new concepts barely with human intervention. The contribution of this paper is two-fold: (1) we propose a novel, unsupervised approach for prediction and recommendation of phrases as candidate atoms in an existing terminology; and (2) we present a cost-effective and effective evaluation methodology to estimate the performance of the terminology enrichment method. Through experimental evaluation on large datasets, we show the accuracy of our recommendation approach, and demonstrate its potential to improve the scalability and throughput of terminology enrichment.

Methods

Glossary and Design Rationale

Before describing our methodology, we clarify the definition of *concept, atom, and n-gram,* which are key to understanding our approach.

- ***Concept*** is the fundamental unit of meaning in ontology. According to the semiotic triangle theory (10), each object in the world has one and only one concept describing it, but the concept may be associated with multiple terms, which are also called atoms. Concepts are assigned a unique identifier, such as the unique reference as an identifier (e.g., Concept Unique Identifier in Unified Medical Language System (UMLS) that can be used to unambiguously identify a concept. Each concept is associated with at least one or more lexical variants or natural language text strings, which is often referred to as atoms in the UMLS.
- ***Atom*** is the smallest unit of meaning contributed by a source. Atoms are phrases (i.e. sequence of terms) that represent a distinguished meaning. A unique identifier is often assigned to an atom, such as the atomic unique identifier (AUI) in UMLS. Atoms are symbols representing a concept, which contain one or more atoms.
- ***N-Gram*** is a contiguous sequence of *n* terms found in text. For example, *lymph node metastasis* is a 3-gram (or tri-gram). An n-gram deemed meaningful by a terminology curator becomes an atom.

The goal of our study is to identify in a collection of written texts of a particular domain, the n-grams that are highly probable of constituting an atom in an existing terminology; and to recommend them as candidate atoms to the curators. It is the decision of the curators to enrich the terminology accordingly by approving the candidates as atoms to an existing concept or a new concept in the terminology.

Our proposed method requires two components as input: seed terminology and a text corpus. We explain each of these components, our proposed approach to recommend new concepts and the evaluation strategy below.

Dataset

Seed Terminology is the terminology to be enriched using the text corpus. It must have at least two components: a set of concepts and a set of lexical variants or terms for each concept. Following de Keizer's definition (11), a concept is a "cognitive construct of objects" that consists of one or more terms, which are also referred as atoms. For example, a concept "breast cancer" might have the following atoms: "Malignant neoplasm of breast", "breast cancer", etc. In UMLS, atoms are often assigned an atomic unique identifier (AUI).

Text Corpus: A large set of unstructured text on a specific domain from which concepts are to be extracted. Text documents and literature are valuable sources for identifying new concepts on a domain. The written text must ideally be a representative sample of the knowledge prevalent in the domain. It is important that the text exists as free-text as our approach relies on occurrence frequency as well as linguistic properties of a concept in the text. The goal in this work is to identify potential atoms from the text to enrich an existing terminology.

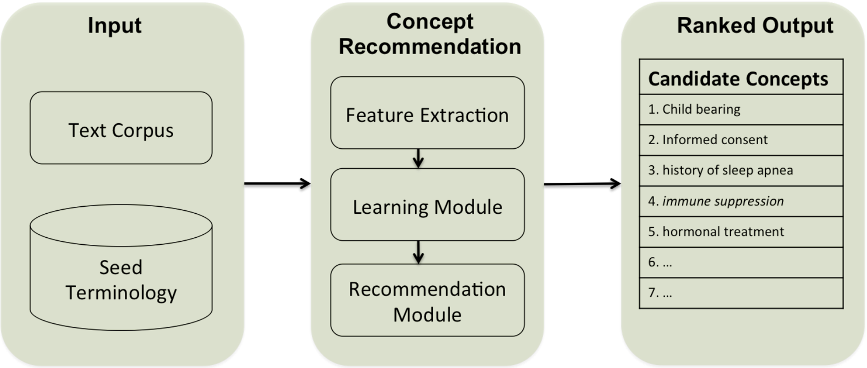


Figure 1 – System outline illustrating various modules in our proposed approach.

Approach

Our proposed approach takes two components as input: seed terminology and text corpora. It identifies as output a sorted list of n-grams (sequence of words) that are most likely to be added to the seed terminology as atoms. The basic idea behind our approach is that we can learn from the characteristics of n-grams that are already atoms in the seed terminology on to recommend new atoms that could potentially be added to the seed terminology.

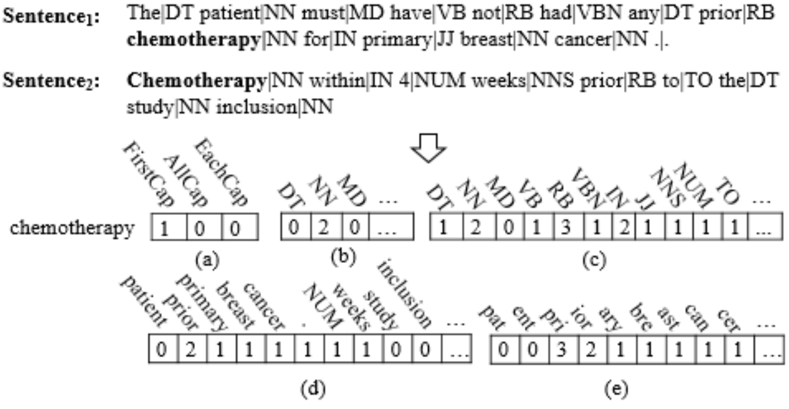
Figure 1 provides an outline of our approach. For a given text corpus, the text is preprocessed by identifying set of sentences. Then, a set of n-grams is extracted from each sentence. The extracted n-grams are aggregated and each unique n-gram is characterized using a set of features. We use a set of morphological, contextual, and syntactic features to capture the characteristics of an n-gram that represents a concept. Of all n-grams identified in the text corpus, a subset of them can be matched to atoms in a seed terminology. We leverage this information to learn a model that characterizes an atom using a set of features. Considering each atom in identified in the text corpus as a candidate atom, the model is used to score these candidate atoms indicating the likelihood of them being an atom in the seed terminology. A summary of various steps involved are listed below:

1. Preprocessing the text corpus
2. Identifying n-grams from text and extraction of features for each n-gram;
3. Learning the characteristics of those n-grams that belong to the seed terminology;
4. Leveraging the learnt characteristics to recommend new atoms to the terminology;

Step 1: Pre-processing **-** The first step is to segment the free-text from the text corpus into sentences We relied on OpenNLP (12) to perform sentence segmentation. Each sentence was tokenized and n-grams were extracted by iterating over the tokens (up to 5-grams are extracted). The extracted n-grams were filtered if the n-gram begins with a number token or a stop-word, or if the n-gram ends with a stop word. We used a Standard English stop-word list in our experiments.

Step 2: Feature Extraction: The problem of identifying meaningful atoms in free text is similar to the named entity recognition (NER) problem (13). Inspired by the research in NER, we engineer a set contextual and syntactic features to represent the n-grams extracted from the text corpus (13, 14). The feature representation of n-gram is not specific to any terminology or text corpora. This is an important property of our method that enables our method to be used across other domains. The engineered features include:

- Capitalization Features (CAP): the capitalization information of the first and all letters, and the first letters of each tokens of the current n-gram;
- Syntactic Features (POS): the part-of-speech (POS) tags of the current n-gram;
- Contextual Features (CONTEXT): the POS-tags, tokens and the prefixes/suffixes of the tokens that co-occur with the current n-gram within a window of 10 tokens;



Sentence$_1$: The|DT patient|NN must|MD have|VB not|RB had|VBN any|DT prior|RB **chemotherapy**|NN for|IN primary|JJ breast|NN cancer|NN .|.

Sentence$_2$: **Chemotherapy**|NN within|IN 4|NUM weeks|NNS prior|RB to|TO the|DT study|NN inclusion|NN

Figure 2 – Feature extraction example for the unigram "chemotherapy". Capitalization Features (a), Syntactic Features (b), and Contextual Features (c, d, and e)

We illustrate the feature extraction process with an example in Figure 2, the example shows two different sentences containing "chemotherapy", which is the unigram under consideration. The feature extraction process starts with the tokenized and POS-tagged sentences. In Figure 2, (a) represents the capitalization features; (b) illustrates the syntactic features of the unigram; and (c, d and e) are the contextual features that include frequencies of POS-tags, tokens, and prefixes/suffixes of the surrounding tokens within a windows size of 10. As suggested in the literature, the length of the prefixes and suffixes was limited to 3 characters. The feature vector of the unigram was obtained by concatenating of the features shown in (a, b, c, d and e).

Step 3: Learning Module: Once the n-grams were characterized by a set of features, the next step was to build a model that accurately represents n-grams that were the concepts belonging to the seed terminology using our feature representation. Thus, we split the n-grams identified into two sets: n-grams that were concepts in the seed terminology (Concept Set) and n-grams that were not in the seed terminology (Candidate Atom Set). Note that the concept set here was automatically generated using the seed terminology and the text corpus.

The n-grams in the candidate atom set are represented using a set of features, and clustered using the K-means algorithm as shown in Figure 3. The rationale behind clustering is that the groupings would represent different characteristics of an n-gram that are atoms in the seed terminology. We hypothesize that the clustering categorizes the characteristics of the atoms in a seed terminology, thereby improving recommendation of candidate atoms. Determining the right number of clusters is not obvious for a given dataset, several methods exist for the finding the optimal *k* in K-means (15, 16). We prune those clusters that contain less than 1% of the total n-gram. The pruning was performed to eliminate less representative clusters that could introduce noise into our model.

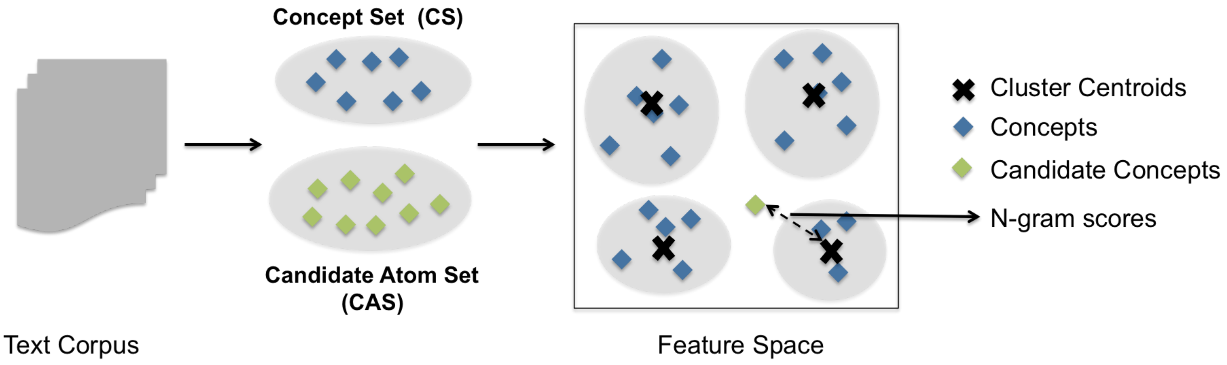


Figure 3 – Learning Module and Recommendation Module

Step 4: Recommendation Module: The final step is to score the n-gram in the candidate atom set (CAS) and obtain a final ranking. The n-gram score indicates the likelihood that the n-gram belongs to the seed terminology. Figure 3 illustrates the scoring of a candidate n-gram against the concept clusters. The n-gram scoring method was determined by the distance metric used to measure the gap between the n-gram and cluster centroid(s). In this study, we used centroid linkage function to determine the distance between an n-gram vector and the cluster centroid.

Centroid linkage function: the minimum Euclidean distance between the feature vector of an n-gram and the cluster centroids.

$$score_i = min\big(dist\big(x_i, c_j \big) \big), \qquad j \in (1, 2, \dots C)$$

Where, i is the i^{th} n-gram in the set; x_i is the feature vector of the i^{th} n-gram; C is the cluster count; c_j feature vector of the centroid in j^{th} cluster, and $dist(x_i, c_j)$ is the function that measures the Euclidean distance between two vectors.

Number of Clusters: The value of *k* in k-means is determined using a cross-validated likelihood criterion. The method follows a cross-validated clustering approach to provide insights about cluster structure. The concept set (CS) was split into 10 folds. For each fold, samples from the training folds were clustered using k-means with different *k* values.

The sum of average differences between the samples in the test set and the nearest cluster centroid were computed. The computed sum is called the average minimum distance. Figure 4 shows the average minimum distance for various *k* values. It can be seen that the minimum average distance stabilizes for *k* values between 30 and 40, which can be used for an indication of the optimum number of the clusters.

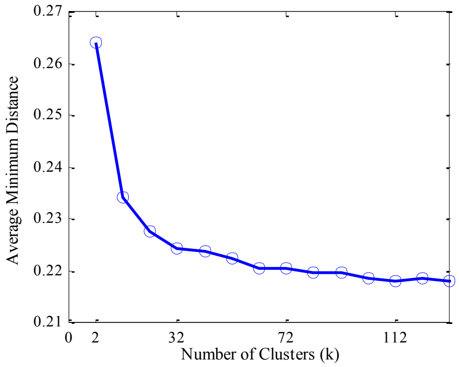


Figure 4 - The average minimum distance scores for various *k*-values.

Evaluation

In this section, we describe the evaluation measures and the details of the dataset used in our study to evaluate our proposed approaches.

*Evaluation Measure**:*** We use precision at various rank levels (*n*) to evaluate our atom recommendation algorithm. Precision at rank *n* is commonly used for various ranking problems to measure the effectiveness. Precision at rank *n* is the fraction of retrieved n-grams that are indeed meaningful atoms.

$$\text{Precision@}n = \frac{\#\text{ of n}-\text{grams identified as meaningful atoms}}{n}$$

Where, *n* is the rank cut-off. We compute precision at *n* = 1000, 5000 and 10,000, as it is reasonable to assume that a terminology curator enriching a terminology such as SNOMED CT (containing about 300,000 active concepts) would be willing to manually go through at least 5000 n-grams. Since, the task at hand is to re-rank the atoms in the candidate atom set (CAS), traditional ranking measures such as recall and F-measure are not informative.

Domain-Specific Text Corpus: We use the ClinicalTrials.gov dataset that is a rich source of information containing a set of eligibility criteria used by clinical researcher (17). The dataset was created by the National Library of Medicine and is publicly available. Each clinical trial consists of various structured fields such as study title, sponsor, etc., and a free text eligibility criteria field that contains the criteria for inclusion and exclusion of a participant. The eligibility criteria define various patient characteristics that make a person eligible or ineligible for a research study. We make use of the eligibility criteria free text field for the purpose of this study. We downloaded 181,356 (as of February 2015) clinical trial summaries for our experiments.

Seed Terminology: For the purpose of this experiment, we assume that SNOMED CT is the only available terminology at hand so that we do not consider the possibility of one term belonging to multiple terminologies simultaneously. SNOMED CT is one of most comprehensive clinical publicly available clinical terminology and is maintained and distributed by International Health Terminology Standards Development Organization (IHTSDO). We obtained the latest version of SNOMED CT available as part of the UMLS Metathesaurus (version 2014AA) from the NLM website. Our algorithm intends to extend SNOMED CT terminology with atoms relevant to the clinical domain using free-text clinical trial eligibility criteria.

Labeled Data: In order to evaluate our proposed methodology, we require a set of n-grams with positive and negative labels, i.e., positive n-grams are meaningful atoms whereas negative refers to a meaningfulness n-gram. As human annotations are extremely expensive and time consuming, we built our labeled data using various terminologies that are part of UMLS, which were curated by experts from different domains. All n-grams identified in the ClinicalTrials.gov dataset that were part of a terminology (other than SNOMED) were labeled positive and the rest were labeled negative. It is certainly possible that a negatively labeled n-gram could still be meaningful, UMLS terminologies are not complete and there may still be meaningful n-grams not part of UMLS. Nevertheless, the simulated labeled data provides a cost-effective, quick, and reasonable lower-bound performance estimate of our proposed methods.

Results

Dataset Statistics

We used the free-text clinical trial eligibility criteria to extract all possible n-grams. The n-grams were preprocessed to remove stop-words and numbers as described in the pre-processing step. We extracted up to 5-grams that occurred at least 5 times in the corpus to filter out misspelled or noisy n-grams. The pre-processed n-grams were categorized into concept set (CS) and candidate atom set (CAS). An n-gram that matched against an atom in SNOMED belonged to the CS and all other n-grams belonged to CAS.

Table 1 shows the total number of unique n-grams extracted. Only 0.05% of those n-grams matched against an atom in the SNOMED. Of the remaining 429,465 n-grams, about 0.06% (25,291 n-grams) matched against an atom in UMLS (other than SNOMED). The 25,291 n-grams constitute our positive labels and are used only for evaluation purpose. The remaining 404,174 n-grams belong to the candidate atoms set (CAS); the n-grams in CAS are to be ranked by our recommendation algorithm.

Table 1 – Statistics of the N-Grams in the text corpus

	Total Unique N-grams (1 to 5 tokens)	**449,978**
Concept Set (CS)	Concept Set (*SNOMED CT*)	20,513
Candidate Atoms Set (CAS)	Labeled Atoms (*UMLS–SNOMED CT*)	25,291
	Number of Unmatched N-grams	404,174

Lower Bound Precision Scores

The candidate atoms in CAS were ranked by our algorithm, which promoted n-grams that are relevant to the seed terminology higher and suppressed meaningless n-grams. An n-gram in CAS that matched against an atom in UMLS (other than SNOMED CT) was considered as true positive. Those n-grams that do not match are considered as true negatives. We computed precision at various rank intervals (1000, 5000, 10000 and 20000), to estimate the effectiveness of our approach.

As described in the methods section, the choice of features used to represent n-grams influences the performance of the recommendation algorithm. Therefore, we experiment with different sets of features. We experimented with different feature sets and their combination of the capitalization (CAP), syntactic (POS) and contextual features (CONTEXT). In order to observe the effect of the syntactic information, we merged all the syntactic information under the feature set named POS (including the syntactic information involved in the contextual feature set).

In addition, we included two baseline methods for the comparison: a random ranking and frequency-based. The random ranking baseline was obtained by randomly sampling *20,000* n-grams from the candidate atom set (CAS). The frequency-based ranking was obtained by sorting n-grams in the candidate atom set (CAS) by frequency of their occurrence in the text corpora.

Table 2 – Lower Bound Precision scores at different rank intervals for various methods.

Methods	**Lower Bound Precision @ n**		
	n=1000	***n=5000***	***n=10,000***
CAPS+POS+CONTEXT	**0.743**	**0.582**	**0.498**
CAPS+POS	0.695	0.531	0.460
CAPS+CONTEXT	0.550	0.436	0.388
Frequency	0.529	0.356	0.292
Random	0.047	0.044	0. 046

Table 2 provides the accuracy scores for various methods at various rank intervals ***n***. The results show that our algorithm performs considerably better than a frequency-based baseline. Note that, especially for larger *n* values the difference between our recommendation algorithm and the frequency baseline is more apparent. It is observed that the inclusion of the POS features further increases the precision by 5%. The best results were obtained using all the features (CAPS+POS+CONTEXT) that represent an atom.

Top-10 n-grams

Table 3 shows the top ranked atoms categorized by 1, 2, 3, 4, and 5 grams. The atoms that belong to a terminology in UMLS are italicized and the rest are highlighted in grey. It is interesting to note that atoms such as “history of sleep apnea”, “hormonal treatment” are meaningful but are not part of any existing UMLS terminology.

Examining the top ranked n-grams lead to a few interesting observations. We found that there existed meaningful n-grams that did not match against UMLS in the top ranked list. Thus, we decided to investigate this further by estimating the number of meaningful n-grams (candidate atoms) that are not part of UMLS. We used the help of a physician to determine if the candidate atoms are meaningful concepts. The physician was provided a list of 100 n-grams that were not matched against a UMLS terminology for manual review.

The top 100 n-grams were obtained using the algorithm with the best precision score (as shown in in Table 2); and 37 of the n-grams in this were labeled as meaningful. The annotations clearly indicate that there exist several n-grams that are currently not part of UMLS. The annotation effort was preliminary step and we plan to perform a comprehensive evaluation in the future.

Table 3 – Top 10 n-grams ($n \in$ [1-5]) ranked by our recommendation algorithm. N-grams that matched to UMLS concepts are italicized, and n-grams that did not match are highlighted in gray.

UNIGRAMS	BIGRAMS	TRIGRAMS
facial	*immune suppression*	*collagen vascular disease*
opioids	hormonal treatment	*estrogen replacement therapy*
activation	medical treatment	*bone marrow suppression*
treatment	*cell therapy*	substance abuse treatment
content	immunosuppressive treatment	*lymph node metastasis*
strength	pain treatment	*lymph node involvement*
coil	*pharmacological treatment*	significant cognitive impairment
genetic	functional impairment	*tyrosine kinase inhibitor*
resistance	*vertebral fracture*	diagnosis of cancer
titration	*cognitive deficit*	severe liver disease
4-GRAMS		**5-GRAMS**
central nervous system involvement		severe chronic obstructive lung disease
stable coronary artery disease		history of congestive heart failure
central nervous system disease		*history of coronary artery disease*
decompensated congestive heart failure		history of ischemic heart disease
active urinary tract infection		*autologous hematopoietic stem cell transplantation*
diagnosis of breast cancer		*history of traumatic brain injury*
multivessel coronary artery disease		evidence of coronary artery disease
history of seizure disorder		*low white blood cell count*
history of sleep apnea		radiographic evidence of disease progression
nonalcoholic fatty liver disease		history of substance use disorder

New Concepts vs. Synonyms

The precision at rank interval is computed by dividing the number of n-gram identified as an atom of the concept by rank interval. The number of n-grams can be categorized into two groups: n-grams that would be added to the terminology as a new concept from the atoms that will be added as synonyms to an existing concept. This enables us to separate atoms of new concepts from atoms of existing concepts in the terminology. The analysis on our results relieved that on average at least 45% of the n-grams were atoms of existing concepts (often referred to as synonyms in the literature) and the remaining were atoms that belong to new concepts in the seed terminology.

Discussion

Cluster Analysis

In the learning module, the n-grams in the concept set (i.e. n-grams that are atoms in SNOMED CT) are clustered based on their distance between the feature vectors. We hypothesized that the clustering would group similar atoms into the same cluster. We analyzed each cluster by manually reviewing 100 n-grams closest to the cluster centroids. The intention was to identify similarities between n-grams in the cluster or manual label.

Table 4 – Representative n-grams from the n-gram clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4
mexico	extent of disease	pta	repeated
illinois	quality of life	mm	improved
canada	range of motion	pd	restricted
europe	lactose intolerance	dt	complicated
denmark	history of anemia	cvd	involved
kenya	history of surgery	hctz	healed
africa	pain at rest	ctx	provoked
vermont	history of depression	ttp	stabilized
indiana	level of consciousness	nsaid	established

Table 4 illustrates a set of 10 n-grams that are closest to the centroids of four clusters. Interestingly, Cluster 1 tends to consist of atoms relating to geographic locations or place names. Cluster 3 contains abbreviations. The atoms in Cluster 2 contain atoms following a specific pattern, this is somewhat expected due to the fact that the atoms are represented using syntactic features. Cluster 4 groups adjectives that are prevalent in the clinical domain.

Further, we tried to align the clusters with the semantic hierarchies present in SNOMED CT. The concepts in SNOMED CT are organized into 19 acyclic taxonomic (is-a) hierarchies. We observed the distribution of the semantic hierarchies in each cluster. For example, 48% of the concepts found within cluster belonged to the hierarchy "Staging & Scales", 72% of the concepts in another cluster belonged to "Situation with Explicit Context". However, in general only a weak correlation was observed between the clusters and SNOMED CT hierarchies.

Modifier Extraction

Modifiers are terms or phrases that provide additional meaning to a concept. For instance, in the phrase *history of heart disease*, the prefix *history of* modifies the meaning of the concept *heart disease.* Modifiers play an important role in tasks such as cohort identification. A researcher looking for cohort of patients with "history of heart disease" might possibly find little use with patients with "*active* heart disease". As an extension to this work, our intention is to identify candidate modifier that enrich the identified concepts.

Table 5 – Candidate modifiers identified from the top ranked n-grams.

history of	severe	significant	active	previous
current	evidence of	diagnosis of	treatment with	prior

We manually analyzed the top ranked n-gram and found that the 3-, 4- and 5-grams tend to contain a modifier followed by a concept. Following the manual analysis, we wanted to identify those n-grams that matched the pattern; prefix followed by UMLS atom. Using regular expressions we filtered all n-grams that matched the <prefix, atom> pattern. We found 14,000 n-grams that contain a prefix followed by a UMLS atom, i.e., partially matched against an atom in UMLS. Further, we ranked the prefix by its frequency; hypothesizing that commonly used modifiers would be ranked higher.

Table 5 shows the top 10 candidate modifiers extracted using this approach. The results are promising and we plan to investigate further in this direction.

Post & Pre Coordination

SNOMED Clinical Terms (SNOMED CT), as the most comprehensive multilingual clinical terminology, is being widely implemented as a standard within IHTSDO member countries. By 2015, SNOMED CT will be the United States standard for encoding diagnoses, procedures, and vital signs in electronic health records (EHRs) under Stage 2 of Meaningful Use (18). Even though SNOMED CT provides rich conceptual content, researchers have advocated greater coverage of common problem statements with improved synonymy and conceptual content (19). A survey among the direct users of SNOMED CT reported that 23% and 17% of the respondents encountered missing concepts and missing synonyms, respectively (2).

Post-coordination of SNOMED CT allows its users to create new meaning by combining existing concepts, which can potentially enhance SNOMED CT's conceptual coverage (20). In spite of post-coordination, researchers reported that some clinical statements with complex and rare clinical scenarios could not be encoded (21). Meanwhile, the same clinical meaning may be represented by different post-coordinated expressions, which hampers their interoperability among different authoring entities. In the setting of this paper, post-coordinated expressions may also pose difficulties in unifying and structuring clinical trial eligibility criteria.

To the best of our knowledge, there is no tool available for automating the creation of post-coordinated expressions. Therefore, they have to be modeled by a clinical expert manually. In this proof-of-concept study, we consider only pre-coordinated SNOMED CT concepts for training and prediction. Computational approaches to filter out suggested concepts that can be constructed with post-coordination need to be developed. Nevertheless, the suggested concepts were identified to be important in clinical trial eligibility criteria, thereby should be considered by SNOMED CT curators as pre-coordinated concepts.

Limitations

Since the recommendation algorithm relies on linguistic features of atoms extracted from the text corpus, the corpus must contain atoms already present in the seed terminology. Also, our method recommends atoms of a concept and does not provide any information about the relationship between the atoms and concepts. For instance, the algorithm would identify "sleep disordered breathing" and "breathing disorder during sleeping" as meaningful atoms although it provides no information w.r.t its relation with the concept *Sleep Apnea.* The problem of finding if an atom belongs to an existing or new concept is still an open research area that warrants more investigation. Another limitation is that, we rely on exact string matching between n-grams and atoms in a terminology, which could be error-prone. Although fuzzy matching of text is possible, we believe exact string matching would provide higher precision. And, we normalize and pre-process the text in both corpus and terminology the same way to improve coverage.

Future work

In the future, we intend to conduct a larger-scale comprehensive evaluation study using domain experts to accurately estimate the performance of our methods. The manual evaluation process would require the experts to review each candidate atom to determine if it is meaningful or not. Also, we plan to compare our approach with previously proposed symbolic and statistical ontology learning methods (3, 6, 7). Another avenue of future work is to recommend sentences that have the highest chance of containing new atoms/concepts. Ranking sentences has an advantage, it provides context to the atoms that enables curators to make better decision about atom's usefulness to a terminology. We also plan to experiment with different datasets from heterogeneous domains to prove the generalizability of our proposed approach.

Conclusions

This paper contributes a similarity-based approach for recommending candidate n-grams to terminology curators for consideration. The method characterizes the n-grams in a text corpus by using a feature set to learn important features of a concept in order to produce a ranked list of n-grams by decreasing order of meaningfulness. The ranked list of n-grams would enable quicker and easier identification of new concepts for an existing terminology. Another contribution of this paper is a cost-effective evaluation methodology to estimate the performance of the terminology enrichment method. Through careful experimental design, we simulate labeled data using various terminologies in UMLS to evaluate effectiveness of our proposed approach. We computed the precision score and observed our method provides about 20% improvement over a frequency-based baseline. We have only scratched the surface of what is possible. There is ample opportunity for work within this framework such as scoring methods for ranking candidate atom, and post filtering of candidate atoms.

Acknowledgments

We would like to thank Dr. Kin Wah Fung from Lister Hill National Center for Biomedical Communications for the informative discussion. We would also like to thank Fendanda Polubriaginof M.D. for her help with manual annotations. This study was sponsored by the U.S. National Library of Medicine grant **R01LM009886** (PI: Weng) and U.S. National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Methods of information in medicine. 1998;37(4-5):394-403.
2. Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. Journal of the American Medical Informatics Association : JAMIA. 2011;18(Suppl 1):i36-i44.
3. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research. 2009;37(Web Server issue):W170-W3.
4. Tudorache T, Noy N, Tu S, Musen M. Supporting Collaborative Ontology Development in Protégé. In: Sheth A, Staab S, Dean M, Paolucci M, Maynard D, Finin T, et al., editors. The Semantic Web - ISWC 2008. Lecture Notes in Computer Science. 5318: Springer Berlin Heidelberg; 2008. p. 17-32.
5. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics - Volume 2; Nantes, France. 992154: Association for Computational Linguistics; 1992. p. 539-45.
6. Liu K, Chapman WW, Savova G, Chute CG, Sioutos N, Crowley RS. Effectiveness of Lexico-Syntactic Pattern Matching for Ontology Enrichment with Clinical Documents. Methods of information in medicine. 2011;50(5):397-407.
7. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Proceedings of the 27th annual meeting on Association for Computational Linguistics; Vancouver, British Columbia, Canada. 981633: Association for Computational Linguistics; 1989. p. 76-83.
8. Liu K, Mitchell KJ, Chapman WW, Savova GK, Sioutos N, Rubin DL, et al. Formative Evaluation of Ontology Learning Methods for Entity Discovery by Using Existing Ontologies as Reference Standards. Methods of Information in Medicine. 2013;52(4):308-16.
9. He Z, Geller J, Elhanan G. Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization. AMIA Summits on Translational Science Proceedings. 2014;2014:48-53.
10. Ogden CK, Richards IA. The Meaning of Meaning1989. 396 p.
11. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JHM. Understanding Terminological Systems I: Terminology and Typology. Methods Archive. 2000;39(1):16-21.
12. OpenNLP. https://opennlp.apache.org.
13. Bikel DM, Schwartz R, Weischedel RM. An Algorithm that Learns What's in a Name. Mach Learn. 1999;34(1-3):211-31.
14. Florian R, Ittycheriah A, Jing H, Zhang T. Named entity recognition through classifier combination. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4; Edmonton, Canada. 1119201: Association for Computational Linguistics; 2003. p. 168-71.
15. Phama D, Dimov S, Nguyen C, editors. Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers,; 2005.
16. Smyth P. Clustering using Monte Carlo cross-validation1996. Medium: X; Size: pp. 126-33 p.
17. NIH. ClinicalTrials.gov. Available from http://www.clinicaltrials.gov [cited 2015 January].
18. CMS. Electronic Health Record Incentive Program—Stage 2. Available from http://www.gpo.gov/fdsys/pkg/FR-2012-09-04/pdf/2012-21050.pdf. [cited 2014 30 December].
19. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clinic Proceedings. 2006;81(6):741-8.
20. Liu H, Wagholikar K, Wu ST-I. Using SNOMED-CT to encode summary level data – a corpus analysis. AMIA Summits on Translational Science Proceedings. 2012;2012:30-7.
21. Campbell WS, Campbell JR, West WW, McClay JC, Hinrichs SH. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings2014 2014-09-01 00:00:00. 885-92 p.