Decompositional Terminology Translation Using Network Analysis

Chintan O. Patel, MS, James J. Cimino, MD
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

Biomedical terminologies often contain composite concepts that cannot be translated into single unique synonymous concepts in a target controlled terminology. Such composite concepts need to be decomposed into sets of component concepts present in the target terminology that can serve as the proxy for applications in information retrieval, decision support or data analysis. Towards this goal, we use a "clustering coefficient" over the UMLS Metathesaurus to traverse the closely clustered neighbors of the composite source concept to generate a ranked list of possible component concepts. Using the MeSH Associated Expression mappings as the gold-standard, we show that the proposed approach generates relevant component concepts as compared to existing semantic locality based methods. The topological connectivity of the concepts in the UMLS Metathesaurus is a useful feature that can be coupled with existing lexical and semantic locality based approaches towards terminology translation.

Introduction

Composite concepts such as "Fracture of unspecified intracapsular section of neck of femur, closed" are often used in biomedical terminologies to support health care processes such as billing and coding. However, for applications in information retrieval,[1] decision support[2] and data analysis/integration, there is often a need to translate and decompose such composite concepts from a source terminology to constituent concept(s) in a given target terminology, for example, "Neck of Femur" and "Fracture of lower limb" would be the decompositions for the aforementioned example.

Existing terminology translation methods attempt to discover the target concept(s) that are synonymous or that have a closest possible meaning to the source concept. The methods to perform such translations range from lexical matching[3] to semantic locality-based approaches.[4] We distinguish a different type of translation that involves identifying a set of component concepts that constitute (or compose the meaning of) a given composite concept; we term this as *decompositional terminology translation*.

Consider an electronic health record application containing the ICD-9-CM code **238.7** ("Neoplasm of uncertain behavior of other lymphatic and hematopoietic tissue"). Suppose that we wish to use this information to perform an automated retrieval from PubMed. Searching for this concept name returns no results. Alternatively, searching PubMed with the corresponding component concepts ("Neoplasm AND lymphoma AND Vascular Neoplasms") retrieves many relevant results.

The proposed notion of decompositional terminology translation is similar to the Associated Expression mappings in the Unified Medical Language System (UMLS),[5] where a composite concept has an associated, manually created Boolean expression composed up of Medical Subject Heading (MeSH) terms, for example, "Neck Pain" has an associated expression "<Neck> AND <Pain>".

In this paper, we investigate whether the connectivity of the concepts in the UMLS Metathesaurus graph can be exploited to perform decompositional terminology translation. We hypothesize that the component concepts occur in the closely clustered neighborhood of the composite concept in the UMLS Metathesaurus graph and that the component concepts tend to have a higher number of relationships to other concepts. Towards this goal, we propose a traversal algorithm that uses the network analysis measure of *clustering coefficient* to traverse only the clustered neighbors of a given concept. We use the Associated Expressions in the UMLS as a gold-standard to evaluate the proposed method against existing semantic locality based method[4] and simple transitive traversal.

Related Work

The issue of decomposing or expanding a multi-word query into Boolean expressions is an important problem in information retrieval. In terminology-driven digital library applications,[1,2,6] the challenge is to translate or map the composite query concepts to the concepts in resource indexing terminologies. Lexical approaches to map terminologies[3,7] use various methods such as stemming, synonyms and so on. On the other hand, semantic locality-based approaches such as the *restrictToMeSH* algorithm[4] explore the UMLS semantic relationships around the source concept to identify the closest neighbor concept in the target terminology.

The UMLS Metathesaurus, when viewed as a graph with concepts as nodes and relationships as edges, allows application of various network analysis-based methods. One approach to network analysis is to consider whether the network is *scale-free*, wherein a few nodes ("hubs") have most of the edges. The scale-free property of the UMLS Metathesaurus allows pruning irrelevant paths[8] for a machine learning approach to terminology translation.

Another important measure used in network analysis is the *clustering coefficient,*[9] which quantifies the connectedness of the neighbors for a given node. For example, a clustering coefficient-based algorithm[10] over protein-protein interactions has been used to identify dense molecular complexes.

To calculate the clustering coefficient (CC) for a given node *n*, let the degree (or number of immediate neighboring nodes) of *n* be *k* and let *t* be the total number of edges between the neighboring nodes, then

$$CC(n) = \frac{t}{½k*(k-1)}$$

i.e., the clustering coefficient is the ratio of number of edges between the neighbors of n and the total possible number of edges between the neighbors (if each neighbor was connected to every other neighbor).

The clustering coefficient essentially gives a measure for how 'densely connected' are the neighbors of a given node, ranging from 0 (no connection) to 1 (all connected). In the context of social networks, for example, the clustering coefficient is used to characterize the clusters of friends who are generally well connected and have few connections outside the circle.

Datasets

UMLS Metathesaurus. An important component of the UMLS[5] is the Metathesaurus (META). The META is organized around concepts that are created by integrating synonymous terms from about 140 different biomedical terminologies. When a UMLS source contains relationships between its terms, those relationships are carried into the META and serve as relationships between the corresponding META concepts. Since concepts can have terms from multiple terminologies, the result is that the relationships, originally linking only terms within the same terminology, can now serve as cross-terminology relationships.

Associated Expression Mappings. The MRMAP table in the UMLS provides a set of associated expression mappings between the composite concepts and manually created Boolean expressions that are used to search PubMed when a user enters the composite term. There are 9,002 such mappings in the UMLS 2005AA version.

Approach

Composite and Primitive Concepts. In biomedical terminologies, a set of concepts is often used as building blocks to create other composite concepts; for example, the condition "Acute pericardial effusion" can be defined as having the morphology of "effusion", the finding site "pericardial cavity structure" and the course "acute". The building block concepts are known as primitive concepts that are related to the composite concept with semantic relationships. In the UMLS, several biomedical terminologies with formal knowledge representation schemes such as description logics (SNOMED-CT, NCI Thesaurus) and frames (FMA, MED) generally define the composite concepts based on primitive concepts. Other biomedical terminologies, such as ICD-9-CM and MeSH, are represented as taxonomies that organize the concepts hierarchically, with no

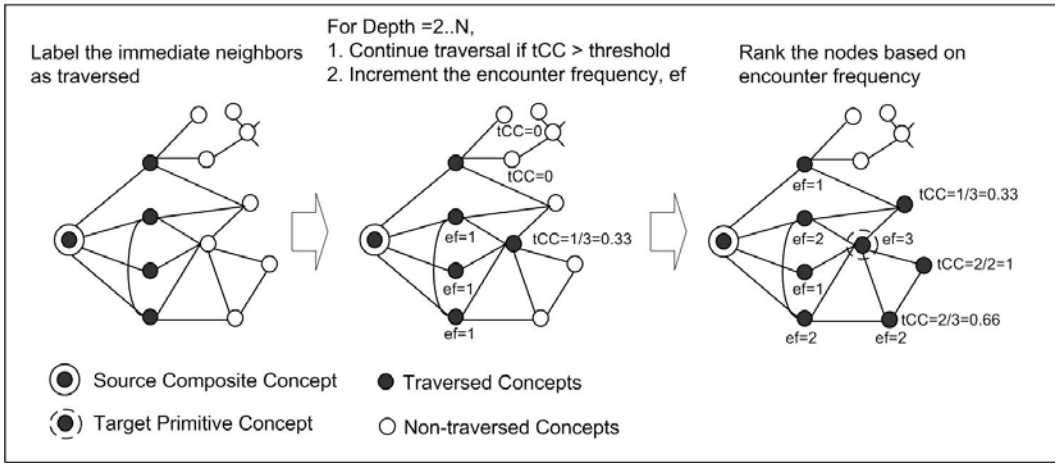

Figure 1. An illustration showing how the *TClustN* algorithm favors more closely clustered neighborhoods over sparsely connected regions of graph.

specific notion of primitive or composite concepts. Nevertheless, due to concept-oriented integration of terms from different terminologies in the UMLS, the composite terms from taxonomical terminologies might acquire indirect (that is, multi-step) relationships to primitive concepts contributed by other formal terminologies.

In this paper, we further distinguish the *complex composite concepts* that contain a preposition ('with', 'of', 'for', etc) or a conjunction ('and', 'or') in the concept name, signifying a higher order of compositional meaning in the concept, for example, "Chronic glomerulonephritis with lesion of membranoproliferative glomerulonephritis". We focus our analysis on such complex composite concepts that generally do not have any direct relationships to primitive (or component) concepts and require a transitive traversal of the graph of the interrelationships among META concepts (which we refer to as the META graph) to reach the relevant component concepts. In the next section, we describe a strategy to perform such traversal that exploits the connectivity in the META graph to identify and rank the relevant component concepts.

Traversing Clustered Neighborhoods. We hypothesize that given a composite source concept, the relevant component target concepts occur close to the source concept in META graph. We build on the notion of the clustering coefficient to characterize the neighborhood connectivity in the META. In order to use the clustering coefficient to drive the transitive traversal from the source concept through closely clustered neighborhoods, we introduce the notion of the *traversed neighborhood* (TN), which constitutes the set of the nodes that have already been traversed. We redefine the clustering coefficient as a traversal-based clustering coefficient (tCC) in which, for a given node n, let k' be the number of immediate neighboring nodes from TN and let t' be the number of edges between the immediate neighboring nodes from TN, then

$$tCC(n) = \frac{t'}{½k'*(k'-1)}$$

which is same as the original definition of clustering coefficient, except that now we only allow the traversed nodes as the neighbors. The direct neighbors for the given composite source concept are initially labeled as "traversed" (added to TN) and the further traversal is determined by the tCC value of the concepts. Next we describe the algorithm (Figure 1) to traverse the META. The algorithm takes as parameters an arbitrary limit to traversal depth (D) and an arbitrary threshold value (T) for tCC to determine the traversal.

Algorithm: **TClustN**

***Input**: Source composite concept SC, Depth D, tCC threshold T, target terminology tSAB*

***Output**: Ranked list of possible component concepts*

1. **[Initialize]**
   a. *TN = {immediate neighboring concepts of SC}*
   b. *INQUEUE= {neighbors of concepts in TN}*
   c. *Initialize encounter frequency, $ef(tn_i)=0$, $tn_i \in TN$*
2. **[Traverse]**
   For Depth, d=2 to D
   For each concept, c in INQUEUE
   If tCC(c) > T
   a. *$TN = TN \cup \{c\}$, add c to TN*
   b. *$ef(tc_i) = ef(tc_i) + 1$, where tc_i are neighbors of c in TN*
   c. *$INQUEUE = INQUEUE \cup \{neighbors\ of\ c\}$*
3. **[Restrict To Target SAB]**
   Remove the concepts from TN that are not from tSAB
4. **[Rank]**
   Sort the concept nodes in descending order in TN based on encounter frequency, ef.

The use of the traversal-based clustering coefficient (tCC) directs the traversal towards the locally clustered concepts around the source concept, whereas the simple clustering coefficient (CC) would consider all global neighbors equally and produce, we suspect, less relevant results.

Methods

1. As a gold-standard, the mappings of type "Associated Expressions" (ATX) were obtained from the MRMAP table in the UMLS 2005AA version. The concepts with source vocabulary as ICD-9-CM and having a conjunction ('and') or preposition ('with') in the concept name were retained. ICD-9-CM was chosen due to its taxonomical structure and large number of complex composite concepts. From this set, 50 mappings were randomly selected and the source concepts were labeled as the composite concepts and the MeSH heading concepts in the associated expression were considered as component concepts.
2. The relationships with source terminology as MTH (Metathesaurus) were removed from the META graph that asserted relationships between the concepts from gold-standard mappings, since these were added explicitly to the UMLS to provide the relationships we were seeking to discover with our approach.
3. Using the proposed *TClustN* algorithm, the possible list of component target concepts were generated for each given gold-standard source composite concept. The *TClustN* algorithm was repeatedly executed with parameter Depth, D = 2,

3, 4; tCC Threshold, T= 0.30, 0.60, 0.90; the target terminology (SAB) was MeSH.

4. The proposed algorithm may generate concepts that are related (or collectively related) to the gold-standard target MeSH concepts but not necessarily the exact same concepts. Consider, for example the case where the gold-standard target MeSH "Hernia, Inguinal" and TClustN produces two concepts "Hernia" and "Inguinal canal structure". In order to capture the similarity between the algorithm results and the gold-standard associated expression we compared the overlap of the PubMed search results as the evaluation criteria. A single query expression was therefore prepared by pairing each of the concepts in the top 5 results with each of the other concepts, using a Boolean 'AND'; the resulting 20 pairs were then aggregated by using a Boolean 'OR'. For each given composite source concept, the precision and recall were calculated as follows:
   GoldSet = {PubMed IDs indexed with the gold-standard associated expression, e.g. "<Acute Disease> AND <Glomerulonephritis>"}
   ResultSet = {PubMed IDs indexed with the permutated Boolean expression of the top 5 results, e.g. "(<Kidney>AND<Nephrotic Syndrome>) OR (<Kidney>AND<Kidney Glomerulus>) OR…"}
   Precision = |GoldSet ∩ ResultSet| / |ResultSet|
   Recall = |GoldSet ∩ ResultSet| / |GoldSet|
   For each given parameter setting (of depth D and threshold T) for TClustN, an average of precision and recall was calculated for the 50 gold-standard source concepts.
5. We compared TClustN results with following:
   a. *Baseline*: The immediate neighbors of the composite source gold-standard concepts in META were considered as the result (D=1, T=0).
   b. *Simple transitive traversal*: TClustN was executed through three iterations without the tCC threshold step to generate results (D=2, T=0).
   c. *RestrictToMeSH*[4]: Using this semantic locality based algorithm, a set of target MeSH concepts were generated for the source concepts.

   The result from these methods were converted to Boolean queries and evaluated against the gold-standard using PubMed search overlap results as above.

Results

The 50 gold-standard composite concepts mapped to 118 target MeSH heading concepts, of which we were able to retrieve an exact match for 12 concepts in top 5 results (Table 1). There was no matching gold-standard MeSH heading in the immediate neighbors of the composite source concept, which validates the need for a transitive traversal-based approach (when the ATX-inspired relationships are unavailable, as they are for most of ICD-9-CM terms). The results of average precision and recall values are shown in Figure 2 for different cases. For different values of TClustN parameters, the optimum results were obtained for depth of 2 and threshold of 0.3. For this case, the f-measure (harmonic mean) of precision and recall result, was significantly higher than the baseline (t[54]=-2.68, p=0.03) and the simple transitive traversal method. Furthermore, the f-measure showed a high standard deviation of 0.28 around the mean of 0.19, indicating a mix of moderately relevant and completely non-relevant retrieved results. A small improvement in precision of TClustN over restrictToMeSH was noted, but it did not reach statistical significance.

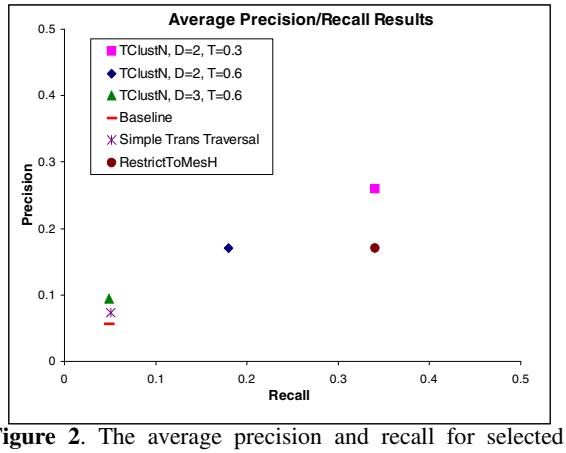

Figure 2. The average precision and recall for selected parameter settings of *TClustN* and other approaches

Source Concept = Associated Expression	TClustN		Simple Trans Traversal	restrictToMeSH
	Top 5 results	ef		
Duodenal ulcer, unspecified as acute or chronic, without mention of hemorrhage or perforation, with obstruction = <Duodenal Ulcer> AND <Intestinal Obstruction>	Duodenum	187	immunologic	Duodenal Ulcer
	Duodenal Diseases	128	metabolic aspects	
	Hemorrhage	60	Peptic Ulcer	
	Duodenal Ulcer	50	Historical aspects qualifier	
	Duodenal Obstruction	6	Taxonomic	
Unilateral or unspecified inguinal hernia, with gangrene, recurrent, = <Hernia, Inguinal> AND <Gangrene> AND <Recurrence>	Hernia	263	Gastrointestinal Hemorrhage	Fibromyalgia
	Abdominal cavity structure	138	immunologic	Gangrene
	Hernia of abdominal cavity	86	In Urine	Hernia,Inguinal
	Testis	82	metabolic aspects	
	Inguinal canal structure	51	Historical aspects qualifier	

Table 1. Results for two example composite source ICD-9-CM concepts using *TClustN* and other approaches

Discussion

Using network analysis over the UMLS Metathesuarus provides a promising solution towards the problem of terminology translation. Based on the results obtained using simple transitive traversal and the proposed approach, we were able to successfully validate our hypothesis that useful decompositions with high encounter frequency occur in the closely clustered neighborhoods of the composite source concept.

In our experiments with different parameter settings for the algorithm, we found an optimum value for depth as 2 and the performance trailed around it. Note that the depth of 2 implies that we examine neighbors that are 3 transitive relationships away the source. For higher values of D (>3), a large set of irrelevant concepts were traversed (about 10,000 possible targets per given source). The reason for such behavior can be explained by the small-world characteristics of the UMLS Metathesaurus, wherein the average distance between any two concepts is very small and hence increasing the depth results in fetching neighbors from different parts of the graph. The threshold parameter showed an optimum value of 0.3; increasing this parameter resulted in decreased precision/recall by eliminating appropriate MeSH terms, while reducing the threshold resulted in an uncontrolled traversal pulling in several irrelevant concepts that only served to dilute the PubMed search string.

We observed a high relevancy of component concepts to the composite source concepts generated by TClustN; however, we did not find many exact concept matches with the gold-standard MeSH heading. For example, for the complex concept "Other malignant neoplasm of skin of ear and external auditory canal", the algorithm identified the component concept "Cancer of Skin" (C0007114), whereas the gold-standard target "Skin Neoplasm" (C0037286) was a different concept in the UMLS. Using PubMed search overlap as the evaluation criteria served as the ultimate gold standard, since one of the important applications in performing the translation is to conduct PubMed searches. Note that the precision-recall results are only good for comparing the performance of different approaches presented. The next logical step in our research is to evaluate the performance of TClustN for top k results for different values of k and on a larger sample of ATX terms.

The proposed notion of network analysis can be used in applications beyond terminology translation such as listing recommended primitives for modeling logic-based concept descriptions or generating sub-network cluster of closely related concepts. We believe that integrating the proposed network analysis based approach with lexical and semantic based methods should lead to even better algorithms for terminology translation.

Conclusions

The connectivity of concepts in the UMLS Metathesaurus was exploited using a network analysis approach for translating composite concepts into component concepts. The results, when compared to the gold-standard of MeSH Associated Expressions, showed a performance comparable to existing semantic relationship driven methods.

Acknowledgements

This work was supported in part by the NLM grant R01LM07593.

References

1. McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. Proc AMIA Symp. 1999;:107-11.
2. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp. 1997;:528-32
3. Marquet G, Mosser J, Burgun A. Aligning biomedical ontologies using lexical methods and the UMLS: the case of disease ontologies. Stud Health Technol Inform. 2006;124:781-6
4. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998;:815-9.
5. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.
6. McCulloch E, Shiri A, Nicholson D. Challenges and issues in terminology mapping: a digital library perspective. The Electronic Library. 2004 23(6):671-7.
7. Johnson HL, Cohen KB, Baumgartner WA et al. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. Pac Symp Biocomput. 2006;:28-39.
8. Patel CO, Cimino JJ. A Scale-Free Network View of the UMLS to Learn Terminology Translations. Proc MedInfo 2007: in press.
9. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998 Jun 4;393(6684):440-2.
10. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf. 2003 13; 4: 2.