MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

A Scale-Free Network View of the UMLS to Learn Terminology Translations

Chintan O. Patel, James J. Cimino

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

Abstract

The UMLS Metathesaurus belongs to the class of scale-free networks with few concept hubs possessing a large number of relationships. The hubs provide useful links between the concepts from disparate terminologies in the UMLS; however, they also exponentially increase the number of possible transitive cross-terminology paths.

Towards the goal of using machine learning to rank cross-terminology translations, we propose a traversal algorithm that exploits the scale-free property of the UMLS to reduce the number of candidate translations. We characterize the concept hubs into "informational" and "noisy" concept hubs and provide an automated method to detect them.

Using gold standard mappings from SNOMED-CT to ICD9CM, we found an average 20-fold reduction in the number of candidate mappings while achieving comparable recall and ranking results. A hub-driven traversal strategy provides a promising approach to generate high quality cross-terminology translations from the UMLS.

Keywords:

UMLS, scale-free networks, terminology, mapping, machine learning

Introduction

The Unified Medical Language System® (UMLS) [1] is a large compendium of biomedical terminologies that integrates synonymous *terms* from disparate source terminologies into a *concept*. As a result, the source-asserted relationships between the terms provide cross-terminology relationships across the concepts in the UMLS. Various informatics applications pertaining to terminology translation, [2, 3, 4] information retrieval [5] and knowledge discovery [6] perform transitive traversal of these relationships to find mappings, links and sub-networks across biomedical terminologies in the UMLS.

One of the critical challenges in performing such transitive traversal of the UMLS relationships is the exponential increase in the number of possible paths. For example, starting with the concept *Hemoglobin* in the UMLS, there are 868 direct relationships to other concepts, which in turn have 450,692 relationships to a different set of UMLS concepts. Identifying the relevant concepts of interest (such as *Anemia*, *Methemoglobinemia*, etc.) with high accuracy among all the candidate paths entails a significant computational burden.

We found that the enormous increase in the number of transitive paths can be attributed to the scale-free property [7] of the UMLS. A graph is said to have a scale-free property when a few nodes (or "hubs") have most of the edges and large number of nodes are each connected by only a small number of edges. An indicator of scale-free characteristics for a network is whether it obeys the *power-law*, which states that the probability p(k) of a given node connected to k other nodes is proportional to k^{-c}, where c is an arbitrary constant. On plotting the power-law for the connectivity distribution of the UMLS 2005AA Metathesaurus (when viewed as a network with concepts as nodes and relationships as edges), we found the constant c=3.0032, strongly indicating the presence of scale-free characteristics for the UMLS.

Another relevant class of network is the small world network wherein the transitive edge traversal distance between any two nodes in the network is very small. For any scale-free network with c=3 (where c is the power-law constant), the average shortest path distance between any two nodes of the network is close to *ln N/ln ln N*, where N is the number of nodes in the network [8]. For the UMLS Metathesaurus, the distance is 5.29 (N=1179179), hence, for a given concept in the UMLS, any other concept can be reached by traversing on average 5.29 transitive relationships.

It has been shown that many real-world networks, such as the World Wide Web, biological networks, and social networks, are also scale-free [7]. However, the meaning of hubs (the nodes with large number of edges) is different across all networks and has important implications in transitive traversal of edges across the hubs. Consider, for example, the p53 molecule, which is an important hub for pathways across regulatory networks. In contrast, for metabolic networks, the water molecule hub acts as noise and produces irrelevant pathways.

In this paper, we characterize the concepts that act as hubs in the UMLS and propose a method to computationally identify the "informational" and "noisy" hubs. We propose a hub-driven algorithm for transitive traversal of relationships in the UMLS to minimize the exponential growth and generate high quality transitive paths. We evaluate the proposed approach with gold-standard mappings from SNOMED-CT to ICD9CM.

Related work

Existing approaches for traversing the UMLS to find mappings, links, and sub-networks have largely relied on expert-designed heuristics that exploit the semantics of relation-

ships. Cimino et al. (1993) [2] presented an approach of using the relationships in the UMLS for translating ICD9 terms to MeSH. The *restrict to MeSH algorithm* [3] starts by traversing hierarchical relationships (parent, broader), followed by associative relationships, until a concept from MeSH is reached. A method [4] to map SNOMED-CT and ICD9CM using the UMLS resulted in recall of 42% and precision of 20%. In our previous work, [5] the exponential rise in the number of links was shown as the critical limitation for applying machine learning approaches to identify cross-terminology translations.

Datasets

UMLS

The Metathesaurus (Meta) of the Unified Medical Language System (UMLS) integrates synonymous terms from different terminologies into concepts. The source terminology relationships are abstracted into 11 top-level relationship types (REL) such as *Parent (PAR), Sibling (SIB) or Child (CHD)*. The REL can have a relationship attribute (RELA) that provides further granular meaning such as *clinically_associated_with, mapped_from etc*. The relationships are provided in the MRREL file of the UMLS. The UMLS 2005AA distribution was used for the experiments reported in this paper.

SNOMED-CT to ICD9CM mappings

The SNOMED-CT vocabulary in the UMLS provides mappings from SNOMED-CT to ICD9CM terms, for example, *Massive Hepatic Necrosis* (SNOMED-CT) to *Acute and subacute liver necrosis* (ICD9CM). These mappings exist as separate relationships in the UMLS with relationship attribute type *mapped_from and mapped_to*. We used 65,417 such concept mappings from the 2005AA distribution of UMLS as our gold-standard.

Approach

Connection definition

Two given terminologies in the UMLS are labeled as the source terminology and target terminology that provide the source concepts and target concepts (respectively). We define a *connection* as a set of transitive relationships between a source concept and a target concept connected by zero or more intermediate terminologies through the relationships in MRREL.

- *0-Step Connection* – The source concept and the target concept are identical.
- *1-Step Connection* – A direct relationship between a source concept and a different, unique target concept.
- *2-Step Connection* – A set of two transitive relationships, wherein the source concept is related to a second concept, which in turn is related to the target concept.

Hub identification

Several different metrics can be used to categorize a given node as a hub or non-hub such as the *degree* (number of edges for a given node), *betweenness centrality* (how often a given node is encountered in shortest paths across the network) and the *"authority" measure* (as determined by algorithms such as the PageRank or HITS) [9]. Given that our goal is to spot the concepts responsible for linking a large number of disparate concepts, the node degree provides a desirable metric for identifying hub concepts in the Meta.

The degree for a given concept in Meta is calculated based on the number of relationships present in MRREL to different unique concepts (i.e. if there are multiple relationships between two concepts, it is considered as a single relationship edge). The concepts are then sorted in descending order based on concept degree to rank the highly connected concepts first. A threshold variable, k (*Hub cutoff*), is used to label the top k concepts as the hubs.

Characterizing concept hubs in meta

In our analysis of the hub concepts in the Meta, we observed a systematic semantic pattern in the type of hubs:

1. *"Noisy" Concept Hubs*: The set of concepts that do not generate meaningful transitive connections across them. Several of these concepts are found at the top of the hub list (see Figure 1a). We can further classify these into:
   a) *"Property" Concept Hubs*: The hub concepts that are used as "attributes" or "properties", for example, *metabolic aspects,* which relates cells, organs and diseases with metabolic biochemical changes

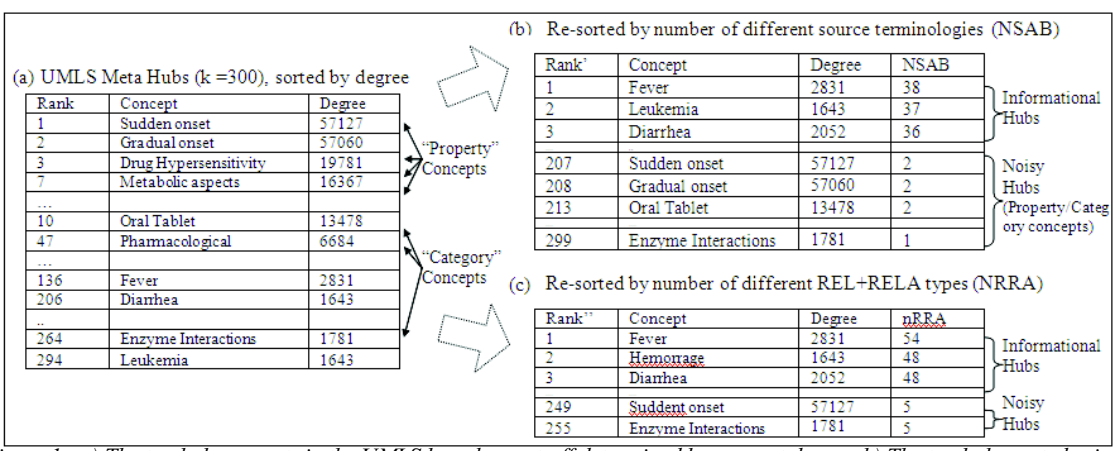

Figure 1 - a) The top hub concepts in the UMLS based on cut-off determined by concept degree. b) The top hubs sorted using the number of source terminologies for a given concept. c) The top hubs sorted using the number of unique relationship types

or *immunologic,* which provides immunological attributes to different concepts.

b) *"Category" Concept Hubs*: The hub concepts that are used as "containers" to group other concepts, for example, *Oral Tablets* (Methotrexate, Nitoman, etc.), or *Microbiological* to group organs, animals, etc used for microbiologic studies.

2. *"Informational" Concept Hubs*: In the concept hub list, we also found several concept hubs that are indispensable for generating useful cross-terminology connections across the Meta such as *Fever, Diarrhea, Edema, Hypertension, etc.* We were able to filter these "informational" concept hubs by re-sorting the hub list based on the number of source terminologies contributing to the Meta hub concept (Figure 1b). Another approach that produced similar results was sorting based on the number of different REL and RELA types associated with the concept e.g. *Fever* has 54 different types of relationships such as *RO_associated_with,RB_inverse_isa, RB_mapped_from* and so on (Figure 1c).

By using a threshold variable, *t (NSAB cutoff)*, the number of different source terminologies, we separate informational hubs from noisy hubs. We consider informational hubs as having number of source terminologies equal to or greater than *t* and the rest are labeled as noisy hubs. The noisy hubs are thrown away during the transitive traversal and the informational hubs are retained. The informational hubs nevertheless have significantly high degree and hence can possibly lead to explosion in number of possible connections. In the next section, we propose a strategy that deals with this explosion problem without eliminating any asserted relationships from the Meta.

Concept hub decomposition

To limit the exponential growth in the number of possible connections generated via hub concepts, we propose an approach to decompose the hubs. The core idea is to traverse relationships only from a single terminology when passing through a hub. Formally, if we have a connection of type $Concept_1 – Relationship_1 – Concept_2$ *(labeled as Hub)* $– Relationship_2 – Concept_3$ then the source terminology contributing the $Relationship_1$ should be same as the source terminology for $Relationship_2$ (Figure 2). The hub decomposition leads to a significant reduction in the number of possible connections, consider if we have a hub concept with *n* relationships all from different source terminologies, a naïve traversal would produce total *n*(n-1)* paths however with proposed approach we generate only *n* paths. An interesting aspect of the proposed approach is that informational hubs generally tend to have a large number of source terminologies, (Figure 1b) which further helps in reducing the number of connections.

Hub driven connection generating algorithm (CoGen)

We summarize here the connection generating algorithm (CoGen) that uses the aforementioned hub-based traversal strategies to generate high-quality connections from the Meta.

Algorithm: CoGen (SCUI, TSAB, n, k, t)

Input: *SCUI: Source Concepts, TSAB: Target Terminology, n: length of connection (n-step), k: Hub cutoff, t: NSAB cutoff*

Output: *Candidate Connections*

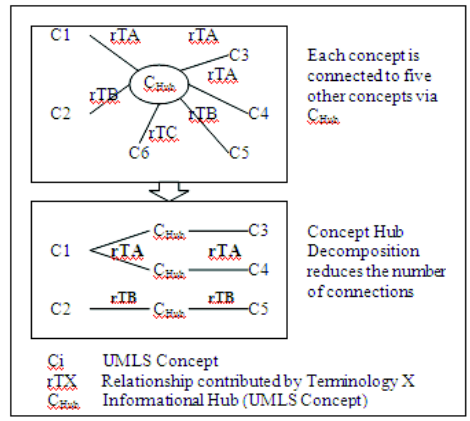

Figure 2 - An illustration showing how decomposing hub concepts reduces the number of connections by traversing only single source asserted terminology relationships via informational hubs

1. **[Clean Meta]**: *Remove all the relationships (REL) of type SIB (sibling), AQ (allowed qualifier) and QB (qualified by).*
2. **[Identify Hubs]**: *Sort the concepts in the cleaned Meta based on their degree. Label top k concepts as hubs.*
3. **[Differentiate Hubs]**: *Re-sort the hubs based on the number of source terminologies (or different relationship types). Label top t concepts as informational hubs and the rest as noisy hubs.*
4. **[Traverse Meta]**: *Generate Candidate Connections by traversing all n-step transitive connections from SCUI to concepts in TSAB, if encounter a concept of type*

   a. Noisy hub: break the traversal.

   b. Informational hub: use Concept Hub Decomposition.

Note that we added a pre-processing step to clean the Meta by removing all relationships of type SIB (sibling), AQ (allowed qualifier) and QB (qualified by) that do not generate meaningful connections in the transitive sense.

Learning features and training set

Towards our goal of using machine learning to identify and rank terminology translations from all possible connections, we describe here strategies for choosing learning features, developing training sets, and ranking the connections (refer to [5] for more details). The Semantic Network and Meta provide broad semantic categories for all concepts (Semantic Types, TUI) and relationships (REL, RELA). A connection can be abstracted into these broad semantic categories, which then creates learning tuples of the form $<REL_1, RELA_1, I_1TUI, REL_2, RELA_2$*:Decision Class*> for 2-step paths where $REL(A)_i$ indicates i^{th} step relationship, I_1TUI is the Semantic Type of intermediate concept and *Decision Class* can be *Positive* or *Negative*.

The positive examples are generated from the gold-standard terminology translations. However to generate negative examples we randomly sample 'strict' N-step connections i.e., the connections where the corresponding source and target concepts do not have any connection that is less than

N steps. The underlying rationale is that when the given concepts are related only by a 'strict' N-step connection, it implies they are relatively "unrelated" or "distant" as compared to the concepts related by less than N-step connection [5]. Once the connections are classified (using a supervised machine learning algorithm), we rank them based on the weighted difference of the number of positive and negative connections between the concepts. For example, given all 2-step connections between *Thyroid_Gland - Hypothyroidism*, 193 connections are classified as positive and 23 connections are classified negative, the connection rank is given by w_p*193-w_n*23, where w_p and w_n represent the average of confidence weight (for positive and negative connections respectively) obtained from a given classifier such as the likelihood of testing instance in Naïve Bayes.

Methods

1. The gold-standard mappings (SNOMED-CT to ICD9CM) were used to create the training sets. We randomly selected 1000 gold-standard mappings as positive training instances and 1000 'strict' 2-step connections as negative instances. The 2-step training connections were generated using the cleaned Meta (without SIB, AQ and QB).
2. Using the proposed CoGen algorithm, all possible 2 step connections were generated for 100 randomly selected source SNOMED-CT concepts to all possible candidate concepts in ICD9CM. The source concepts that were used for generating training sets were not included. Different set of connections for same source concepts were generated by varying each of the following thresholds
   a) The top k hub threshold *(Hub cutoff)* with values k=1000, k=5000 and k=9000
   b) The informational hub threshold *(NSAB cutoff)* with values t=5, t=10 and t=∞ (all hubs are considered as noisy)
3. Using simple transitive traversal (without scale-free view) over cleaned Meta, a different set of 2-step connections to ICD9CM were generated for the same 100 source SNOMED-CT concepts.
4. The training connections were used to learn a Naïve Bayes classifier. The test (candidate) connections were evaluated using the classifier and the likelihood was recorded for each connection. The connection rank was calculated for the classified connections.
5. The following measures were used to evaluate the results of the proposed approach.
   a) Recall: Number of source concepts, for which the correct (gold-standard mapped) target concept was *retrieved*.
   b) Top Rank: The number of correct target concepts ranked in the first position (top1), within top 5 and within top 10.

Results

For the randomly chosen 100 source SNOMED-CT concepts, 31 had the same target concept (0-Step Connection) and were excluded from further analysis. Considering the number of candidate connections generated (Figure 3a), the simple transitive traversal produced 315,329 connections, which was about 21.1 times more than the number of connections generated using CoGen (average 14,967.44). Subsequent analysis of connections produced by simple transitive traversal revealed 11,120 unique candidate ICD9CM concepts that represent 58.2 % of all ICD9CM concepts in the Meta. Within CoGen results, the increase in Hub cutoff parameter led to an overall reduction in the number of candidate connections and the reduction was more prominent for t=∞ (about 80% decrease from k=1000 to 9000), implying the importance of hub concepts in keeping the network connected. On removing all hubs (i.e. t=∞), 4,309 average connections were generated (62.3 connections per source concept) as compared to average 22,343 connections for t=5 (323.8 connections per source concept).

The 69 source SNOMED-CT concepts under analysis had gold standard mappings to 72 ICD9CM concepts. The simple transitive traversal retrieved a marginally higher number (57) of gold-standard ICD9CM concepts as compared to the

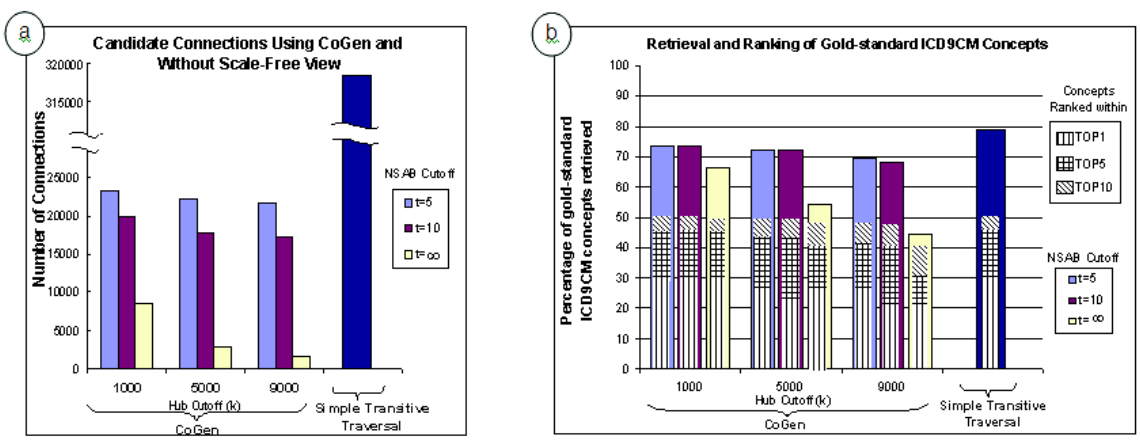

Figure 3 - a) the number of candidate 2-step connections generated for 69 source SNOMED-CT concepts to target ICD9CM using CoGen (with different parameters, k, t) and using simple transitive traversal. b) Percentage of total gold-standard target ICD9CM concepts (72) retrieved and the corresponding ranking after applying connection learning

CoGen results (54) (Figure 3b). The machine learning training set gave 10-fold cross-validation AUC (area under ROC curve) score of 0.97 for Naïve Bayes. No significant difference was observed in the ranking of gold-standard targets, despite of the large reduction in the number of candidate connections, indicating the high specificity of hub-based approach towards irrelevant translations. On average, 25.2% of time the gold-standard target ICD9CM concept was ranked in the top position and 48.9% of time in top 10 positions.

Discussion

Existing research on the UMLS has explored the semantic and linguistic aspects of the Meta at a greater depth, however, little analysis has been done to understand the large-scale structure and network of the Meta. Our results indicate that the hub concepts play a critical role in the Meta for applications that perform a transitive traversal of the relationships. We observed a 20-fold reduction in the number of candidate connections generated by proposed hub-based algorithm when compared to simple transitive traversal while the gold-standard connection ranking produced similar results; implying a significant decrease in the computational effort to traverse and rank the candidate connections. An argument can be made that all the connections corresponding to gold-standard mappings do not generally traverse through hubs. However, when we excluded all hubs ($t=\infty$) in the candidate connections, there was a significant decrease in recall for k=5000 and k=9000 and a marginal decrease for k=1000 (Figure 3b). This finding reflects the essential characteristic of all scale-free networks, where the few top hubs form the backbone of the network and removing them can cause the network to break apart.

Interestingly, the machine learning based ranking of connections was found to be independent of the size of the candidate set and the recall. This suggests that our machine learning approach is resilient to noisy connections. Secondly, it implies an inherent limitation in learning and detecting specific types of gold-standard connections. A possible solution would be to use different training sets to mine different types of connections.

Choosing an optimal value of cutoff parameters for the CoGen algorithm requires further analysis. Consider the following gold-standard connection,

Aircraft accidents – PAR - accidents (hub) – RN - Accident to powered aircraft, other and unspecified, injuring other person.

The hub concept *accidents* was removed when the hub cutoff changed from k=1000 to k=5000. Similarly, here is an example where change in NSAB cutoff led to disconnection of the gold-standard target

Acrokeratosis – RO – Hyperkeratosis (hub) – associated_with – Acquired Keratoderma

Hyperkeratosis is an informational hub at t=5 but becomes a noisy hub at t=10. Our experiments indicate that as more concepts are considered as hubs, (increasing k) a higher variation in the size of candidate connections is obtained for different values of NSAB cutoff parameter. We found that the optimal values of candidate size and ranking occur at around t=5.

In comparing our approach to the method by Fung et al. [4] over the same gold-standard dataset, our approach would obtain a higher recall given that we fetch all possible n-step connections for all relationship types and intermediate terminologies, instead of terminating the traversal when the first target terminology concept is found. Furthermore, our approach produces a ranked list of connections, thereby making it difficult to perform an exact comparison with the results obtained by Fung et al.

Our future work includes validating the approach on LOINC to CPT mappings and investigating the CoGen parameter settings for other terminological applications pertaining to knowledge discovery and information retrieval. Coupling other network properties of the UMLS (such as clustering coefficient, cohesion, and betweenness) to its semantic aspects might result in further improvements to our approach.

Conclusions

The inherent scale-free topology of the UMLS Metathesaurus provides a powerful feature to reduce the number of candidate cross-terminology connections without sacrificing the recall and ranking performance. Our experiments on gold-standard mappings from SNOMED-CT to ICD9CM showed a 20-fold decrease in the number of candidate translations by using a scale-free network view of the Metathesaurus. Generalizing the approach for terminology applications (other than translation) requires further investigation into optimal parameter settings of the proposed algorithm.

Acknowledgements

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

References

[1] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

[2] Cimino JJ, Johnson S, Peng P, Aguirre A. From ICD9-CM to MeSH using the UMLS: a how-to guide. Proc.Annu.Symp.Comput.Appl.Med.Care. 1993:730-734

[3] Bodenreider O, Nelson S, Hole W, Chang H. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc.AMIA.Symp. 1998:815-819.

[4] Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. Proc AMIA Annu Symp 2005:266-270.

[5] Patel CO, Cimino JJ. Mining Cross-Terminology links in the UMLS. Proc AMIA Annu Symp 2006:624-628.

[6] Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol. 2006 Jan;24(1):55-62.

[7] Barabasi A, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509-512.

[8] Cohen R, Havlin S. Scale-free networks are ultrasmall. Phys Rev Lett. 2003 Feb 7;90(5):058701.

[9] Wu AY, Garland M, Han J. Mining scale-free networks using geodesic clustering. Proc of KDD 2004:719-724.

Address for correspondence

Chintan O. Patel
Email: chintan.patel@dbmi.columbia.edu