Mining Cross-Terminology Links in the UMLS

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

OBJECTIVE: *To explore link mining approaches over transitive relationship paths in the Unified Medical Language System (UMLS). The goal is to classify relevant and 'interesting' cross-terminology links/paths for integration of Electronic Health Records (EHRs) and information resources.*
METHODS: *We present approaches for using the link semantics as learning features, sampling the UMLS to create training examples, and ranking the classified links. We use the clinical query and MEDLINE pairs in the OHSUMED dataset to extract 'gold-links' between SNOMED-CT and MeSH respectively, and compare them against corresponding two-step transitive links generated from the UMLS.*
RESULTS: *a). 75.7% increase in reachable MeSH concepts with two-step links as compared to direct one-step links b). 94.08% recall after link classification.*
CONCLUSION: *Using link mining with the UMLS is a promising approach for inter-terminology translation; further research is needed to handle the exponential link growth.*

1. INTRODUCTION

Integrating health related knowledge resources into electronic health records (EHR) enables clinicians to access relevant information at the point of care [1]. Towards this goal various solutions, such as KnowledgeLink [2] and Infobuttons [3] have been deployed and their impacts have been evaluated. The methods [3] used for integrating EHR and knowledge resources range from simple keyword-based searches to intelligent Web agents that interact with the resource. Terminology-based integration approaches [4] generally attempt to find matching concepts between the terminology used in the EHR and the indexing thesaurus/terminology used by the knowledge resource e.g. *Hgb* (LOINC:LP14449) and *Hemoglobin* (MeSH:D006454).
The Unified Medical Language System (UMLS) [5] provides such mappings across the terms in different terminologies by integrating terms with the same meanings into the same concepts; the relationships derived from source terminologies hence can relate terms across different terminologies. We can traverse these UMLS relationships in a transitive fashion and "hop" across different terminologies to discover relevant links or paths linking the source terminology and the indexing terminology. Consider the following examples from the UMLS for links emanating from Hemoglobin (C0019046) in LOINC

$Link_1$: **Hemoglobin** – RO – *has_component* – Hemoglobin measurement:CPT – RQ – *clinically_associated_with* – **Anemia:**SNOMED-CT

$Link_2$: **Hemoglobin** – SIB – Thyroxine:RxNORM – RO – *may_be_treated_by* – **Goiter:**MeSH

$Link_3$: **Hemoglobin** – CHD –Methemoglobin:MeSH – RQ – *mapped_to* – **Methemoglobinemia:**MeSH

(where RO=Related Other, RQ=Related/Synonymous, SIB=Sibling, CHD=Child, CPT=Current Procedural Terminology, MeSH=Medical Subject Headings)[1]

In context of performing information retrieval for clinically relevant concepts, it is evident that $Link_1$ relating *Hemoglobin-Anemia* is 'stronger' or more relevant than *Hemoglobin-Goiter*($Link_2$), whereas $Link_3$ relating *Hemoglobin-Methemoglobinemia* has relatively 'medium' relevance (in general) as compared to $Link_{1,3}$. In this paper, we present methods for classifying such cross-terminology links (in the UMLS) based on their relevance in order to integrate knowledge resources into EHRs.
The transitive traversal of relationships leads to an exponential increase in the total number of links. For example, Hemoglobin has 868 one-steps links or direct relationships in the MRREL table of UMLS[2] and 450,692 two-step transitive links. Consider a more global example: there are 613,058 direct relationships (in MRREL) between LOINC and SNOMED-CT concepts, 3,149,617 direct relationships between SNOMED-CT and MeSH concepts, however, there are 146,994,468 possible transitive paths from LOINC to SNOMED-CT to MeSH. Not all of these paths necessarily have a clinically relevant or meaningful relationship between the source and target concepts, hence the proposed link mining and ranking techniques become important.
Although direct relationships may exist between two given terminologies in the UMLS, we believe that a transitive traversal approach through different

[1] See UMLS documentation for details at: www.nlm.nih.gov/research/umls/documentation.html
[2] As per UMLS2005AA dataset.

intermediate terminologies is advantageous due to following reasons:

- We achieve higher coverage in terms of number of concepts than could be reached directly from the given concept in the source terminology.
- Transitivity leads to a network effect - for the purposes of mining interesting or novel links, the links that were not directly asserted (and were previously unknown) may be more desirable.

In [6], Bodenreider et al. use a deterministic graph traversal algorithm that exploits the 'semantic locality' in UMLS to map any concept in UMLS to MeSH. Zang et al [7] present an approach of using associative relationship patterns to map different ontologies. In this paper, we propose a novel learning approach that uses the semantics of relationships and intermediate concepts in the UMLS to classify relevant links. We evaluate our approach using the 'gold-links' extracted from the OHSUMED dataset [8].

2. DATASET

We used the following datasets to implement and evaluate the approaches presented in this paper

2.1. UMLS

The Unified Medical Language System (UMLS) [5] is an effort by the National Library of Medicine to create biomedical information resources amenable to computational inference and processing. In this paper, we use the Metathesaurus (META) and the Semantic Network (SN) component of the UMLS. Currently, META integrates about 140 biomedical terminologies by assigning the same concept identifier to synonymous terms from different source terminologies. The SN provides top-level concepts (e.g. Clinical Drug, Laboratory Test) to categorize the concepts in META. The relationships present in the source terminologies are asserted between the concepts (and not between the terms) in META, hence we obtain relationships across different terminologies.

2.2. OHSUMED

OHSUMED [8] is a widely used test collection of biomedical query-document pairs for evaluating information retrieval systems. The collection was prepared by capturing 106 clinical information needs/queries along with the history and context of patients while physicians were using a decision support system. Based on the query and context, a set of medical librarians and physicians searched MEDLINE articles from given time period (1987-1991, 348,566 articles) and classified them in to three relevancy classes – definitely(d), probably(p) and not relevant(n). Here is an example of a record from the OHSUMED

Context/History – *60 year old menopausal woman without hormone replacement therapy*
Query – *Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy*
MEDLINE UI – *87097544(d), 87157536(p), 87157537(n),…*
In our experiment, we use the entire query dataset and consider only the MEDLINE UIs that had definite (d) relevance.

3. APPROACH

3.1. Link Definition - N-Step Relationships

We define a *link* as a set of transitive relationships between a given concept in a source terminology and a set of concepts in a target terminology connected by zero or more intermediate terminologies in the UMLS. Further, a link may be characterized by the source and target Semantic Type e.g. links from Sodium (*Element or Ion*) to all *Findings* in MeSH.
0-Step Link – A 'loop' relationship between terms from different terminologies that are 'synonymous' and belong to a single unique concept in the UMLS.
1-Step Link – A direct relationship(s) asserted in the UMLS file MRREL between two concepts in different terminologies.
2-Step Link – A set of two transitive relationships, wherein a concept in a given source terminology is related to a concept in an intermediate terminology which in turn is related to a concept in a given target terminology.
Similarly, one can envision n-Step links that consist of n transitive relationships across n+1 terminologies.

3.2. Semantic Learning Features

To classify the links, we used following link attributes as the learning features -

3.2.1. Relationship Types (REL_i) – The relationship type (REL) for the i^{th} step relationship in a given link e.g. 2-step link having REL_1 = CHD, REL_2 = SIB. The RELs correspond to 11 broad semantic categories in the UMLS to which all source terminology relationships are abstracted viz. hierarchical such as Parent (PAR), Child (CHD), Narrower (RN), Broader (RB) and 'lateral' such as Sibling (SIB), Other (RO), synonym (SY), Allowed-Qualifier (AQ), Qualified-By (QB) etc.

3.2.2. Relationship Attributes ($RELA_i$) – Various qualifiers are assigned to REL that provide a more specific semantic meaning for the relationship e.g. REL_1=RO, $RELA_1$=clinically_associated_with. Only a few REL types have a corresponding RELA.

3.2.3. Intermediate Semantic Types (I_iTUI) - The Semantic Type of an intermediate concept at a given depth. The Semantic Types are derived from

the SN as described earlier. Some concepts can have more than one Semantic Type - for such cases we create different independent links for each corresponding Semantic Type.

These features provide 'link abstraction', which is needed for machine learning tasks. We experiment with 2-step links in this paper, so the instance tuples are of the form <REL_1, $RELA_1$, I_1TUI, REL_2, $RELA_2$, 'Class={Positive, Negative}'>

3.3. Training Set – Sampling the UMLS

In the context of information retrieval in EHRs, an important challenge lies in objectively determining what constitutes a relatively 'Relevant'(Positive) or a relatively 'Irrelevant'(Negative) link in order to create training examples for link classification. With the UMLS, a training set must be representative of the underlying distribution of the dataset and hence using training examples outside of UMLS would create models overfitted to the training set. We propose here methods for sampling the UMLS to create training sets that exploit specific properties of the UMLS structure to determine relative link relevance.

3.3.1. *Link Distance Approach* - The presence of a relationship between two concepts in the UMLS signifies the fact these concepts are 'relevant' or 'related' based on some source biomedical terminology. Hence, we assume that concepts related by 1-Step links have a 'stronger' relationship as compared to concepts related only by greater than 1-Step links. This approach is similar to the notion of Data Processing Inequality [9]. In general,

N-step links $>_{relevant}$ Strict (N+1)-step links

By 'Strict' (N+1)-step links we mean that the corresponding concepts don't have any N or less than N step links between them. Using this inequality, we can label learning class for the concept pairs, such as *N-step=Positive* and *strict (N+1)-step=Negative*. Then we find M-step links for these concept pairs to train a classifier for classifying M-step links (note that M has to be greater than or equal to N+1). Below is an example for creating a training set for 3-step link mining,

Sampling: C1 – C2 (1-step)
C3 – C4 (strict 2-step)
Training: <C1-X-Y-C2>: Positive
(3-step links) <C3-X-Y-C4>: Negative

where C1,3 are concepts in the source terminology and C2,4 are concepts in the target terminology and X,Y are arbitrary concepts in the intermediate terminologies of the 3-step link.

3.3.2. *Link Voting Approach* - In the UMLS, several source terminologies can assert the same or different relationships between two given concepts, all of which are stored in the UMLS. Using a simple voting approach based on number of links between two concepts in the UMLS, we can determine the strength of 'relatedness' for given concepts. The higher the link count between two concepts (for N-step links), we assume the higher link 'strength' or relevancy. The learning classes for concept pairs are created by using a threshold value (k) for link count i.e. *if N-Step link count > k* then *Positive* else *Negative*. We present a simple but effective strategy to select the value of k in Section 4.2.2.

3.4. Link Ranking

Here we present a method to rank the links classified by the learning algorithm. Given that two concepts can have more than one link between them, we calculate the link rank as the difference of number of links that were positively classified and negatively classified, e.g. for 2-step links between *Thyroid Gland – Hypothyroidism*, 193 links were classified as positive and 23 links were classified negative, hence link rank = 193-23 = 170. The link count difference is not normalized because we want to factor-in the measure of Link Voting evidence.

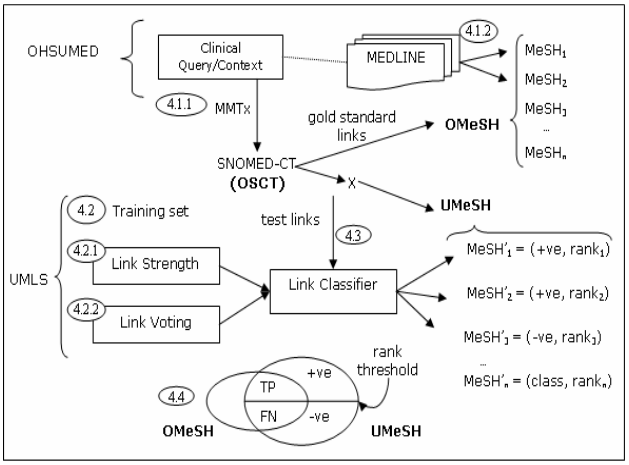

Figure 1. The experimental steps performed to evaluate the proposed link mining approaches. The details for each step are described in Section 4.

4. METHODS

In this section, we describe the methods that were implemented to evaluate the proposed link mining approaches. The steps of the experiment are depicted sequentially in Figure 1 and corresponding descriptions are given below:

4.1 Generating gold-standard links

4.1.1. The clinical queries and context descriptions in the OHSUMED dataset were processed using the MMTx tool[3] to extract UMLS concepts. Only the concepts having MMTx score above 900 (out of 1000) and source terminology as SNOMED-CT were retained (and labeled as set **OSCT**).

[3] MetaMap, http://mmtx.nlm.nih.gov/

4.1.2. The MeSH headings for MEDLINE UIs in OHSUMED were fetched using the Entrez Utilities[4] and the MeSH qualifiers/sub-headings were filtered out. Based on the query/document pairing (Section 2), gold standard links were assumed between the SNOMED-CT and MeSH concepts. These MeSH concepts were labeled as the set **OMeSH**.

4.2. Preparing Training sets – Another set of links were randomly sampled from the UMLS across SNOMED-CT and MeSH using the training strategies proposed in Section 3.3:

4.2.1. Link Distance Approach: 1-step links=Positive and strict 2-step links=Negative.

4.2.2. Link Voting: The value of k was determined by sorting links in descending order of link counts and thresholding k for top 50% links.

Different sizes of training examples were sampled and evaluated by performing a 10-fold cross-validation with Naïve Bayes (using Weka[5]) and Support Vector Machines (using libSVM[6]).

4.3. Link Classification– Using the UMLS, all possible 2-step test-links were generated from the source concepts in OSCT to MeSH as the target terminology (and these MeSH concepts were labeled as the **UMeSH** set). The 1-step links for concepts in OSCT were also generated. The 2-step test-links were passed through a Naïve Bayes classifier trained using the best performing (on cross-validation) training set generated from step 2. The semantic attributes of link (viz. $REL_{1,2}$ $RELA_{1,2}$ and I_1TUI) were used as the features for the classifiers.

4.4. Overlap/Evaluation – For each clinical query in OHSUMED, an overlap set of OMeSH concepts and positively classified MeSH links ($UMeSH^+$) in step 3 was calculated. The sensitivity/recall per query was calculated as follows:

$$\text{Recall} = \frac{|UMeSH^+ \cap OMeSH|}{|OMeSH|}$$

Then we tuned the $UMeSH^+$ membership by using the link rank measure i.e. top x% of ranked links were considered positive and plotted against the average recall.

5. RESULTS

From 106 clinical queries/context in OHSUMED, MMTx extracted 385 unique SNOMED-CT concepts (average 3.6 concepts per query). The 2252 MEDLINE (definitely (d) relevant) articles in OHSUMED mapped to 3046 unique MeSH concepts (average 28.7 per query). The results of cross-validation on training sets are shown in table 1.

Table 1. Area Under ROC (AUC) for cross-validation on training sets

Method	Size	Naïve Bayes	SVM
Link Dist Approach	500	0.927	0.924
	1000	0.934	0.938
	3000	0.929	0.925
Link Vote Approach	500 (k=44)	0.871	0.752
	1000 (k=38)	0.903	0.812
	3000 (k=38)	0.895	0.792

For 385 OSCT query concepts, we found 9,450 unique MeSH concepts with 1-step links and 63,770 unique MeSH concepts (UMeSH) with 2-step links in the UMLS. The overlap of 1-step MeSH concepts with OMeSH was 47.4% and for UMeSH, the overlap was 83.3% (i.e. there was a **75.7%** increase in overlap for 2-step links as compared to 1-step links).

The 2-step links were classified with a Naïve Bayes classifier using the training sample with 1000 examples (Link Dist Approach). The average recall was **94.08%** across all OHSUMED query records. Following are the examples of 2-step links (related to the OHSUMED query shown in Section 2.2) discovered by our approach:

Lipids→Hyperlipidemia→Osteoporosis
Progesterone→Inhibit_Ovulation→Medroxyprogesterone

The result of varying the $UMeSH^+$ set based on link rank is shown in Figure 2.

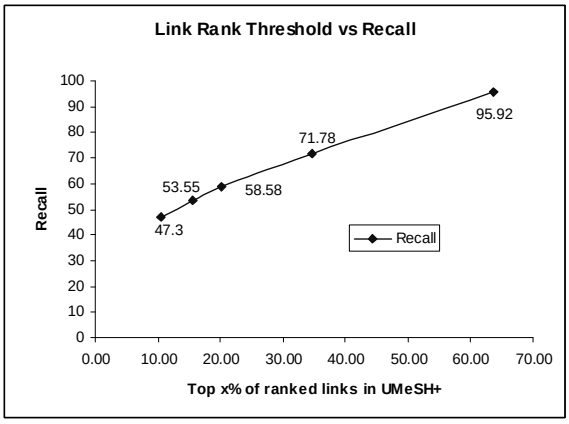

Figure 2. Average recall versus variation of $UMeSH^+$ size using the link rank

6. DISCUSSION

Most of the existing research using the UMLS has focused on finding mappings between similar or synonymous concepts across different terminologies. We use a broader meaning for 'mapping' that includes all types of 'links' (synonymous and other). The motivation to consider all links types was driven by the fact that we wanted to link clinical concepts (such as Hemoglobin) in an EHR terminology to not

[4] Entrez Utils, http://eutils.ncbi.nlm.nih.gov/
[5] Weka, http://www.cs.waikato.ac.nz/ml/weka/
[6] libSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

only synonymous concepts (of 0-step links) but all possible relevant concepts in an indexing terminology (Anemia, Methemoglobinemia) using n-step links. Given the extensive 'latent' knowledge hidden in the UMLS due to integration of biomedical terminologies, we believe that machine learning approaches such as the one proposed in this paper provide a key to unlock the hidden knowledge.

To understand the mechanics of why link semantics are good features for link classification, consider the following rules that were generated by a decision tree-based algorithm (C4.5) on our experimental training dataset of 2-step links:

If REL_1 = SY Then Positive
If REL_1= RN and $RELA_2$ = 'induces' Then Positive
If REL_1=PAR and REL_2 = CHD Then Negative
If REL_1= AQ and REL_2 = QB Then Negative

These rules make an intuitive sense if we consider the implied meaning of the relationships e.g. if first REL was Synonym (SY) relationship then the next step relationship becomes irrelevant, however, having first REL as Parent (PAR) and second as Child (CHD) becomes equivalent to a sibling relationship, which is a weak relationship. An interesting implication of cross-validation results is that although the training samples were generated using pure statistical properties of links, we obtained high classification accuracy with the semantic features.

We observed an exponential rise in number of links as we increased the link depth – which can be correlated to 'small-world' properties (few hub nodes with many connections) for some terminologies in the UMLS. We found 63,770 UMeSH concepts with 2-step links (for 385 source SNOMED-CT concepts) and using the link ranking threshold we obtain 47% of relevant OMeSH concepts in the top 10% of links (Figure 2). However, 10% of UMeSH is still 6377 concepts, which is a large number. Consider if we were to use these links to retrieve PubMed literature in EHR, we would still have 60 MeSH headings per query. This conclusion is not entirely accurate, as we treated each SNOMED-CT to MeSH 'gold-link' as independent regardless of other links present in the given query, which would have restricted the 'search-space'. Another possible solution to avert the problem of exponential increase would be to combine graph-theoretic measures (node degree, clustering coefficients, etc.) with link traversal and mining approaches.

Further interesting link analysis can be performed by restricting other parameters such as allowing only specific Semantic Types for source, intermediate and target concepts e.g. find all 2-step links to *Disease or Syndrome* concepts related to *Hemoglobin.* In addition, the intermediatory terminologies can be restricted; e.g. all 2-step paths from SNOMED-CT to MeSH that pass through RxNORM wherein we want to retrieve drug related literature.

7. CONCLUSION

The proposed link mining methods allowed discovery of (implicit and) transitive cross-terminology relationships in the UMLS (e.g. Lipids → Hyperlipidemia → Osteoporosis). The methods were evaluated using the OHSUMED gold-standard. Our results indicate that link mining is a promising approach for terminology-based integration of EHRs and knowledge resources, however further research is needed to handle the exponential link growth.

Acknowledgments

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

References

1. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE Button. Proc Annu Symp Comput Appl Med Care. 1992;:81-5.
2. Maviglia SM, Yoon CS, Bates DW, Kuperman G. KnowledgeLink: impact of context-sensitive information retrieval on clinicians'information needs.J Am Med Inform Assoc. 2006 Jan-Feb;13(1):67-73.
3. Mendonca EA, Cimino JJ, Johnson SB, Seol YH. Accessing heterogeneous sources of evidence to answer clinical questions.J Biomed Inform. 2001 Apr;34(2):85-98.
4. Cimino JJ, Elhanan G, Zeng Q. Supporting infobuttons with terminological knowledge.Proc AMIA Annu Fall Symp. 1997;:528-32.
5. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91
6. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp. 1998;:815-9.
7. Zhang S, Bodenreider O. Comparing associative relationships among equivalent concepts across ontologies. Medinfo. 2004;11(Pt 1):459-66.
8. Hersh WR, Buckley C, Leone TJ, Hickam DH. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research, 'Proc of the 17th Annual International ACM-SIGIR Conference. Dublin, Ireland, 3-6 July 1994, ACM/Springer, pp. 192-201.
9. Cover T, Thomas J. Elements of Information Theory. Wiley and Sons, New York, 1991.