# Semantic Distance of Concepts within a Unified Framework in the Biomedical Domain

Hisham Al-Mubaid
University of Houston-Clear Lake,
Houston, TX 77058, USA
hisham@uhcl.edu

Hoa A. Nguyen
University of Houston-Clear Lake,
Houston, TX 77058, USA
nguyenh3308@uhcl.edu

## ABSTRACT
This paper presents a cross-ontology approach, as an extension of the Cluster-Based approach, to measure semantic distance between concepts within single ontology or between concepts dispersed in multiple ontologies in a unified framework in the biomedical domain. The approach was evaluated in the biomedical domain within the UMLS framework with two biomedical ontologies (MeSH and SNOMED-CT). We used two datasets of biomedical terms scored for similarity by human experts. The experimental results (with ~0.81 correlation with human scores) confirmed that the proposed approach is effective and has great potential in measuring semantic distance using multiple ontologies in a unified framework.

## 1. INTRODUCTION
Ontology-based semantic distance (inverse of semantic similarity) techniques, also called similarity measures, can estimate the semantic similarity between two terms/concepts according to a given ontology or taxonomy. The pure ontology-based semantic distance/similarity measures use IS-A relations in ontology as the primary information source to determine the semantic similarity between concepts [1].

The Metathesaurus in UMLS (Unified Medical Language System) [3] is built from the electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging of biomedical literature. These are referred to as the "source vocabularies" of the Metathesaurus. The control vocabularies or terminologies in these resources are expressed hierarchically with the major relations between concepts are IS-A relations, therefore, these sources are also called ontology, taxonomy, etc. The ontologies in UMLS Metathesaurus overlap in a set of UMLS concepts.

In this paper, we propose an ontology-based semantic distance approach that can measure semantic distance in single ontology as well as in cross-ontology in UMLS framework. The proposed measure is adapted from (and is an extension of) the Cluster-Based approach proposed by Al-Mubaid & Nguyen [2] which was developed to compute the semantic distance/similarity between two terms across multiple clusters within a single ontology .

## 2. THE CROSS-ONTOLOGY APPROACH
In this section, we focus on extending and adapting the Cluster-Based approach [1] for measuring semantic distance of concept nodes in cross-ontology. We will treat ontology as a cluster, *i.e.*, the cluster here is one ontology in a unified framework and two ontologies overlap in set of controlled/unified concepts. The key point of this approach is the mapping of the secondary ontology into primary ontology doesn't deteriorate the semantic distance/similarity scale of the primary ontology according to different granularities of ontologies.

### 2.1 Cross-Ontology Semantic Distance
The four cases of semantic distance/similarity of concepts depending on whether the concept nodes occur in primary or in secondary ontologies. These four cases are also the same in the Cluster-Based approach. However, Case-2 of the cross-ontology approach is slightly different from Case-2 of the Cluster-Based approach according to the difference of the ontology mapping approach. **Case 1: Semantic Distance within the Primary Ontology:** If the two concept nodes occur in the primary ontology then the semantic distance (Dist) between two concept nodes is given as follows:

$$\mathrm{CSpec}(C_1,C_2) = D_1 - \mathrm{Depth}(\mathrm{LCS}(C_1,C_2)) \quad (1)$$

$$\mathrm{Dist}(C_1,C_2) = \log\left((\mathrm{Path}-1)^{\alpha} \times (\mathrm{CSpec})^{\beta} + k\right) \quad (2)$$

where $\alpha>0$ and $\beta>0$ are contribution factors of two features (Path and CSpec); $D_1$ is the depth of the ontology; k is a constant; and Path is the shortest path length between the two concept nodes.

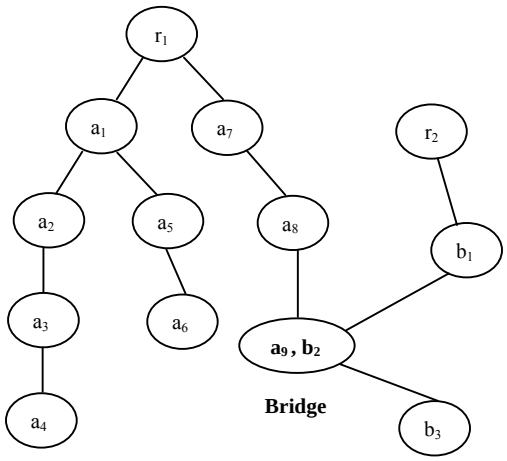


**Figure 1. Ontology mapping.**

**Case 2: Cross-Ontology Distance (Primary-Secondary):** In this case, the two concept nodes belong to two different ontologies. We connect the secondary ontology to the primary ontology by joining the associate/common nodes (*e.g.*, $a_9$ and $b_2$ in Figure 1 belong to two ontologies having roots of $r_1$ and $r_2$, respectively) of two ontologies. However, two ontologies may


*SAC'07*, March 11-15, 2007, Seoul, Korea.

have many common or equivalent concept nodes. Two concept nodes in two ontologies are equivalent if they refer to the same concept. For example, in Figure 1, suppose that $b_2$ and $a_9$ refer to the same concept ($b_2 = a_9$), then we merge $b_2$ and $a_9$ into one node called *Bridge* as in Figure 1. Thus, Figure 1 shows how the two ontologies are mapped and how the *Bridge* appears. As there can be more than one Bridge node when mapping two ontologies, there can be more than one LCS node ($\{LCS_n\}$) for the two concept nodes. The LCS node of two concept nodes ($C_1$, $C_2$) belonging to two ontologies is the LCS of the first node $C_1$ in primary ontology and the Bridges node, that is:

$$LCS_n(C_1,C_2) = LCS(C_1, Bridge_n) \quad (3)$$

such that $C_1$ belongs to the primary ontology $a_i$ while $C_2$ belongs to the secondary ontology $b_i$. The path length between two concept nodes in two ontologies passes through the Bridge node and goes through two ontologies having different granularities. The part of path length in secondary ontology is then converted into primary ontology's scale of path length feature, and the cross-distance of two concept nodes is given in Eq.(5). Finally, the semantic distance between two concept nodes are given as follows:

$$CSpec_n(C_1,C_2) = D_1 - Depth(LCS(C_1,Bridge_n)) \quad (4)$$

(5)

(6)

$$Dist(C_1, C_2) = \min\{Dist_n(C_1, C_2)\} \quad (7)$$

where $Path_n$ is the path length of two concept nodes calculated via $Bridge_n$; $d_1$, $d_2$ are parts of the path length distance via $Bridge_n$ between two concept nodes in primary ontology and secondary ontology, respectively; $D_1$, $D_2$ are depths of primary ontology and secondary ontology, respectively. The semantic distance between two concept nodes is finally chosen as the minimum among all possible semantic distances, Eq. (7).

## 2.2 Choosing the Secondary Ontologies

In the biomedical domain within the UMLS framework, as there are many ontologies overlapping in set of UMLS concepts, therefore, one problem stands out: which ontology is chosen as the secondary ontology? For that, we proposed a metric to measure the "*goodness"* of choosing a secondary ontology. The higher the *goodness* value, the better it is chosen as the secondary ontology for mapping for semantic distance/similarity. The metric is as follows:

(8)

where:

- Op is primary ontology and Os is a source ontology that is examined the goodness for choosing as secondary ontology.
- is the set of common concepts of two ontologies.
- is the union of two sets of concepts of two ontologies.
- Ds and Dp are depths of primary ontology and secondary ontology, respectively.

# 3. EXPERIMENTS AND RESULTS

## 3.1 Evaluation Method

To evaluate the approach in cross-ontology, we should have a dataset containing term pairs dispersed in multiple ontologies. For example, in Case-2, we need in one concept pair ($C_1$, $C_2$): one concept ($C_1$) belongs to primary ontology and the other concept ($C_2$) belongs to a secondary ontology and both the ontologies are in the unified framework. We do not have such dataset with human ratings; we, therefore, combined datasets from two domains: general English domain and biomedical domain. For that, we used a general English ontology, WordNet [1], and it does not belong to UMLS framework, therefore, two same concepts may have different names in two ontologies. For general English dataset, we used the well-known standard RG dataset containing 65 term pairs rated by human experts for semantic similarity. In biomedical domain, there are two datasets. The first one (Dataset 1) [2] contains 30 biomedical term pairs evaluated by 9 experts and 3 physicians, and the second one (Dataset 2) contains 36 biomedical term pairs evaluated by reliable doctors [2]. We used WordNet [1] as primary ontology and MeSH [3] as secondary ontologies. The default parameters ($\alpha$=1, $\beta$=1, k=1) of the cross-ontology approach are used in experiments this paper.

## 3.2 Experimental Results

We combined the RG dataset with the two biomedical datasets in three combinations as follows:

(a) RG (65 pairs) + Dataset 1 (30 pairs): total 95 pairs.
(b) RG (65 pairs) + Dataset 2 (36 pairs): total 101 pairs.
(c) RG + Dataset 1 + Dataset 2: total 131 pairs.

In these experiments, we used WordNet [1] as primary general ontology and MeSH as secondary ontology. We conducted three experiments using the three dataset combinations (a), (b) and (c). We only used 25 found pairs (out of the 30 pairs in Dataset 1) in MeSH in experiments. We further calculated semantic distance of 65 pairs in WordNet as in single WordNet ontology (Case-1) and other term pairs as cross-ontology technique (Case-3).

**Table 1. Absolute correlation results**

| Dataset combination | a | b | c |
|---|---|---|---|
| Correlation | 0.808 | 0.804 | 0.814 |
| *Average number of tested pairs* | ***105.7*** | | |
| *Average correlation* | ***0.809*** | | |

The results in Table 1 shows that the cross-ontology approach is very promising and efficient with very good correlations with human ratings in combined datasets using a primary ontology (WordNet) and the secondary ontology MeSH on three combined datasets.

# 5. CONCLUSION

We have presented a cross-ontology semantic distance approach in a unified framework. For example, in biomedical IR, there is a great need for measuring the semantic similarity between biomedical terms/concepts and documents and there are several potential ontologies. The experimental results show that the approach is very promising in computing semantic distance/similarity of concepts dispersed in multiple ontologies.

# 6. REFERENCES


[1] Budanitsky, A. and Hirst, G. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, vol.32, 1, March 2006.

[2] Al-Mubaid, H. and Nguyen, H. A. A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain. In *Proc. The 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006.

[3] UMLS. Available: http://umlsinfo.nlm.nih.gov