UMLS::Similarity: Measuring the Relatedness and Similarity of Biomedical Concepts

Bridget T. McInnes* **& Ying Liu**
Minnesota Supercomputing Institute
University of Minnesota
Minneapolis, MN 55455

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812

Genevieve B. Melton
Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455

Serguei V. Pakhomov
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455

Abstract

UMLS::Similarity is freely available open source software that allows a user to measure the semantic similarity or relatedness of biomedical terms found in the Unified Medical Language System (UMLS). It is written in Perl and can be used via a command line interface, an API, or a Web interface.

1 Introduction

UMLS::Similarity[1] implements a number of semantic similarity and relatedness measures that are based on the structure and content of the Unified Medical Language System. The UMLS is a data warehouse that provides a unified view of many medical terminologies, ontologies and other lexical resources, and is also freely available from the National Library of Medicine.[2]

Measures of semantic similarity quantify the degree to which two terms are similar based on their proximity in an *is-a* hierarchy. These measures are often based on the distance between the two concepts and their common ancestor. For example, *lung disease* and *Goodpasture's Syndrome* share the concept *disease* as a common ancestor. Or in general English, *scalpel* and *switchblade* would be considered very similar since both are nearby descendents of the concept *knife*.

However, concepts that are not technically similar can still be very closely related. For example, *Goodpasture's Syndrome* and *Doxycycline* are not similar since they do not have a nearby common ancestor, but they are very closely related since *Doxycycline* is a possible treatment for *Goodpasture's Syndrome*. A more general example might be *elbow* and *arm*, while they are not similar, an *elbow* is a *part-of* an *arm* and is therefore very closely related. Measures of relatedness quantify these types of relationships by using information beyond that which is found in an *is-a hierarchy*, which the UMLS contains in abundance.

2 Related Work

Measures of semantic similarity and relatedness have been used in a number of different biomedical and clinical applications. Early work relied on the Gene Ontology (GO)[3], which is a hierarchy of terms used to describe genomic information. For example, (Lord et al., 2003) measured the similarity of gene sequence data and used this in an application for conducting semantic searches of textual resources. (Guo et al., 2006) used semantic similarity measures to identify direct and indirect protein interactions within human regulatory pathways. (Névéol et al., 2006) used semantic similarity measures based on MeSH (Medical Subject Headings)[4] to evaluate automatic indexing of biomedical articles by measuring the similarity between their recommended terms and the gold standard index terms.

UMLS::Similarity was first released in 2009, and since that time has been used in various different applications. (Sahay and Ram, 2010) used it in a

*Contact author : bthomson@umn.edu.

[1] http://umls-similarity.sourceforge.net

[2] http://www.nlm.nih.gov/research/umls/

[3] http://www.geneontology.org/

[4] http://www.ncbi.nlm.nih.gov/mesh

Proceedings of the NAACL HLT 2013 Demonstration Session, pages 28–31,
Atlanta, Georgia, 10-12 June 2013.

health information search and recommendation system. (Zhang et al., 2011) used the measures to identify redundancy within clinical records, while (Mathur and Dinakarpandian, 2011) used them to help identify similar diseases. UMLS::Similarity has also enabled the development and evaluation of new measures by allowing them to be compared to existing methods, e.g., (Pivovarov and Elhadad, 2012). Finally, UMLS::Similarity can serve as a building block in other NLP systems, for example UMLS::SenseRelate (McInnes et al., 2011) is a word sense disambiguation system for medical text based on semantic similarity and relatedness.

3 UMLS::Similarity

UMLS::Similarity is a descendent of WordNet::Similarity (Pedersen et al., 2004), which implements various measures of similarity and relatedness for WordNet.[5] However, the structure, nature, and size of the UMLS is quite different from WordNet, and the adaptations from WordNet were not always straightforward. One very significant difference, for example, is that the UMLS is stored in a MySQL database while WordNet has its own customized storage format. As a result, the core of UMLS::Similarity is different and offers a great deal of functionality specific to the UMLS. Table 1 lists the measures currently provided in UMLS::Similarity (as of version 1.27).

The Web interface provides a subset of the functionality offered by the API and command line interface, and allows a user to utilize UMLS::Similarity without requiring the installation of the UMLS (which is an admittedly time–consuming process).

4 Unified Medical Language System

The UMLS is a data warehouse that includes over 100 different biomedical and clinical data resources. One of the largest individual sources is the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT), a comprehensive terminology created for the electronic exchange of clinical health information. Perhaps the most fine–grained source is the Foundational Model of Anatomy (FMA), an ontology created for biomedical and clinical research. One of the most popular sources is MeSH (MSH), a terminology that is used for indexing medical journal articles in PubMed.

Table 1: UMLS::Similarity Measures

Type	Citation	Name
Similarity	(Rada et al., 1989)	path
	(Caviedes and Cimino, 2004)	cdist
	(Wu and Palmer, 1994)	wup
	(Leacock and Chodorow, 1998)	lch
	(Nguyen and Al-Mubaid, 2006)	nam
	(Zhong et al., 2002)	zhong
	(Resnik, 1995)	res
	(Lin, 1998)	lin
	(Jiang and Conrath, 1997)	jcn
Relatedness	(Banerjee and Pedersen, 2003)	lesk
	(Patwardhan and Pedersen, 2006)	vector

These many different resources are semi-automatically combined into the Metathesaurus, which provides a unified view of nearly 3,000,000 different concepts. This is very important since the same concept can exist in multiple different sources. For example, the concept *Autonomic nerve* exists in both SNOMED CT and FMA. The Metathesaurus assigns synonymous concepts from multiple sources a single Concept Unique Identifier (CUI). Thus both *Autonomic nerve* concepts in SNOMED CT and FMA are assigned the same CUI (C0206250). These shared CUIs essentially merge multiple sources into a single resource in the Metathesaurus.

Some sources in the Metathesaurus contain additional information about the concept such as synonyms, definitions,[6] and related concepts. Parent/child (PAR/CHD) and broader/narrower (RB/RN) are the main types of hierarchical relations between concepts in the Metathesaurus. Parent/child relations are already defined in the sources before they are integrated into the UMLS, whereas broader/-narrower relations are added by the UMLS editors. For example, *Splanchnic nerve* has an *is-a* relation with *Autonomic nerve* in FMA. This relation is carried forward in the Metathesaurus by creating a parent/child relation between the CUIs C0037991 [Splanchnic nerve] and C0206250 [Autonomic nerve].

[5] http://wordnet.princeton.edu/

[6] However, not all concepts in the UMLS have a definition.

Table 2: Similarity scores for *finger* and *arm*

Source	Relations	CUIs	path	cdist	wup	lch	nam	zhong	res	lin	jcn
FMA	PAR/CHD	82,071	0.14	0.14	0.69	1.84	0.15	0.06	0.82	0.34	0.35
SNOMED CT	PAR/CHD	321,357	0.20	0.20	0.73	2.45	0.15	0.16	2.16	0.62	0.48
MSH	PAR/CHD	26,685	0.25	0.25	0.76	2.30	0.18	0.19	2.03	0.68	0.55

5 Demonstration System

The UMLS::Similarity Web interface[7] allows a user to enter two terms or UMLS CUIs as input in term boxes. The user can choose to calculate similarity or relatedness by clicking on the **Calculate Similarity** or **Calculate Relatedness** button. The user can also choose which UMLS sources and relations should be used in the calculation. For example, if the terms *finger* and *arm* are entered and the **Compute Similarity** button is pressed, the following is output:

```
View Definitions
View Shortest Path

Results:
The similarity of finger
(C0016129) and arm (C0446516)
using Path Length (path) is
0.25.

Using:
  SAB :: include MSH
  REL :: include PAR/CHD
```

The **Results** show the terms and their assigned CUIs. If a term has multiple possible CUIs associated with it, UMLS::Similarity returns the CUI pair that obtained the highest similarity score. In this case, *finger* was assigned CUI *C0016129* and *arm* assigned CUI *C0449516* and the resulting similarity score for the path measure using the MeSH hierarchy was 0.25.

Additionally, the paths between the concepts and their definitions are shown. The **View Definitions** and **View Shortest Path** buttons show the definition and shortest path between the concepts in a separate window. In the example above, the shortest path between *finger* (C0016129) and *arm* (C0446516) is **C0016129 (Finger, NOS)** => **C0018563 (Hand, NOS)** => **C1140618 (Extremity, Upper)** => **C0446516 (Upper arm)**, and one of the definitions shown for *arm* (C0446516) is **The superior part of the upper extremity between the shoulder and the elbow.**

SAB :: **include** and **REL** :: **include** are configuration parameters that define the sources and relations used to find the paths between the two CUIs when measuring similarity. In the example above, similarity was calculated using PAR/CHD relations in the MeSH hierarchy.

All similarity measures default to the use of MeSH as the source (SAB) with PAR/CHD relations. While these are reasonable defaults, for many use cases these should be changed. Table 2 shows the similarity scores returned for each measure using different sources. It also shows the number of CUIs connected via PAR/CHD relations per source.

A similar view is displayed when pressing the **Compute Relatedness** button:

```
View Definitions
View Shortest Path

Results:
The relatedness of finger
(C0016129) and arm (C0446516)
using Vector Measure (vector)
is 0.5513.

Using:
  SABDEF :: include
       UMLS_ALL
  RELDEF :: include
       CUI/PAR/CHD/RB/RN
```

Relatedness measures differ from similarity in their use of the SABDEF and RELDEF parameters. **SABDEF** :: **include** and **RELDEF** :: **include** define the source(s) and relation(s) used to extract definitions for the relatedness measures. In this example, the definitions come from any source in the UMLS and include not only the definition of the concept but

[7] http://atlas.ahc.umn.edu/

Table 3: Relatedness scores for *finger* and *arm*

Source	Relations	lesk	vector
UMLS_ALL	CUI/PAR/CHD/RB/RN	10,607	0.55
UMLS_ALL	CUI	39	0.05

also the definition of its PAR/CHD and RB/RN relations. Table 3 shows the relatedness scores returned for each of the relatedness measures using just the concept's definition (CUI) from all of the sources in the UMLS (UMLS_ALL) and when the definitions are extended to include the definitions of the concept's PAR/CHD and RB/RN relations.

6 Acknowledgments

This work was supported by the National Institute of Health, National Library of Medicine Grant #R01LM009623-01. It was carried out in part using computing resources at the University of Minnesota Supercomputing Institute.

The results reported here are based on the 2012AA version of the UMLS and were computed using version 1.23 of UMLS::Similarity and version 1.27 of UMLS::Interface.

References

S. Banerjee and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August.

J.E. Caviedes and J.J. Cimino. 2004. Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics*, 37(2):77–85.

X. Guo, R. Liu, C.D. Shriver, H. Hu, and M.N. Liebman. 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–973.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on Intl Conf on Research in CL*, pages pp. 19–33.

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

D. Lin. 1998. An information-theoretic definition of similarity. In *Intl Conf ML Proc.*, pages 296–304.

PW Lord, RD Stevens, A. Brass, and CA Goble. 2003. Semantic similarity measures as tools for exploring the gene ontology. In *Pacific Symposium on Biocomputing*, volume 8, pages 601–612.

S. Mathur and D. Dinakarpandian. 2011. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, 45(2):363–371.

B.T. McInnes, T. Pedersen, Y. Liu, S. Pakhomov, and G. Melton. 2011. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 895 – 904, Washington, DC.

A. Névéol, K. Zeng, and O. Bodenreider. 2006. Besides Precision & Recall: Exploring Alternative Approaches to Evaluating an Automatic Indexing Tool for MEDLINE. In *AMIA Annu Symp Proc.*, page 589.

H.A. Nguyen and H. Al-Mubaid. 2006. New ontology-based semantic similarity measure for the biomedical domain. In *Proc of the IEEE Intl Conf on Granular Computing*, pages 623–628.

S. Patwardhan and T. Pedersen. 2006. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proc of the EACL 2006 Workshop Making Sense of Sense*, pages 1–8.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *The Annual Meeting of the HLT and NAACL: Demonstration Papers*, pages 38–41.

R. Pivovarov and N. Elhadad. 2012. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of Biomedical Informatics*, 45(3):471–481.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th Intl Joint Conf on AI*, pages 448–453.

S. Sahay and A. Ram. 2010. Socio-semantic health information access. In *Proceedings of the AAAI Spring Symposium on AI and Health Communication*.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Meeting of ACL*, pages 133–138, Las Cruces, NM, June.

R. Zhang, S. Pakhomov, B.T. McInnes, and G.B. Melton. 2011. Evaluating measures of redundancy in clinical texts. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1612.

J. Zhong, H. Zhu, J. Li, and Y. Yu. 2002. Conceptual graph matching for semantic search. *Proceedings of the 10th International Conference on Conceptual Structures*, pages 92–106.