UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity

Bridget T. McInnes, MS[1], Ted Pedersen, PhD[2], and Serguei V.S. Pakhomov, PhD[1]

[1] **University of Minnesota, Minneapolis, MN, USA**
[2] **University of Minnesota, Duluth, MN, USA**

Abstract

A number of computational measures for determining semantic similarity between pairs of biomedical concepts have been developed using various standards and programming platforms. In this paper, we introduce two new open-source frameworks based on the Unified Medical Language System (UMLS). These frameworks consist of the UMLS-Similarity and UMLS-Interface packages. UMLS-Interface provides path information about UMLS concepts. UMLS-Similarity calculates the semantic similarity between UMLS concepts using several previously developed measures and can be extended to include new measures. We validate the functionality of these frameworks by reproducing the results from previous work. Our frameworks constitute a significant contribution to the field of biomedical Natural Language Processing by providing a common development and testing platform for semantic similarity measures based on the UMLS.

Introduction

Automated discovery of groups of semantically similar biomedical concepts is critical to improving the sensitivity of document/information retrieval[1] of scientific journals and clinical reports, the development of biomedical terminologies and ontologies[2], and the clustering[3] of biomedical documents. These applications are used to address important clinical and research problems ranging from the identification of patients for clinical studies and finding articles with similar content in PubMed to clustering symptoms and disorders found in the text of clinical reports for post-marketing medication safety surveillance.

Semantic similarity measures quantify how "alike" (or similar) two concepts are by determining their closeness in a hierarchy. Currently, a number of semantic similarity measures for the biomedical domain have been developed by various investigators and groups. One of the challenging issues is that these efforts are typically independent of each other and rely on different standards, programming languages and interfaces to ontological resources. These differences make it difficult to implement published measures consistently and systematically compare the results obtained with various approaches. To address this problem, we developed two open-source, extensible frameworks called UMLS-Similarity[4] and UMLS-Interface[5]. UMLS-Similarity is designed to support efforts aimed at developing and testing new semantic similarity measures and comparing them to existing methods based on the Unified Medical Language System (UMLS). UMLS-Interface is the foundation of UMLS-Similarity and is designed to extract path and concept information from the UMLS.

In this paper, we introduce the UMLS, and existing packages that have been used to extract information from the UMLS. We discuss related semantic similarity work and then introduce the UMLS-Interface and UMLS-Similarity packages. Our experimental evaluation shows that UMLS-Similarity can reliably reproduce results from previous work.

Background

The Unified Medical Language System

The UMLS is a knowledge representation framework designed to support broad scope biomedical research. It includes over 100 controlled medical terminologies such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH). These terminologies are semi-automatically integrated into the UMLS Metathesaurus which organizes knowledge based on concepts labeled with Concept Unique Identifiers (CUIs). A CUI may refer to multiple concepts from the individual terminologies. These concepts are labeled with Atomic Unique Identifiers (AUIs). For example, the AUI Cold Temperature [A15588749] from MeSH and the AUI Low Temperature [A3292554] from SNOMED-CT are mapped to the CUI Cold Tem-

perature [C0009264]. The AUIs are encoded with relation information such as is-a relations. If there exists a relation between two AUIs, a relation is created between their associated CUIs.

The Metathesaurus contains two table that provide relation information between CUIs: MRHIER and MRREL. MRHIER contains the sources whose relations are identified by the source as hierarchical. MRREL contains both hierarchical and non-hierarchical information with the two main hierarchical relations being parent/child (PAR/CHD) and broader/narrower (RB/RN). MRHIER represents the full path-to-root from the sources whereas MRREL only represents pairwise relations. For most sources, it is possible to generate MRHIER from MRREL because the full path-to-root is a transitive closure of the pairwise PAR/CHD relation. There exist sources where this is not the case: AIR, OMS, SNM, USPMG, and MeSH (including its various translations).

The UMLS glossary defines a hierarchy to be "any source-asserted multi-level organization of a source vocabulary's content."[6] The nature and purpose of these hierarchies may differ between vocabularies. When a source is incorporated into the UMLS, the UMLS editors determine whether an explicit hierarchical structure exists. If there is, the source is included in MRHIER and MRREL as PAR/CHD relations; otherwise, it is not included in MRHIER but may be included either algorithmically or manually in MRREL as RB/RN relations by the UMLS editors. Due to the National Library of Medicine's (NLM) focus on source transparency both tables indicate where the relations originated.

Existing Interface Packages

There exist two other freely available packages that provide APIs to allow interaction with the UMLS: UMLS Knowledge Source Server (UMLSKS) API[7], and UMLS-Query[8]. The UMLSKS allows remote access to different versions of the UMLS installed and maintained by NLM on their servers.

UMLS-Query is a Perl module that allows access to the UMLS installed locally in a MySQL database. UMLS-Query was designed to map terms to CUIs, AUIs, String Unique Identifiers (SUIs) and Lexical Unique Identifiers (LUIs). It also extracts path information between AUIs using MRHIER very efficiently since MRHIER contains the complete paths of all the concepts to the root of any of its sources.

Related Work

Rada, et. al.'s[1] measure of Conceptual Distance was the first to quantify the similarity between concepts in the UMLS. Their measure uses RB/RN relations in MeSH. Caviedes and Cimino[9] also use this measure but use the PAR/CHD relations in MeSH, ICD-9-CM and SNOMED-CT.

Lord, et. al.[10] measure the similarity between concepts in the Gene Ontology (GO) using *is-a* relations. They experiment with the similarity measures proposed by Resnik[11], Lin[12], and Jiang & Conrath[13].

Pedersen, et. al.[14] measure the similarity between concepts in SNOMED-CT using *is-a* relations. They experiment with the similarity measured proposed by Leacock & Chodorow, Lin, Jiang & Conrath, Resnik, the Path measure and their relatedness measure.

Nguyen and Al-Mubaid[15] proposed a new path-based measure using *is-a* relations in MeSH. They compare this with measures introduced by Leacock & Chodorow, Wu & Palmer, Choi & Kim, and the Path measure.

Lee, et. al.[16] investigate using the similarity measures described by Al-Mubaid and Nguyen[17] and two distance measures described by Melton, et. al.[18] using SNOMED-CT.

Methods

UMLS-Interface is a Perl package that provides an API to a local installation of the UMLS in a MySQL database, as well as command line programs to allow a user to interactively explore the UMLS. As of version 0.23, there exist 13 utility programs, five of the utility programs return information about a CUI such as its definition or semantic type. The remaining eight programs return path information such as all possible paths of a concept to the root. These programs require the source and relation information be specified in a configuration file.

UMLS-Interface obtains the path information about a CUI from MRREL. This allows for paths from multiple sources and multiple relations to be returned. A path may be limited to a single source and just the PAR/CHD relations replicating UMLS-Query except for the sources whose path-to-root is not a transitive closure of the pairwise PAR/CHD relation such as MeSH. It also allows for the RB/RN relations and relations between multiple sources to be included in the path-to-root information.

UMLS-Similarity is a Perl package that provides an API and a command line program to obtain the semantic similarity between CUIs in the UMLS given a specified set of source(s) and relations. As of version 0.13, UMLS-Similarity contains five semantic similarity measures proposed by Rada, et. al.[1], Wu & Palmer[19], Leacock & Chodorow[20], and Nguyen & Al-Mubaid[15], and the Path measure.

The Conceptual Distance (Cdist) measure proposed by Rada, et. al. determines the similarity between two

Table 1: UMLS-Similarity Results

				SNOMED-CT				MeSH							
Medical Term Pair		CUIs		path		lch		path		lch		wup		nam	
Term 1	Term 2	CUI1	CUI2	score	rank	score	rank	score	rank	score	rank	score	rank	score	rank
Renal failure	Kidney failure	C0035078	C0035078	1.00	1	4.22	1	1.00	1	3.64	1	1.00	1	1.00	1
Abortion	Miscarriage	C0156543	C0000786	0.50	2	3.53	2	0.10	20	1.34	20	0.47	20	7.10	20
Heart	Myocardium	C0018787	C0027061	0.25	3	2.83	3	0.50	2	2.94	2	0.93	2	3.81	2
Metastasis	Adenocarcinoma	C0027627	C0001418	0.25	3	2.83	3	0.14	7	1.69	7	0.63	9	6.43	8
Pulmonary fibrosis	Lung cancer	C0034069	C0242379	0.25	3	2.83	3	0.33	3	2.54	3	0.89	3	4.70	3
Brain tumor	Intracranial hemorrhage	C0006118	C0151699	0.25	3	2.83	3	0.25	4	2.25	4	0.88	4	5.13	4
Rheumatoid arthritis	Lupus	C0003873	C0409974	0.20	7	2.61	7	0.11	12	1.44	12	0.56	13	6.83	13
Pulmonary embolus	Myocardial infarction	C0034065	C0027051	0.17	8	2.43	8								
Antibiotic	Allergy	C0003232	C0020517	0.17	8	2.43	8	0.10	20	1.34	20	0.47	20	7.10	20
Depression	Cellulitis	C0011581	C0007642	0.17	8	2.43	8	0.10	20	1.34	20	0.47	20	7.10	20
Diarrhea	Stomach cramps	C0011991	C0344375	0.14	11	2.27	11								
Multiple sclerosis	Psychosis	C0026769	C0033975	0.14	11	2.27	11	0.10	20	1.34	20	0.47	20	7.10	20
Mitral stenosis	Atrial fibrillation	C0026269	C0004238	0.14	11	2.27	11	0.20	5	2.03	5	0.78	5	5.64	5
Congestive heart failure	Pulmonary edema	C0018802	C0034063	0.14	11	2.27	11	0.14	7	1.69	7	0.67	6	6.32	7
Lymphoid hyperplasia	Laryngeal cancer	C0333997	C0007107	0.13	15	2.14	15	0.14	7	1.69	7	0.67	6	6.43	8
Diabetes mellitus	Hypertension	C0011849	C0020538	0.13	15	2.14	15	0.17	6	1.85	6	0.67	6	6.17	6
Carpal tunnel syndrome	Osteoarthritis	C0007286	C0029408	0.13	15	2.14	15	0.11	12	1.44	12	0.56	13	6.83	13
Xerostomia	Alcoholic cirrhosis	C0043352	C0023891	0.11	18	2.02	18	0.11	12	1.44	12	0.59	12	6.73	12
Peptic ulcer disease	Myopia	C0030920	C0027092	0.11	18	2.02	18	0.14	7	1.69	7	0.63	9	6.43	8
Appendicitis	Osteoporosis	C0003615	C0029456	0.11	18	2.02	18	0.11	12	1.44	12	0.56	13	6.83	13
Hyperlipidemia	Metastasis	C0020473	C0027627	0.11	18	2.02	18	0.11	12	1.44	12	0.56	13	6.83	13
Cortisone	Total knee replacement	C0010137	C0086511	0.09	22	1.82	22	0.08	24	1.15	24	0.42	24	7.38	24
Acne	Syringe	C0702166	C0039142	0.08	23	1.65	23	0.11	12	1.44	12	0.50	18	6.93	18
Stroke	Infarct	C0038454	C0021308	0.07	24	1.58	24	0.11	12	1.44	12	0.56	13	6.83	13
Varicose vein	Entire knee meniscus	C0042345	C0224701	0.07	24	1.58	24								
Rectal polyp	Aorta	C0034887	C0003483	0.07	24	1.58	24								
Delusion	Schizophrenia	C0011253	C0036341	0.07	27	1.51	27	0.13	11	1.56	11	0.63	9	6.64	11
Cholangiocarcinoma	Colonoscopy	C0206698	C0009378	0.07	27	1.51	27	0.07	25	1.00	25	0.38	25	7.62	25
Calcification	Stenosis	C0175895	C0009814	0.00	29	0.00	29	0.11	12	1.44	12	0.50	18	6.93	18

concepts by counting the number of edges between them. Its range is between zero and twice the depth of the taxonomy. The similarity measure proposed by Wu & Palmer (wup) is twice the depth of the two concepts least common subsumer (LCS) divided by the product of the depths of the individual concepts. The LCS is the most specific concept two concepts share as an ancestor. Its range is between zero and one. The similarity measure proposed by Leacock & Chodorow (lch) is the negative log of the shortest path between two concepts divided by twice the total depth of the taxonomy. Its range is unbounded. The similarity measure proposed by Nguyen & Al-Mubaid (nam) is the log of two plus the product of the shortest distance between the two concepts minus one and the depth of the taxonomy minus the depth of the concepts LCS. Its range depends on the depth of the taxonomy. The Path measure (path) is the reciprocal of the number of nodes between two concepts and its range is between zero and one.

Results and Discussion

We evaluate UMLS-Similarity by comparing its results to those described by Pedersen, et. al.[14], Nguyen and Al-Mubaid[15], and Caviedes and Cimino[9] showing the package can reliably reproduce their results.

Comparison with Pedersen, et. al.

Pedersen, et. al.'s dataset contains 30 medical term pairs whose semantic similarity was determined by nine medical coders and three physicians from the Mayo Clinic. The semantic similarity of each term pair was annotated based on a four point scale: (4.0) practically synonymous, (3.0) related, (2.0) marginally related and (1.0) unrelated. Out of the 30 medical term pairs, Pedersen, et. al. use 29 of them - one was excluded because it did not exist in SNOMED-CT version 2004.

We evaluate UMLS-Similarity on the same 29 term pairs using the PAR/CHD relations in SNOMED-CT from the UMLS version 2008AB. If multiple CUIs were associated with a term we chose the CUI that obtained the highest similarity score. We use the Path measure (path) and the measure introduced by Leacock & Chodorow (lch) in this evaluation. For each measure, we rank the terms based on their similarity scores and calculate the correlation between our rankings and the Physicians and Coders using the non-parametric Spearman rank correlation coefficient which is used because the scales used by the annotators for the dataset are different from the scores that are returned by the measures. The difference between the scales make it infeasible to conduct a direct comparison.

Table 1 shows the term pairs, their associated CUIs from SNOMED-CT, and the similarity of the terms

determined by UMLS-Similarity for path and lch. Table 2 shows the correlation results between the ranking obtained by UMLS-Similarity and the Physicians and Coders, and the results reported by Pedersen, et. al. The results show that UMLS-Similarity obtains the same correlation as Pedersen, et. al. demonstrating that UMLS-Similarity can be used to reliably reproduce Pedersen, et. al.'s results.

Table 2: Correlation Results

Measure		Physician	Coder
path	Pedersen, et. al.	0.36	0.51
	UMLS-Similarity	0.35	0.50
lch	Pedersen, et. al.	0.35	0.50
	UMLS-Similarity	0.35	0.50
path	Nguyen and Al-Mubaid	0.627	0.852
	UMLS-Similarity	0.486	0.581
lch	Nguyen and Al-Mubaid	0.672	0.856
	UMLS-Similarity	0.486	0.581
wup	Nguyen and Al-Mubaid	0.652	0.794
	UMLS-Similarity	0.453	0.535
nam	Nguyen and Al-Mubaid	0.666	0.862
	UMLS-Similarity	0.448	0.551

Comparison with Nguyen and Al-Mubaid

Nguyen and Al-Mubaid use 25 out of the 30 terms in the dataset created by Pedersen, et. al.[14] seen in Table 1 - five terms were excluded because they did not exist in MeSH version 2006. The mappings of the terms to CUIs in MeSH were obtained by first, using the online tool provided by Nguyen and Al-Mubaid[21] to obtain the MeSH identifiers and then mapping them to the appropriate CUIs using the UMLS MRCONSO table. There are nine terms that have a different MeSH mapping than the SNOMED-CT mapping: Acne (C0001144), Antibiotic (C0279516), Depression (C0011570), Lung Cancer (C0024121), Laryngeal cancer (C0023055), Lupus C0024143), Lymphoid hyperplasia (C0020507), Stenosis (C1261287), and Congestive heart failure (C0018801).

We evaluate UMLS-Similarity on the 25 term pairs using the PAR/CHD relations in MeSH from the UMLS version 2008AB. We use the Path measure (path) and the measures described by Leacock & Chodorow (lch), Wu & Palmer (wup), and Nguyen & Al-Mubaid (nam) in this evaluation. For each measure, we rank the terms based on the similarity scores and calculate the correlation between our rankings and the Physicians and Coders using the Spearman rank correlation coefficient. We compare our correlation results with those reported by Nguyen, et. al.

Table 1 shows the term pairs in the dataset and the similarity of the terms determined by UMLS-Similarity for path, lch, wup and nam using MeSH. Table 2 shows the correlation results between the ranking of UMLS-Similarity and the Physicians and Coders, and the results reported by Nguyen and Al-Mubaid.

These results show that UMLS-Similarity obtains a lower correlation with the Physician and Coder judgements than the results reported by Nguyen and Al-Mubaid. We believe that the reason is because different versions of MeSH were used to conduct the experiments, and the path information used by Nguyen and Al-Mubaid comes directly from MeSH while UMLS-Similarity obtained the information from MRREL. As previously stated, for most sources, it is possible to generate MRHIER from MRREL because the full path-to-root is a transitive closure of the pairwise PAR/CHD relations, this does not hold true for MeSH because a MeSH descriptor may have different children depending on its tree position.

Comparison with Caviedes and Cimino

Caviedes and Cimino evaluate the Conceptual Distance[1] for ten term pairs using a combination of the following terms: Digestive system diseases (C0012242), Peptic esophagitis (C0014869), Psychotherapy (C0033968), Thirst (C0039971), and Thoracic duct (C0039979).

We use UMLS-Similarity to calculate the Conceptual Distance for the term pairs using the PAR/CHD relations in MeSH from the UMLS version 2008. We rank the terms in the dataset based on the Conceptual Distance scores and calculate the correlation between our rankings and those reported by Caviedes and Cimino using Spearman rank correlation coefficient.

Table 3: Conceptual Distance (CDist) Results

Term Pair		Caviedes and Cimino		UMLS-Sim	
CUI1	CUI2	Cdist	Rank	Cdist	Rank
C0012242	C0014869	3	1	3	1
C0012242	C0033968	5	2	5	2
C0033968	C0039971	6	3	6	3
C0012242	C0039971	7	4	7	4
C0012242	C0039979	7	4	6	3
C0033968	C0039979	8	5	9	6
C0014869	C0033968	8	5	8	5
C0014869	C0039971	10	6	10	7
C0014869	C0039979	10	6	11	8
C0039971	C0039979	10	6	11	8

Table 3 shows the Conceptual Distance (Cdist) reported by Caviedes and Cimino and the Cdist obtained using UMLS-Similarity (UMLS-Sim). Six out of the ten scores are the same and four differ by one. We believe that the difference is due to the different version of MeSH used to conduct the experiments. The correlation between the results is 0.9576 showing that UMLS-Similarity can reliably reproduce the results described by Caviedes and Cimino.

Conclusion

In this paper, we introduced the UMLS-Interface and UMLS-Similarity packages that measure the semantic similarity between concepts in the UMLS. We showed that they can be used to reliably reproduce the previous results described by Caviedes and Cimino, Pedersen, et. al. and Nguyen and Al-Mubaid.

In the future, we plan to explore other semantic similarity measures such as information content based measures which are based on the probability of the concept occurring in the taxonomy such as the measures proposed by Resnik[11] and Lin[12] as well as semantic relatedness measures. Semantic relatedness is a more general form of semantic similarity, for example, $foot$ and $pedal\ edema$ are not similar but are related, where as $foot$ and $hand$ are both similar and related. The semantic similarity framework described in this paper constitutes the first step to providing a publicly available common development and testing platform for semantic similarity and relatedness measures for biomedical NLP.

Acknowledgements

The authors thank Lan Aronson, Kin Wah Fung, Olivier Bodenreider and Jan Willis for their help in our understanding of the UMLS.

The first author was supported in part by an appointment to the NLM Research Participation Program sponsored by the National Library of Medicine and by the Graduate Assistance in Area of National Needs award sponsored by the US Department of Education. The second and third authors were supported in part by grant 1R01LM009623-01A2 from the National Library of Medicine, National Institutes of Health. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Health.

References

1. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans Syst Man Cybern Proc.*, 19(1):17–30, 1989.
2. O. Bodenreider and A. Burgun. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In *Medinfo Proc.*, pages 327–331, 2004.
3. Y. Lin, W. Li, K. Chen, and Y. Liu. A document clustering and ranking system for exploring medline citations. *J Am Med Inform Assoc.*, 14(5):651–661, 2007.
4. http://search.cpan.org/dist/UMLS-Similarity/ .
5. http://search.cpan.org/dist/UMLS-Interface/ .
6. http://www.nlm.nih.gov/research/umls/glossary.html .
7. http://umlsks.nlm.nih.gov/ .
8. N. Shah and M.A. Musen. UMLS-Query: A Perl module for querying the UMLS. In *AMIA Annu Symp Proc.*, pages 652–653, 2008.
9. J.E. Caviedes and J.J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *J of Biomed Inf.*, 37(2):77–85, 2004.
10. P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Pac Symp Biocomput Proc.*, volume 8, pages 601–612, 2003.
11. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Int J Comput AI Proc.*, 1:448–453, 1995.
12. D. Lin. An information-theoretic definition of similarity. In *Intl Conf ML Proc.*, pages 296–304, San Francisco, CA, 1998.
13. D. Conrath J. Jiang. Semantic similarity based on corpus statistics and lexical taxonomy. In *Comp Linguist Proc.*, pages pp. 19–33, Taiwan, 1997.
14. T. Pedersen, S.V. Pakhomov, S. Patwardhan, and C.G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *J of Biomed Inf.*, 40(3):288–299, 2007.
15. H.A. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *IEEE Eng Med Biol Proc.*, pages 623–628, 2006.
16. W.N. Lee, N. Shah, K. Sundlass, and M. Musen. Comparison of ontology-based semantic-similarity measures. In *AMIA Annu Symp Proc.*, pages 384–388, 2008.
17. H. Al-Mubaid and H.A. Nguyen. A cluster-based approach for semantic similarity in the biomedical domain. In *IEEE Int Conf GR Proc.*, pages 2713–2717, 2006.
18. G.B. Melton, S. Parsons, F.P. Morrison, A.S. Rothschild, M. Markatou, and G. Hripcsak. Inter-patient distance metrics using SNOMED-CT defining relationships. *J of Biomed Inf.*, 39(6):697–705, 2006.
19. Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Assoc Comput Linguist Proc.*, pages 133–138, Las Cruces, NM, 1994.
20. C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press, 1998.
21. http://sce.uhcl.edu/biomedsim/ .