Evaluating Semantic Relatedness and Similarity Measures with Standardized MedDRA Queries

Robert W. Bill[1], Ying Liu[2], PhD, Bridget T. McInnes[2], PhD, Genevieve B. Melton[1], MA, MD, Ted Pedersen[3], PhD, Serguei Pakhomov[1,2], PhD

[1]Institute for Health Informatics, University of Minnesota, Twin Cities, MN, USA
[2]College of Pharmacy, University of Minnesota, Twin Cities, MN, USA
[3]Department of Computer Science, University of Minnesota, Duluth, MN, USA

Abstract

A potential use of automated concept similarity and relatedness measures is to improve automatic detection of clinical text that relates to a condition indicative of an adverse drug reaction. This is also one of the purposes of the Medical Dictionary for Regulatory Activities (MedDRA) Standardized Queries (SMQ). An expert panel evaluates SMQs for their ability to detect a condition of interest and thus qualifies them as a reference standard for evaluating automated approaches. We compare similarity and relatedness measurement methods on rates of correctly identifying intra-category and inter-category concept pairs from SMQ data to create ROC curves of each method's sensitivity and specificity. Results indicate an information content measure, specifically the Resnik method, achieved the highest results as measured by area under the curve, but using two different measures as predictors, Resnik and Lin, obtained the highest score. Overall, using SMQ data resulted in a productive method of evaluating automated semantic relatedness and similarity scores.

Introduction

Multiple tasks in biomedicine, clinical care, and public health such as biosurveillance, clinical trials recruitment, and other secondary uses of clinical data rely on identifying groups of terms associated with a medical condition of interest. However, data mining methods using only preferred terms and their synonyms have limited sensitivity. The sensitivity of data mining and information retrieval efforts from medical texts can be improved by including nearly synonymous or semantically related terms in the search query along with the preferred term.

The Medical Dictionary for Regulatory Activities (MedDRA) Standardized Queries (SMQ) are created to improve adverse drug reaction signal detection from the data accumulated in the Adverse Event Reporting System (AERS) maintained by the U.S. Food and Drug Administration. SMQs are formed by defining clusters of MedDRA terms that are highly associated with a medical condition[1]. Using glaucoma as an example, the concepts *optic nerve cupping* and *eye pain* are listed among a number of other terms contained within the SMQ category for glaucoma. Therefore, using the cluster of concepts within the glaucoma SMQ for record retrieval should be helpful for detecting records that mention these related terms regardless of whether the term glaucoma is present. Concepts within SMQ categories are divided into broad and narrow terms. Within the glaucoma category, *eye pain* is categorized as a broad term and *optic nerve cupping* is a narrow term. Narrow terms are designed to result in fewer false positives than the use of broad terms. A 2009 study that compared data mining with preferred terms, high level terms and SMQs confirmed that the largest signal detection rate (highest sensitivity) was obtained with using SMQs[2]. An important property of SMQs is that they are developed and maintained by clinical experts[1]. The combination of clinical validation and evidence of beneficial effects when used in data mining qualifies SMQs as a reference standard against which automated semantic relatedness methods can be evaluated.

In contrast to manually created SMQs, automated semantic similarity and relatedness measures provide a numeric score for the similarity or relatedness between two terms or concepts. Accordingly, record retrieval could include a preferred term and related terms that fall within a predetermined threshold of relatedness according to one of the automated relatedness scoring methods. This approach could accomplish an outcome similar to the use of SMQs. There are various means of creating such scores that include, for example, the distance between term pairs in an ontology, or the amount of text overlap in the two terms' definitions. Because semantic relatedness methods seem to be task dependent, we are interested in examining MedDRA SMQ data as a potential standard for evaluating automated similarity and relatedness scores for their use in signal detection in the pharmacovigilance domain.

An assumption of the concepts in the SMQ data is that concepts pairs in the same SMQ category are more related than concepts pairs drawn from different categories. Therefore, we evaluate the various semantic similarity and relatedness measures on how well they predict if concept pairs are drawn from a single SMQ category (intra-category) or across different SMQ categories (inter-category). Results are reported as the area under the curve using receiver operator characteristic (ROC) curves.

Background

Unified Medical Language System (UMLS[16])

The UMLS is the aggregation of structured health and biomedical vocabularies and software. It contains the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus contains the terms and identifiers from over 100 different source terminologies for which the UMLS::Similarity software used in this study can use all, or a subset of the source vocabularies to calculate semantic similarity and relatedness measures. Terms that are related in the hierarchy of the Metathesaurus have a parent/child (PAR/CHD) relationship. During integration of vocabularies, editors also note those terms that are related by a broader/narrow (RB/RN) relationship. The UMLS::Similarity software optionally uses these relationships in calculating the similarity and relatedness of term pairs. The Lesk and vector measures used in this study used the CUI definition itself in addition to those concepts related by either a PAR/CHD or RB/RN relationship across all source terminologies available in UMLS.

Standardized MedDRA Queries

SMQs are a hierarchical collection of terms related to a clinical condition that experts have evaluated for consistency, the ability to detect the condition of interest, overlap and clinical relevance[1]. SMQ categories exist for over 80 conditions, and more are being developed. We labeled each term in 67 of these categories with a Unified Medical Language System (UMLS) Concept Unique Identifier (CUI), creating a total of 7,031 CUIs. Labeling was accomplished with an automated lookup of the SMQ term in the STR field of the MRCONSO relation within UMLS (if the optional MedDRA source has been added) and returning the UMLS CUI field associated with that record.

Table 1. SMQ categories having the greatest overlap.

Category 1	Overlap	Category 2
Ovarian neoplasms, malignant and unspecified	100.00%	Malignancies
Prostate neoplasms, malignant and unspecified	100.00%	Malignancies
Uterine and fallopian tube neoplasms, malignant and unspecified	98.48%	Malignancies
Torsade de pointes/QT prolongation	95.45%	Cardiac arrhythmias
Breast neoplasms, malignant and unspecified	93.48%	Malignancies
Noninfectious meningitis	83.12%	Noninfectious encephalitis
Noninfectious encephalitis	69.29%	Noninfectious encephalopathy/delirium
Peripheral neuropathy	67.27%	Guillain-Barre syndrome
Thrombophlebitis	64.00%	Embolic and thrombotic events
Torsade de pointes/QT prolongation	63.64%	Shock
Noninfectious encephalopathy/delirium	55.63%	Noninfectious encephalitis
Noninfectious meningitis	53.25%	Noninfectious encephalopathy/delirium
Dementia	52.56%	Noninfectious encephalopathy/delirium
Cerebrovascular disorders	52.08%	Embolic and thrombotic events
Noninfectious encephalitis	51.18%	Noninfectious meningitis

Each SMQ category contains varied numbers of concepts, and it is possible that a term/CUI appears in multiple categories. For example, the CUI C0085631 represents the terms *agitation, restlessness and psychomotor hyperactivity* that collectively appears in 14 different SMQ categories. 1,600 CUIS appeared in multiple SMQ categories in the set we examined. Some categories are fully contained in others in that all of the terms in said category also appear in a different category. Table 1 shows the SMQ category pairs with the greatest percentage of CUIs from category 1 appearing in category 2.

A basic aggregate test of automated similarity scores is that the average relatedness of terms within a category is greater than the average relatedness of that category's terms to the terms in a different category. This proved to be the case for 98% of the category pairs; however, categories with substantial overlap created instances where the highest similarity was between concepts from two different categories instead of between intra-category concepts. The overlap noted in Table 1 creates complications for automated clustering of such categories; however, the largest majority of categories have only 1 or 2 CUIS in common. Figure 1 is a histogram of the degree to which each possible pair of SMQ categories overlap. This figure is included to confirm that the degree of overlap is most frequently small (1-3 CUIs) between categories. Difficulty with clustering approaches due to category overlap led to the proposed method of distinguishing same-category pairs from different-category pairs because the results of such remains meaningful regardless of the number of categories in which a term exists.

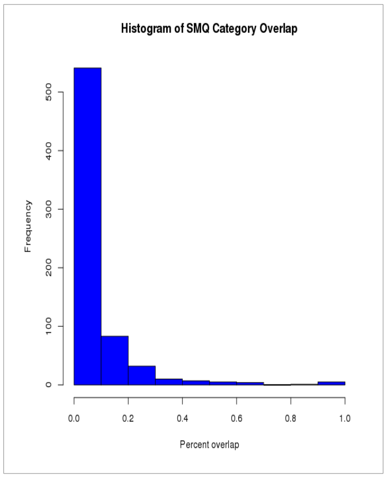

Figure 1.

Many SMQ categories contain subcategories. Additionally, preferred terms in a category are subdivided into broad and narrow types. It is possible to use the combination of hierarchies and the distinction of broad and narrow terms to create many potential tests for automated methods. However, for the purposes of this preliminary study, we limited these possibilities to using all terms within each category and did not consider the hierarchical information or the differences between broad and narrow terms; leaving that for future work.

Reference standards for evaluating the success of automated measures compared to human judgements are limited. In two studies, Pakhomov et al. propose a set of term pairs rated for similarity by a group of raters [17,19]. Consequently, these term pairs serve as a reference standard for evaluating automated semantic similarity and relatedness measures. However, such term pairs do not represent extensive collections of terms related to specific medical conditions of interest, such as is the case with SMQs.

Automatic Semantic Similarity and Relatedness

The different measures presented here can be roughly divided into similarity measures and relatedness measures. The distinction between similarity and relatedness measures is loosely based on whether ontological information was used in calculating the score with similarity having a unidirectional entailment relationship to relatedness [17,19]. The similarity measures include the path measure[3], Leacock-Chodorow[4], Resnik[5], Jiang-Conrath[6], and Lin[7]. The relatedness measures used include co-occurrence vectors [8,9] and Lesk[10].

The formulas for each measure can use path distance, information content or contextual information. The path measure came from an approach using the shortest distance between two concepts in a taxonomy originally proposed by Rada et al[11]. The current implementation uses the reciprocal of the shortest distance[3]. The Leacock-Chodorow[4] method is also a path-based measure but adds taxonomy/ontology depth information to the measure.

The Lin[7], Resnik[5] and Jiang-Conrath[6] methods utilize path information but also incorporate information content (the negative log of the probability of a concept) with the probability sources being drawn from the AERS database in this study. The Resnik measure uses the information content (IC) of the least common subsumer (LCS) of two concepts. Jiang-Conrath uses the sum of IC of the two concepts minus twice the IC of their LCS. The Lin method is the IC of the LCS of two concepts (multiplied by 2) divided by the sum of IC for the two concepts themselves. It is interesting to note that the Resnik measure appears in the numerator of the Lin measure. The similarity measures are formally defined in Table 2.

Table 2. Formulas for Similarity Measures[3]

Method	Formula	Variables
Path	$sim_{path} = \frac{1}{\text{minpath}} \left(c_{1,} c_2\right)$	**c** Concept **D** Depth of path
Leacock-Chodorow	$sim_{lch} = -\log \frac{\text{minpath}\left(c_{1,} c_2\right)}{2 * D}$	**IC** Information Content
Resnik	$sim_{res} = IC\left(lcs\left(c_{1,} c_2\right)\right) = -\log\left(P\left(lcs\left(c_{1,} c_2\right)\right)\right)$	**lcs** Least common subsumer
Jiang-Conrath	$sim_{jcn} = \frac{1}{IC\left(c_1\right) + IC\left(c_2\right) - 2 * IC\left(lcs\left(c_{1,} c_2\right)\right)}$	**minpath** Distance of the shortest path between two concepts **P** Probability **sim** Similarity
Lin	$sim_{lin} = \frac{2 * IC\left(lcs\left(c_{1,} c_2\right)\right)}{IC\left(c_1\right) + IC\left(c_2\right)}$	

The Lesk method counts the number of overlaps between two definitions. However, extended glosses that include the definitions of related terms, as proposed by Banerjee and Pedersen[12], are used in this study. We used the combination of definitions from the concepts themselves plus the definitions of concepts having a PAR/CHD and RB/RN relationship noted in the UMLS.

The co-occurrence vector method compares term definitions similar to the Lesk method but does not compare them directly. Instead, a context vector replaces each word in a definition. The context vector is the frequency of co-occurrence with other terms in a co-occurrence matrix built from the AERS database in this study. Frequencies of each definition word create a vector for each concept, and the cosine of the two vectors is the measure of their relatedness[9].

Methods

We used a sample consisting of a random selection of 10,000 term pairs that exist in the same SMQ category combined with a random selection of 10,000 term pairs that exist in different SMQ categories. We assigned UMLS CUIs to each term using the mapping information available in the UMLS MRCONSO table and then used the UMLS::Similarity package[13] to create automated similarity and relatedness scores for each of these 20,000 CUI pairs. The data set for each similarity and relatedness measure included two CUIs, the score itself, and a 0/1 for whether the two CUIs exist in the same category or not. Undefined similarity and relatedness measures were discarded, and the results should be interpreted as applying only to concept pairs with defined relatedness values. Different measurement methods have different reasons for undefined scores. For example, the Lesk method is undefined if there is no definition, while path-based measures are undefined if a path does not exist between the two concepts in the taxonomy being used or if one of the terms does not exist in that terminology. Consequently, undefined values are not always predictive of unrelated concepts and not do not imply anything similar across different measurement methods. This is the reason for the exclusion of undefined values. The UMLS::Similarity package has options to specify the source vocabularies and relationships that should be included in the final measurement. For the vector and Lesk methods, we used all the UMLS source vocabularies and we included definition data from related CUIs defined in the UMLS has having a PAR, CHD, RB, RN relationship. For similarity measures, the MedDRA source vocabulary was used with the PAR/CHD relationships.

Using R[14] and the pROC[15] library, ROC curves were created for each of the measures. The specific set of CUI pairs varied between measurement methods because all meaningful scores for a particular measure were included while excluding all undefined results. The similarity or relatedness score was used as the predictor in creating ROC curves, and the outcome variable was 1 or 0 respectively for whether the CUI pair exists in the same category or not. The score for the CUIs for *eye pain* and *optic nerve cupping* should be high enough to predict these concepts are both in the same SMQ category (glaucoma). Repeating this process over all possible thresholds and all CUI pairs creates a curve indicating a measurement method's sensitivity and specificity. We took the area under that curve as a measure for the overall effectiveness of a particular automated relatedness method. After repeating this process for the seven different relatedness measures, we can compare the differences in each area under the curve. The resulting ROC curves appear in Figure 2. The area under the curve (AUC) from the ROC plot was used to compare the different measures.

In addition to the methods mentioned, we tried combining multiple methods in a generalized linear model (GLM) with a logit link function. Creating a model that combines the scores from a similarity method and a relatedness method, or fitting two methods with different areas of strength in the ROC curve may have the potential to outperform independent methods.

Results

After creating relatedness scores, we calculated the average relatedness within a category and between each combination of categories. The mean score of all CUI pairs within a single category was generally higher than the mean similarity/relatedness between the terms of one category and other categories. Automated similarity/relatedness scores produced intra-category averages higher than the inter-category averages in 98% of the cases. Even the categories with substantial overlap generally reported greater mean within-category (intra-category) over between-category (inter-category) relatedness scores.

Concept coverage is the percent of concept pairs for which a measurement method has a meaningful score. Each method has conditions for which a relatedness score is not possible to create. A path-based relatedness score is undefined if concept pairs have no connecting path. Similarly, a definition-based relatedness score is undefined for concepts whose definitions do not appear in the available corpus. We did not use additional vocabularies and did not use concept pairs that have undefined relatedness for a particular method. Table 3 lists the number of undefined values in a random sample of 10,000 CUI pairs from the same category and 10,000 CUI pairs coming from different categories.

Table 3. Number of undefined values in a random sample of N=20,000

	Total	Intra-Category	Inter-Category
Jiang-Conrath	5,913 (29.6%)	3,576 (17.6%)	2,337 (11.7%)
Leacock-Chodorow	0	0	0
Lin	5,913 (29.6%)	3,576 (17.6%)	2,337 (11.7%)
Path	0	0	0
Resnik	36 (0.2%)	8 (0.0%)	28 (0.1%)
vector	68 (0.3%)	29 (0.1%)	39 (0.2%)
Lesk	68 (0.3%)	29 (0.1%)	39 (0.2%)

We chose to analyze each method independently using all CUI pairs defined for that particular measurement method. Corpus-based methods have lower coverage due to the limited number of concepts with definitions in the UMLS. Approximately 5% of the CUIs appearing in the UMLS MRCONSO table have definitions in the MRDEF table. It is apparent that augmenting the corpus with different dictionary resources has value for future studies[9]. Coverage is presented here as a consideration of the different measurement methods, but it only affects results in so much as it creates selection bias by limiting the random sample to concepts for which a measurement has a meaningful value. Concepts pairs without meaningful values were discarded.

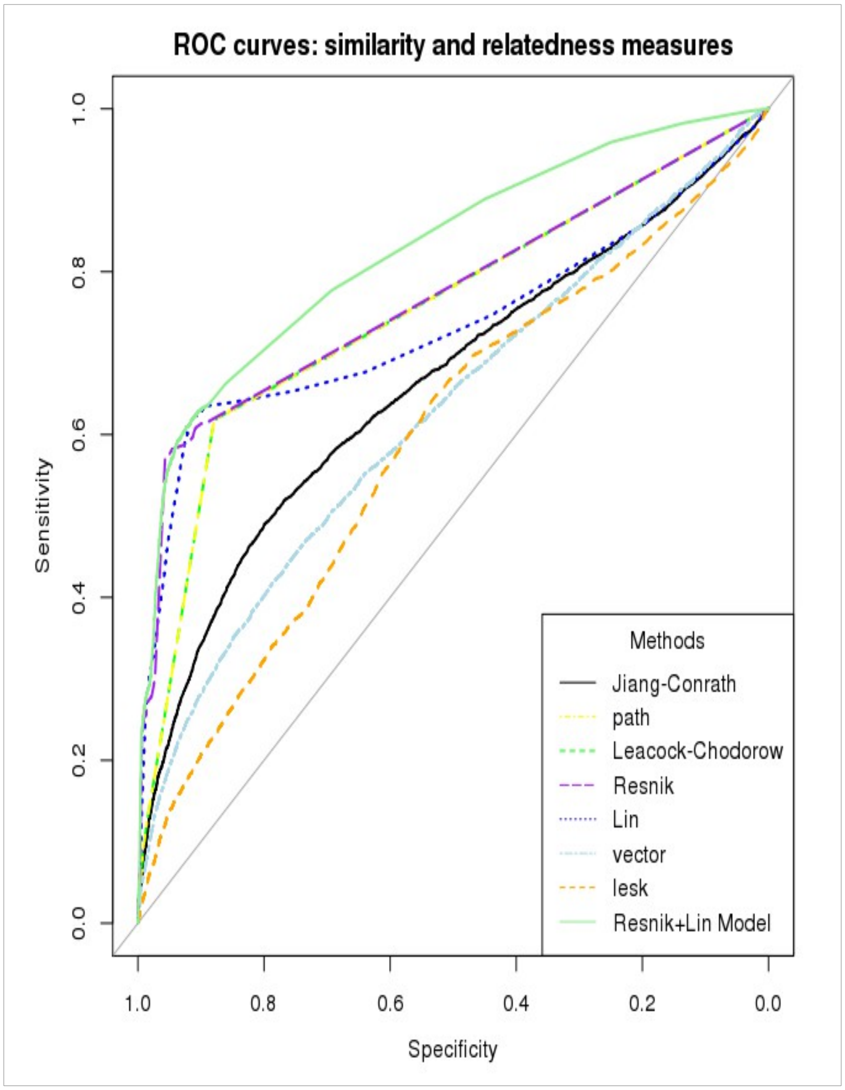

Figure 2.

Table 4 shows the AUC for the different relatedness scores. The data included 10,000 randomly selected term pairs from the same category and 10,000 randomly selected pairs from different categories. The corresponding ROC curves appear in Figure 2. The Resnik method achieved the highest AUC of the studied automated measurement methods. Fitting the combined Resnik and Lin measures in a logistic model and then using that model to predict outcomes resulted in a significantly higher AUC of 0.827 (95% CI 0.82-0.83). The Lin and path measures both had an AUC of 0.752. The Lesk method produced the lowest AUC. The only confidence intervals that overlapped were those for the Leacock-Chodorow and the path measures.

These results indicate that the combined use of Resnik and Lin measurements in a GLM can achieve significantly better category membership distinctions than each method can separately. Additionally, the use of SMQs combined with the AUC of ROC curves is an effective means of evaluating automated similarity and relatedness measures for use in pharmacovigilance. Models combing the Resnik with either the path or Leacock-Chodorow methods did not improve the AUC, but combining the Resnik and Lin methods did increase the AUC.

Table 4. Area under the curve by measurement method

Method	AUC	Confidence Interval
Resnik	0.772	0.766-0.777
Leacock-Chodorow	0.752	0.746-0.758
Path	0.752	0.746-0.758
Lin	0.734	0.725-0.743
Jiang-Conrath	0.662	0.653-0.671
Vector	0.626	0.618-0.634
Lesk	0.598	0.589-0.608
Combined Methods		
Resnik + Lin Model	0.827	0.820-0.834
Resnik + Leacock-Chodorow	0.772	0.767-0.778
Resnik + Path	0.772	0.767-0.778

Discussion

Evaluating automated similarity and relatedness scores is task dependent, and using SMQ terms and category information as a means of evaluating measures for the task of information retrieval or signal detection in the pharmacovigilance domain appears promising. The proposed evaluation framework is useful for optimizing relatedness scoring methods. The results from this study indicate that models of combinations of scores can currently achieve higher performance (measured by AUC). A measure's ability to predict inter-category and intra-category pairs produced ROC curves and AUC values with confidence intervals unique to all but the path-based and Leacock-Chodorow measures. Using those values to determine which combinations of methods to fit in GLMs led to significantly higher AUC values with both the Resnik and Lin scores used as predictors in the model. The reason for the higher AUC when using the combined Resnik and Lin model is not clear. Those measures are similar in their use of IC, and the formula for the Resnik measure appears in the numerator of the Lin measure. The higher performance may indicate the weighting of IC for the concepts themselves should be less than the weighting of the IC for their LCS. We do not think the combined benefit is related to concept coverage due to the fact that undefined similarity measures were discarded before evaluation; however, the effect of undefined scores is unclear.

Evaluating semantic similarity and relatedness measures based on their ability to distinguish intra-category concept pairs from inter-category pairs is easily generalizable to future SMQ categories as long as the terms used in these future SMQs are either drawn from the UMLS or can be mapped to the UMLS. However, selection of term pairs was random without regard for category in this study. Differences in performance of specific categories is unknown. This might be a limit generalization and is good to explore in the future.

The higher AUC when combining the Resnik and Lin measures is an important finding because it shows the utility of SMQs as a reference standard to evaluate how combinations of multiple semantic relatedness measures can perform better than individual measures. It demonstrates a productive way of testing various combinations of scores produced by semantic relatedness packages. This addresses an important issue in research and development of automated semantic relatedness measures that has to do with combining various scores produced by these measures. The scores produced by different measures are not always on the same scale – some vary between -1 and +1, others may vary between 0 and 1. Furthermore, it is not currently entirely clear how to compare scores computed on different scales or combine them in productive ways. Using generalized linear modeling of the kind described in this paper directly addresses this issue and offers a viable solution to this problem. However, having a large manually curated dataset as a reference standard is critical to being able to estimate the parameters of such multivariate statistical models. This use of the methods we have proposed in this paper, to evaluate and to identify incremental improvements, constitutes a valuable framework for evaluating automated measurement methods.

Conclusions

We conclude that the framework of using SMQ data to examine the ability of automated similarity or relatedness measures to differentiate within-category and between-category concept pairs is an effective means of evaluation. Use of this framework leads to the conclusion that IC-based similarity measures in general perform better than path or corpus-based relatedness measures when limited to using the UMLS. Specifically, the Resnik method achieved the highest AUC for existing methods. However, combining the Resnik and Lin measures in a GLM was able to achieve a higher AUC and invites testing of further combinations of similarity and relatedness scores.

Future directions indicated by these results include using this method of evaluation to identify or confirm optimal parameters and corpora for existing similarity and relatedness measurement methods, including methods not listed in this paper and sources beyond the UMLS. The higher results created by combinations of scores used in generalized linear models indicate additional experiments with different combinations of methods has potential for new and improved models for similarity and relatedness measures.

Acknowledgements

This work was supported by the National Library of Medicine grant R01LM009623.

References

1. Mozzicato P. MedDRA: An overview of the medical dictionary for regulatory activities. Pharmaceutical Medicine. 2009;23(2):65–75.
2. Pearson RK, Hauben M, Goldsmith DI, Gould AL, Madigan D, O'Hara DJ, et al. Influence of the MedDRA® hierarchy on pharmacovigilance data mining results. International Journal of Medical Informatics. 2009;78(12):e97–e103.
3. McInnes BT, Pedersen T, Liu Y, Melton GB, Pakhomov SV. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. AMIA Annual Symposium Proceedings. 2011:895.
4. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database. 1998;49(2):265–83.
5. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995:448-453.
6. Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the International Conference on Research in Computational Linguistics. 1997:19-33.
7. Lin D. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning. 1998:296-304.
8. Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. 2006:1–8.
9. Liu Y, McInnes BT, Pedersen T, Melton-Meaux G, Pakhomov S. Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet. Proceedings of the 2nd ACM SIGHIT International HealthInformatics Symposium. 2012:363-72.
10. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th Annual International Conference on Systems Documentation. 1986:24-6.
11. Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics. 1989;19(1):17-30.
12. Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. International Joint Conference on Artificial Intelligence. 2003:805-10.
13. McInnes B, Pedersen T, Pakhomov S. UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity. Proceedings of the Annual symposium of the American Medical Informatics Association. 2009:431-435.
14. R Development Core Team. R: A language and environment for statistical computing. 2011. http://www.R-project.org (accessed Jan 2012).

15. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):77.
16. Unified medical language system® (UMLS®). National Library of Medicine (US); 1996. http://www.nlm.nih.gov/archive/20040831/pubs/cbm/umlscbm.html (accessed Nov 2011).
17. Pakhomov S, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. Journal of Biomedical Informatics. 2011;44(2):251–65.
18. UMLS::Similarity. CPAN. Available from: http://search.cpan.org/dist/UMLS-Similarity/ (accessed Oct 2011).
19. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annual Symposium Proceedings. 2010:572-6.