Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study

Serguei Pakhomov, PhD[1, 3]; Bridget McInnes, PhD[1]; Terrence Adam, MD, PhD[1, 3]; Ying Liu, PhD[1]; Ted Pedersen, PhD[2]; Genevieve B. Melton, MD[3,4]

[1] *College of Pharmacy, University of Minnesota, MN, USA*

[2] *Department of Computer Science, University of Minnesota, Duluth, MN, USA*

[3] *Institute for Health Informatics, University of Minnesota, MN, USA*

[4] *Department of Surgery, University of Minnesota, MN, USA*

Abstract

Automated approaches to measuring semantic similarity and relatedness can provide necessary semantic context information for information retrieval applications and a number of fundamental natural language processing tasks including word sense disambiguation. Challenges for the development of these approaches include the limited availability of validated reference standards and the need for better understanding of the notions of semantic relatedness and similarity in medical vocabulary. We present results of a study in which eight medical residents were asked to judge 724 pairs of medical terms for semantic similarity and relatedness. The results of the study confirm the existence of a measurable mental representation of semantic relatedness between medical terms that is distinct from similarity and independent of the context in which the terms occur. This study produced a validated publicly available dataset for developing automated approaches to measuring semantic relatedness and similarity.

Introduction

The combination of clinical and biomedical terms organized into controlled vocabularies contained in the Unified Medical Language System (UMLS) and the use of large repositories of clinical and biomedical text provide a rich resource for developing automated approaches to measuring semantic similarity and relatedness among concepts. Querying electronic health record (EHR) systems for patients with a particular syndrome often requires using a variety of medical terms that not only denote diagnoses but also symptoms, treatments, conditions, and other concepts closely related to the syndrome. Automated measures of similarity and relatedness may be used to compile groups of terms to enhance querying of EHRs.

An established body of work in psycholinguistics focuses on lexical semantics and semantic relatedness.[1, 2] Two types of relatedness have been identified and studied in detail – associative and semantic. Associative relatedness refers to the probability that one word calls to mind another word (e.g., needle-thread), while semantic relatedness by psycholinguistic definition reflects the degree of semantic feature overlap between words (e.g., whale-dolphin). This distinction is based on the results of priming experiments in which, for example, a prime word that is either semantically related or unrelated to the target is shown to the subject first and the reading time or another type of response (e.g., eye movements) to the presentation of the target word is measured. These experiments indicate that subjects respond faster to targets primed with words that have common semantic features (i.e., are semantically similar) rather than those that have an associative relationship to the target (i.e., are semantically related).[3, 4] In addition to behavioral priming experiments, neuroimaging studies also demonstrated that semantically related words elicit clearly detectable differences in neural response from unrelated words.[5, 6]

Currently, several research groups, including ours, are investigating computerized methods for determining the strength of similarity and relatedness between medical terms.[7-12] One of the critical prerequisites in this work is the availability of validated reference standards that may be used to assess the performance of automated algorithms relative to human judgments. Furthermore, in order to advance this area of research, a more detailed understanding of the notions of semantic relatedness and similarity in medical language is needed.

The objectives of the present study were to a) confirm that relatedness is distinct from similarity of concepts in the biomedical domain, b) determine if context-free semantic relatedness can be measured behaviorally, and c) to create a publicly available dataset that may be used as a reference to develop and test approaches to measuring semantic relatedness and similarity.

Materials and Methods

Participants: Eight medical residents (2 women and 6 men; mean age - 30) at the University of Minnesota Medical School were invited to participate for a modest monetary compensation. Participants represented a convenience sample of all medical residents at the University of Minnesota. This study was approved by the University of Minnesota Institutional Review Board.

Data set: The term pairs dataset was compiled by first selecting all concepts from the UMLS (unrestricted by source) with one of three semantic types: disorders, symptoms and drugs. Subsequently, only concepts with entry terms containing at least one single-word term were further selected to control for potential differences in similarity and relatedness responses due to differences in term complexity. After this automated selection, a practicing physician (Adam) manually selected pairs of the single-word terms to contain approximately 30 term pairs in each of the four relatedness categories (completely unrelated, somewhat unrelated, somewhat related, and closely related) and in 6 semantic type categories of term pairs (DISORDER-DISORDER, DISORDER-SYMPTOM, DISORDER-DRUG, SYMPTOM-SYMPTOM, SYMPTOM-DRUG, DRUG-DRUG). This resulted in a dataset illustrated in Figure 1[i].

The order of presentation of the term pairs and the order of the terms in each pair were randomized. With terms denoting medications, we used brand names in most cases because generic names for drugs with similar chemical composition and/or function tend to have similar orthography and pronunciation, presenting a potential source of bias.

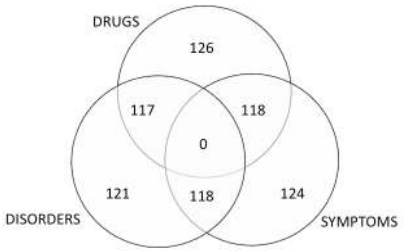

Figure 1 Data set of 724 medical term pairs and their counts by category.

Instruments and procedures: Subjects were seated in a quiet room two feet away from a 22" computer monitor and asked to make relatedness and similarity judgments for each pair of medical terms that were shown in the center of the monitor (HP TouchSmart IQ506). Subjects were instructed to respond by touching the screen to indicate how similar the two terms are, on a scale from left (low similarity/relatedness) to right (high similarity/relatedness). We imposed a 4 seconds time limit in order to elicit an intuitive reaction. This was particularly important for relatedness judgments, as we wanted to prevent the subjects from examining unlimited chains of relationships between concepts and thus possibly biasing the results towards greater degree of relatedness than justified by their initial reaction. To minimize practice effects, subjects were given a practice session consisting of 30 term pairs that were not included in the main dataset.

Each subject performed only one of the two tasks (similarity or relatedness) resulting in two groups of 4 subjects (3 men and 1 woman in each). The tasks were explained to the subjects by providing examples of the two phenomena (e.g., "pulmonary edema" and "heart failure" are related but not similar). We relied on examples rather than detailed rules, as we were interested in eliciting the subjects' intuitive responses.We recorded the X and Y coordinates (in pixels offset from the left edge of the screen (range: 0-1600)) of the location where the subject pressed the screen, as well as the response latency measured in milliseconds lapsed from the stimulus presentation. The X coordinate was used as a measure of relatedness/similarity with lower values indicating less related/similar judgments.

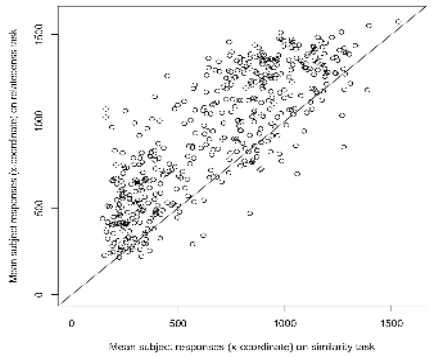

Figure 2 Scatter plot illustrating the correlation between similarity and relatedness tasks (averaged across all subjects)

Computerized measures: We expected to find a number of disagreements between the raters and hypothesized that disagreements may be associated with the very notions of the strength of semantic relatedness and similarity. To test this hypothesis we used measures and relatedness that were derived independently from the raters' judgments. The measure of similarity consisted of a simple path-based approach[13] where the degree of similarity between concepts is a function of the path length between them in the UMLS. The measure of relatedness comprised a vector-space model approach[10] that represents each term as a second-order vector of frequencies of words in the term's definition with frequencies computed over a large corpus of medical text. We used a relatedness measure derived from a corpus of ~500,000 inpatient reports (admission, progress and discharge notes) from the University of Minnesota Fairview system.

Statistical analyses: To measure inter-rater reliability on continuous scale responses, we used the Intra-Class Correlation Coefficient (ICC) as defined by Shrout and Fleiss.[14] Due to the apparent multimodal distribution of the raters' responses, we could

[i]Freely available from http://www.rxinformatics.umn.edu

not meet the normality assumption and therefore relied on non-parametric Spearman's rank correlation to test for linear associations between variables in this study. Variables that indicated non-linear relationship were analyzed using polynomial models. All statistical analyses were conducted using R (v 2.9.1) package.

Results

Due to the time limit placed on making the similarity and relatedness judgments, we expected that the raters would not be able to complete the assessment of some of the term pairs. On the relatedness task, all raters succeeded on 587 (81%) of 724 samples. On the similarity task, 566 (78%) of 724 pairs were successfully completed by all. On the similarity task, the raters failed more on the DRUG-DRUG category (22% of all failures), followed by SYMPTOM-SYMPTOM category (20% of all failures). On the relatedness task, the most failures were observed on the SYMPTOM-DRUG category (21%) followed by the DISORDER-DRUG category (20%).

Similarity vs. Relatedness: On both tasks, all eight raters succeeded to respond on 457 (63%) of 724 pairs. The responses on the relatedness and similarity tasks for these 457 pairs were highly correlated ($r=0.80$, $p < 0.0001$) indicating a strong relationship between similarity and relatedness. The plot in Figure 2 shows that most of the term pairs that were judged as dissimilar were also judged as unrelated and vice versa. A number of term pairs were also judged as dissimilar but somewhat related (upper left corner of the plot), whereas no pairs were judged as similar and also marked as unrelated (lower right corner). The lower success rate on the similarity task was likely due to the difficulty in assessing similarity across different semantic types evident from lower agreement on these judgments in Table 1.

	DIS-DIS	SYM-SYM	DR-DR	DIS-DR	SYM-DR	DIS-SYM
Rel. ICC	0.56	0.49	0.47	0.49	0.45	0.50
Sim. ICC	0.56	0.58	0.63	0.33	0.24	0.34

Table 1 Inter-rater agreement on medical term pairs separated by the semantic types of the terms in the pair

Inter-rater agreement: The agreement on the 587 relatedness and 566 similarity pairs on which responses from all four raters were obtained was in the moderate range (ICC=0.50 and 0.47 respectively). Table 1 illustrates inter-rater agreement separately for each semantic type of the terms in the pairs.

To determine if the disagreements were uniformly distributed throughout the dataset or limited to a specific subset, we used the standard deviation of the responses provided by the four raters to reduce the set of pairs. We determined the number and the distribution of the pairs with good agreement (ICC > 0.7) across the relatedness and similarity continua showing that good agreement was reached on over 70% of the term pairs. However, in order to use this subset as a reference standard it is important to ensure that the distribution of the pairs across the relatedness and similarity continua after the reduction remains similar to the distribution prior to the reduction.

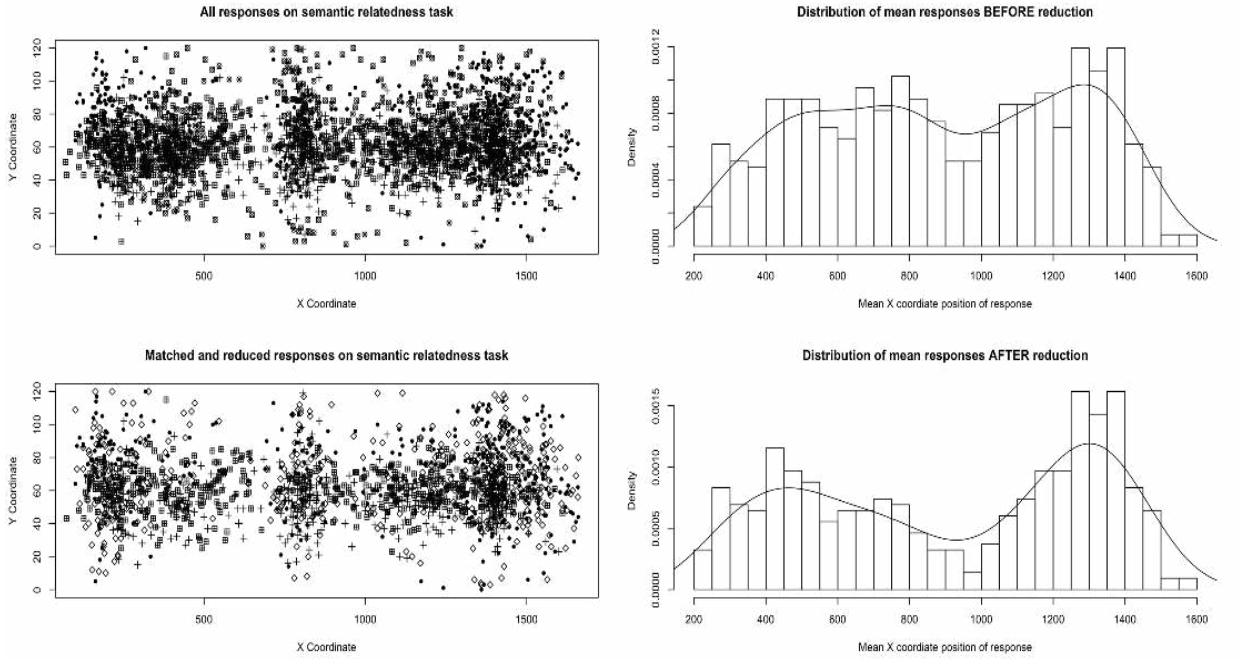

Figure 3: Distribution of responses on the relatedness task BEFORE and AFTER reducing the dataset to samples with good agreement

Results of the comparisons between the two distributions are shown in Figures 3 and 4. Panels on the left in Figures 3 and 4 show the locations where the raters touched the screen in response to the term pair stimuli. The panels on the right show histograms and probability density distributions for the responses depicted in the plots on the left.

The plots in Figure 3 show that the raters' responses are clustered into 3 groups roughly corresponding to "unrelated", "related", and "somewhat related" categories. The removal of pairs with large standard deviations in responses resulted in slightly reduced density of the "somewhat related" category. This indicates that the majority of the disagreements came from the "somewhat related" category; however, the overall shape of the distribution across these categories remains similar before and after the reduction with a sufficiently large proportion of samples in the space between the "unrelated" and "related" clusters.

The distribution of responses on the similarity task (Figure 4) shows a more diffuse pattern in the space between the "dissimilar" and "similar" end of the scale. Similar to the relatedness task, removal of pairs on which the raters disagreed the most decreased the number of "somewhat similar" pairs without substantially changing the overall shape of the distribution.

Response Latency: The latency of raters' responses was distributed in a U-shaped pattern with faster responses on unrelated/dissimilar and related/similar pairs and slower responses in between these two extremes (Figure 5). The plots in Figure 5 indicate that the association between the response latency and relatedness judgments is relatively strong (r=0.42), but is much weaker (r=0.26) on the similarity task.

Analysis of Disagreements: First, as shown in Table 1, the semantic type of the terms in the pairs evaluated by the raters clearly plays a role. Not surprisingly, the agreement was the lowest on judgments of similarity between disorders and drugs and between symptoms and drugs. If the notion of semantic similarity between two concepts relies on comparing sets of features that define the concepts, then one would expect the raters to have a harder time comparing abstract concepts such as disorders and symptoms with concrete chemical substances. By the same token, one would expect similarity judgments for pairs of such concrete concepts to be easier to make than for pairs of disorders and symptoms, which is consistent with the higher ICC (0.63) for DRUG-DRUG pairs than for other categories in Table 1.

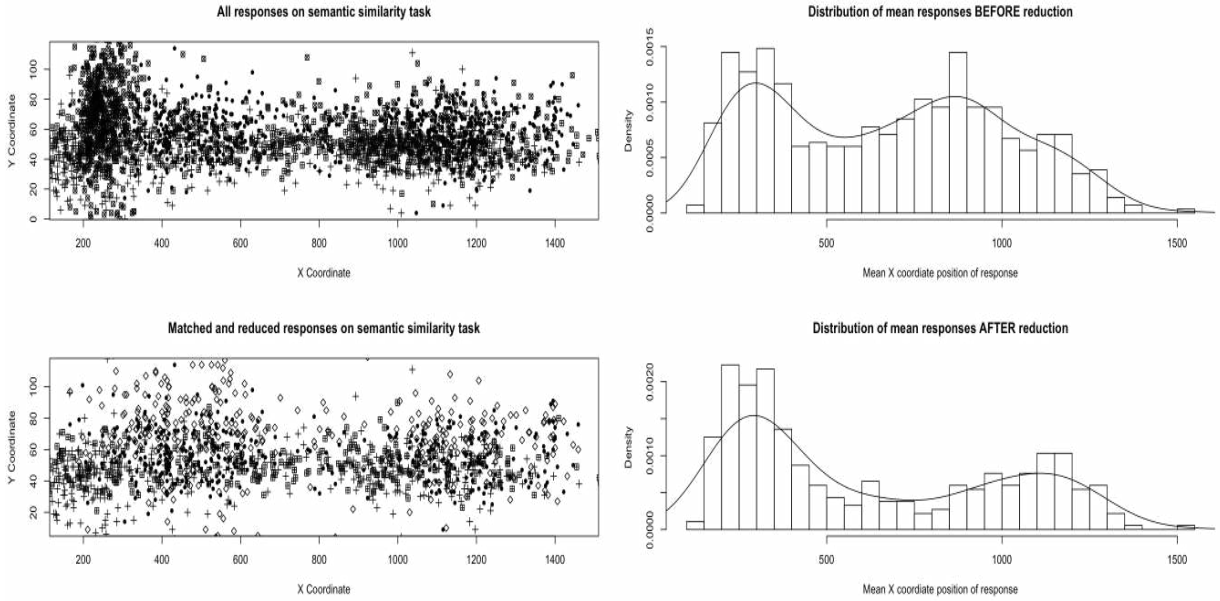

Figure 4: Distribution of responses on the similarity task BEFORE and AFTER reducing the dataset to samples with good agreement

Testing for correlations with path-based and vector-based similarity and relatedness scores resulted in a weak but statistically significant negative correlation between the vector-based relatedness score ($r = -0.13$, $p = 0.004$) and the standard deviation in the raters' relatedness judgments, but not their similarity judgments ($r = 0.02$, $p = 0.604$). We also found a weak but significant correlation between the path-based measure of similarity and the standard deviations in rater's judgments on the similarity task ($r = 0.14$, $p = 0.002$) as well as the relatedness task ($r = 0.1$, $p = 0.027$). These correlations explain only a small part of the disagreements; however, the fact that there is a difference between the relatedness and similarity tasks for the vector-based approach but not for the path-based approach is consistent with the asymmetrical relation-

ship between similarity and relatedness judgments evident in Figure 2.

Discussion

Our results demonstrate that the relationship between semantic similarity and relatedness is that of unidirectional entailment – pairs of terms that are similar are also likely to be related but not vice versa. We also found a tendency for higher relatedness relative to similarity scores for any given pair of terms (see Figure 2). Our results also suggest that the semantic associations between clinical terms are largely independent of the possible context. The distribution of response latencies provided additional evidence that raters were performing as expected.

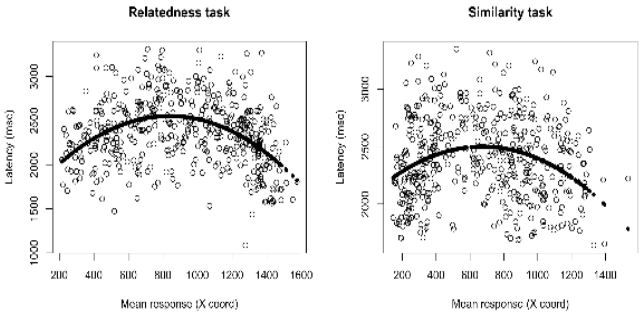

Figure 5 Quadratic models for latency of responses on relatedness (r= 0.42, p<0.0001) and similarity (r=0.26, p < 0.001) task.

We did find a difference, however, between the relatedness and similarity tasks with respect to the degree of correlation. This difference is likely due to similar reasons to the finding that agreement on some of the semantic types within the similarity task was much lower than on the relatedness task (see Table 1). Manual examination of disagreements (sd. > 400 pixels) showed that 48% of them were due to one of the 4 raters providing a response different from the other three raters with responses similar to each other. Thus, we believe that a greater proportion of the data may be used as a reference standard if the responses are combined using "majority vote" rather than the mean.

The main difference between the our approach to generating the relatedness/similarity dataset and some of the other previously reported approaches[15, 16] is the use of a continuous scale as well as having a time limit on the judgments. Thus our approach aims to elicit implicit relations between concepts in the minds of the raters.

Conclusion

This study furthers our understanding of semantic similarity and relatedness between biomedical terms and will enable the development of automated approaches to their measurement. Having distinct ratings of similarity and relatedness on the same set of term pairs is particularly important to enable testing and comparison of path-based and vector-based automated measures. Using the semantic similarity dataset is more appropriate for the former, whereas the relatedness dataset is better suited for the latter based on how these two types of relationships are defined.

Acknowledgements

This work was supported by the National Library of Medicine (R01 LM009623-01).

References

1. Collins AM, Loftus EF. Spreading Activation Theory of Semantic Processing. Psychol. Review 1975;82:407-28.

2. Ferrand L, Ric F, Augustinova M. Affective priming: A case of semantic priming? Annee Psychologique 2006;106:79-104.

3. Thompson-Schill SL, Kurtz KJ, Gabrieli JDE. Effects of semantic and associative relatedness on automatic priming. Journal of Memory and Language 1998;38:440-58.

4. Tversky A. Features of Similarity. Psychological Review 1977;84:327-52.

5. Weber M, Thompson-Schill SL, Osherson D, Haxby J, Parsons L. Predicting judged similarity of natural categories from their neural representations. Neuropsychologia 2009;47:859-68.

6. Mitchell TM, Shinkareva SV, Carlson A, et al. Predicting human brain activity associated with the meanings of nouns. Science 2008;320:1191-5.

7. Bousquet C, Jaulent M, Chantellier G, Degoulet P. Using Semantic Distance for the Efficient Coding of Medical Concepts. In Proc. AMIA Symposium 2000:96-100.

8. Bousquet C, Lagier G, Lillo-Le LA, Le Beller C, Venot A, Jaulent MC. Appraisal of the MedDRA Conceptual Structure for Describing and Grouping Adverse Drug Reactions. Drug Safety 2005;28:19-34.

9. Caviedes J, Cimino J. Towards the Development of a Conceptual Distance Metric for the UMLS. J Biomed Inform 2004;37:77-85.

10. Pedersen T, Pakhomov SV, Patwardhan S. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. J of Biomed Inform 2006;40:288-99.

11. Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. Conf Proc IEEE Eng Med Biol Soc 2006;1:2713-7.

12. Lee WN, Shah N, Sundlass K, Musen M. Comparison of ontology-based semantic-similarity measures. In Proc. AMIA Symposium 2008:384-8.

13. McInnes B, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In Proc. AMIA Symposium San Francisco, CA; 2009.

14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-8.

15. Rubenstein H, Goodenough J. Contextual Correlates of Synonymy. Communications of the ACM 1965;8:627-33.

16. Miller G, Charles W. Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 1991;6:1-28.