Journal of Biomedical Informatics 44 (2011) 251–265

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Towards a framework for developing semantic relatedness reference standards

Serguei V.S. Pakhomov [a,c,*], Ted Pedersen [b], Bridget McInnes [a], Genevieve B. Melton [c], Alexander Ruggieri [e], Christopher G. Chute [d]

[a] College of Pharmacy, University of Minnesota, Twin Cities, MN, USA
[b] Computer Science, University of Minnesota, Duluth, MN, USA
[c] Institute for Health Informatics, University of Minnesota, Twin Cities, MN, USA
[d] Biomedical Informatics, Mayo College of Medicine, Rochester MN, USA
[e] WellPoint, Inc., Woodland Hills, CA, USA

ARTICLE INFO

Article history:
Received 19 February 2010
Available online 31 October 2010

Keywords:
Semantic relatedness
Reference standards
Reliability
Inter-annotator agreement

ABSTRACT

Our objective is to develop a framework for creating reference standards for functional testing of computerized measures of semantic relatedness. Currently, research on computerized approaches to semantic relatedness between biomedical concepts relies on reference standards created for specific purposes using a variety of methods for their analysis. In most cases, these reference standards are not publicly available and the published information provided in manuscripts that evaluate computerized semantic relatedness measurement approaches is not sufficient to reproduce the results. Our proposed framework is based on the experiences of medical informatics and computational linguistics communities and addresses practical and theoretical issues with creating reference standards for semantic relatedness. We demonstrate the use of the framework on a pilot set of 101 medical term pairs rated for semantic relatedness by 13 medical coding experts. While the reliability of this particular reference standard is in the "moderate" range; we show that using clustering and factor analyses offers a data-driven approach to finding systematic differences among raters and identifying groups of potential outliers. We test two ontology-based measures of relatedness and provide both the reference standard containing individual ratings and the R program used to analyze the ratings as open-source. Currently, these resources are intended to be used to reproduce and compare results of studies involving computerized measures of semantic relatedness. Our framework may be extended to the development of reference standards in other research areas in medical informatics including automatic classification, information retrieval from medical records and vocabulary/ontology development.

1. Introduction

Querying electronic health records (EHR) for patients with a particular syndrome often requires using a variety of medical terms that not only denote the diagnosis itself but also symptoms and conditions closely related to the syndrome. This task is especially important for effective clinical trial recruitment and quality assurance functions where identifying these patients can assist with the process of care. For example, when searching for cases of heart failure, it is desirable to include terms for "pulmonary edema", "volume/fluid overload", "shortness of breath," and possibly others to achieve the maximum sensitivity in finding relevant medical records [1]. The terms "heart failure", "pulmonary edema", "volume/fluid overload" and "shortness of breath" are clearly not synonymous but are semantically related because they denote different aspects of the same underlying condition. The terms "cardiomegaly" and "splenomegaly" may be considered semantically similar, but not necessarily related, because they share a semantic feature – both refer to an enlargement of an internal body organ. These examples illustrate that semantic relatedness is a more general notion that subsumes semantic similarity and synonymy as special cases of semantic relatedness. This hierarchical relationship between similarity and relatedness is a theoretically well established notion in lexical semantics work on general English [2,3] that we were able to confirm experimentally in a recent psycholinguistic study in the clinical sublanguage domain. In this study, two groups of physicians were asked to rate the same 724 pairs of clinical concepts (drugs, disorders, symptoms) for the degree of either relatedness or similarity. The results of correlating the two groups' judgments clearly showed a relationship of unidirectional entailment – the pairs of terms that were judged as similar were also judged as related but not vice versa [4].

Apart from searching EHR systems for patients with related conditions, clusters of semantically related terms are also currently

* Corresponding author. Address: 7-125F Weaver-Densford Hall, 308 Harvard Street S.E., Minneapolis, MN 55455, USA. Fax: +1 612 625 9931.
E-mail address: pakh0002@umn.edu (S.V.S. Pakhomov).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.10.004

used for post-marketing drug safety surveillance. In this context, drug safety reports comprising FDA Adverse Events Reporting System (AERS) may be searched using queries consisting of groups of terms (Standardized MedDRA Queries) that are indicative of the adverse event of interest [5]. For example, the adverse event of "demyelination" is represented by a Standardized MedDRA Query (SMQ) consisting of a total of 35 other terms including "optic neuritis", "multiple sclerosis" and "saccadic eye movement." A recent study found that using SMQs for automated identification of potential adverse drug event signals in AERS yields more sensitive (albeit less specific) results than using ungrouped terms [6]. Developing tools to support querying of clinical data for drug safety surveillance is particularly important in light of the recent FDA Sentinel initiative[1] launched in May 2008 and designed to develop a national electronic system to utilize electronic health records for safety monitoring of FDA regulated medications. Electronic health records include large amounts of highly variable free text in addition to structured data and their use for drug safety surveillance will require aggregation of semantically related concepts to reduce this variability.

Currently, several research groups, including ours, are investigating computerized methods for determining the strength of similarity and relatedness between medical terms and for grouping the terms based on the strength of relatedness [7–12]. One of the critical prerequisites in these investigations is the availability of publicly accessible, validated reference standards that may be used to assess the performance of automated algorithms relative to human judgments. In previous work, we have used and made publicly available a reference standard of 30 medical term pairs [10]. These medical term pairs represent a subset of a larger set of 120 term pairs and were selected because the majority of the human annotators agreed on the strength of relatedness between the terms. In the current paper, we build on our previous work by providing a detailed analysis of the entire dataset and describing a methodology for analyzing human semantic relatedness ratings.

The main objective of this paper is to propose a framework for creating reference standards for evaluating computerized measures of semantic relatedness. Towards this objective, we validate a reference standard of medical concept pairs manually rated by professional medical coding experts for the degree of semantic relatedness between terms. We evaluate the reliability of human ratings and, based on these results, provide suggestions for future public use of this dataset to test computerized measures of semantic relatedness. We also provide both the reference standards and the statistical software for their analysis as open-source. Furthermore, we demonstrate the use of the proposed framework to evaluate two automatic computerized approaches to measuring semantic relatedness between biomedical terms.

2. Background

2.1. Semantic relatedness in general English language text

Creating computer programs that can automatically judge if one pair of concepts is more closely related than another is still an unmet goal of natural language processing (NLP) and Artificial Intelligence (AI). Most existing methods base relatedness judgments on knowledge sources such as concept hierarchies or ontologies, which may be augmented with statistics from text corpora. A number of proposed solutions to NLP problems rely on these measures. For example, Budanitsky and Hirst [13] identify when a correctly spelled word is used in the wrong context (i.e., malapropisms) using various measures of relatedness. Both Resnik [14] and Patwardhan et al. [15] have shown that these measures can be used to assign the meaning of a word in context, and perform word sense disambiguation.

Most research on measuring semantic relatedness in general English text has relied on WordNet [16], a freely available dictionary that can also be viewed as a semantic network. WordNet groups related words into synonym sets or *synsets* that represent a particular concept. Each synset has a definition or gloss that characterizes its meaning and is connected to other synsets via links that represent relations such as *is-a, has-part*, and *is-a-way-of-doing*. WordNet is best known for its noun hierarchies, which are made up of *is-a* relations. The most recent version (3.0) includes a noun hierarchy of 82,000 concepts. By comparison, there are 14,000 verb concepts arranged in more than 600 *is-a-way-of-doing* hierarchies (e.g., *to run* is a way of *moving*) and networks of 19,000 adjectives and 4000 adverbs. An evaluation of the automated measures of semantic relatedness in general English has been successfully conducted based on a corpus of 30 English words found in WordNet manually rated by over 50 native English speakers for semantic similarity with high correlation [17].

2.2. Semantic relatedness in the biomedical sublanguage domain

The discipline of biomedical informatics has generated a significant amount of experience in terminology and ontology resource development along with medical knowledge representation. The Unified Medical Language System (UMLS) developed and maintained by the National Library of Medicine represents the most comprehensive effort to date in unifying over 150 disparate terminologies and ontologies into a single system. The UMLS has a significantly wider and deeper coverage of the biomedical domain with a significantly richer set of relations between medical concepts than WordNet [18,19]. Current research on measures of relatedness for biomedical text includes adaptations of existing WordNet-based measures or variations specific to tasks such as MEDLINE document retrieval or Gene Ontology (GO) searching. For example, Lord et al. [20] adapted three Information Content measures of similarity based on WordNet that were discussed previously (Lin, Resnik and Jiang-Conrath) to the GO ontology. Guo et al. [21] demonstrated the utility of Information Content measures of semantic similarity derived from the Gene Ontology for identifying direct and indirect protein interactions within human regulatory pathways. In earlier work, Rada et al. [22] notably devised a measure based on path lengths between concepts found in MeSH. Rada used this measure to improve information retrieval by ranking documents retrieved from MEDLINE.

Several other examples of the development of semantic relatedness approaches can be found in the biomedical informatics literature. Bousquet et al. [7,8] explored the use of semantic distance (the inverse of similarity) for coding of medical diagnoses and adverse drug reactions. In contrast, Rodriguez and Egenhofer [23] used semantic similarity to integrate various ontologies into a unified resource for subsequent use in information retrieval from medical documents. To investigate the feasibility of semantic similarity metrics based on the UMLS framework, Caviedes and Cimino [9] developed a measure called CDist based on the shortest path between two concepts and demonstrated that even such relatively simple approaches tend to yield reliable results. Work by Wilbur and Yang [24] defined a strength metric used to retrieve relevant articles using lexical techniques. The described metric uses the correlation between the occurrences of a term in documents with the subjects of the documents to define the weight of each term. Research by Spasic and Ananiadou [25] defined a new similarity measure based on a variation of edit distance applied at the word level. In short, semantic similarity between two terms is the cost associated with converting one term to another, using insert, delete and

[1] http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm.

replace operations on words (instead of letters). This method additionally uses the UMLS taxonomy to minimize the effect of word variants. It also varies the costs associated with the operations based on the "semantic load" of the word being edited. For example, deleting a known term present in the UMLS has a higher cost than deleting a conjunction.

A hybrid approach that incorporates measures of semantic similarity based on hierarchical relations with measures of relatedness based on the vector-space model has also been developed and demonstrated for computing semantic similarity between genes and gene products [26]. More recently, Al-Mubaid and Nguyen developed a measure of similarity between biomedical concepts in the Medical Subjects Heading (MeSH) taxonomy, based on a combination of existing ontology-only based approaches and two additional characteristics – the specificity and the branching factor of the hierarchies of the concepts being compared [11]. In a subsequent study, Lee et al. compared Al-Mubaid and Nguyen's ontology-only and Information Content based approaches [12]. We have also adapted a number of ontology-based measures that were originally developed for WordNet to use the UMLS as the source of hierarchical relations information [27]. Our prior experiments with a number of ontology-based and corpus-based measures of semantic relatedness found preliminary evidence indicating that physicians' assessments of semantic relatedness correlated better with corpus-based measures derived from dictated clinical reports, while medical coding experts' assessments of relatedness correlated better with ontology-based measures [10].

Apart from the work described above on developing the semantic relatedness measures themselves, several efforts have been directed towards using automatically computed semantic relatedness in information extraction [28] and textual inference [29]. Bodenreider and Burgun [30] used the notions of lexical and conceptual similarity to align the UMLS Metathesaurus and the UMLS Semantic Network. Low-level measures of semantic relatedness between individual medical concepts have also been used to determine the similarity between patient cases based on the text of the patient's EHR [31] and similarity between biomedical articles for subsequent indexing [32] and document clustering [33]. In another subsequent study of patient similarity based on EHR, the approach to computing patient case similarity was enhanced by adding specific features such as diagnoses, findings and procedures abstracted from the medical record [34].

2.3. Reference standards to evaluate semantic relatedness measures

Current efforts to develop measures of semantic similarity and relatedness in medical informatics have traditionally relied on reference standards created specifically for each individual study using various scales, human annotators and annotation instructions. The existing reference standard generation approaches are thus limited in a number of important aspects. In addition to limited generalizability due to the utilization of different scales, annotators and guidelines, reference standards created to evaluate the tools in individual studies of semantic relatedness/similarity are rated only by 2–3 human raters. Having larger numbers of raters is particularly advantageous in domains such as medical terminology where inter-rater agreement on semantic relatedness tends to be low as we show in the current study. By comparison, reference standards published on the general English domain consist of ratings by over 50 participants [17,35], which constitutes a more representative sample and enables techniques aimed at reducing inter-rater variability (e.g., majority voting). Studies that use or create reference standards for semantic relatedness published in biomedical literature typically do not provide in-depth information on the methods and characteristics of the reference standards and thus are difficult to replicate. In addition to differences in reference standard creation methods, investigators may use a variety of statistical methods to assess the reliability of their standards, thus limiting the comparability of the results obtained in different studies. As demonstrated by Krippendorf, inferring the reliability of a reference standard from a specific inter-annotator agreement coefficient depends on the purpose for which the reference standard is used [36]. Individual researchers report the agreement statistics that are appropriate for the purposes of their study that may or may not be appropriate if one were to use the reference standard under different conditions and for a different purpose. Having publicly available reference standards requires a variety of purpose-sensitive methods to assess their reliability.

Furthermore, the majority of the reference standards reported in the biomedical literature are not publicly available. The importance of releasing software used by researchers to obtain and publish their results so that they can be reproduced and improved upon by others is abundantly clear [37]. An equally important counterpart to publicly available software are the numerous reference standards that are also sometimes referred to as "test beds", "gold standards", "test sets" and "held-out sets." Having reliable and publicly available reference standards in addition to software for semantic relatedness research is critical to enabling reproducible and easily comparable results. In the current study, we address these issues with the availability of validated reference standards for semantic relatedness in medical informatics by proposing and piloting a standardized framework for creating, validating and disseminating reference corpora for the assessment of computerized semantic relatedness measures. We demonstrate our approach on a set of heterogeneous medical term pairs rated by 13 medical coding experts and two computerized automated measures.

3. Methods

3.1. Datasets

While random sampling is a widely used technique to generate reference standards, we did not want to rely on this methodology to create medical term pairs because random sampling was likely to result in a disproportionate number of unrelated pairs. To have a more balanced distribution across the relatedness spectrum, we asked a practicing Mayo Clinic physician (AR, a rheumatology specialist also formally trained in health informatics) to generate a list of 120 pairs of medical terms that would roughly correspond to four categories: closely related, somewhat related, somewhat unrelated and completely unrelated. The physician was instructed to rely on his intuition in selecting pairs without having to define explicitly the nature of the relationship between the terms. The original list of 120 pairs was further revised to remove duplicates and items that the physician felt unsure about. The resulting set consisted of 101 pairs.

The corpus of 101 pairs was subsequently manually rated on a scale of 1–10 (1-closely related, 10-unrelated) by 13 medical coding experts. All experts were at the time of the study a part of the Mayo Medical Index group that continues to support multiple epidemiologic studies at the Mayo Clinic and beyond, including the Rochester Epidemiology Project, a legacy of Henry Plummer [38]. These coding experts were previously trained to classify the diagnoses contained in the Mayo Clinic medical records using a Hospital Adaptation of the International Classification of Diseases [39]. They had varying degrees of experience ranging from 2 years to over 15. All of the experts participating in this study were female. Subsequent to evaluating the corpus of 101 pairs, 9 of the 13 experts[2] were also asked to rate 30 pairs of general English words in

[2] The remaining four coders were not available for reasons unrelated to the study.

the Miller and Charles corpus [17]. All experts were instructed to rate the term pairs for semantic relatedness rather than similarity for both biomedical and general English terms. To focus the ratings on relatedness, the raters were provided with a number of examples; however, the raters were not formally trained to distinguish cases of similarity from relatedness.

3.2. UMLS-similarity package

We recently implemented several ontology-based approaches to computing semantic similarity based on the relational information contained in the UMLS (version 2008 AB) [27]. For this study, we correlated the automatic ratings produced by two ontology-based approaches originally developed by Wu and Palmer [40], and Leacock and Chodorow [40] in order to demonstrate the intended use of our framework. Both of these approaches rely on computing the length of the paths in the UMLS hierarchies between concepts and thus represent measures of similarity, a special case of semantic relatedness.

3.3. Statistical analysis

Within our framework, we propose two types of statistical analysis – inter-rater reliability with subsequent exploration of the internal structure that may be present in the use of the scales by the raters. In the spirit of generating publicly available research resources, we rely on the open-source R package for all statistical computations and provide the code used for these computations in Appendix C, available online at http://rxinformatics.umn.edu.

3.3.1. Inter-rater reliability

We report several inter-annotator agreement coefficients. First, we computed a number of pair-wise (between raters) inter-annotator agreement coefficients including Cohen's kappa [41], Krippendorff's alpha [42] and mean Spearman's rank correlation (rho) appropriate for ordinal scales. To account for the fact that our data are ordinal, we used the squared distance weighting scheme for Cohen's kappa whereby disagreements are weighted according to their squared distance from perfect agreement. Second, to assess more than two raters at a time, we used Cronbach's alpha [43], Kendall's coefficient of concordance [44] and the Intra-class Correlation Coefficient (ICC) [45]. Shrout and Fleiss define six types of ICC depending on whether: (a) one-way random effects, two-way random effects or two-way mixed effects ANOVA model is used, (b) ICC is used to measure consistency of absolute agreement, and (c) ICC is computed over single or averaged measures. Since our study focuses on the raters as much as on the term pairs, we used a two-way random effects ANOVA model in ICC computations (i.e., ICC(2,k) in Shrout and Fleiss's notation). Following McGraw and Wong [46], we report on consistency measures rather than absolute agreement because we expect to see some systematic variation in raters' judgments and we are more interested in determining whether the raters' judgments consistently point in the same direction even if they are using slightly different scales. We report both single and average ICC measures – the former is useful in assessing the reproducibility of the reference standard based on individual raters' judgments, while the latter is useful in assessing the reliability of the means of the ratings by multiple raters. In Shrout and Fleiss's notation, the single measures ICC based on a two-way model is represented by ICC(2,1), while the average measures ICC based on a two-way model is represented by ICC(2,k), where k is the number of raters. We should also note that the latter type of ICC(2,k) for consistency is equivalent to Cronbach's alpha [46].

Following Krippendorff's [36] conditions for measuring inter-rater reliability, we do not infer reliability of our reference standard based on correlation (Spearman's rho) and consistency (Chronbach's alpha, Kendall's coefficient of concordance) coefficients as these measures do not correct for random chance. We use these coefficients only as indicators of the raters' behavior rather than reproducibility of their ratings. Instead, we rely on the Krippendorf's alpha, Cohen's kappa and ICC as indicators of the reference standard reliability. While Cohen's kappa is defined only for two raters, Fleiss's kappa [45] is defined for multiple raters but requires nominal/categorical rating scales and is thus not appropriate for assessing semantic relatedness judgments made on an ordinal scale. However, for nominal/categorical scales, Fleiss's kappa has been shown to be equivalent to ICC [45].

3.3.2. Internal structure of the ratings

We conducted subgroup analyses on the group of 13 experts to determine if human ratings had any latent internal structure. For example, it is likely that there may be subgroups within the 13 experts whose ratings are similar within the subgroups but not across the subgroups. To conduct these analyses we used both a top-down partitioning (k-means) and a hierarchical agglomerative clustering based on the Ward's method of minimum variances [47] available as part of the *irr* library for the R statistical package available from Comprehensive R Archive Network (CRAN) repository.[3] To determine the number of clusters for the k-means clustering approach, we used a sense discrimination approach based on the open-source SenseClusters package[4] as well as the sum-of-squares approach [48] available as part of R's *irr* library.

SenseClusters contains algorithms for four cluster stopping rules (PK1, PK2, PK3 and Gap) described in detail elsewhere [49]. Briefly, the cluster stopping rules are based on the clustering criterion function. PK1–3 algorithms attempt to determine a point in the list of successive cluster criterion function values after which the values stop improving significantly. The PK 2 method is similar to the Hartigan's sum-of-squares approach. The Gap measure is an adaptation of the Gap Statistic [50], which relies on detecting the greatest difference between the criterion function values and a null reference distribution. In the current study, the determination of the optimal number of clusters was made by averaging the number of clusters predicted by each of the four cluster stopping rules. The sum-of-squares method consists of computing the within-cluster sums-of-squares over all variables. The sum-of-squares declines as more clusters are added to the solution and can be plotted for visual examination to determine a sharp decline to estimate the optimal number of clusters supported by the data.

Clustering analyses were followed by a factor analysis based on the principal components analysis (PCA). Similarly to the sum-of-squares method, we used Scree plots to determine the number of factors. The loadings on factors were rotated using Varimax rotation to obtain the final PCA solutions.

3.3.3. Correlations between automatic and manual measures

We used non-parametric correlation methods including Spearman rank correlation and Kendal's tau to measure the degree to which automated measures of semantic relatedness represent manually established relatedness judgments.

3.4. Mapping to the UMLS

Since the measures of semantic relatedness used in this study rely on the information on the location of the concepts in an ontology or a hierarchical vocabulary, it was necessary to map the terms in our medical term pairs dataset to an ontology of medical con-

[3] http://cran.r-project.org/web/packages/irr/irr.pdf.
[4] http://www.d.umn.edu/~tpederse/seneclusters.html.

cepts. Our reference standard was initially generated by a physician without any reference to an ontology. Although in previous work we have used SNOMED CT, the UMLS was used in this study as the source of ontological relationships between concepts, as its overall coverage of concepts includes over 2 million biomedical concepts, and relationships between concepts (as defined in the MRREL file). SNOMED CT is one of the vocabularies in the UMLS.

The mapping process could not be fully automated due to significant orthographic, syntactic and semantic variation of terms in the dataset. For this study, we use a semi-automatic approach to determine the appropriate concept for each of the terms using the concept mapping system MetaMap [51] which is specifically designed to map terms in biomedical text to concepts in the UMLS to support indexing of MEDLINE citations. MetaMap is a freely available natural language processing system developed by the National Library of Medicine. It operates by identifying simple noun, verb and prepositional phrases with the help of a minimal commitment parser [52] and bringing lexical and morphological variants of medical terms to a standard form. It also uses linguistic principles to map the different types of phrases to the UMLS Metathesaurus, a compendium of over 100 medical vocabularies. The system has found multiple applications beyond MEDLINE indexing and has been shown to perform "out-of-the-box" at 72% sensitivity and 56% specificity on the task of identifying respiratory findings from hospital discharge summaries. The error analysis of these results showed that MetaMap was responsible for less than 1% of the errors. Most of the errors had to do with inherent problems with manual annotation of the reference standard [53].

In our study, mappings suggested by MetaMap were subsequently manually verified to ensure that the most appropriate concept unique identifier was selected to represent each of the terms. The terms for which MetaMap was unable to find any suitable mappings, were matched to the closest UMLS concept manually using the UMLS Knowledge Server on-line interface.

4. Results

4.1. Descriptive statistics of the datasets

The dataset consisting of 101 medical term pairs and individual ratings is shown in Table 1. The distributions of the mean ratings for the datasets consisting of 101 medical pairs and 30 general English word pairs are shown in Fig. 1. The ratings are not uniformly distributed for either of the datasets. For the medical term pairs, a greater proportion of ratings on average are found in the "related" (lower values) than the "unrelated" end of the scale. The distribution for the general English word pairs is bimodal suggesting that the raters tended to make binary decisions.

4.2. Inter-rater reliability analysis

4.2.1. Pair-wise comparisons

The results of pair-wise comparisons between raters on both datasets are shown in Tables 2 and 3. Each cell in these tables displays three values corresponding to different coefficients of inter-rater agreement (Spearman's rho, Cohen's weighted kappa and Krippendorff's alpha for ordinal scales). In Table 2, all three values for raters 5 and 7, as compared to most of the other raters, are low (rho < 0.5, Kappa < 0.3 and alpha < 0.2) with the exception of raters 2, 3, and 11. These values are also somewhat higher (rho = 0.46, Kappa < 0.54 and alpha < 0.43) in the cells that represents the agreement between these two raters. This indicates that raters 2, 3, 5, 7 and 11 agree with each other but not the other raters. Rater 13 is borderline and has a mix of large and small coefficients. These results indicate the presence of subgroups in the data.

The distribution of coefficients is more homogeneous for the general English pairs than for medical term pairs (Table 3) with the majority of raters having high coefficients in pair-wise comparisons with the exception of rater 9. The latter has markedly lower coefficients in most pair-wise comparisons across all three statistical measures.

4.2.2. Multi-rater agreement

The results of multi-rater analysis are summarized in Table 3. The intra-class correlation coefficients for consistency on both single measures (ICC(2,1)) and average measures (ICC(2,13)) were lower for the medical term pairs than for the general English pairs. This was particularly evident on single measures where the difference between ICC's was 0.27. The difference between the datasets based on the Cronbach's alpha coefficient was identical to the difference between ICC's on average measures – 0.04, while the difference between Kendall's coefficients of concordance was similar to the difference between ICC's on single measures – 0.23.

4.3. Analysis of internal structure of ratings

4.3.1. Determining the number of clusters

We used two different methods for determining the number of clusters in the two datasets – sum-of-squares [48] implemented in R and the cluster stopping rules available through the SenseClusters package [49]. The sum-of-squares plots (Fig. 2) for both the medical and the general English datasets indicate a sharp decline in the sum-of-squares with a two cluster solution.

The four SenseClusters cluster stopping rules produced the following results: PK1 predicted four clusters, PK2 predicted three clusters, PK3 predicted two clusters and the Gap predicted one cluster averaging to 2.5 clusters. Based on SenseClusters results and those obtained with the sum-of-squares method, we decided to take a conservative approach and used two clusters for further clustering analysis.

4.3.2. Partitioning clustering

The results of the top-down two-cluster *k*-means clustering solutions based on correlations between raters are displayed in Fig. 3. The solution for the medical terms dataset consists of a cluster with four members (raters 4, 6, 12 and 13) and another cluster with the rest of the raters as members. The solution for general English word pairs singles out rater 9 into a cluster all of his/her own, suggesting that this rater may be an outlier.

4.3.3. Hierarchical agglomerative clustering

The hierarchies resulting from the bottom-up agglomerative clustering based on the Ward's algorithm are shown in Fig. 4. In Fig. 4a showing clustering results for the medical terms dataset, the two branches at the top level separate the raters into a group consisting of raters 4, 6, 12 and 13, and another group consisting of the rest of the raters. In Fig. 4b, the results suggest that rater 9 is an outlier. Furthermore, the results of both clustering solutions in Fig. 4 provide additional grouping information that is not as apparent in the non-hierarchical solutions in Fig. 3. For example, the hierarchical solution for the medical term pairs dataset suggests that experts 1–2 form a group distinct from the group consisting of experts 3, 5, 7 and another group consisting of experts 9, 8 and 11. Similarly, the hierarchy for the general English words dataset suggests that raters 5 and 6 form a distinct group Table 4.

4.3.4. Factor analysis

Further examination using factor analysis confirmed the clustering results as illustrated in Fig. 5a. The plot in Fig. 5a shows a clear separation between two subgroups of raters. Group 1 consists of raters 1–3, 5, 8, 9, 10, and 11, while Group 2 consists of raters 4,

Table 1
Medical term pairs corpus.

CUI1	CUI2	TERM1	TERM2	P1	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	Mean
C0003243	C0443146	Antinuclear antibody (ANA)	Autoimmune	10	8	9	9	1	10	1	6	1	3	1	4	8	7	5.23
C0003873	C0035450	Rheumatoid arthritis	Rheumatoid nodule	10	7	8	9	9	10	3	5	7	8	5	10	3	8	7.08
C0003873	C1396859	Rheumatoid arthritis	Joint erosion	10	7	10	1	8	10	2	6	5	6	7	8	2	8	6.15
C0004093	C0442874	Weakness	Neuropathy	10	5	5	5	1	10	1	2	2	5	5	5	2	5	4.08
C0004604	C0037944	Back pain	Spinal stenosis	10	7	9	5	1	8	2	2	5	8	9	6	2	5	5.31
C0006121	C1269897	Brainstem	Cranial nerve	10	7	5	8	1	8	2	2	3	2	4	5	3	5	4.23
C0009676	C0011253	Confusion	Delusion	10	5	5	7	9	9	5	4	6	3	7	7	5	7	6.08
C0011167	C0031133	Swallowing	Peristalsis	10	5	8	8	1	5	1	2	4	3	9	5	1	5	4.38
C0011168	C0679317	Dysphagia	Hypomotility	10	5	8	4	1	5	5	1	3	2	5	1	2	7	3.77
C0011644	C0036421	Scleroderma	Systemic sclerosis	10	7	8	7	1	8	1	6	5	2	1	10	1	8	5
C0011991	C0009319	Diarrhea	Colitis	10	5	9	7	5	8	1	6	4	4	8	8	6	1	5.54
C0013395	C0030920	Dyspepsia	Peptic ulcer disease	10	5	8	7	4	10	1	3	2	4	8	8	1	7	5.23
C0013404	C0231835	Dyspnea	Tachypnea	10	5	8	3	1	5	5	1	6	7	1	8	2	1	4.08
C0018524	C0033975	Hallucination	Psychotic	10	7	7	7	1	9	3	8	6	5	7	8	3	7	6
C0020877	C0010346	Ileitis	Crohns Disease	10	5	9	9	5	8	8	7	5	7	8	7	3	8	6.85
C0022650	C0041956	Kidney stone	Ureteral obstruction	10	7	8	7	1	10	1	1	4	3	7	8	1	3	4.69
C0026848	C0011633	Myopathy	Dermatomyositis	10	7	7	8	1	3	4	1	3	3	8	6	2	5	4.46
C0027540	Cl547030	Necrosis	Liquefaction	10	7	1	5	1	2	1	1	2	1	1	1	1	1	1.92
C0027627	C0205699	Metastasis	Carcinomatosis	10	9	10	7	10	10	8	9	8	5	7	8	6	10	8.23
C0029408	C0221434	Osteoarthritis	Bone sclerosis	10	5	5	1	1	3	1	3	3	3	3	5	2	2	2.85
C0030472	Cl306459	Paraneoplastic	Malignancy	10	7	9	9	8	4	3	2	4	5	3	5	1	9	5.31
C0032285	C0332448	Pneumonia	Infiltrate	10	7	9	7	3	5	1	5	6	4	3	7	1	5	4.85
C0034065	C0019079	Pulmonary embolus	Hemoptysis	10	7	5	3	1	4	1	1	3	2	1	5	2	1	2.77
C0034735	C0022116	Raynauds phenomenon	Digital ischemia	10	7	8	1	1	8	2	3	6	4	1	9	5	8	4.85
C0038362	C0149745	Stomatitis	Mouth ulcer	10	9	8	8	5	10	5	9	6	2	7	7	3	1	6.85
C0038454	C0018989	Stroke	Hemipareisis	10	5	9	8	1	10	1	8	5	5	1	6	5	1	5
C0231736	C0231749	Drawer sign	Knee pain	10	5	6	3	1	8	1	3	6	5	1	8	1	1	3.77
C0311394	C0231685	Difficulty walking	Antalgic gait	10	7	9	8	5	9	8	2	6	3	10	8	7	5	6.69
C0332536	C0024796	Laxity	Marfan Syndrome	10	5	5	3	1	5	1	1	1	1	1	1	1	1	2.08
C0457086	C0003873	Morning stiffness	Rheumatoid arthritis	10	5	8	8	1	10	1	8	2	7	8	6	2	8	5.69
C0458343	C0016053	Trigger point	Fibromyalgia	10	5	5	5	1	8	1	1	2	3	8	5	1	1	3.54
C1510420	C0041296	Cavitation	Tuberculosis	10	5	9	4	1	10	1	10	2	1	1	3	5	1	4.08
C1956089	C0018862	Osteophyte	Heberdens node	10	7	2	9	1	9	3	8	4	2	1	1	4	1	4
C1956391	C0018681	temporal arteritis	Headache	10	5	9	7	1	8	1	4	3	4	8	9	1	1	4.69
C2267026	C0020473	HMG Co A reductase inhibitor	Hyperlipidemia	10	8	1	5	1	2	1	1	2	1	1	1	2	3	2.23
C0003507	C0006660	Aortic stenosis	Calcification	5	5	9	8	5	5	1	1	5	8	8	5	4	9	5.62
C0003811	C0026264	Arrythmia	Mitral valve	5	3	6	7	1	4	2	2	3	7	1	5	1	1	3.31
C0003873	C0003904	Rheumatoid arthritis	Arthroscopy	5	7	5	5	1	5	1	1	1	5	9	7	1	1	3.77
C0009378	C0032584	Colonoscopy	Polyp	5	7	9	8	5	8	1	2	7	7	8	7	1	1	5.46
C0010137	Cl563292	Cortisone	Osteoporosis	5	7	8	3	1	5	1	1	2	2	4	1	1	1	2.85
C0013378	C0011167	Dysgeusia	Swallowing	5	1	1	3	1	5	2	1	4	2	1	2	1	1	1.92
C0013394	C0029965	Dysparunia	Ovulation	5	3	1	2	1	3	1	3	3	2	3	6	1	1	2.31
C0013604	C0017654	Edema	Glomerular filtration rate	5	7	6	5	1	3	1	1	1	2	1	1	1	1	2.38
C0017196	C0162429	Gastrostomy	Malnutrition	5	7	9	4	1	3	1	1	4	4	9	9	3	1	4.31
C0018802	C0020541	Congestive heart failure	Portal hypertension	5	1	1	2	1	2	1	2	2	3	2	3	1	1	1.69
C0019054	Cl561562	Hemolysis	Hemoglobin	5	5	8	8	5	9	3	9	6	2	3	6	2	5	5.46
C0020971	C0021051	Immunization	immunodeficient	5	1	5	1	1	2	1	1	2	4	1	6	2	1	2.15
C0023223	C0042345	Leg ulcer	Varicsoe vein	5	5	8	3	1	8	1	1	3	5	6	6	3	1	3.92
C0023418	C0038250	Leukemia	Stem cell	5	7	9	8	1	10	1	7	3	7	4	7	1	1	5.08
C0024530	C0002438	Malaria	Amebiasis	5	5	2	8	1	9	3	1	1	1	1	5	1	3	3.15
C0030326	C0023798	Panniculitis	Lipoma	5	5	7	8	1	1	1	1	3	1	5	5	2	1	3.15
C0030842	C0020517	Penicillin	Allergy	5	5	8	5	1	4	1	1	3	3	1	5	1	1	3
C0034065	C0032285	Pulmonary embolus	Pneumonia	5	5	7	5	1	4	1	1	4	2	3	7	1	1	3.23
C0035450	C0034079	Rheumatoid nodule	Lung nodule	5	5	7	2	1	1	1	1	4	3	3	1	1	1	2.38
C0036202	C0042866	Sarcoidosis	Vitamin D	5	7	1	1	1	1	1	1	1	1	1	1	1	1	1.46
C0036572	C0018681	Seizure	Headache	5	5	5	5	1	6	1	1	2	4	4	7	1	1	3.31
C0042109	C0277942	Urticaria	Buterfly rash	5	5	5	7	1	8	7	4	5	6	1	8	4	8	5.31
C0042164	C0019740	Uveitis	HLA B27	5	7	7	1	1	4	1	1	1	1	1	1	2	1	2.23
C0042384	C0040053	Vasculitis	Thrombus	5	5	5	3	1	5	1	1	3	5	1	4	4	1	3
C0080331	C0432601	Walking	Stair climbing	5	5	4	5	5	8	6	4	2	3	3	4	3	5	4.38
C0085649	C0034063	Peripheral edema	Pulmonary edema	5	5	7	5	1	5	3	5	5	5	1	1	1	7	3.92
C0224498	C0029408	Meniscus	Osteoarthritis	5	1	1	7	1	3	1	1	2	4	5	6	1	1	2.62
C0243026	C0020649	Sepsis	hypotension	5	5	5	1	1	1	1	1	1	2	1	1	1	1	1.69
C0333997	C0043352	Lymphoid hyperplasia	Xerostomia	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0376358	C0001109	Prostate cancer	Acid phosphatase	5	1	1	1	1	3	1	1	1	1	8	1	1	1	1.69
C0429103	C0027051	T wave	Myocardial infarction	5	7	9	8	1	6	1	4	5	5	9	7	1	5	5.23

Table 1 (*continued*)

CUI1	CUI2	TERM1	TERM2	P1	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	Mean
C0442874	C0522224	Neuropathy	Paralysis	5	5	2	5	1	5	1	4	2	5	3	1	2	1	2.85
C1533685	C1253936	Injection	Synovial effusion	5	7	4	3	1	8	2	1	2	1	7	5	1	1	3.31
C0005587	C0232208	Bipolar depression	Junctional rhythm	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0007286	C0549493	Carpal tunnel syndrome	Alveolitis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0009871	C1444657	Contraceptive	Contraindicated	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0010043	C0011127	Corneal ulcer	Decubitus ulcer	1	1	1	5	1	2	3	1	4	2	3	1	1	1	2
C0011849	C0032584	Diabetes	Polyp	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0011849	C0035450	Diabetes	Rheumatoid nodule	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0013404	C0231528	Dyspnea	myalgia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0018801	C0033975	Heart failure	psychosis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0018926	C0043352	Hematemesis	Xerostomia	1	1	1	2	1	1	2	1	1	1	1	5	2	1	1.54
C0020541	C0027962	Portal Hypertension	Nevus	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0024117	C0018520	Chronic obstructive pulmonary disease	Haletosis	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1.08
C0026269	C0030920	Mitral stenosis	Peptic ulcer disease	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0026764	C0011581	Multiple myeloma	Depression	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1.08
C0029408	C0392525	Osteoarthritis	Nephrolithiasis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0033706	C0039142	Prothrombin	Syringe	1	5	5	1	1	3	1	1	1	1	1	1	1	1	1.77
C0034069	C0003615	Pulmonary fibrosis	Appendicitis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0034887	C0023055	Rectal polyp	Laryngeal cancer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0035222	C0007642	Acute respiratory distress syndrome	Celluitis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0037199	C0442041	Sinusitis	Sinusoid	1	1	1	8	1	4	1	8	2	1	1	1	3	5	2.85
C0037473	C0024485	Sodium	Mri	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0037926	C0886052	Spinal cord compression	Wound compress	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0040583	C0007102	Tracheal stenosis	Colon cancer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0041834	C0029456	Erythema nodusum	Osteoporosis	1	1	1	1	1	5	1	1	1	1	4	1	1	1	1.54
C0149925	C0011860	Small cell carcinoma of lung	Type 2 diabetes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0158280	C0023891	Cervical spinal stenosis	Alcoholic cirrhosis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0162595	C0702166	Antiphospholipid antibody	Acne	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1.08
C0220982	C0409974	Ketoacidosis	Lupus	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1.23
C0233651	C1527356	Perseveration	Venous stasis ulcer	1	1	1	1	1	5	1	1	1	1	1	1	1	1	1.31
C0241910	C0022876	Autoimmune hepatitis	Premature labor	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0262538	C0040479	Medial collateral ligament tear	Torsade de pointes	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1.08
C0267809	C0376180	Cryptogenic cirrhosis	Gastrin	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0409162	C0333286	Hand splint	Splinter hemorrhage	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C0917996	C0034065	Cerebral aneurysm	Pulmonary embolus	1	1	1	1	1	4	1	1	1	4	1	1	1	1	1.46

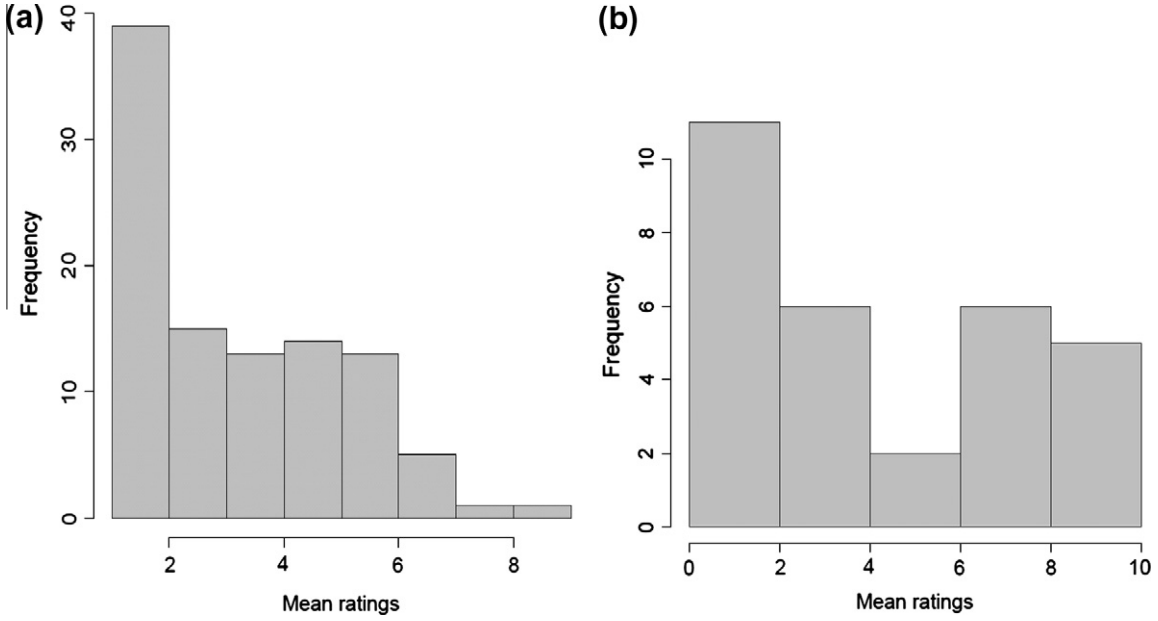

Fig. 1. Distributions of mean ratings by 13 medical coding experts on the set of 101 medical concept pairs (a) and nine medical coding experts on 30 general English word pairs (b).

Table 2
Pair-wise inter-rater agreement coefficients (Spearman's rho, Cohen's weighted kappa and Krippendorff's alpha) for 13 coding experts on the dataset of 101 medical term pairs.

Raters (statistic)	1	2	3	4	5	6	7	8	9	10	11	12
2 (Spearman)	0.76											
2 (Kappa)	0.65											
2 (Krippendorff)	0.73											
3 (Spearman)	0.65	0.68										
3 (Kappa)	0.68	0.57										
3 (Krippendorff)	0.64	0.66										
4 (Spearman)	0.35	0.48	0.43									
4 (Kappa)	0.14	0.27	0.27									
4 (Krippendorff)	0.03	0.09	0.07									
5 (Spearman)	0.66	0.73	0.71	0.44								
5 (Kappa)	0.58	0.62	0.68	0.25								
5 (Krippendorff)	0.67	0.72	0.69	−0.02								
6 (Spearman)	0.35	0.32	0.43	0.46	0.44							
6 (Kappa)	0.19	0.19	0.24	0.54	0.21							
6 (Krippendorff)	0.07	0.08	0.16	0.43	0.04							
7 (Spearman)	0.47	0.59	0.62	0.49	0.71	0.39						
7 (Kappa)	0.35	0.38	0.49	0.38	0.50	0.31						
7 (Krippendorff)	0.34	0.41	0.47	0.40	0.44	0.36						
8 (Spearman)	0.61	0.73	0.68	0.51	0.66	0.54	0.64					
8 (Kappa)	0.49	0.50	0.49	0.51	0.47	0.46	0.51					
8 (Krippendorff)	0.52	0.59	0.62	0.23	0.53	0.33	0.57					
9 (Spearman)	0.53	0.74	0.62	0.40	0.67	0.35	0.56	0.74				
9 (Kappa)	0.41	0.51	0.47	0.34	0.48	0.25	0.36	0.65				
9 (Krippendorff)	0.48	0.61	0.57	0.18	0.54	0.21	0.50	0.74				
10 (Spearman)	0.48	0.59	0.56	0.46	0.54	0.25	0.35	0.54	0.57			
10 (Kappa)	0.46	0.53	0.52	0.33	0.47	0.21	0.22	0.42	0.48			
10 (Krippendorff)	0.45	0.52	0.53	0.30	0.44	0.18	0.33	0.54	0.56			
11 (Spearman)	0.58	0.74	0.62	0.45	0.71	0.37	0.56	0.76	0.74	0.61		
11 (Kappa)	0.60	0.72	0.63	0.30	0.68	0.21	0.37	0.60	0.59	0.59		
11 (Krippendorff)	0.59	0.70	0.63	0.17	0.68	0.19	0.45	0.69	0.68	0.59		
12 (Spearman)	0.44	0.49	0.44	0.38	0.52	0.47	0.52	0.50	0.41	0.26	0.43	
12 (Kappa)	0.23	0.30	0.24	0.39	0.27	0.48	0.45	0.38	0.25	0.17	0.23	
12 (Krippendorff)	0.20	0.26	0.23	0.30	0.16	0.45	0.50	0.39	0.31	0.21	0.30	
13 (Spearman)	0.49	0.53	0.56	0.51	0.56	0.54	0.58	0.54	0.49	0.41	0.49	0.48
13 (Kappa)	0.42	0.39	0.42	0.55	0.46	0.46	0.48	0.27	0.33	0.38	0.45	0.40
13 (Krippendorff)	0.37	0.37	0.41	0.43	0.33	0.51	0.58	0.49	0.45	0.39	0.41	0.47

6, 7, 12 and 13. Comparing these results to the groupings obtained with clustering methods we found that rater # 7 is the only rater on which the three methods disagree. *K*-means clustering, similarly to PCA, assigned this rater to Group 2 while Ward's agglomerative clustering method assigned this rater to Group 1. Using the subgroups identified based on the factor analysis results, we recomputed inter-rater agreement with and without rater 7 for the subgroups determined based on the groups discovered by the clustering and factor analyses. The agreement results for the subgroups are shown in Table 5. Without rater 7, the group representative of the first component (Group 1) for both datasets had a higher ICC(2,1) (single measures) – 0.61 vs. 0.51 for medical terms and 0.83 vs. 0.78 for general English terms. However, the ICC(2,*k*) (average measures) remained about the same. Adding rater 7 to either of the groups slightly decreased the agreement within the groups. For example, ICC(2,1) for Group 1 decreased from 0.61 to 0.59 and for Group 2 – from 0.47 to 0.45.

4.4. Mapping to the UMLS

Of the 186 unique terms in the set of 101 pairs, MetaMap was able to map 136 (73%) terms correctly to a single concept in the UMLS; 12 (6%) terms had no matching concept; 8 (4%) terms were mapped to inappropriate concepts; and 30 (15%) terms were mapped to more than one concept. Spelling variations resulted in failing to assign concepts to 12 terms, for example, no concept was found for the term "haletosis", which should map to the concept "C0018520: Halitosis".

The mapping of 30 terms to more than one concept was due to word sense ambiguity. For example, the term "diabetes" has two possible concepts "C0011849: Diabetes mellitus" and «C0011860: Non-insulin dependent diabetes mellitus.» In the cases with word sense ambiguity, we chose the concept that was most similar to its term pair counterpart based on the similarity score between

Table 3
Pair-wise inter-rater agreement coefficients (Spearman's rho, Cohen's weighted kappa and Krippendorff's alpha) for nine coding experts on the dataset of 30 general English word pairs.

Raters (statistic)	1	2	3	4	5	6	7	8
2 (Spearman)	0.83							
2 (Kappa)	0.60							
2 (Krippendorff)	0.81							
3 (Spearman)	0.84	0.85						
3 (Kappa)	0.85	0.55						
3 (Krippendorff)	0.84	0.79						
4 (Spearman)	0.86	0.90	0.89					
4 (Kappa)	0.79	0.77	0.87					
4 (Krippendorff)	0.83	0.88	0.86					
5 (Spearman)	0.81	0.83	0.82	0.88				
5 (Kappa)	0.82	0.74	0.83	0.89				
5 (Krippendorff)	0.82	0.81	0.81	0.86				
6 (Spearman)	0.83	0.84	0.85	0.89	0.89			
6 (Kappa)	0.49	0.71	0.76	0.62	0.86			
6 (Krippendorff)	0.85	0.74	0.83	0.86	0.85			
7 (Spearman)	0.80	0.82	0.77	0.79	0.74	0.82		
7 (Kappa)	0.72	0.68	0.71	0.73	0.64	0.66		
7 (Krippendorff)	0.75	0.72	0.74	0.74	0.70	0.81		
8 (Spearman)	0.85	0.88	0.84	0.88	0.81	0.77	0.84	
8 (Kappa)	0.68	0.48	0.64	0.68	0.25	0.32	0.74	
8 (Krippendorff)	0.52	0.48	0.57	0.55	0.51	0.55	0.61	
9 (Spearman)	0.58	0.68	0.59	0.64	0.77	0.65	0.66	0.70
9 (Kappa)	0.30	0.60	0.33	0.34	0.39	0.61	0.52	0.40
9 (Krippendorff)	0.38	0.33	0.27	0.42	0.42	0.41	0.51	0.48

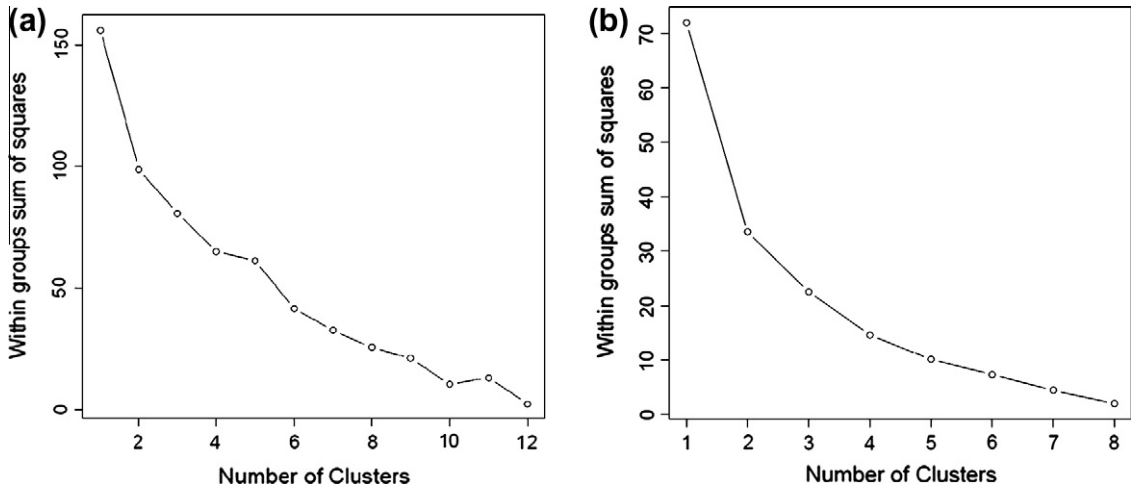

Fig. 2. Hartigan's sum-of-squares plot for different numbers of clusters based on the ratings by 13 coding experts on 101 medical term pairs (a) and nine coding experts on 30 general word pairs. The steep drop in the sum-of-squares after adding the second cluster indicates that the data supports two clusters.

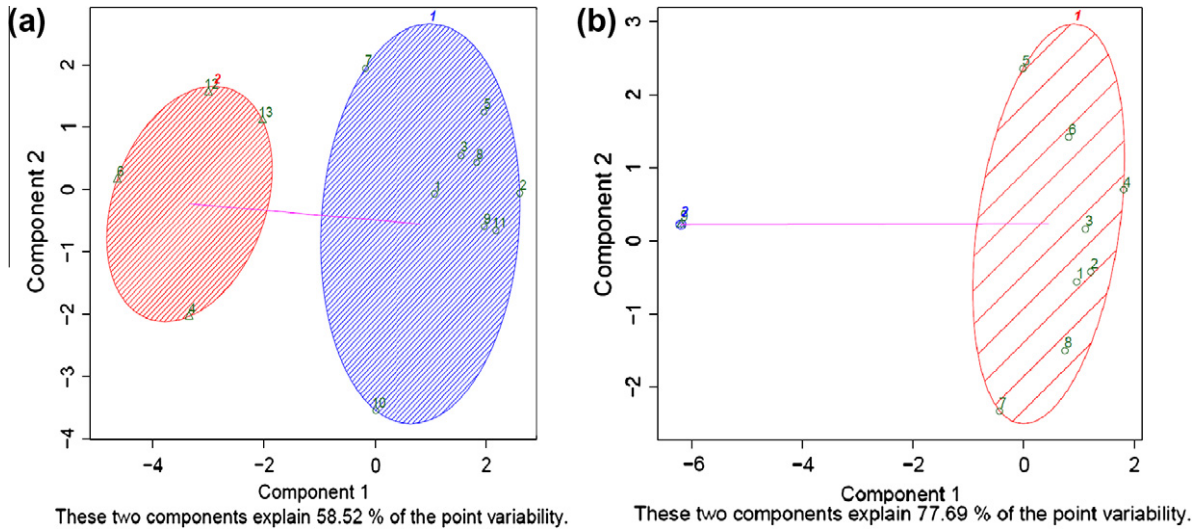

Fig. 3. K-means clustering solution obtained on the semantic relatedness ratings by the13 medical coding experts on 101 medical term pairs (a) and nine medical coding experts on 30 general English term pairs (b). In (a), the ratings on the term pairs are clustered into two groups – cluster 1 consisting of raters 4, 6 12 and 13 and cluster two consisting of raters 1–3, 5, 7, 8–11. In (b), the ratings are clustered into a group of raters 1–8 and an outlier rater 9 in cluster 2.

the two terms computed with UMLS-Similarity package.[5] In this case, the term pair counterpart for "diabetes" is "polyp." The latter maps to the concept: "C0032584: polyp" and the similarity between C0011849 and C0032584 is 0.1111 using the path measure while the similarity between C0011860 (non-insulin dependent diabetes mellitus) and C0032584 (Polyp) is 0.1000 therefore the term "diabetes" is disambiguated (albeit somewhat arbitrarily) to the concept "C0011849: Diabetes mellitus".

Based on the prior work reported in the literature on semantic relatedness both in the general English and biomedical domains, as well as the lessons learned from this study reported in this paper, we propose the following framework for the development of semantic relatedness reference standards.

4.5. Framework for reference standard development

The framework consists of the following components:

4.5.1. Dataset

Compile a set of concept pairs balanced across the semantic relatedness spectrum. Balancing the pairs across the relatedness spectrum is difficult. As indicated by Lee et al.'s study [12], selecting a list of medical terms and using all of their pair-wise combinations leads to heavy bias towards unrelated or dissimilar pairs. The quasi-random stratified sampling that we used in the current study leads to the opposite effect, as illustrated in Fig. 1. Both Lee et al.'s [12] and our findings with respect to relatedness distributions are not surprising as we would expect most pair-wise combinations on a random list of terms to be unrelated, while we would also expect people to be biased towards thinking of pairs of terms that are related. These findings are consistent with prior work in priming in lexical semantics demonstrating that showing a prime (first word in a sequence of two words) that is semantically related to the target (second word) results in shorter reaction times than in pairs of unrelated words [54,55]. A number of neuroimaging studies have also provided evidence that semantically related words elicit clearly detectable differences in neural response from semantically unrelated words. Weber and colleagues [56] used functional Mag-

[5] http://search.cpan.org/dist/UMLS-Similarity/.

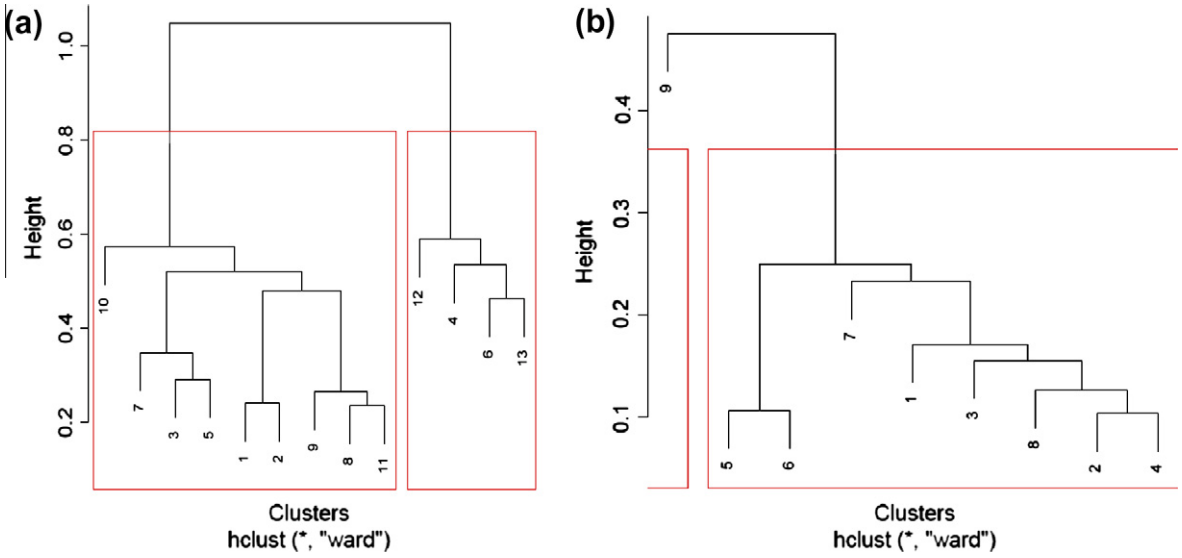

Fig. 4. Hierarchical agglomerative clustering solution (WARDS) obtained on the semantic relatedness ratings by 13 medical coders on 101 medical term pairs (a) and nine medical coders on 30 general English term pairs (b). In (a), the ratings on term pairs are clustered into two groups – cluster 1 consisting of raters 4, 6 12 and 13 and cluster two consisting of raters 1–3, 5, 7, 8–11. In (b), the ratings are clustered into a group of raters 1–8 and an outlier rater 9 in cluster 2.

Table 4
Multi-rater statistics for the two reference standards (101 medical term pairs and 30 general English word pairs).

Agreement statistic	Reference standard	
	101 Medical term pairs dataset (k = 13)	30 General English pairs dataset (k = 9)
ICC(2,1) (two-way, consistency, single measures)	0.50	0.78
ICC(2,k) (two-way, consistency, average measures)	0.93	0.97
Chronbach's alpha	0.93	0.97
Kendall's coefficient of concordance	0.57	0.82

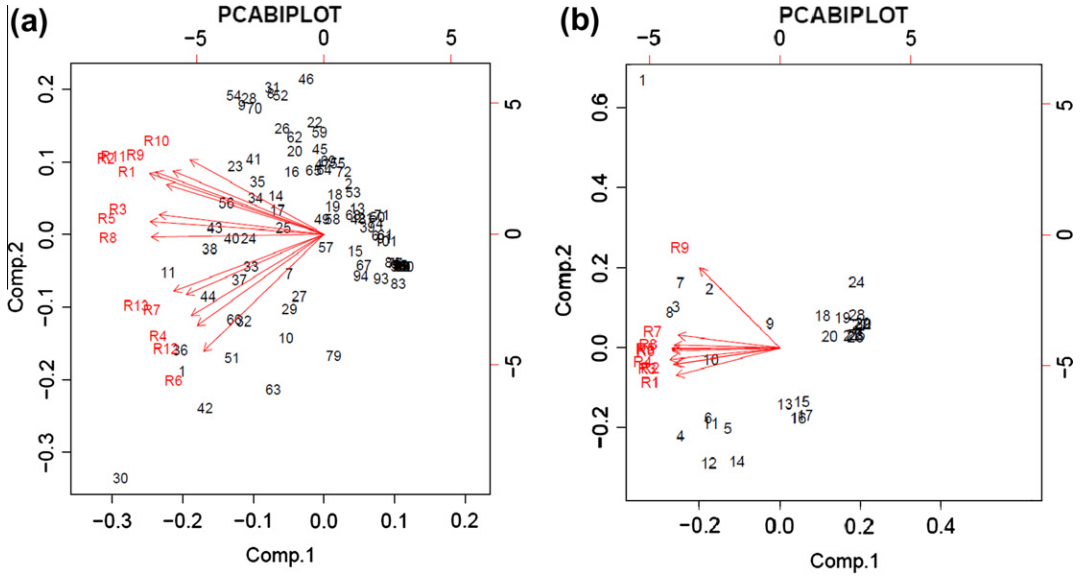

Fig. 5. Plots representing 2-component PCA solutions with Varimax rotation for the medical pairs dataset (a) and the general English word pairs (b). The plot in (a) shows a split between two major groups: raters 1–3, 5, 8, 9, 10 and 11 in Group 1 and raters 4, 6, 12 and 13 in Group 2. The plot in (b) indicated a separation of rater 9 from the rest.

netic Resonance Imaging (fMRI) to demonstrate that neural similarity computed by comparing the location and intensity of the signal on fMRI images correlates with behavioral ratings of similarity between pairs of concepts representing a semantic category of mammals. In another recent study, Mitchell and colleagues [57] used a large word corpus of general English text to train a neural network model able to predict the neural activation patterns detected with fMRI imaging for common English nouns (e.g., celery, airplane).

We propose that an approach to compiling reference standards that are balanced across the relatedness spectrum should consist of

Table 5
Agreement coefficients calculated for subgroups of raters based on the groups of raters discovered by all three clustering analysis methods (*k*-means, Ward's and PCA).

Agreement coefficient	13 Raters on 101 medical pairs Subgroups of raters		Nine raters on 30 general English pairs Subgroups of raters	
	G1: 1–3, 5, 8, 9, 10, 11	G2: 4, 6, 12, 13	G1: 1–8	G2: 9
ICC(2,1)	0.61	0.47	0.83	–
ICC(2,*k*)	0.93	0.78	0.98	
Chronbach's	0.93	0.78	0.98	
Kendall	0.68	0.36	0.79	–

a combination of random sampling from an ontological resource (e.g., UMLS, SNOMED CT) and using an expert (e.g., a medical professional).

4.5.2. Annotator training

The annotators may be trained on existing general English or medical reference standards iteratively until acceptable agreement has been reached on these corpora. This step ensures that the annotators understand the task (e.g. difference between similarity and relatedness) and have internalized a common rating scale. Previously published mean ratings on these corpora should be used during the training only as a guide for the study investigators but not the annotators to avoid possible bias. The protocol for training may include the presentation of term pairs to the raters independently of each other at first with subsequent interim analysis of the disagreements and discussion. It is also important to either ensure that the medical terms used for the reference standard are either not polysemous or have a single clearly dominant meaning that is readily understood by the annotators.

4.5.3. Annotation process

At the most basic level, the annotators are presented the concept pairs one at a time and are asked to rate them on a pre-defined Likert scale. Variations of this process may include presenting the pairs multiple times at random intervals, counterbalancing the order of presentation of the terms in the pairs, or using reaction time in priming experiments instead of the Likert scale.

4.5.4. Corpus reliability analysis

At the most basic level, one should analyze the ratings produced as a result of the annotation process using standard statistical measures of inter-rater reliability including intra-class correlation coefficient or weighted Kappa statistic. Variations may include transformations of the rating scales (e.g. reducing the dimensionality of the rating scale, shifting the scales used by individual raters). The choice of reliability coefficients and their interpretation strategies greatly depend on the context of the study and the purpose for which the reference standard is being generated. We refer the reader to previously published work regarding the choice of the coefficients and interpretation [36,42,46,58]; however, we suggest that it is important to report several appropriate reliability coefficients, as long as the data meet their assumptions.

4.5.5. Analysis of the ratings

From the standpoint of the end user of the reference standard, it is important to know not only what the inter-rater reliability is but also if the ratings in the reference standard are homogeneous. Standard clustering or factor analysis techniques may be used to identify subgroups, if any, among the human annotators. From the standpoint of the reference standard developer, these techniques may help identify the existence of any relationships between the annotators and the ratings. In-depth systematic analysis of the latent internal structure of the ratings may help discover this group and any outliers, and facilitate appropriate data partitioning. Arguably raters 4, 6, 12 and 13 in Table 1 (also available at http://rxinformatics.umn.edu) could be spotted right away because their ratings are substantially different from the rest even on visual examination. However, using clustering and factor analyses that can help find patterns in ratings based on their correlations and variance is a more systematic approach to identify raters that are consistently different from the majority. These raters' judgments need to be carefully examined as they may represent a valid interpretation of semantic relatedness.

4.6. Demonstration of intended use of the framework

The comparison of correlations between the reference standard and the two ontology-based approaches resulted in a Spearman correlation coefficient 0.29 ($p = 0.006$) for the Leacock and Chodorow method and 0.30 ($p = 0.006$) for the Wu and Palmer method. We also correlated these measures with the two subgroups within the reference standard: Group1 consisting of raters 1–3, 5, 8, 9, 10, and 11, Group 2 consisting of raters 4, 6, 7, 12 and 13). The Leacock and Chodorow method resulted in a coefficient of 0.27 ($p = 0.008$) on Group 1 and 0.24 ($p = 0.019$) on Group 2. The Wu and Palmer method also resulted in a coefficient of 0.27 ($p = 0.010$) on Group 1 but a slightly higher coefficient of 0.29 ($p = 0.005$) on Group 2 than the Leacock and Chodorow method.

4.7. Tools and resources in support of the framework

In order to make the proposed framework for developing and validating semantic relatedness reference standards easier to examine, adopt and further develop by other investigators in the community, we are also releasing the datasets (see Appendices A and B, available at http://rxinformatics.umn.edu) as well as the R program (see Appendix C, available at http://rxinformatics.umn.edu) that were used in the development of this framework and the writing of this manuscript. The UMLS-Similarity and UMLS-Interface packages that enable the user to experiment with some of the existing measures of semantic relatedness have been released as open-source [27].

5. Discussion

Automatic assessment of the degree to which biomedical concepts are semantically related is important in a number of different primary and secondary application contexts. The primary applications include information retrieval from clinical records and biomedical literature, and drug safety surveillance, whereas secondary applications are no less important and include support for such natural language processing tasks as automatic word sense disambiguation. In the context of drug safety surveillance, measures of semantic relatedness may be used to group together concepts that constitute an adverse effect of a medication currently defined by Standardized MedDRA Queries. Measures of semantic relatedness with subsequent clustering may enhance the process

of creating Standardized MedDRA Queries by adding a data driven component. Another potential use for clusters of terms defined using semantic relatedness is to improve the sensitivity of electronic searches for patients that meet clinical research study eligibility criteria.

Our study is the first to make publicly available a validated reference standard for testing computerized measures of semantic relatedness in the medical domain. A previous study by Lee et al. [12] reported using averaged ratings on 190 pairs of medical terms from 25 practicing primary care physicians as the reference standard. However, their corpus is limited by the fact that both the physicians and the term pairs were split for efficiency of data collection into non-overlapping groups of physician and term pair blocks effectively resulting in seven smaller (20 term pairs) datasets each annotated by at most four physicians. In creating our reference standard, we asked the raters to rely on their intuition about the terms rather than any explicit knowledge of the assertions underlying the term's meaning. The main objective of the current manuscript is to introduce a systematic approach valuable for exploring implicit semantic relatedness judgments to determine if the judgments contain an internal structure indicative of differences in conceptualization of medical knowledge. The results of this work to date show that, on the task in which the raters are given unlimited time to make the judgments, groups of raters as well as term pairs emerge with relatively high agreement even if the overall agreement for the entire group is low. We hypothesized that the disagreements may be partly due to the raters trying to define the relationship between the terms explicitly and thus resulting in variability across groups of raters. To test this hypothesis, we designed a subsequent psycholinguistic experiment [4] in which we limited the time the raters were allowed to judge each pair thus trying force them to tap into their implicit knowledge of the meanings of the terms. The results of that experiment indicated that, in this time-limited context, the physician raters did tend to agree with each other better on the majority of term pairs thus providing additional evidence for the reality of the general notion of semantic relatedness that holds between the "core" meanings of the medical terms.

Assessing the validity of reference standards for natural language processing and computational linguistic tasks has been a matter of continued debate in the computational linguistics community [36,58]. This debate is highly relevant to the medical informatics community as it engages in research involving language and cognition that often requires the presence of valid reference standards. Several points of general consensus on acceptable practices in the assessment of reference standards can be identified. One of the key points relevant to the current study is the distinction between reliability and validity. According to Artstein and Poesio [58], reliability refers to the reproducibility of the annotated reference standard – it indicates how reliable the *process* of the data collection is, independently of how close the human raters are to capturing the actual phenomenon (e.g., semantic relatedness between concepts). The validity, on the other hand, is said to reflect the "*trustworthiness*" of the conclusions from subsequent analyses conducted by using the reference standard.

These two notions are often treated as indistinguishable by researches; however the distinction is critical because, as clearly demonstrated by Krippendorff, and Artstein and Poesio, the inter-rater agreement coefficients are sensitive to factors that determine what questions can be answered by using a particular reference standard. When the research question is whether results obtained on a small manually annotated reference standard can be extrapolated to a larger dataset, one wants to be assured that if another data sample were taken from the larger dataset and manually annotated, the same (or at least similar) results would be obtained. Hence in this scenario, the reliability of the reference standard is of utmost importance. If the research question is whether the results obtained on a reference standard with some algorithm A are the same or different from the results obtained on the same standard with algorithm B, the validity of the reference standard takes precedence. Here the researcher is concerned with whether the manual ratings on a specific reference standard accurately reflect the underlying phenomenon the raters were asked to measure. Fortunately, the differences between coefficients that are more "sensitive" to reliability and validity become smaller as the level of agreement and the number of raters increase. Therefore, having more than two raters for creating a reference standard is more likely to improve both the reliability and the validity of the reference standard.

Apart from selecting the right coefficient to measure reliability, choosing an appropriate interpretation of the coefficients is also a matter of debate. The medical community relies on the interpretation scales consisting of the following values: 0–0.2 – poor; 0.2–0.4 – fair; 0.4–0.6 – moderate; 0.6–0.8 – substantial; 0.8–1.0 – perfect [45,59]. The scales proposed in computational linguistics research are much more stringent where the coefficient has to exceed 0.8 for the reference standard to be considered reliable or valid [36,42]. One must keep in mind that these scales are established by convention and, like *p*-values in statistical tests, their use greatly depends on the context and the purpose for which they are being used. Also, Krippendorff's guidelines for interpreting agreement coefficients were discussed in the context of discrete nominal scales where absolute agreement on category names rather than consistency in their relative ordering is important.

The overall consistency of our dataset of 101 medical term pairs measured with average measures ICC(2,13) is in the "perfect" range; however the reliability measured with single measures ICC(2,1) is in the "moderate" range and clearly falls short of the 0.8 threshold proposed by Krippendorff and others in computational linguistics. The former coefficient is an equivalent of Cronbach's alpha and is indicative only of whether the means of the ratings in the dataset may be used as an index representative of the raters. It does not indicate that we would obtain similar results if we used a completely different set of raters. Thus, we would recommend that anyone using the entire dataset provided in Table 1 (also available as at http://rxinformatics.umn.edu) in their research ought to keep in mind that this dataset (in its entirety) is more appropriate for conducting pilot studies rather than for drawing definitive conclusions. For example, given the high consistency of the ratings, it would be appropriate to use the means for all 13 raters provided with this dataset to compare different algorithms for computing semantic relatedness with the caveat that the results of the comparison may not be used to establish the relative superiority of the algorithms in general. Any conclusions drawn from such a comparison would be limited to this particular dataset and may not be readily generalizable; however, despite this limitation, this reference standard will be useful in examining the differences between algorithms and understanding their relative strengths and weaknesses with respect to this particular dataset.

The detailed analysis of the internal structure of the medical coding experts' ratings on the medical pairs dataset suggested that the dataset contains two subgroups – one with higher inter-rater agreement than that of the entire group and one with lower agreement. The inter-rater agreement of the first group is in the "substantial" range (albeit still short of Krippendorff's 0.8 threshold). This agreement is calculated on eight raters, thus we propose that this subset of raters (1–3, 5, 8, 9, 10, 11) represents a reference standard with higher validity (but not reproducibility) than the set of all 13 raters.

5.1. Analysis of disagreements

A typical source of poor agreement stems from variable understanding of the task by the raters. The fact that the medical coding experts that rated the corpus of medical terms had relatively high agreement coefficients on the general English words dataset indicates that they understood the general requirements of the task. The differences in reliability statistics for these two datasets are more likely to be due to the fundamental differences between the datasets. One possible explanation is that general English words are learned by native speakers of English at a very early age and are continually reinforced through contextual co-occurrence throughout life. The meanings of general English words are thus more stable and better internalized than those of medical terms learned much later in life and represent part of a professional jargon prompting greater variability of interpretations. Below, we provide an account of the possible sources for disagreements among human raters on semantic tasks and relate these sources to the specific example of the medical term pairs dataset.

The examination of the scores provided by the medical coding experts suggest the following reasons that likely contributed to the lower inter-rater agreement on the corpus of medical term pairs as compared to the agreement on the general English pairs.

1. *Differences in the use of rating scales*. The coding experts were instructed to use a 10 point scale; however, not all raters used the whole range. For example, the highest rating for raters 6, 8, 9 and 12 was 8. The rating of 9 was the highest for raters 1 and 3. Subsequent experiments with physician raters [4] confirmed that a 10 point scale is too wide and that a coarser scale of at most four points is more optimal to use. Contrary to our initial expectation, a smaller scale is perhaps more suitable for relatedness judgments in a specialized domain such as medicine because the terminology is learned by health care professionals much later in life than general English words making it more difficult to draw fine-grained distinctions.
2. *Differences in understanding of the notion of relatedness*. It is evident from the ratings provided by some of the experts that they clearly interpreted the notion of relatedness differently from other raters. For example, raters 5 and 7 consistently rated most of the term pairs (86 out of 101 for rater 5 and 75 out of 101 for rater 7) as completely unrelated (score of 1).
3. *Differences in conceptual models of the medical terminology*. How coding experts conceptualize the medical terminology space is inevitably going to influence their rating decisions. For example, raters 4, 6, 12 and 13 assigned a score of 1 to the pair "temporal arteritis – headache." The mean rating for this pair among the other experts is 6.33 with a standard deviation of 2.34. Thus the score of 1 is clearly below two standard deviations for this pair. It is possible that these four raters may have interpreted the relationship between the disease "temporal arteritis" and one of its possible manifestations "headache" as belonging to two different categories unrelated to each other in any direct way. This example is indicative of a systematic difference in interpretation by these four experts from the rest of the group rather than just neglect or carelessness on the part of the experts. This view is also supported by the clustering and factor analysis findings placing these four experts in a distinct group that is systematically different from the rest.
4. *Differences in experience*. Medical coding experts that participated in this study had a variety of experience with medical coding. All of them were trained to use the Hospital Adaptation of ICD coding nomenclature; however, they had variable levels of expertise. We should note here that, in this case, greater amount of experience does not necessarily mean better ability to determine the degree of relatedness between medical concepts.

The four reasons we outlined above as possible sources of disagreements between the raters are clearly not independent of each other. It is likely that any given disagreement was caused by a combination of these factors in addition to chance. These and other possible sources for disagreements on semantic relatedness tasks, particularly in specialized areas like medicine, need to be systematically examined in order to develop reliable reference standards.

5.2. Demonstration of intended use of the framework

The reference standard developed under the proposed framework was used to compare two ontology-based approaches. The low overall correlation between the ontology-based measures and the manual reference standard is not surprising as it is consistent with the notion that semantic similarity captured with ontology-based approaches is different from semantic relatedness judgments of the medical coding experts. For example, both ontology-based measures assigned a low rank of 92 to the pair C0003811 (arrythmia) – C0026264 (mitral valve), whereas the medical coders gave it a mid-range rank of 17. Another example, is the pair C0009676 (confusion) – C0011253 (delusion) that received a very low rank of 87 by the ontology-based measures but a very high rank of 5 by the medical coders. These examples show that concept pairs judged by humans as related are not semantically similar based on theirs relative locations in an ontology. The converse is also true – concepts that appear to be semantically similar based on their location in an ontology, may not be judged as related. Examples of this type of discrepancy include the pair of concepts C0010043 (corneal ulcer) and C0011127 (decubitus ulcer) that was ranked number 10 by the ontology-based measures but number 88 by the medical coding experts. In the UMLS, the two types of ulcers are classified under the same parent node; however, clinically, a typical corneal ulcer is arguably very different in its properties and etiology from a typical decubitus ulcer, depending on the context. The latter observation would be better captured with a measure of relatedness rather than ontology-based similarity. A third type of discrepancy that we observed between the automated ontology-based measures and human ratings has to do with either possible problems in the relational organization of the ontology or the path traversal algorithm (or both). For example, the pair C0011849 (diabetes) – C0032584 (polyp) was ranked very low by the human raters but relatively high by the ontology-based measures – rank 25. In this particular case, we were able to identify that this discrepancy was due to the ambiguity of the term "diabetes" as was discussed in Section 4 (Mapping to the UMLS). One of the intended uses of the corpus of 101 medical pairs presented in this article is to enable this type of analysis to guide the process of developing automated measures of semantic relatedness and semantic similarity.

6. Limitations

Several limitations must be discussed to facilitate the interpretation of the results of this study. First, the size of the reference standard is relatively small compared to all possible pairs of biomedical concepts, which limits its generalizability. However, this reference standard may be useful in guiding software development efforts and pilot studies aimed at comparing various measures of semantic relatedness to each other as long as the results of such

comparisons take into account the limitations of this dataset. Second, the pairs were initially chosen by a rheumatology specialist, which may have biased the set towards that specialty. In the future, it will be important to explore other methods for selecting pairs. Third, the concept space in the biomedical domain is very large (e.g. the UMLS contains close to 1 million concepts represented by 2 million names). Thus, annotating even a 10% sample of concept pairs from this space is not feasible. A more feasible approach would consist of partitioning this space into more manageable subdomains. One possible subdivision may consist of medical specialties, for example. Another approach that we are currently pursuing is to subdivide medical concepts according to their semantic types and address a subset that is most relevant for a specific purpose – drug safety surveillance. As part of this approach we have selected pairs of concepts (following the approach presented in this paper) from three semantic types and their intersections: drugs, disorders and symptoms. The resulting corpus contains 724 pairs and is currently being annotated for semantic relatedness. A pilot version of the corpus rated by five physicians for similarity and another five for relatedness is available at rxinformatics.umn.edu. Fourth, the nested nature of the relationship between the notions of semantic similarity and semantic relatedness introduces an asymmetry in the usability of reference standards developed to assess the tools measuring these two notions. A reference standard developed specifically for measuring the strength of semantic relatedness based on associative relationships may be used to some extent to test algorithms for determining semantic similarity, as semantic similarity as a special case of semantic relatedness. Fifth, due to time constraints we were unable to provide extensive training to the medical coders on making distinctions between relatedness and similarity. We were only able to provide a number of examples of term pairs that are related but not necessarily similar; however, now that we have a corpus of terms that have been rated for both relatedness and similarity, we will be able to create a training set for further annotation to address this limitation going forward.

7. Conclusion

While the notions of semantic relatedness are highly subjective, particularly in a narrow terminological domain such as medicine, our work to date shows that it is possible to compile a dataset that represents a generalized notion of relatedness to be used as a reference for developing computerized tools for measuring the degree of association between medical terms. However, our work also shows that considerable disagreements may exist and a systematic approach is necessary to exploring the human ratings of similarity and relatedness. We have proposed a framework and a starting set of publicly available tools and resources to support the efforts aimed at developing reference standards for testing automated approaches to measuring semantic relatedness between medical concepts. Compiling reliable and valid reference standards is clearly a complicated process. Detailed information about the methods used to generate the dataset as well as the motivation behind the use of reliability coefficients and their interpretation should be included by study investigators in publications reporting experimental results. Without this information, the results may be uninterpretable. Our framework attempts to lay the foundation for organizing the efforts of investigators involved in research on semantic relatedness of biomedical concepts by promoting open-source tools and resources. Currently, these resources are intended to be used to reproduce and compare results of different studies on semantic relatedness; however, the framework we propose may be extended to the development of reference standards in other research areas in medical informatics including automatic classification, information retrieval from medical records and vocabulary/ontology development. Our framework has a number of limitations that must be taken into account by anyone using the framework or the reference standards that are generated by using it.

Acknowledgments

We would like to thank the medical coding experts at the Mayo Clinic for participating in developing the reference standards referred to in this study. This work was supported in part by the National Library of Medicine Grants T 15 LM07041-19 and R01 LM009623-01A2.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2010.10.004.

References

[1] Pakhomov SV, Buntrock J, Chute CG. Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 2005;38:145–53.
[2] Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artificial Intell Res 1999;11:95–130.
[3] Rada R, Mili F, Bicknell E, Blettner AM. Development and application of a metric on semantic nets. IEEE Trans Syst, Man, Cybernet 1989;19:17–30.
[4] Pakhomov SV, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. In: Proceedings of American medical informatics association symposium. Washington (DC); 2010. in press.
[5] Mozzicato P. Standardised MedDRA queries: their role in signal detection. Drug Saf 2007;30:617–9.
[6] Pearson RK, Hauben M, Goldsmith DI, et al. Influence of the MedDRA((R)) hierarchy on pharmacovigilance data mining results. Int J Med Inform 2009.
[7] Bousquet C, Jaulent M, Chantellier G, Degoulet P. Using semantic distance for the efficient coding of medical concepts. In: Proceedings of the American medical informatics association symposium; 2000. p. 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 Saf 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 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: Proceedings of the American medical informatics association symposium; 2008. p. 384–8.
[13] Budanitsky A, Hirst G. Evaluating WordNet-based measures of semantic distance. Comput Linguist 2006;32:13–47.
[14] Resnik P. WordNet and class-based probabilities. In: Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998. p. 239–63.
[15] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th international conference on intelligent text processing and computational linguistics; 2003. p. 241–57.
[16] Fellbaum C, editor. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998.
[17] Miller G, Charles W. Contextual correlates of semantic similarity. Language Cognitive Process 1991;6:1–28.
[18] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the unified medical language system. In: Proceedings of the workshop on WordNet and other lexical resources: applications, extensions, and customizations; 2001. p. 77–82.
[19] Bodenreider O, Burgun A. Characterizing the definitions of anatomical concepts in WordNet and specialized sources. In: Proceedings of the first global WordNet conference; 2002. p. 223–30.
[20] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. Bioinformatics 2003;19:1275–83.
[21] Guo X, Liu R, Shriver CD, Hu H, Liebman MN. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 2006;22:967–73.
[22] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst, Man Cybernet 1989;19:17–30.

[23] Rodriguez M, Egenhofer M. Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowledge Data Eng 2003;15:442–56.
[24] Wilbur W, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med 1996;26:209–22.
[25] Spasic I, Ananiadou S. A flexible measure of contextual similarity for biomedical terms. In: Proceedings of the Pacific symposium on biocomputing, vol. 10; 2005. p. 197–208.
[26] Riensche R, Baddley B, Sanfilippo A, Posse C, Gopalan B. XOA: web-enabled cross-ontological analytics. In: IEEE SCW; 2007. p. 99–105.
[27] McInnes B, Pedersen T, Pakhomov S. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: Proceedings of the American medical informatics symposium, San Francisco (CA); November 2009. p. 431–35.
[28] Stevenson M, Greenwood M. A semantic approach to IE pattern induction. In: Proceedings of the 43rd annual meeting of the association for computational linguistics; 2005. p. 379–86.
[29] Raina R, Ng A, Manning C. Robust textual inference via learning and abductive reasoning. In: Proceedings of the twentieth national conference on artificial intelligence; 2005. p. 1099–105.
[30] Bodenreider O, Burgun A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. In: Proceedings of Medinfo symposium; 2004. p. 327–31.
[31] Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006;39:697–705.
[32] Neveol A, Zeng K, Bodenreider O. Besides precision & recall: exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. AMIA Annu Symp Proc 2006:589–93.
[33] Zhu S, Zeng J, Mamitsuka H. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics 2009;25:1944–51.
[34] Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. J Biomed Inform 2008;41:882–8.
[35] Rubenstein H, Goodenough J. Contextual correlates of synonymy. Commun ACM 1965;8:627–33.
[36] Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Human Commun Res 2004;30:411–33.
[37] Pedersen T. Empiricism is not a matter of faith. Comput Linguist 2008;34:465–70.
[38] Melton LJ. History of the Rochester epidemiology project. Mayo Clin Proc 1996;7:266–74.
[39] H-ICDA. H-ICDA Hospital adaptation of ICD-A-HICDA, 2nd ed. Ann Arbor, MI: CPHA (Commission on Professional and Hospital Activities); 1973.
[40] Wu Z, Palmer M. Verb semantics and lexical selection. In: 32nd annual meeting of the association for computational linguistics; 1994. p. 133–8.
[41] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960;20:37–46.
[42] Krippendorff K. Content analysis: an introduction to its methodology. 2nd ed. Thousand Oaks (CA): Sage; 2004.
[43] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrica 1951:16.
[44] Kendall MG, Babington Smith B. The problem of m rankings. Ann Math Stat 1939;10:275–87.
[45] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
[46] McGraw KO, Wong SP. Forming inferences about some intraclass correlations coefficients (vol. 1, 1996. p. 30). Psychol Methods 1996;1:390.
[47] Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–44.
[48] Hartigan J. Clustering algorithms. Wiley and Sons; 1975.
[49] Pedersen T, Kulkarni A. Automatic cluster stopping with criterion functions and the gap statistic. In: Proceedings of the demonstration session of the human language technology conference and the sixth annual meeting of the north American chapter of the association for computational linguistics; 2006. p. 276–9.
[50] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Roy Stat Soc Ser B – Stat Method 2001;63:411–23.
[51] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc/AMIA Annu Symp 2001:17–21.
[52] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pacific Symposium on Biocomput 2000:517–28.
[53] Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Stud Health Technol Inform 2004;107:487–91.
[54] Ferrand L, Ric F, Augustinova M. Affective priming: a case of semantic priming? Annee Psychol 2006;106:79–104.
[55] Thompson-Schill SL, Kurtz KJ, Gabrieli JDE. Effects of semantic and associative relatedness on automatic priming. J Memory Language 1998;38:440–58.
[56] 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.
[57] Mitchell TM, Shinkareva SV, Carlson A, et al. Predicting human brain activity associated with the meanings of nouns. Science 2008;320:1191–5.
[58] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist 2008;34:555–96.
[59] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.