Published in final edited form as:
J Biomed Inform. 2012 October ; 45(5): 992–998. doi:10.1016/j.jbi.2012.04.010.

Ontology-Guided Feature Engineering for Clinical Text Classification

Vijay N. Garla, MS[1,*] and **Cynthia Brandt, MD, MPH**[2,3]

[1]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06520-8009

[2]Connecticut VA Healthcare System, Bldg. 35A, Room 213 (11-ACSLG), 950 Campbell Avenue, West Haven, CT 06516

[3]Yale Center for Medical Informatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06520-8009

Abstract

In this study we present novel feature engineering techniques that leverage the biomedical domain knowledge encoded in the Unified Medical Language System (UMLS) to improve machine-learning based clinical text classification. Critical steps in clinical text classification include identification of features and passages relevant to the classification task, and representation of clinical text to enable discrimination between documents of different classes. We developed novel information-theoretic techniques that utilize the taxonomical structure of the Unified Medical Language System (UMLS) to improve feature ranking, and we developed a semantic similarity measure that projects clinical text into a feature space that improves classification. We evaluated these methods on the 2008 Integrating Informatics with Biology and the Bedside (I2B2) obesity challenge. The methods we developed improve upon the results of this challenge's top machine-learning based system, and may improve the performance of other machine-learning based clinical text classification systems. We have released all tools developed as part of this study as open source, available at http://code.google.com/p/ytex

Keywords

Natural Language Processing; Document Classification; Semantic Similarity; Feature Selection; Kernel Methods; Information Gain; Information Content

1. Introduction

Feature engineering plays an important role in many clinical text classification approaches. Feature engineering involves the selection of a subset of informative features and/or the combination of distinct features into new features in order to obtain a representation that enables classification. In the text classification domain, features typically include all distinct terms – words and/or concepts - present in a text corpus. Even small corpora may possess

[*]Corresponding Author: vijay.garla@yale.edu, Phone: +1 (203) 350-9571, Fax: +1 (203) 737-5708.

tens of thousands of features, potentially necessitating feature engineering for a given classification task. Domain knowledge is often used to guide the feature engineering process; for example, to identify notes that assert the presence of a disease, experts manually define dictionaries of terms related to the disease, e.g. symptoms and medications [1]. However, manual feature engineering may require considerable effort, and the selected features and feature groups are problem and domain-specific [1,2]. In this paper, we present novel methods that utilize the domain knowledge encoded in the taxonomical structure of the Unified Medical Language System (UMLS) to automate the feature engineering process.

Machine-learning based text classification approaches often utilize automated feature selection techniques. These techniques rank features based on statistics derived from the distribution of features within a corpus, or their joint distribution with document classes [3]. One drawback to most automated feature ranking methods is that they are univariate: each feature is considered separately, ignoring dependencies between features [4]. In the clinical text mining domain, free text is often mapped to concepts from an ontology that encodes semantic relationships between concepts. We hypothesize that the dependencies between concepts encoded in biomedical ontologies can be utilized to improve feature ranking.

Automated techniques for feature grouping include clustering, dimensionality reduction, and semantic similarity methods [5–9]. Clustering and dimensionality reduction are unsupervised methods that group features based upon their co-occurrence within documents. Semantic similarity measures utilize the taxonomic structure of an ontology to compute the similarity between pairs of concepts. Semantic similarity measures can be used to assign concepts to groups [10], or to project text into a feature space that effectively combines similar but distinct concepts [11]. One potential drawback to these techniques is that they are unsupervised or context-independent: they do not utilize the class labels assigned to text. Our intuition is that similarity, or the optimal grouping of concepts, is context-dependent. For example, in one context, the cardiovascular diseases congestive heart failure (CHF) and hypertension may be perceived as similar; in a different context, CHF and hypertension would be perceived as dissimilar, as they have different clinical presentation and treatments. We hypothesize that supervised semantic similarity measures that leverage class information can be used to sculpt a feature space that enables better discrimination between classes.

This paper is organized as follows: we provide an overview of related work; this is followed by a description of the proposed ontology-guided feature engineering methods. We then present the results of an empirical evaluation of these methods, followed by a conclusion.

2. Background

2.1. Feature Ranking and Selection

In the popular ‘bag-of-words’ document representation, documents occupy a feature space with one dimension for each term; terms may be words from a natural language, or may be technical identifiers such as a concept id. This feature space typically contains thousands of dimensions, posing a problem for many machine learning algorithms that suffer from overfitting when the number of features greatly exceeds the number of training examples [12]. To address this issue, a subset of relevant features may be selected. A relevant feature is one that increases performance when included in the set of features utilized by a particular machine learning algorithm [13]. In the text mining domain, ‘filter’ feature selection methods such as information gain and chi-squared are commonly used [3]. These methods rank features by measuring their correlation with the target class, and select the top features for use with a machine learning algorithm.

A different but related issue is isolation of passages of text relevant to a classification task. The entire body of a clinical note may not be relevant to a classification task [14–16]. For example, radiology reports often contain findings relevant to multiple organ systems. If the goal of a classification task is to identify reports that assert the presence of liver masses, findings pertaining to the lungs or other organ systems may have little or no relevance to the classification task. Rule-based systems often apply domain knowledge to isolate relevant passages from a clinical note [1,15]. Automated Isolation of Hotspot Passages (AutoHP) is an automated method for isolating passages relevant to a classification task [14]. AutoHP ranks all features by information gain; it designates the top features 'hotspots'; and it generates a bag-of-words using only the text surrounding hotspots. Documents that do not have any hotspots (zero-vectors) are simply assigned the most frequent class. Empirical evaluations have shown that classifiers trained on document vectors generated with the AutoHP method outperform those trained on a bag-of-words derived from the entire document [14,16].

2.2. Semantic Similarity

One potential problem with the bag-of-words document representation is that it does not explicitly express the similarity between related concepts. For example, when attempting to classify hypercholesterolemic patients it may be advantageous to express the similarity shared by anticholesteremic medications such as Lovastatin and Lipitor. A common manual feature engineering approach is to group related features under a single feature; this however requires expert domain knowledge. Semantic similarity measures are automated methods for assigning pairs of concepts a measure of similarity, and can be derived from an ontology or taxonomy of concepts arranged in is-a relationships; e.g. Lovastatin is-a Statin [8]. Semantic similarity measures can be roughly divided into the following classes: similarity measures based on taxonomical structure; measures based on both taxonomical structure and the distribution of concepts in a corpus; and the context vector measure based on solely on the distribution of concepts in a corpus [8,17–20]. A full discussion of various semantic similarity measures exceeds the scope of this paper; refer to Pedersen et al [8] for an excellent overview.

The UMLS Metathesaurus is a compendium of biomedical vocabularies including SNOMED-CT, ICD-9, and RXNORM [21]. The UMLS Metathesaurus enumerates concepts, assigns them unique identifiers (CUI), and encodes relationships between concepts, including parent-child (PAR-CHD) and broader than-narrower than (RB-RN) relationships. The PAR and RB relationships denote 'is-a' relationships, i.e. that one concept is a generalization of another concept. Concepts that are generalizations of other concepts are referred to as parents or hypernyms; specifications of a concept are referred to as children or hyponyms. A taxonomy suitable for use with semantic similarity measures can be derived from the UMLS by taking a subset of 'is-a' relations, and removing relations that induce cycles.

Measures based on taxonomical structure calculate the shortest path between two concepts; this path traverses the least common subsumer (LCS), i.e. the closest common parent concept. One limitation of purely taxonomy-based measures is that they give equal weight to all links [8]. Links between specific concepts, e.g. Lovastatin *is-a* Statin, should be weighted more heavily than links between general concepts, e.g. Lovastatin *is-a* Enyme Inhibitor. Information content (IC) based measures attempt to correct for this by weighting edges based on concept specificity. The IC of a concept is based on its frequency and the frequency of all its children in a corpus of text; frequent concepts are more 'general', whereas infrequent concepts are 'specific'. For example, the IC of the concept Statin is based on the frequency with which this term and all of its children (Lovastatin, Lipitor,…) appear in a corpus of text. The Lin measure is based on IC, and in empirical evaluations

demonstrated a high correlation with expert judgments of concept similarity [8]. The Lin measure is defined as follows:

$$\begin{aligned} sim_{lin}(c_1, c_2) &= \frac{2 \cdot IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \\ IC(c) &= -\log(freq(c)) \\ freq(c) &= freq(c, C) + \sum_{c_s \in children(c)} freq(c_s) \end{aligned} \quad (1)$$

The information content *IC(c)* of a concept is defined as the inverse of the log of the concept's frequency. The frequency of a concept is recursively defined using the taxonomic structure of the UMLS: it is based on the number of times ($freq(c,C)$) the concept c occurs within a corpus C, together with the number of times its children occur.

Semantic similarity measures have been applied to several domain-independent NLP tasks [11,22]. In the clinical text classification domain, Aseervatham and Bennani used semantic similarity measures in their entry to the 2007 Computational Medicine Center (CMC) NLP challenge [23]. Lu et al used a semantic similarity measure to assign symptoms to symptom groups in a system to classify free-text chief complaints into syndrome categories [10].

3. Methods

We hypothesized that 1) we could utilize the UMLS ontology to improve feature ranking methods, and 2) we could develop context-sensitive semantic similarity measures by combining these feature ranking methods with semantic similarity methods.

3.1. Ontology-Guided Feature Ranking

In this study, we use information gain (IG), also known as Kullback–Leibler divergence, a popular feature ranking method in the text classification domain [3]. IG measures the correlation between a feature and document class in bits of information. The IG is computed from the contingency table representing the joint distribution of the feature with the document class. The table below illustrates the typical construction of the contingency table for a binary (Y/N) classification task. D_0 and D_1 represent the sets of documents where a given feature is absent or present respectively. D_Y and D_N represent the sets of documents that are assigned the classes 'Y' and 'N' respectively. The contingency table is constructed by intersecting these sets: the cells of the contingency table represent the cardinality of the intersected sets.

One drawback to this approach is that it ignores relationships between concepts. All children of a UMLS hypernym may be relevant to a classification task. However, the children are ranked independently of one another, potentially obfuscating their value as features. We address this issue by propagating the contingency table of concepts to their hypernyms, using the taxonomical structure of the UMLS as a guide. To compute the propagated contingency table of a hypernym, we modify the joint distribution to account for taxonomical relationships as follows: we modify D_1 to include all documents that contain the concept or any of its children. For example, to compute the propagated contingency table for the statin concept, we assign any document that contains the concept statin or any child (e.g. Lipitor) to the set D_1. We construct the propagated contingency table for every concept in the taxonomy and compute the information gain; we refer to this as the propagated information gain (IG_{prop}) of a concept.

Intuitively, if a hypernym is relevant to a classification task, then specifications of this concept – its children – should be relevant as well. After computing the propagated

information gain, we assign each concept in the UMLS the highest propagated information gain of any hypernym; we refer to this as the imputed information gain (IG_{imp}):

$$IG_{imp}(c)=\max\left\{IG_{prop}(p_1)\ldots IG_{prop}(p_n)\right\} \quad (2)$$

Where $p_1 \ldots p_n$ includes the concept c and all of its hypernyms, up to and including the root of the taxonomy.

We use the imputed information gain for the final ranking of concepts for feature selection. Noteworthy aspects of this approach include: it is applicable to any taxonomical organization of concepts; a concept need not even appear in the training data to be ranked using this method, thereby potentially increasing the generalization capability of a system based on this method; this can be adapted to other feature binning strategies besides concept present/absent; and any feature evaluation method that can be computed from a contingency table (such as chi-squared) can be applied.

3.2. Supervised Semantic Similarity Measures

Semantic similarity measures may accurately measure the similarity between concepts, but this similarity might be irrelevant to a particular classification task. A relevant similiarity is one that increases performance when utilized by a particular machine learning algorithm. For example, Norvasc and Simvastatin, medications used to treat hypertension and hypercholesteremia respectively, are both children of the 'Cardiovascular Agent' concept. A semantic similarity measure may assign these medications a high similarity, but utilizing this similarity in a machine learning algorithm may reduce performance when classifying hypertension or hypercholesteremia.

Recall that taxonomy and information content based semantic similarity measures use the distance between the concepts and their least common subsumer (LCS). Our method is based on the assumption that the similarity between two concepts is relevant only if the imputed information gain (IG_{imp}) of their LCS is high. The intuition underlying this is: if a concept is not relevant to a classification task, then the similarity between this concept's children is also not relevant. We define the supervised semantic similarity of a pair of concepts with respect to a classification task as follows: we find the LCS of a pair of concepts; if the IG_{imp} of the LCS exceeds a configurable threshold we compute the Lin measure, else we assign the concepts a similarity of 0. The optimal LCS IG_{imp} threshold can be identified via cross validation. For example, if the IG_{imp} of Cardiovascular Agent does not exceed the LCS threshold, we assign Simvastatin and Norvasc a similarity of 0 (completely unrelated), else we compute the similarity using the Lin measure. A notable aspect of this approach is that it is applicable to any semantic similarity measure based on the LCS of a pair of concepts.

3.3. Kernel Methods

Kernel methods provide a principled mechanism for integrating domain-specific similarity measures with powerful machine learning algorithms. We utilize kernel methods to incorporate semantic similarity measures with machine learning algorithms to classify clinical text. A kernel can be thought of as a symmetric function that computes the pair-wise similarity between instances [11]. Technically, a kernel projects instances $x, y \in X$ not necessarily from a vector space, into a (potentially unknown) vector space using a map ϕ, and computes the inner products of the images in this space: $\langle \phi(x), \phi(y) \rangle$ [24]. The computational attractiveness of kernels derives from the fact that instead of projecting instances into a high-dimensional feature space and taking their inner product, kernels can compute the similarity directly – this is the renowned 'kernel trick'. Kernel machines such

as support vector machines (SVMs) are algorithms that are trained on the matrix of pair-wise kernel evaluations [25].

Given document vectors x,y indexed by concepts, a semantic similarity kernel can be defined as:

$$k(x, y)=xSy \quad (3)$$

The matrix S has 1's along the diagonal. The off-diagonal (i,j) elements represent the similarity of concept i and concept j. In our implementation, we use the Lin or supervised Lin measure to compute pairwise concept similarity. The semantic similarity kernel allows an intuitive geometric interpretation. In the standard bag-of-words vector representation, all terms represent perpendicular dimensions. The semantic similarity kernel effectively 'bends' the dimensions of this space so that similar concepts are no longer perpendicular, thereby pushing documents that contain similar concepts closer together. Refer to [11] for a more rigorous treatment of this geometric intuition.

3.4. Evaluation Method

We chose the I2B2 2008 Challenge dataset to evaluate our methods because it provides a benchmark to which we can compare our results [1]. For this challenge, domain experts reviewed 1237 discharge summaries from overweight or diabetic patients and classified these documents as asserting the presence of obesity and 15 related diseases, including Hypertension, Coronary Artery Disease (CAD), Congestive Heart Failure (CHF), and Hypercholesterolemia. The I2B2 Obesity challenge is a multi-label, multi-class classification task, and comprised a textual and an intuitive task; the labels correspond to diseases, and the classes correspond to judgments regarding the presence/absence of the disease. For the intuitive task, annotators applied clinical intuition to determine if a disease was present (Y), absent (N), or questionable (Q) based on information contained in the discharge summary. For example, annotators interpreted laboratory values or drug administration to infer the presence or absence of a disease.

This dataset is especially suited to the evaluation of feature selection methods in the clinical domain because of the high correlation between the distinct disease classes: a feature relevant to one cardiovascular disease, e.g. Hypercholesterolemia, may be highly correlated with other cardiovascular diseases, e.g. CAD, but may not be relevant to the classification of other diseases. In this study, we build upon the methods used by the top-scoring machine-learning based system.

Manually-developed rule-based systems achieved the highest performance in the I2B2 2008 challenge. The top 4 systems for the intuitive task were purely rule-based; hybrid systems that applied machine-learning methods to features obtained via manually developed rules occupied ranks 8 and 9. The only purely machine-learning based system among the top 10 for the intuitive task was the submission by Ambert et al (rank 5) [14]. The central feature of this system was the AutoHP method to automatically isolate passages from clinical text relevant to the classification task. Empirical evaluations on the I2B2 challenge data showed that document preprocessing with AutoHP provided the most significant contribution to the system's performance. Central to the AutoHP method is the ranking of features by information gain. We hypothesized that ranking features by imputed information gain would improve the identification of hotspot passages.

To quantify the contribution of our methods, we built upon the system of Ambert et al. in several iterations, adding our proposed techniques. We evaluated our methods on the

intuitive task of the I2B2 2008 challenge. We annotated all notes in the I2B2 2008 challenge using the Yale cTAKES Extensions (YTEX)[26,27]. YTEX identifies sentence boundaries, tokenizes text, performs negation detection, and maps text to concepts from the UMLS. We generated a directed acyclic object graph that represents the UMLS taxonomy using the SNOMED-CT and RXNORM source vocabularies, and the PAR-CHD and RB-RN relations. The I2B2 2008 challenge dataset comprises a training set (n=730) and test set (n=507). Participating teams were provided with the labeled training set three months prior to the challenge; the unlabeled test set was released two days before system output was due. We used the training set for parameter tuning and supervised feature ranking: to compute the raw, propagated, and imputed information gain. We used the entire I2B2 corpus for unsupervised feature ranking: to compute term frequencies and information content. To test semantic similarity measures, we implemented kernels and evaluated them on the I2B2 2008 dataset. We optimized all parameters via a 5×2 fold stratified cross validation on the training dataset; i.e. we performed 2-fold stratified cross validation 5 times. For the final evaluation, we trained an SVM on the training set for each disease (label) using the optimal parameters identified via cross validation, and evaluated the SVM on the test set. Systems submitted to the I2B2 2008 Challenge were ranked by the macro-averaged F1-Score on the test set. The F1-Score is the harmonic mean of positive predictive value and sensitivity; the macro-averaged F1-Score is the average of F1-scores across all classes. We used the Libsvm version 3.1 SVM implementation [28], YTEX v0.5, and UMLS version 2010AB. All computations were performed on CentOS release 5.4 running Intel Xeon 64-bit processors. Data was stored in a MySQL version 5.1.41 database. We have released all tools and scripts required to reproduce our results as open source.

4. Results and Discussion

4.1. Feature Selection

To illustrate the differences between raw, propagated, and imputed information gain (IG), we list the top ranked features for the Hypertension label (Table 2). Recall that a feature is considered relevant if it improves the performance of a given machine learning algorithm. In the context of this evaluation, we consider a feature relevant if it improves classifier performance when used to identify hotspots, or improves classifier performance when used with the supervised semantic similarity measure. To identify the relevant features, we performed a cross-validation using a range of IG thresholds; all features with an IG above the specified threshold were used for AutoHP, or with the supervised semantic similarity measure.

All methods assigned hypertension the highest rank. The imputed IG assigned Hypertension and its subtypes (e.g. renal hypertension, omitted for brevity) the highest ranks.

The ranking of Hyperlipidemia illustrates the shortcomings of raw information gain. Hyperlipidemia is a disease strongly correlated with hypertension, but represents a distinct label in the I2B2 challenge. In our evaluations, adding hyperlipidemia to the list of hotspot features reduced classifier performance. However, raw information gain ranks hyperlipidemia high, assigning it 0.033 bits, making it the 5^{th} highest feature. In contrast, propagated IG reduces hyperlipidemia's rank, assigning it 0.025 bits, reducing its rank to 36.

The ranking of calcium channel and beta blockers illustrates the power of propagated and imputed information gain. The propagated IG assigns the general concepts 'calcium channel blockers' and 'beta blockers' high information (0.046 and 0.051 bits): the children of these classes, distinct drugs relevant to Hypertension, contribute to the IG of their parents. The imputed IG assigned medications that belong to these drug classes high ranks. In contrast,

the raw IG ranks the beta blocker metoprolol – a relevant feature - below hyperlipidemia – an irrelevant feature.

For the classification of hypertension, the top-ranked rule-based systems compiled thesauri of terms informative for this disease class [1]. These included medications used to treat the hypertension, e.g. calcium channel blockers and beta blockers, and diagnostic findings related to hypertension (e.g. Blood pressure) [15]. The propagated and imputed IG assigned these terms high ranks, suggesting that this feature ranking method reflects expert human judgments more closely.

4.2. Semantic Similarity

To illustrate the differences between the Lin and supervised Lin semantic similarity measures we present the semantic similarity for selected pairs of biomedical concepts, their least common subsumer (LCS), and their propagated information gain (IG_{imp}) with respect to selected diseases (Table 3). Semantic similarity measures assign pairs of concepts a similarity that ranges between 0 (unrelated) to 1 (identical). The supervised semantic similarity measure assigns a concept pair a similarity of 0 if the IG_{imp} of their LCS is below a configured threshold. Our assumption is that, if the similarity between a pair of concepts is relevant to a classification task, their least common subsumer (LCS) will exhibit a high IG_{imp}. We consider the similarity of a pair of concepts to be relevant if utilizing the similarity in the semantic kernel improves classifier performance.

The statins Simvastatin and Pravastatin are relevant to the Hypercholesterolemia classification task; these features and hence their similarity may be irrelevant to other classification tasks, e.g. Hypertension. The propagated IG for statin, the LCS of these concepts, is relatively high for Hypercholesterolemia (0.441 bits, Table 3) and much lower for Hypertension (0.033 bits, Table 3). The ‘relevance’ of these similarities is confirmed by empirical evaluations (see below).

The similarities between the concept pairs Heart Failure-Hypertension and Simvastatin-Norvasc illustrate the shortcomings of unsupervised semantic similarity metrics: these similarities are not relevant to certain classification tasks, as demonstrated by empirical evaluations. Heart Failure and Hypertension represent two distinct disease labels for the I2B2 challenge: using their semantic similarity does not improve the classification of these diseases. Simvastatin and Norvasc are both used to treat cardiovascular diseases, and hence are similar in this respect. However, using their semantic similarity does not improve classification of either Hypertension or Hypercholesterolemia: conflating these distinct concepts harms classifier performance. The imputed IG for the LCSs of these pairs is relatively low. With the appropriate parameters (IG cutoff), the supervised semantic similarity metric can compute a ‘context-dependent’ similarity that tailors the ‘perception’ of similarity to a specific classification task.

4.3. Classification Results

We performed five ‘experiments’ in which we progressively built upon the methods of Ambert et al, adding our proposed techniques. We report the macro-averaged F1 score, the metric by which submitted systems were ranked in this challenge (Table 4). We compare the results of these experiments to each other and to those of the best I2B2 2008 submissions to quantify the contribution our methods.

Bag-of-Words (*word*)—This experiment represents a baseline with which we attempted to reproduce Ambert’s results. Our approach differs from Ambert in the following respects: instead of a simple whitespace and punctuation tokenizer we used the YTEX tokenizer;

instead of using a window of 100 characters on either side of hotspot features, we use a window that includes all sentences within 100 characters on either side of hotspot features - this was done primarily because it simplified the implementation; instead of inverse-class frequency weighted SVMs, we used unweighted SVMs.

We represented text as a binary feature vector that included all words within these sentences and applied SVMs with a linear kernel. The performance of this system was slightly lower than that of Ambert (macro-f1 0.6399 vs. 0.6344); this may be due to the use of wider windows around hotspots, tokenizer differences, or use of unweighted SVMs.

Bag-of-Words + Imputed Hotspots (*imputed*)—This experiment was designed to measure the contribution of imputed feature ranking. For this experiment, we retain all hotspots from the *word* experiment, using the optimal cutoffs as determined by cross-validation. We then apply AutoHP using concepts as hotspot features, and imputed information gain to rank concepts. We represented text as a binary feature vector that included all words within these sentences and applied SVMs with a linear kernel. We identified the optimal imputed hotspot cutoff via cross validation. The performance of this system was better than *word* and slightly higher than *ambert*.

Bag-of-Words + Imputed Hotspots + CUIs (*cui*)—This experiment was designed to measure the contribution of enriching the feature set with UMLS Concept Unique Identifiers (CUIs); previous experiments used only words as features. We represented text as a binary feature vector that included all words and CUIs within these sentences and applied SVMs with a linear kernel. As in the *imputed* experiment, we identified the optimal imputed hotspot cutoff via cross validation. The overall performance of this system was better than *word, imputed*, and *ambert* indicating that adding CUIs improved performance. Interestingly, the optimal imputed hotspot cutoff was different in the *cui* experiment than in the *imputed* experiment, indicating that the contribution of CUIs as features is not simply an additive improvement.

Semantic Kernel (*lin*)—This experiment was designed to measure the contribution of unsupervised semantic similarity measures. For this experiment, we used the binary feature vector from *cui* and applied SVMs with the semantic kernel using the Lin measure. In general, performance on all diseases decreased relative to *cui*, suggesting that unsupervised semantic similarity measures harm performance.

Supervised Semantic Kernel (*superlin*)—This experiment was designed to measure the contribution of supervised semantic similarity measures. For this experiment, we used the binary feature vector from *cui* and applied the supervised semantic kernel using the Lin measure and a range of LCS IG_{imp} thresholds. We used LCS thresholds corresponding to the IG_{prop} with ranks 1, 3, 5, and 10. For example, for Hypertension we used the thresholds 0.191, 0.066, 0.047, and 0.042 (Table 2). We identified the optimal threshold via cross-validation. In general, *superlin* outperformed other experiments and *ambert*.

Our system's performance does not differ significantly from the top challenge systems [1]. Of the top 10 systems, only the techniques employed by ourselves and Ambert can be fully automated. However, the comparison to challenge systems may not be fair, as these systems were developed under strict time constraints (two months).

Using imputed information gain to identify hotspot passages yielded the greatest performance improvement: in general, performance improved in the *imputed* and *cui* experiments relative to *word*. This improvement is attributable to the use of additional hotspot features identified via imputed infogain (IG_{imp}). For many diseases, IG_{imp} identified

members of relevant drug classes, e.g. Statins for Hypercholesteremia or Fibrate antihyperlipidemics for Hypertriglyceridemia. Infrequently administered drugs in these classes, or brand names of drugs usually referred to by their generic name, have a low raw IG (IG_{raw}) but a high IG_{imp}. For example, for the Hypertriglyceridemia label, IG_{imp} ranked highly the term 'Lopid', a brand name for Gemfibrozil with few mentions in the corpus. The identification of Lopid as a hotspot feature improved the performance of the *imputed* experiment on the Hypertriglyceridemia label: this was the only other Fibrate antihyperlipidemic (aside from those identified by IG_{raw}) that was present in this corpus.

For some diseases, the IG_{imp} identified relevant classes of clinical findings. For example, for Venous insufficiency, IG_{imp} ranked highly members of the "decreased vascular flow" class of clinical findings. For the Venous Insufficiency label, enriching the feature vectors with CUIs also greatly improved performance. This is probably due to the relevance of multi-word terms such as "Postthrombotic Syndrome" and "Abnormal vascular flow": the proximity of words belonging to these terms is lost in the bag-of-words representation. Representing these terms as CUIs rectifies this. For some diseases, performance of the *cui* experiment is lower than that of *imputed*; this may be due to noise introduced by named entity recognition errors.

IG_{imp} did not improve performance on the Coronary Artery Disease, Congestive Heart Failure (CHF), and Gallstones labels. For these diseases, the IG_{raw} 'found' all the relevant hotspots; expanding the hotspot feature set did not help. This was somewhat surprising for CHF, as IG_{imp} highly ranked loop diuretics, a class of drugs used to treat CHF, and which we expected would improve the identification of relevant passages.

The supervised Lin measure yielded only a small improvement relative to *cui/imputed*. This was somewhat surprising, as we expected that the semantic similarity metric would improve classification by making explicit the 'proximity' of related concepts. The unsupervised Lin measure in general harmed performance; this supports our hypothesis that, for this dataset, semantic similarity is context-dependent.

One limitation of our study was the limited corpus size used to compute the semantic similarity measures. Recall that the Lin measure is based on the frequency with which concepts occur in a corpus. Empirical evaluations have shown that the accuracy of the Lin measure is dependent on corpus size [8]. The poor performance of the unsupervised semantic kernel may have been an artifact of the limited corpus size. However this is a fundamental limitation of corpus-based semantic similarity measures: large, publicly available, annotated corpora do not exist from which concept frequencies can be computed. Recently, semantic similarity measures that estimate the information content of concepts solely from the taxonomical structure of an ontology have been developed [19]. These measures eliminate the dependency on a large, annotated corpus, and surpass corpus-based measures in accuracy. As part of future work, we will evaluate these measures on the I2B2 2008 challenge and other datasets.

The use of a single corpus to evaluate our methods is another limitation of our study. However, the I2B2 Obesity challenge represents a common use case in clinical NLP: the identification of patients with specific diseases from the narrative text of the medical record. We believe that the feature engineering methods presented here have general applicability. In particular, imputed information gain may be useful for the automated identification of features relevant to clinical NLP tasks.

Barriers to NLP development in the clinical domain include the formidability of reproducing the results of other systems, and the limited collaboration within the NLP community [29]. We addressed these issues in this study: we released all tools and code developed as part of

this study as open source and provided detailed documentation enabling others to reproduce our results. This study unifies the work of several disparate groups in the clinical NLP field: we used cTAKES, an NLP pipeline developed at the Mayo Clinic [27]; the AutoHP method, developed at Oregon Health & Science University [16]; and the semantic kernel developed at the Université Paris [23]. We combined these methods in an integrated, open-source software framework. There has been much research dedicated to the development of semantic similarity methods for the biomedical domain, but few applications of these methods to clinical NLP problems. The tools we developed facilitate the application of semantic similarity methods to problems in the clinical NLP domain.

5. Conclusions

Feature engineering approaches that leverage domain knowledge can improve the performance of machine-learning based classifiers. In this study, we presented a novel feature ranking method that utilizes the domain knowledge encoded in the taxonomical structure of the UMLS, and we developed a novel context-dependent semantic similarity measure. Semantic similarity measures quantify the relatedness between pairs of concepts. Our 'context-dependent' semantic similarity measure tailors the 'perception' of similarity to a specific classification task. We improved the performance of the top-ranked machine learning-based system from the I2B2 2008 challenge by extending it with our methods. The methods we have developed may improve the performance of other machine-learning based clinical text classification systems. We have released all tools developed as part of this study as open source, available at http://code.google.com/p/ytex.

Acknowledgments

We would like to thank Kyle Ambert whose advice enabled us to reproduce his results on the I2B2 2008 challenge. We would like to thank the Yale University Biomedical High Performance Computing Center (NIH grant RR19895). We would especially like to thank Sujeevan Aseervatham for providing the source code from his entry to the CMC 2007 challenge, which formed the basis of our semantic kernel. We appreciate the feedback from our anonymous reviewers, whose insightful comments greatly improved the manuscript. This work was supported in part by NIH Grant T15 LM07056 from the National Library of Medicine, CTSA Grant Number UL1 RR024139 from the NIH National Center for Advancing Translational Sciences (NCATS), and VA Grant HIR 08-374 HSR&D: Consortium for Health Informatics.

References

1. Uzuner O. Recognizing Obesity and Comorbidities in Sparse Data. Journal of the American Medical Informatics Association. 2009; 16:561–570. [PubMed: 19390096]
2. Pestian, JP.; Brew, C.; Matykiewicz, P.; Hovermale, DJ.; Johnson, N.; Cohen, KB.; Duch, W., editors. ACL. Prague: Proceedings of ACL BioNLP; 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text.
3. Yang, Y.; Pedersen, JO. A Comparative Study on Feature Selection in Text Categorization. Morgan Kaufmann Publishers; 1997. p. 412-420.
4. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23:2507–2517. [PubMed: 17720704]
5. Miller, S.; Guinness, J.; Zamanian, A. Proceedings of HLT. 2004. Name tagging with word clusters and discriminative training; p. 337-442.
6. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. Journal of the American Medical Informatics Association. 2011; 18:557–562. [PubMed: 21565856]
7. Dumais ST. Latent semantic analysis. Ann. Rev. Info. Sci. Tech. 2004; 38:188–230.
8. Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform. 2007; 40:288–299. [PubMed: 16875881]

9. Bekkerman R, El-Yaniv R, Tishby N, Winter Y, Guyon I, Elisseeff A. Distributional Word Clusters vs. Words for Text Categorization. Journal of Machine Learning Research. 2003; 3:1183–1208.
10. Lu, Hsin-Min; Zeng, Daniel; Chen, Hsinchun. Ontology-Based Automatic Chief Complaints Classification for Syndromic Surveillance. Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on. 2006; vol. 2:1137–1142.
11. Bloehdorn, S.; Moschitti, A. Proceedings of the 29th European conference on IR research. Rome, Italy: Springer-Verlag; 2007. Combined syntactic and semantic Kernels for text classification; p. 307-318.
12. Joachims T, Informatik F, Informatik F, Informatik F, Informatik F, Viii L. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. 1997
13. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial InTelligence. 1997; 97:245–271.
14. Ambert KH, Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. J Am Med Inform Assoc. 2009; 16:590–595. [PubMed: 19390099]
15. Farkas R, Szarvas G, Hegedűs I, Almási A, Vincze V, Ormándi R, Busa-Fekete R. Semi-automated Construction of Decision Rules to Predict Morbidities from Clinical Texts. Journal of the American Medical Informatics Association. 2009; 16:601–605. [PubMed: 19390097]
16. Cohen AM. Five-way Smoking Status Classification Using Text Hot-Spot Identification and Error-correcting Output Codes. Journal of the American Medical Informatics Association. 2007; 15:32–35. [PubMed: 17947623]
17. Patwardhan, S. Proceedings of the EACL. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts; p. 1-8.
18. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy; Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995. p. 448-453.
19. Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics. 2011; 44:749–759. [PubMed: 21463704]
20. Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform. 2010
21. National Library of Medicine. UMLS® Reference Manual - NCBI Bookshelf. 2009.
22. Bloehdorn, S.; Hotho, A. Ontologies for Machine Learning. In: Staab, S.; Studer, R., editors. Handbook on Ontologies. Springer; 2009.
23. Aseervatham S, Bennani Y. Semi-structured document categorization with a semantic kernel. Pattern Recognition. 2009; 42:2067–2076.
24. Shawe-Taylor, J. Kernel methods for pattern analysis. Cambridge UK; New York: Cambridge University Press; 2004.
25. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998; 2:121–167.
26. Garla V, Re VL, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, Justice A, Brandt C. The Yale cTAKES extensions for document classification: architecture and application. Journal of the American Medical Informatics Association. 2011
27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010; 17:507–513. [PubMed: 20819853]
28. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27.
29. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011; 18:540–543. [PubMed: 21846785]

Highlights

- Leveraged the structure of the UMLS to rank concepts for text classification
- Used document class labels to define a context-sensitive semantic similarity measure
- Classified clinical text with SVMs using a semantic kernel
- Improved the top machine-learning system from I2B2 2008 challenge

Table 1

Feature-Class contingency table

	Feature Absent D_0	**Feature Present D_1**
Class Y D_y	$D_y \cap D_0$	$D_y \cap D_1$
Class N D_n	$D_n \cap D_0$	$D_n \cap D_1$

Table 2

Comparison of Feature Ranking Methods for the Hypertension Label

Imputed IG feature rank corresponds to that of the parent concept's propagated IG feature rank.

Information Gain			Propagated Information Gain			Imputed Information Gain		
Feature	**Score**	**Rank**	**Concept**	**Score**	**Rank**	**Concept**	**Score**	**Rank**
Hypertension	0.191	1	Hypertension	0.191	1	Hypertension	0.190	1
Htn	0.114	2	Essential Hypertension	0.191	2	…		
Norvasc	0.037	3	Systemic arterial finding	0.066	3	Femoral bruit	0.066	3
Lisinopril	0.034	4	Beta Blocker	0.051	4	…		
Hyperlipidemia	0.033	5	Calcium Channel Blocker	0.047	5	Metoprolol	0.051	4
…			…			…		
Metoprolol	0.022	33	Hyperlipidemia	0.025	36	Norvasc	0.047	5

Table 3

Semantic Similarity Measures

Concepts	Lin	LCS	Imputed IG of LCS	
			Disease	Score
Simvastatin-Pravastatin	0.876	Statin	Hypercholesterolemia	0.441
			Hypertension	0.044
Simvastatin-Norvasc	0.705	Cardiovascular agent	Hypercholesterolemia	0.033
			Hypertension	0.037
Heart failure-Hypertension	0.695	Cardiovascular Diseases	Hypertension	0.007
			Congestive Heart Failure	0.021

Table 4

Macro-Averaged F1 on I2B2 2008 Test Set

Disease	Word	Imputed	Cui	Lin	superlin	Ambert
Asthma	**0.970**	**0.970**	**0.970**	**0.970**	**0.970**	0.970
CAD	**0.631**	0.624	0.621	0.618	0.618	0.630
CHF	**0.615**	0.601	0.607	0.601	0.612	0.612
Depression	0.963	0.976	0.979	**0.982**	0.979	0.935
Diabetes	0.945	0.959	0.957	0.958	**0.960**	0.915
Gallstones	**0.954**	0.950	0.950	0.942	0.950	0.961
GERD	0.578	**0.579**	**0.579**	**0.579**	**0.579**	0.579
Gout	**0.982**	**0.982**	**0.982**	**0.982**	**0.982**	0.981
Hypercholesterolemia	0.901	0.903	0.901	0.896	**0.908**	0.912
Hypertension	0.926	0.920	0.926	**0.929**	**0.929**	0.899
Hypertriglyceridemia	0.902	**0.939**	0.913	0.913	0.928	0.876
Osteoarthritis	0.613	0.608	**0.617**	0.604	0.604	0.631
Obesity	**0.972**	**0.972**	**0.972**	**0.972**	**0.972**	0.973
OSA	**0.659**	**0.659**	**0.659**	0.653	0.656	0.653
PVD	0.586	0.603	**0.606**	0.587	**0.606**	0.623
Venous Insufficiency	0.757	0.788	**0.816**	0.801	**0.816**	0.725
Macro-F1	0.6339	0.6345	0.6354	0.6330	**0.6355**	0.6344
Micro-F1	0.9573	0.9581	0.9593	0.9561	**0.9594**	0.9558

Best scores from our study indicated in bold.

Abbreviations: CAD-Coronary Artery Disease, CHF-Congestive Heart Failure, GERD-Gastroesophogeal Reflux Disease, OSA-Obstructive Sleep Apnea, PVD-Peripheral Vascular Disease.

Table 5

Top Systems from I2B2 2008 Challenge

System	Micro-F1	Macro-F1
Ware	0.9654	0.6404
Szarvas	0.9642	0.6727
Superlin	**0.9594**	**0.6355**
Solt	0.9590	0.6745
Childs	0.9582	0.6696
Yang	0.9572	0.6336
Meystre	0.9566	0.6343
Ambert	0.9558	0.6344
DeShazo	0.9524	0.6292
Matthews	0.9509	0.6288
Jazayeri	0.9508	0.6287

sorted by Micro-Averaged F1