Enhancing clinical concept extraction with distributional semantics

Siddhartha Jonnalagadda a,b,*, Trevor Cohen c, Stephen Wu b, Graciela Gonzalez a

a Department of Biomedical Informatics, Arizona State University, Phoenix, AZ, USA
b Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
c School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA

A R T I C L E  I N F O

Article history:
Received 23 May 2011
Accepted 16 October 2011
Available online 7 November 2011

Keywords:
NLP
Information extraction
NER
Distributional semantics
Clinical informatics

A B S T R A C T

Extracting concepts (such as drugs, symptoms, and diagnoses) from clinical narratives constitutes a basic enabling technology to unlock the knowledge within and support more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment. The recent release of annotated training sets of de-identified clinical narratives has contributed to the development and refinement of concept extraction methods. However, as the annotation process is labor-intensive, training data are necessarily limited in the concepts and concept patterns covered, which impacts the performance of supervised machine learning applications trained with these data. This paper proposes an approach to minimize this limitation by combining supervised machine learning with empirical learning of semantic relatedness from the distribution of the relevant words in additional unannotated text.

The approach uses a sequential discriminative classifier (Conditional Random Fields) to extract the mentions of medical problems, treatments and tests from clinical narratives. It takes advantage of all Medline abstracts indexed as being of the publication type “clinical trials” to estimate the relatedness between words in the i2b2/VA training and testing corpora. In addition to the traditional features such as dictionary matching, pattern matching and part-of-speech tags, we also used as a feature words that appear in similar contexts to the word in question (that is, words that have a similar vector representation measured with the commonly used cosine metric, where vector representations are derived using methods of distributional semantics). To the best of our knowledge, this is the first effort exploring the use of distributional semantics, the semantics derived empirically from unannotated text often using vector space models, for a sequence classification task such as concept extraction. Therefore, we first experimented with different sliding window models and found the model with parameters that led to best performance in a preliminary sequence labeling task.

The evaluation of this approach, performed against the i2b2/VA concept extraction corpus, showed that incorporating features based on the distribution of words across a large unannotated corpus significantly aids concept extraction. Compared to a supervised-only approach as a baseline, the micro-averaged F-score for exact match increased from 80.3% to 82.3% and the micro-averaged F-score based on inexact match increased from 89.7% to 91.3%. These improvements are highly significant according to the bootstrap resampling method and also considering the performance of other systems. Thus, distributional semantic features significantly improve the performance of concept extraction from clinical narratives by taking advantage of word distribution information obtained from unannotated data.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Information contained in clinical records is of value for both clinical practice and research; however, text mining from clinical records, particularly from narrative-style fields (such as discharge summaries and progress reports), has proven to be an elusive target for clinical Natural Language Processing (NLP) [1], due in part to the lack of availability of annotated corpora specific to the task. Yet, the extraction of concepts (such as mentions of problems, treatments, and tests) from clinical narratives constitutes the basic enabling technology that will unlock the knowledge contained in them and drive more advanced reasoning applications such as diagnosis explanation, disease progression modeling, and intelligent analysis of the effectiveness of treatment.

A baseline approach to concept extraction typically relies on a dictionary or lexicon of the concepts to be extracted, using string comparison to identify concepts of interest. Clinical narratives contain drug names, anatomical nomenclature, other specialized...
names and phrases that are not standard in everyday English such as “benign positional vertigo”, “I should inj”, “po pain medications”, “a c5-6 acdf”, “st changes”, “resp status” and others. There is also a high incidence of abbreviation usage, and many of the abbreviations have a different meaning in other genres of English. Descriptive expressions (such as “coil embolization of bleeding vessel”, “a large bloody bowel movement”, a tagged RBC scan and R intracerebral hemorrhage drainage) are commonly used to refer to concepts rather than using canonical terms. The specialized knowledge required and the labor-intensive nature of the task make it difficult to create a lexicon that would include all such expressions, particularly given that their use is often non-standard and varies across institutions and medical specialties, or even from one department to another in the same hospital, rendering dictionary-based approaches less adaptable in this domain. While supervised machine learning approaches offer a promising alternative, a reliable system usually needs a large annotated corpus with as many relevant examples as possible. However, because of privacy concerns and the domain expertise needed to annotate them, large corpora of clinical text are not generally available for research purposes.

This work proposes the design of a semi-supervised machine learning approach for extracting clinical concepts. The idea of combining supervised and unsupervised learning is not without precedent, although prior approaches are not as scalable and portable as the one proposed here. An earlier system in the biomedical domain owned by IBM Watson Research Center [2] used a large corpus of text (5 million Medline abstracts) to improve concept extraction, and achieved the best performance in the BioCreative II gene mention shared task. IBM’s system used a computationally expensive machine learning algorithm called Alternating Structure Optimization (ASO). This limits the extent to which their methods can be utilized in settings where computational resources are more limited. For example, Liu and Ng [3], who applied ASO to Semantic Role Labeling, acknowledge that the extensive computing resources required by ASO limited their experiments. In this paper, we discuss a different approach to the use of unannotated data through construction of a vector-based similarity model using Random Indexing [4], which offers computational advantages over previous methods such as ASO. Our approach is also scalable to large unannotated corpora and performs within the limits of computational resources available to most researchers. Consequently, we anticipate the dissemination of our method will promote widespread use of unannotated data for the task of clinical concept extraction.

The following sections detail our approach, providing first a detailed background on the distributional semantics algorithms we used, followed by a discussion of our methods and results. The methods section elucidates how distributional semantic measures are adapted to concept extraction, what kind of distributional models are most effective, how their usefulness is estimated and how they have been integrated into an existing supervised machine learning system.

2. Background

The problem of extracting the relevant concepts automatically from text is also known as “Named Entity Recognition and Classification (NERC)”, “Named Entity Recognition (NER)”, or simply, “concept extraction”. It has been studied in Computer Science for almost two decades [5], with significant progress in the field. Early approaches to extracting concepts from text tended to use dictionary or rule-based approaches, while later generation systems tend to use variants of supervised machine learning. In this latter category, generative models (Naive Bayes Classifier and Hidden Markov Models) and instance-based classifiers (Logistic Regression and Naive Bayes Classifier) are reasoned to be less accurate for extracting concepts or named entities from text than sequence-based discriminative models such as Conditional Random Fields [6,7]. Thus, machine learning based methods vary on the specific technique or implementation details and the features used for it. Most of the contemporary high-performing tools use non-semantic features such as parts of speech, lemmata, regular expressions, prefixes and n-grams. The high computational cost associated with using deep syntactic and semantic features largely restricted the NERC systems to the orthographic, morphological and shallow syntactic features. Other advances in machine learning, such as active learning and feature selection, still need to be thoroughly explored for concept extraction.

The first application of concept extraction in clinical domain can be attributed to Hirschman et al. [8] for converting a corpus of X-ray reports on patients with breast cancer into a structured database using a theory of sublanguage of grammars [9]. Medical Language Extraction and Encoding System (MedLEE [10]) automatically generates coded information from general clinical notes using a rule-based approach in addition to finding modifiers. As is the case with MedLEE, MetaMap [11], an initiative of NLM to map text to the UMLS Metathesaurus using a large lexicon, has been widely adopted for research purposes. For example, a direct application of MetaMap for detecting medical problems [12] shows that it has an F-score accuracy of 75% for that task. A recent open-source tool HITEX [13] uses MetaMap to map concepts to UMLS strings. More recently, hospital systems are moving forward in developing open source clinical information extraction systems, such as eTAKES [14]. Some other clinical extraction projects include [15–24]. However, as pointed out in a comprehensive review of clinical concept extraction [1], clinical NLP is currently lagging behind because of insufficient exposure to clinical text and rarity of annotated corpora. For a systematic review on automatic clinical coding and concept classification systems, refer to [25].

The focal point of this paper is the use of distributional semantics for concept extraction, and how it can improve other machine-learning based approaches. We first introduce Random Indexing, a scalable way to assign meaningful vectors to words in the unannotated text, before we discuss different ways of empirically inferring the semantics of the words based on their distribution. This discussion covers the distinction between syntagmatic and paradigmatic relations, two different kinds of distributional relationships with different semantic implications. We then discuss methods to model different kinds of paradigmatic relationships.

Distributional semantics is based on the tenet that the semantics of a piece of text (discourse) can be inferred from the distribution of the elements across multiple contexts. Methods of distributional semantics derive measures of semantic relatedness between words and text passages from large bodies of unannotated natural language text (for a review, see Cohen and Widdows, 2009 [26]). These measures of relatedness have been shown to correlate well with human estimates of relatedness; however, little is known about how best to apply these relations to support structured prediction tasks (such as parsing), or sequence labeling tasks such as concept extraction [27,28].

Geometric models of distributional semantics represent each word as a vector in high-dimensional space. The dimensions of this space each correspond to a context in the corpus, such as co-occurrence near another word or co-occurrence within a document. Consequently, the distribution of words is represented as a word-by-context matrix. However, since distributional semantic models constructed from millions of documents and/or millions of words would be unmanageable in size, distributional models approaching corpora of this magnitude tend to reduce dimensionality. Traditional dimensionality reduction techniques such as singular value decomposition (SVD) are computationally expensive.
Distributional models derive representations for words such that words occurring in similar contexts across a set of documents in a corpus will have similar representations. However, the definition of what constitutes a context differs across models. For example, Latent Semantic Analysis (LSA) [33], perhaps the best-established distributional model, generally uses an entire document as a context. In LSA, a word-document matrix is built, and the semantic representation for words and documents is found through singular value decomposition (SVD). This type of context discovers syntagmatic relations by capturing similarity according to word collocation.

In contrast, other models such as the Hyperspace Analog to Language (HAL) [35] use a sliding window that is moved through the text corpus as the context. Two words are said to be in vicinity of each other if and only if the number of words separating them is

---

**Table 1**

Characteristics of various distributional semantic models.

<table>
<thead>
<tr>
<th>Reference(s)</th>
<th>Name</th>
<th>Context</th>
<th>Semantic model</th>
<th>Dim. reduction/inference</th>
</tr>
</thead>
</table>
less than an integer parameter known as the sliding window radius. Hence when a small sliding window is used, this model captures paradigmatic relationships, as two words that can substitute for one another would be expected to occur within similar (if not identical) narrow sliding windows. Sahlgren argues that using a small sliding-window rather than an entire document as a context is better suited to extracting paradigmatic relations, and supports this argument with empirical results. For example, narrow sliding windows are shown to generate associations preferentially between words of the same part of speech, and between synonymous words, when compared to word-document based models or wider sliding windows [48].

### 2.3.1. Directional models

Directional models, such as HAL, use a sliding window to generate a word-word matrix, \( T \), where \( T[i,j] \) is one if the word representing the \( j \)th column appears within the radius of the word representing the \( i \)th column and zero otherwise. Directional models take into account the direction in which one word occurs with respect to the other by having two columns for each word, with one column representing the number of occurrences to the left and the other column representing the number of occurrences to the right. Table 2 shows a matrix computed for “she was increasingly her right side” using a window radius of six words. The sliding window is moved through the entire corpus, such that words occurring in similar sliding windows will have similar (row) vector representations.

The elemental vectors for each word are created using Random Indexing. A random permutation function \( p \), a bijection that maps each element in a collection to a different element, is applied to each elemental vector to obtain a permuted vector. The original elemental vectors of each word occurring to the right in the sliding window are added to the permuted elemental vectors of each word occurring to the left in the sliding window to build the contextual vectors for each occurrence of a word. For example, the contextual vector for the word “using” in the sliding window is calculated as follows:

\[
c(\text{using}) = e(\text{her}) + e(\text{right}) + e(\text{side}) + p(e(\text{she})) + p(e(\text{was})) + p(e(\text{increasingly})),
\]

where \( c(t) \) is the contextual vector for word \( t \) in that sliding window, \( e(t) \) is the elemental vector of word \( t \) and \( p(\nu) \) is the permuted vector of the vector \( \nu \).

The word-word matrix such as the one in Table 2 is used for this calculation. The semantic vector for a word is obtained by adding the contextual vectors obtained at each occurrence of that word and normalizing the sum. A one in the first seven columns represents that the word in the corresponding row is to the left of the word in the column in the sliding window. A one in the last seven columns represents that the word in the corresponding row is to the right of the word in the column in the sliding window.

### 2.3.2. Positional permutation-based sliding window model

Sahlgren’s permutation-based method [45] goes a step further and encodes word-order, thus accounting for the sequential structure of language. Representing the word order may capture some of the grammatical role and hence the meaning of the word. This method is an alternative implementation to the convolution operations used by BEAGLE [49] (based on Tony Plate’s Holographic Reduced Representation (HRR) [50]) to encode word-order information in word spaces. Sahlgren’s method captures word information by permutation of vector coordinates which is a computationally convenient alternative to BEAGLE’s convolution operation, although similar performance improvements have recently been achieved using HRR in the frequency domain [51].

To achieve this, Sahlgren et al. first randomly generate sparse high-dimensional elemental vectors for each word, and use permutation, specifically shuffling of coordinates (shifting of all of the non-zero values of a sparse elemental vector to the left or right according to the relative position of words) to replace the convolution operation. In this way, a different close-to-orthogonal elemental vector is generated for each word depending on its position within the sliding window. A semantic word vector for each word is then generated as the linear sum of the permuted elemental vectors for each word co-occurring with this word in a sliding window. An example is shown in Fig. 1 below. This permutation function is bijective (for every member in the range, there is exactly one element from domain mapped to it and vice versa), allowing for construction of order-based queries (such as a query to find words occurring one position to the right of the word “president”).

### 3. Methods

The purpose of our work is to apply the methods of distributional semantics we have discussed to enhance the performance of machine learning approaches to concept extraction. For concept extraction using supervised machine learning methods, the main challenges are the scarcity of annotated examples, the level of expertise required to generate these examples, and the fact that no such large corpus of clinical text can practically be shared without raising privacy concerns. Harris’ sublanguage theory [52] suggests that the constraints of language in a specialized domain such as that used in clinical narratives include domain semantics in addition to purely grammatical constraints, suggesting the applicability of a semantic grammar in which domain semantics are an integral component of syntax. In highly constrained genres of texts such as biomedical literature and clinical narratives, there are inherent inequalities of the likelihood of certain patterns of words occurring [53]. As a corollary, high probability combinations convey implicit information. For example, sentences such as Patient \( X \) received drug \( Y \) are more common than sentences such as Doctor \( X \) received drug \( Y \). Thus, a sentence pattern \( A \) received \( B \) could be decoded to assume that \( A \) might refer to patient and \( B \) might refer to treatment. Therefore, and this is the premise of our approach, it is

### Table 2

Example for directional model matrix computation.

<table>
<thead>
<tr>
<th>Focus word</th>
<th>Occurrence within six words following the focus word</th>
<th>Occurrence within six words preceding the focus word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>She</td>
<td>Was</td>
</tr>
<tr>
<td>She</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Was</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Increasingly</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Using</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Her</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Right</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Side</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
likely that information derived from the distribution of words in a larger, unannotated corpus could be used to compensate for the limited vocabulary present in a smaller annotated corpus and allow for more accurate concept recognition. We hypothesize that the concept extraction task can be seen as finding a class of words that could conceivably represent the word to be labeled without disturbing the surrounding syntactic structure. To evaluate this hypothesis, distributional semantic features are added in this work to commonly used features (Table 2) which were then used as a basis for supervised machine learning to extract medical problems and treatments from clinical narratives. Improvements in accuracy after adding distributional semantic features (using i2b2/VA NLP shared task corpus as gold standard) would validate the utility of these additional features.

Building a sliding-window based model involves:

- Constructing elemental word vectors of pre-determined dimension \(N\) and seed \(S\), where \(N - 2 + S\) dimensions are zeros, \(S\) dimensions are +1s, and \(N\) dimensions are −1s. To ensure that these elemental vectors have a high probability of being mutually orthogonal, or close-to-orthogonal, \(S \ll N\).
- Computing the semantic vector representation of the word based on the words surrounding each occurrence of the word.

This process is performed using the open source Semantic Vectors package [54,55], which supports the generation of basic, direction-aware and permutation-based sliding window models using Random Indexing for dimension reduction.

Medline abstracts indexed as clinical trials help in creating semantic vector spaces. Different distributional semantic algorithms with different parameters respectively are employed to create various models. The best model, determined by their ability to discern words from different categories, is used to generate a similarity matrix for the words or tokens present in the corpus. The CRF algorithm uses features such as lexicons and linguistic features from the corpus, in addition to the distributional semantic features. The CRF model created during the training phase is used to tag the input sentences with the concepts “medical problem”, “treatment” and “test”.

As shown in Fig. 2, features from a state of the art biomedical NER system, BANNER [56], were adapted to create the baseline. BANNER uses the Conditional Random Fields algorithm implemented in MALLET [57]. However, any concept extraction system using a supervised machine learning algorithm can easily adopt the proposed set-up. We trained BANNER on i2b2/VA NLP training corpus using (a) sentence-level features (last two rows in Table 2); (b) a lexicon derived from the UMLS, DrugBank, Drugs@FDA and MedDRA knowledge resources; and (c) distributional semantic features based on a large unannotated corpus of Medline abstracts categorized as “clinical trials” (\(n = 447,000\), considering those would be the abstracts with vocabulary closer to clinical narratives. Selecting the distributional semantic features required tuning the parameters and finding the better model among different distributional semantic models, based on performance on a related task that will be described in detail in sections to follow. We used the best model with parameters that led to best performance in the preliminary sequence labeling task to create a similarity matrix for adding distributional semantic features. After the systems are trained and tuned offline and a CRF (Conditional Random Fields) machine learning model was created, processing of a sentence was accomplished using the same set of features used for training.

3.1. i2b2/VA NLP shared task corpus

To the best of our knowledge, the i2b2/VA NLP shared task corpus is the only public corpus of deidentified clinical documents that has annotation for medical problems, treatments and tests. A complete detail of the corpus is made available by Uzuner et al. [58], especially in Table 1. The corpus consists of 394 documents for training and 877 documents for testing. All the documents are either discharge summaries from Partners Healthcare, Beth Israel Deaconess Medical Center and the University of Pittsburgh Medical Center or progress reports from the University of Pittsburgh Medical Center. Overall, there are 20,267 test named entities, 22,056 treatment named entities and 30,517 problem named entities. The overall Inter-Annotator agreement distributed uniformly across the three types of concepts is 85% for exact match and 91% for inexact match. The annotation guidelines used, the evaluation metrics that were suggested (followed in this paper) as well as three example documents along with the annotations are available at http://www.i2b2.org/NLP/Relations/Documentation.php.

3.2. Automatic word categorization test

The 447,000 Medline abstracts that are indexed as pertaining to “clinical trials” are used as unannotated data. We used Semantic Vectors software [54] built on the top of Apache Lucene [59] to find the vector model of each word in the abstracts that appears at least twice in the entire unannotated corpus and has not more than three non-alphabet characters. We are not using a stop word list. The different paradigmatic model algorithms considered were the positional model, the directional model and the positional + basic model. The positional + basic model is the combination of the positional and basic sliding window models where the semantic vector of a word is the normalized vector sum of the corresponding semantic vectors in the positional model and the basic model.

We can customize the following parameters of the distributional semantic model: (a) dimensions of the final vector space, (b) seed length, and (c) the sliding window radius. We used the following methodology for deciding among parameter values:
134

Fig. 2. Overall architecture of the system.

- Experiment with dimensions of sizes 100, 500, 1000, 1500 and 2000 to choose the dimensions that resulted in better categorization with respect to semantic type.
- Experiment with seeds of sizes 5, 10, 15, and 20; chose the seed length that resulted in better categorization with respect to semantic type.
- Experiment with window radii 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 separately for each paradigmatic model algorithm; and chose the window radius that resulted in better categorization with respect to semantic type.

In order to select parameters, we tested the distributional model on the related task of automatic word categorization test before annotating all the clinical concepts As our system has limited training examples, it will need to infer the semantic type of terms that have not been encountered in the training set, in order to determine whether or not they should be extracted. The purpose of including distributional information is to provide a basis for this inference. We would expect that the choice of parameters that leads to accurate classification with respect to semantic type would best support concept extraction. Consequently, we have chosen to explore the parameter space in the context of a related task so that we do not over fit the parameter selection for the specific task.

To model paradigmatic relationships, we choose the model that performed best in the experiments in the automatic word categorization test. This test is similar to part-of-speech test performed by Sahlgren [48] to study the differences between different wordspace models. Sahlgren studied the extent to which different distributional models retrieved nearest neighboring words that shared a part of speech with a cue word. He used two evaluation metrics that he referred to as “strict” and “lax”. Both of them are variants of the k-nearest neighbors algorithm – strict has k = 1 (i.e., the closest neighbor has the same part of speech) and lax has k = 10 (i.e., at least one of the ten closest neighbors has the same part of speech).

We divided all the UMLS single word phrases into four categories: problem (n = 122,264), treatment (n = 189,283), test (n = 19,149) and none (n = 476,424) according to their UMLS semantic type. According to the definitions from the 2010 i2b2/VA task guidelines,1 a word is a:

1. Medical problem if its UMLS semantic type is pathologic function, disease or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, congenital abnormality, acquired abnormality, injury or poisoning, anatomic abnormality, neoplastic process, virus/bacterium, sign or symptom.
2. Medical treatment if its UMLS semantic type is therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, or drug delivery device.
3. Medical test if its semantic type is laboratory procedure or diagnostic procedure.

We performed the automatic word categorization test for each of the paradigmatic models discussed in the previous section. For each model, we found the 10 nearest neighbors for each word and removed the word itself from the nearest neighbor list. Each of the 4 categories had a weight equal to the sum of the cosines of the neighbors that belong to the category. We assigned the category with the highest score to the word. The evaluation compares this automatically determined word category to the actual word category. For example, thoracotomy has 6 “treatment” neighbors (sum of cosines = .79 + .86 + .82 + .79 + .81 = 4.89) and 3 “test” neighbors (sum of cosines = .82 + .80 + .82 = 2.44). Hence, the test classified (correctly) thoracotomy as treatment. Thus, we find whether the assignment is a true positive, false positive, false negative, or true negative with respect to each category. In each case, we recorded the individual recall and precision as well as micro-averaged recall, precision, F-score and accuracy.

### 3.3. Distributional semantic features for clinical concept extraction

We now constructed the similarity matrix K over the words in the i2b2/VA corpus, where

\[
K(w_1, w_2) = \begin{cases} 
\text{cosine of the semantic vector representations of} & \\
\text{ exist in the i2b2/VA corpus} & \text{if both the words } w_1 \text{ and } w_2 \\
\text{Zero, otherwise} & \text{exist in the i2b2/VA corpus} 
\end{cases}
\]

The purpose of the similarity matrix is to build a thesaurus of words automatically. Each entry in the thesaurus consists of a token from the i2b2/VA NLP corpus and N most similar words.
(from the i2b2/VA corpus) based on the distributional semantics. Computing the similarity matrix \((K)\) scales linearly in the number of dimensions of each vector and quadratic in the number of words in the i2b2/VA corpus. Computing the thesaurus (with a pre-determined number of similar words from the similarity matrix) scales linearly with the number of words. Overall, our approach scales linearly with respect to the number of dimensions of each vector and quadratic in the number of word vectors. Using the pre-computed matrix instead of directly computing thesaurus online saves the time spent in computing the cosines \((O(N)\) time complexity) during the construction of the thesaurus.

We used a 1st order Conditional Random Fields (CRF) algorithm, as implemented by MALLET [57]. The time complexity of the CRF algorithm is \(O(L^2N + M + F)\), where \(L\) is the number of labels, \(N\) is the number of sequences (sentences), \(M\) is the average length of the sequences, and \(F\) is the average number of the features. It is observed [56] that the accuracy is almost the same for all label types such as – IO, IOB and IOBEW, where \(I\) stands for labeling a token to be Inside, \(O\) for Outside, \(B\) for Beginning, \(E\) for End and \(W\) for Within. We chose the IO notation for labeling to minimize time complexity. Thus, we used four labels – Iproblem, Itest, Itreatment and O since there are three annotated concepts. In addition to all the features used in BANNER [56], three additional feature types based on: (1) thesaurus, (2) vector representation of the token, and (3) dictionaries are added (see Table 3).

### 4. Results

#### 4.1. Automatic word categorization results

The evaluation metrics, as defined in Section 3.2, consist of micro-averaged precision, recall, \(F\)-score and accuracy. Both \(F\)-score and accuracy were measured while adjusting each parameter discussed in Section 3.2. The advantage of using \(F\)-score is that in certain concept extraction tasks, true positives are more important than true negatives. The advantage of using accuracy is that it is invariant to label switching, which means the value does not change when a label changes from positive to negative. Thus, the goal was to optimize both \(F\)-score and accuracy. Figs. 3–8 show the results of the experiments. We used the positional vector model as baseline to determine the better number of dimensions and seed length to serve the parameters for the Random Indexing algorithm. These parameters are general to the application of Random Indexing for the different models in this task. Once these desirable parameters were found, we used them for evaluating different sliding window algorithms against each other.

![Fig. 3. Number of dimensions vs. performance. The accuracy and \(F\)-score of automatic word categorization initially increases but does not increase dramatically after 1000 dimensions.](image)

![Fig. 4. Number of seeds vs. performance. The accuracy and \(F\)-score of automatic word categorization do not change with the number of seeds used in the model.](image)

![Fig. 5. Window radius in positional model vs. performance. The performance is the highest at the window radius of 2 corroborating with results from similar experiments by Sahlgren [45].](image)

![Fig. 6. Window radius in directional model vs. performance. The performance is the highest at the window radius of 6.](image)

Table 3

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary</td>
<td>Semantic</td>
<td>UMLS, DrugBank, Drugs@FDA and MedDRA</td>
</tr>
<tr>
<td>Distributional</td>
<td>Semantic</td>
<td>Distributional thesaurus and dimensions of words (word embeddings)</td>
</tr>
<tr>
<td>Section</td>
<td>Pragmatic</td>
<td>Name of the section in which the sentence appears</td>
</tr>
<tr>
<td>Part-of-speech</td>
<td>Syntactic</td>
<td>Part of speech of the token in the sentence</td>
</tr>
<tr>
<td>Others</td>
<td>Lexical</td>
<td>Lower case token, lemma, prefixes, suffixes, (n)-grams, Matching patterns such as beginning with a capital, etc.</td>
</tr>
</tbody>
</table>

These are the different features used in the machine learning algorithm. Roughly, the features could be lexical (at the level of word), syntactic (at the level of grammar), pragmatic (at the level of context) and semantic (at the level of meaning).
F-score beyond 1000 dimensions were only gradually increasing and could be considered almost constant (Fig. 3). The accuracy and F-score were constant as the number of seeds was increased (Fig. 4). Therefore, we used five seeds.

We then changed the window radius for the three models – the positional permutation vector model, the directional permutation vector model, and the positional + basic model. As expected, the performance varies differently in each case with increase in the window radius. The positional model had the highest accuracy and F-score at a window radius of 2 (Fig. 5). This corroborates similar experiments by Sahlgren [45, 48]. For directional model, the performance increased until we increased the window radius to 6 and it decreased after that (Fig. 6). The positional + basic model performed better than the positional model, and the system achieves the best accuracy and F-score at the window radius of 4 (Fig. 7). As the word categorization test can be considered as a test of our models’ ability to find associations between words that serve the same semantic role (paradigmatic associations), we decided on the basis of these evaluations to obtain a better model for concept extraction.

4.2. Concept extraction results

Two of the parameters involved in use of the model that lead to the best performance in the previous section are: (a) the number of most similar words to consider, and (b) the minimum similarity between the similar word and the original token. The experiments showed that the best number of most similar words to use is 20. With a threshold on the cosine similarity between the similar words and the original token, the performance decreased slightly. Consequently, the threshold was set to 0, i.e., we used all the 20 nearest words irrespective of their cosine value. Table 4 compares the results using different models with these settings.

Table 5 shows the results for a system without any distributional semantic features and that of the best system with distributional semantic features (direction-based, 2000 dimensions, 5 seeds and window radius = 6). Both were trained on the competition training corpus and tested on the testing corpus.

It is encouraging to see that addition of distributional semantic features increases the recall and precision in all cases. Bootstrap Resampling [60] with 1000 repetitions on the test corpus showed the improvement after adding the distributional semantic features derived from clinical trial abstracts is highly significant (confidence = 100%). This metric, which involves calculating f-scores over different versions of corpora obtained by allowing repetitions over original corpora and averaging the f-scores, was also used in comparing state of the art systems that extract concepts in biomedical literature [61] and general English [62]. However, the improvement because of adding local distributional semantic based features after adding distributional semantic features from clinical trials was insignificant (confidence = 56.6%). Hence, we might also conclude that addition of distributional semantic features derived from large unannotated corpus is sufficient and features from smaller corpora need not supplement it. We also found that using each dimension of the vector representation of tokens as features does not significantly affect the accuracy of the system. This could be due to curse of dimensionality, which exponentially increases the size of the training set needed based on the number of features used. The distributional semantics model for Medline and the thesaurus of distributionally similar words for the words in the i2b2/VA corpus are available at http://diego.asu.edu/downloads/AZCCE/, as well as is the code to generate a thesaurus and integrate this with a CRF system.

Further, we asked clinical researchers at the Mayo Clinic, Rochester and the University of Texas Health Sciences Center at Houston to use 1.5 million clinical narratives each and report the evaluation results using the i2b2/VA shared task testing corpus. Both the centers reported significant increases in the accuracy of concept extraction (overall F-scores of 82.0% and 82.3% for exact

Table 4

Comparison of concept extraction performance with different models.

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>Exact R</th>
<th>Exact P</th>
<th>Exact F</th>
<th>Inexact R</th>
<th>Inexact P</th>
<th>Inexact F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 Non distributional semantic features</td>
<td>78.10</td>
<td>82.65</td>
<td>80.31</td>
<td>88.66</td>
<td>90.84</td>
<td>89.74</td>
</tr>
<tr>
<td>2</td>
<td>2 + Distributional Semantic features using i2b2 corpus</td>
<td>78.71</td>
<td>83.20</td>
<td>80.89</td>
<td>89.02</td>
<td>91.27</td>
<td>90.13</td>
</tr>
<tr>
<td>3</td>
<td>3 - Position-based features using Medline clinical trials</td>
<td>79.85</td>
<td>83.74</td>
<td>81.75</td>
<td>89.84</td>
<td>91.61</td>
<td>90.72</td>
</tr>
<tr>
<td>4</td>
<td>4 - Direction-based features using Medline clinical trials</td>
<td>79.92</td>
<td>83.50</td>
<td>81.67</td>
<td>90.26</td>
<td>91.63</td>
<td>90.94</td>
</tr>
<tr>
<td>5</td>
<td>5 + Position + basic based features using Medline clinical trials</td>
<td>79.81</td>
<td>83.55</td>
<td>81.64</td>
<td>90.05</td>
<td>91.61</td>
<td>90.83</td>
</tr>
</tbody>
</table>

The highest scores are in bold. R_exact is the micro-averaged recall for exact match. P_exact is the micro-averaged precision for exact match. F_exact is the micro-averaged F-score for exact match. R_inexact is the micro-averaged recall for inexact match. P_inexact is the micro-averaged precision for inexact match. F_inexact is the micro-averaged F-score for inexact match. The direction-based model has the highest impact.
match and 91.3% and 91.3% for inexact match respectively. An increase in F-score of 2% over the state of the art (80.3%) is considered very significant.

5. Discussion

Through this work on concept extraction, we amalgamated two diverse research areas – distributional semantics and information extraction. Firstly, we evaluated different distributional semantic models based on their ability to predict the semantic types of words. The parameters we found to be better might be useful in future experiments by other researchers. A novel finding of this work is that semantic vector models are helpful in extracting named entities. The method proposed here using distributional semantic achievements a performance improvement of 2.0% in the task of extracting medical problems, tests, and treatments from clinical narratives when measured against a state of the art system. To be able to generalize the conclusion, we performed a similar experiment for biomedical literature [63]. We used the settings found to yield better vector representation using UMLS (Section 3.2) to generate semantic vector representation for the words in the BioCreative II task [61] corpus, one of the most commonly used corpus for benchmarking gene tagging. These semantic vectors are used to generate distributional semantic features which are integrated into BANNER [56], one of the best gene tagger [64]. The F-score of gene tagging on the BioCreative training and testing set increased by 1.9%. This suggests that the approach proposed in this work to use unlabeled text for concept extraction could be adapted to a different domain.

5.1. Comparison with other systems

Our evaluation showed that the accuracy of the system significantly improves after adding distributional semantic features. However, how much it contributes toward improving the state of the art determines the practical significance of the improvement. Thus, we compared the performance of the system to the top systems in the i2b2/VA 2010 concept extraction task in which 22 international teams have submitted multiple runs as shown in Fig. 9. Our baseline system (without distributional semantic features) ranked 7th both in F-scores measured using exact match and inexact match. The best system uses a relatively complex variant of Hidden Markov Models algorithm. Systems with ranks 2–5 use proprietary (non-public) components. System 6 uses Conditional Random Fields algorithm (as does our baseline). After adding distributional semantic features as described in this paper, the F-score as measured by inexact match is better than systems 3–6 and equal to system 2. The F-score as measured by exact match is also better than systems 3–6. This suggests that adding distributional semantic features to a supervised machine learning system significantly improves the state of art.

In addition, this system is the only one among this list that adds a semi-supervised component on the top of an existing supervised system. Our semi-supervised component (distributional semantic features) can integrate with other machine learning algorithms such as Hidden Markov Models (used by NRC, the best system submission for i2b2/VA concept extraction task) and Naïve Bayes. This may lead to improved performance with these systems also.
models that incorporate both symbolic and distributional information (for example [65]) attempt to address this limitation. Another limitation in using distributional semantics is that the meanings of words often differ with the passage of time and for specialized disciplines. This phenomenon is semantic change or semantic shift. For example, the word “guy” was originally used to mean “grotesque person” (as named after Guy Fawkes hanged for an assassination attempt in 1605), and now is a common expression for a male person. The meaning of the word “case” changes with specialization [65]. In the general domain, it means “circumstance”, while in the medical domain, it means “a patient”, and in the legal domain, “a lawsuit”. The reliability of the empirically derived semantics needs to be analyzed in order to determine the extent to which domain specificity and temporal synchronization of the training corpus affect results. Further investigation of this effect is beyond the scope of the current paper (however, see for example [66,67]).

5.4. Future work

Automatic coding of concepts in clinical narratives: Since systems for extracting concepts are still being explored, automatic coding is an emerging technology. In text mining pipeline, there is a further step after extracting concepts known as normalization. This step involves disambiguating polysemy and synonymy. Polysemy is the event of the same entity having different semantic meanings in two different contexts. For example, PPI could mean Inorganic Pyro Phosphate when the subject of discussion is about enzymes; it could mean Peptideyl prolyl cis–trans-isomerase when the discussion is about proteins; it could also mean protein–protein interaction in other situations. The first step in normalization will be to assign different senses to polysemous entities. Synonymy is the event of two different entities having the same semantic meanings in all contexts. For example, “AD” and “SDAT”, and “poliomyelitis” and “polio” are synonyms. Automatic coding systems not only need
concept extraction, but also require normalization using different schemas, such as the International Classification of Diseases (ICD) or Current Procedural Terminology (CPT), and use-case-specific, non-standardized schemas, such as the presence or absence of a given condition. The entities in the schemas could be pre-coordinated (coding as a single field) or post-coordinated (reassembled after storing individual components separately). Two problems associated with pre-coordination [68] are: (a) updates to the cross index could be slow because of the knowledge resources needed and (b) computation ambiguity of the reuse of the concept. Given the ability of distributional semantics to detect words appearing in similar context and to predict the appropriateness of a concept to a context, it is possible that we could automatically address both these problems. Distributional semantics may also help in automatically mapping “pre-coordinated” terms with “post-coordinated” expressions and vice versa.

Secondary use of clinical narratives: Hospital systems could use the clinical narratives to create a conceptual thesaurus for the words in the narratives automatically. This thesaurus would be useful for annotating concepts in a machine learning system or a rule-based system. This thesaurus would be more useful for annotating the clinical narratives within the hospital system, since it is customized for that genre of text. However, as this distributional thesaurus is created without human intervention, the system could be conveniently transferred to a new institution given the availability of free-text records in this domain.

6. Conclusion
The approach we have used in this research is agnostic to the specific machine learning algorithm that is used to recognize entities in text once distributional semantic features have been generated. Other semi-supervised approaches such as ASO [2] tend to be restricted to a specific framework and algorithm. This is disadvantageous considering most state-of-the-art concept extraction systems use supervised machine learning algorithms, such as CRF, and it is of interest to the scientific community to determine the upper limits of performance these algorithms can achieve. Our approach was to use NLP features generated through unsupervised means within a supervised machine learning system. While this approach renders all the advantages offered in other semi-supervised machine learning systems, it also permits maximum flexibility in the choice of the basic framework of the set-up. In sum, the results indicate that distributional semantic features aid clinical concept extraction. In our current research, we are exploring the use of clinical records, rather than Medline records, as a source of unannotated data. We anticipate further increases in accuracy, since these corpora contain the language used by physicians in practice.

Acknowledgments
We thank NLM for making us available the Medline abstracts to create semantic resources and for partially funding all authors for this work (Contract# HHSN27620100003C1, Elmer V Bernstam’s group at the School of Biomedical Informatics at University of Texas Houston (NCRR Grant 3UL1RR024148, NCRR 1RC1RR028254, NSF 0964613 and the Brown Foundation) and Mayo Clinic for testing the process using their respective clinical notes, the developers of BANNER, MALLET and Semantic Vectors for the software packages and the organizers of the i2b2/VA 2010 NLP challenge for sharing the corpus.

Appendix A. Supplementary material
Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.10.007.