Published in final edited form as:
J Biomed Inform. 2009 April ; 42(2): 390–405. doi:10.1016/j.jbi.2009.02.002.

Empirical Distributional Semantics: Methods and Biomedical Applications

Trevor Cohen[1] and **Dominic Widdows**[2]

[1] Center for Decision Making and Cognition. Department of Biomedical Informatics, School of Computing and Informatics, Arizona State University

[2] Google, Inc

Abstract

Over the past fifteen years, a range of methods have been developed that are able to learn human-like estimates of the semantic relatedness between terms from the way in which these terms are distributed in a corpus of unannotated natural language text. These methods have also been evaluated in a number of applications in the cognitive science, computational linguistics and the information retrieval literatures. In this paper, we review the available methodologies for derivation of semantic relatedness from free text, as well as their evaluation in a variety of biomedical and other applications. Recent methodological developments, and their applicability to several existing applications are also discussed.

Keywords

Distributional semantics; methodological review; latent semantic analysis; natural language processing

1.1 Introduction

Within the field of biomedical informatics, many applications are supported by representations of knowledge constructed by humans, such as ontologies and controlled vocabularies. These methods are able to represent meaning, or semantics, in a manner that is sufficiently precise to support a range of computational applications including information retrieval, information extraction, data-mining and rule-based reasoning for clinical decision support. However, the construction and maintenance of these representational models is time-consuming and demands a great deal of human effort. In addition, these constructs are often insensitive to context and may not correspond to the way in which clinicians and health care consumers understand concepts in this domain [1]. This paper reviews complementary methodological approaches to the representation of meaning, in which the meanings (or semantics) of terms within a domain are determined empirically from the way in which these terms are distributed across a large body of domain-relevant text. From these distributional statistics, it is possible to derive meaningful estimates of the semantic similarity (or closeness in meaning) between

Communicating Author: Trevor Cohen, Department of Biomedical Informatics, Arizona State University, 425 N. 5th Street, Phoenix, AZ 85004-2157, 602-827-2530.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

terms in an unannotated corpus of text without human intervention. The nature of these similarities is often better determined by other means. For example, pattern-based extraction has been used to identify specific types of relationships [2], and relationship extraction methods that draw on a domain-specific knowledge resource such as a semantic grammar [3] or the semantic relations of the Unified Medical Language System (UMLS) [4] have also been applied to this problem in the biomedical domain. In contrast to these methods for the extraction of specific types of relationships, distributional methods provide a quantitative estimate of the semantic similarity between terms. While these methods do not provide the precise formal definitions contained in an ontology, they have been successfully applied to a range of problems that require a general measure of semantic similarity between terms or whole passages of text. They have also proved useful in several biomedical applications, and show promise as a method to support ontology construction and customize constructed knowledge resources for particular tasks. Recent advances in the implementation of these methods have reduced the processing time and computational power required to conduct distributional semantics research, removing what had been a primary barrier to researchers with limited resources, and making it possible to rapidly prototype new models for the purpose of exploratory research. In this article we discuss the major methodological approaches to deriving semantic similarity from unannotated free text, as well as important general and biomedical applications of these methods. In addition, recent methodological advances and some of the tools available with which to conduct this research are also discussed.

In the rest of this introduction, we outline some of the important principles and variants of distributional semantic models, consider the historical influence of information retrieval and neural network models and outline some of the contrasting and complementary properties of rule-based methods. In section 2, we describe some of the main applications of distributional semantics including document retrieval, word-sense disambiguation, bilingual information extraction and visualization of relationships between terms. Section 3 discusses applications in the biomedical domain, with a focus on applications that are based on an automated measure of textual similarity. Some applications in modeling gene and biological sequences are discussed, but only those where the methodology falls within the scope of the article as a whole. Section 4 discusses some recent advances in distributional models, and Section 5 describes some of the currently available software packages that provide practical implementations of distributional models.

1.2 Methodological Approaches

Over the past fifteen years, a number of methods have emerged that enable computers to derive an automated measure of the semantic similarity between terms from the distributional statistics of electronic text. The first of these models were spatially motivated, such as Lund and Burgess' Hyperspace Analogue to Language (HAL) [5], Schütze's Wordspace [6] and Landauer and Dumais' Latent Semantic Analysis (LSA) [7], representing terms as vectors in a high-dimensional space. These were followed by probabilistic models such as Hofman's Probabilistic LSA (pLSA) [8], Blei and colleagues' Latent Dirichlet Allocation (LDA) [9], and Griffith's and Steyvers' Probabilistic Topics Models [10] These probabilistic models view documents as mixtures of topics, allowing terms to be represented according to the likelihood of their being encountered during the discussion of each topic. Both of these methodological approaches allow for the estimation of the similarity between terms: spatial models compare terms using distance metrics in high-dimensional space, while probabilistic models measure similarity between terms according to the degree to which they share the same topic distributions.

The following similarities are derived from the freely available Oregon Health Sciences University MEDLINE (OHSUMED) [11] corpus of MEDLINE abstracts using the Infomap

NLP software package [12], and give an example of the sort of semantic relations these models can infer from the distribution of terms in text. In this example, the nearest neighbors of (or most closely related terms to) the term “psychosis” have been extracted (Table I).

The neighboring terms refer to a range of concepts that are related to the term “psychosis”, such as symptoms of psychosis (“paranoid”, “psychotic”, delusions”, “hallucinations”), treatment for psychosis (“antipsychotic”), side-effects of this treatment (“parkinsonism”, a common side-effect of anti-psychotic drugs) and other psychiatric or neurological disorders (“mania”, “headaches”, (possibly panic) “attacks”). In the following section we will explain some of the common ways in which such semantic similarities can be derived from unannotated free text.

1.2.1 Spatial Models

Spatial semantic representations define terms as vectors in a high-dimensional space, according to the frequency with which they occur within a particular context. Different approaches to generating this multidimensional space differ in their definition of what constitutes a context. In LSA [8], each document in a text collection is considered as a context. The HAL model [5] uses as its context a smaller neighborhood of words surrounding the target term. The matrix generated is a term-term matrix rather than a term-document matrix, and term frequencies are calculated according to the degree with which they co-occur within a sliding window, usually consisting of a small number of terms only. In contrast, Schütze’s Wordspace [6],[13] defines a sliding window of around 1000 frequently-occurring four-grams (combinations of characters such as “psyc” and “osis”) as a context, resulting in a term-four-gram matrix. Other approaches incorporate word order information [14] or dependency relations produced by a parser [15]. While these approaches differ somewhat in their choice of context, they are all based on the idea of representing terms as vectors (or ordered lists of coordinates) within a high-dimensional space according to the frequency with which they occur within a set of defined contexts. Certain transformations, or weighting functions, when applied to the per-context count for terms have been shown to improve the accuracy of term-term associations when using these methods. These are of two types: local weighting functions, which are based on the occurrence of a particular term within a particular context, and global weighting functions, which are based on the occurrence of a particular term across the entire set of contexts. Statistical weighting techniques have been shown to improve the accuracy of both term-term associations [16] and document retrieval when LSA is used for this purpose [17]. A number of other statistical weighting functions have proved useful in other distributional semantics implementations (for example [18]).

The term-by-context matrices produced by all of the methods we have discussed up to this point are often large. For example, the Touchstone Applied Science Associates (TASA) corpus which has been used frequently in LSA research contains 37,600 documents. Consequently, each term vector in the initial term-by-document matrix will have 37,600 dimensions. A matrix of this size carries significant computational and storage overhead, particularly when performing tasks such as nearest neighbor search which are commonly used to assess the accuracy of term-term associations. Aside from the issue of computational overhead, dimension reduction has been shown to improve the accuracy of term-term associations considerably in LSA applications [7]. One reason for this is that with most commonly employed similarity metrics, terms will only be viewed as similar if they occur in the same document in the full-term document matrix. A reduced-dimensional matrix is obtained by projecting the terms in the original term-by-document matrix into a smaller (100–300 dimensions in most LSA applications) space, while approximately preserving the distance between them. Within this reduced-dimensional space, it is possible for a term to be viewed as similar to another term if they occur in similar contexts, regardless of whether or not they appear together. In most

LSA and related applications, dimension reduction is performed using the Singular Value Decomposition (SVD), an established technique of linear algebra. SVD can be viewed as a multidimensional analogue of finding the approximate best fit line to points on a plane: a reduced-dimensional representation that best approximates the original data is generated. A detailed discussion of the SVD is beyond the scope of this article, but can be found in most linear algebra textbooks, for example Strang [19]. For the purposes of this discussion we note that (1) the SVD produces the reduced-dimensional matrix that best captures the variance between points in the original matrix, and (2) this computation carries significant computational and memory overhead. Recently, Random Indexing [16], [20] (see section 5.1 for a detailed description) has emerged as a scalable alternative to models of distributional semantics that depend on the SVD, supporting the derivation of semantic distance from large corpora at minimal computational expense.

1.2.2 Probabilistic Methods

The early successes of methods such as LSA and HAL led more researchers to consider using latent variable models to represent semantic content learned from distributional data, many of which used primarily probabilistic rather than geometric approaches. Some of the traditional applications of distributional language modeling use a number of statistical and probabilistic methods, such as using clustering for automatic thesaurus construction [21]. The idea that latent variables in statistical models could also correspond to semantic concepts underlies probabilistic Latent Semantic Analysis (pLSA) [8]. In pLSA these latent variables represent topics, with the probability of a word being associated with a particular topic, and the probability of a particular document referring to a given topic related under a probabilistic model. The probability of a word occurring in a given document can then be defined in terms of the probability of this word occurring when a particular topic is discussed and the probability that this topic is a subject of the document under consideration. These two probability distributions (term-topic and topic-document) can be estimated from term-by-document statistics across a corpus of text. Discussion of the algorithms used to estimate these distributions requires considerable background knowledge of Bayesian methods of parameter estimation, and as such falls outside the scope of this review. For further details on these methods, the interested reader is referred to Heinrich’s introduction [22], which covers the prerequisite background knowledge as well as the algorithms that underlie pLSA and similar models. Subsequent models to pLSA such as Latent Dirichlet Allocation (LDA) [9], and the Probabilistic Topics Model [23] add to this approach by introducing prior probabilities on either one or both of the term-topic and topic-document distributions. The introduction of a prior probability distribution before estimating parameters from the data tends to result in a distribution with a similar form to this prior distribution, which allows for the incorporation of prior knowledge (for example of term-document distribution) and prevents over-fitting to the data set used to train the model. The Topics Model work is further distinguished by the algorithm (Gibbs sampling rather than Expectation Maximization) used to train the model. While we are not aware of a published evaluation comparing the computational efficiency of these algorithms to one another (or to LSA) in the context of parameter estimation from text, in our experience the use of Gibbs Sampling in the Topics Model appears to offer some advantage in scalability over previous probabilistic models. However, the computational demands of both of these methods limit their applicability to large data sets.

From a theoretical point of view, it is claimed that probabilistic models have two principle advantages over geometric approaches such as LSA. First, it is claimed that the topic variables have a semantic interpretation that is lacking in the basic coordinate axes produced by matrix methods like SVD. High quality clusters of topical words are cited in support of this claim, though comparable results may result from other distributional thesaurus applications as well. Secondly, the probabilistic models are generative, so they can be used to estimate the likelihood of hitherto unseen documents, or to create completely new documents by random sampling. This latter has been an important traditional challenge in language processing, at least since Shannon [24] introduced language generation from n-gram models as a test-case for the exposition of information theory. However, as Shannon himself pointed out when discussing the entropy of English, such models give gradually closer approximations to natural language given more training data and greater memory. He made no claim that these approximations would asymptotically approach “real language” closely enough for the difference between human and machine generated text not to be obvious, and in practice, automatically generated text of any considerable length still clearly reveals “the voice of the computer” to a human reader. For this reason the claim that probabilistic models give a theoretical model which adequately explains document creation by humans should probably be treated with some caution.

1.3 Summary of Distributional Methods

Various methods exist that allow computers to derive meaningful estimates of semantic similarity from corpora of electronic text. These range from local n-gram models and their various smoothed alternatives to high-dimensional vector models derived from term-document matrices. While a given method is often pioneered with a particular theory or application in mind, its capabilities are often expanded and changed as it finds its technological niche. For example, Support Vector Machines are now a widely used technique employing vector space models for binary classification, an application previously thought of as being uncharacteristic of vector models. All methods advanced so far appear to be influenced at least as much by the nature of their training data as by the nature of their mathematical theory, and the nature of the available data and the purpose of the application should always influence a choice of appropriate models. Given its desirable scaling properties, random projection, which has emerged recently as a scalable alternative to LSA (and is described in more detail in the discussion of recent methodological advances) is probably a sound choice of methodology for very large training sets, and for applications where a general sense of semantic relatedness is desired. For applications such as spam-filtering, which use smaller training sets and need to perform clear classifications, probabilistic methods may be more appropriate.

At the time of writing, probabilistic methods in general are the most reliable distributional techniques available for a number of language processing and data mining applications, such as machine translation and named entity recognition. A theoretical question being raised at the moment is whether or not the traditional distinctions between logical, geometric, and probabilistic models (most characteristic in discussion about search engines that follow a Boolean, vector, or probabilistic design) are differences in substance, or different ways of describing models with many features in common. Given the geometric description of probability distributions as geometric simplexes [9], and the availability of a probabilistic interpretation of vector models using the probability amplitudes familiar to quantum mechanics [25] a more unified approach is a desirable possibility [26].

What is clear by now is that there are numerous ways of finding semantically related groups of objects, by distributional analysis using large collections of data such as text corpora. For this technology to mature, methods and results will need to keep improving in their sophistication and the quality of their output, for which many promising avenues are opening up.

1.4 Historical Roots

1.4.1 Information Retrieval

The roots of certain spatial and probabilistic models can be traced back to models originally used for information retrieval. In the case of LSA, this connection is particularly easy to trace, as LSA was originally conceived as Latent Semantic Indexing (LSI) which was intended to address the problem of synonymy. This problem limits the ability of information retrieval systems based on exact keyword matching to find documents about topics similar to a search keyword when this keyword is not mentioned explicitly. LSI in turn is a derivative of the well-known Vector Space model of information retrieval proposed by Gerald Salton [27], in which both queries and documents are represented as points within a vector space which has terms as its dimensions. Retrieval is based on distance between them, usually measured using the cosine metric (also called the normalized scalar product), which is also the most commonly used distance metric to measure distance between terms in LSA and many other distributional applications. Distributional semantics in the probabilistic tradition can be traced at least to the IR work of Sparck-Jones in the 1960s [28], and its development has been intertwined with that of probabilistic search engines and of statistical machine learning algorithms in general.

1.4.2 Neural Networks

Research in neural networks/associative networks and connectionist models generally have also contributed to the current state-of-the-art in distributional semantics. In a neural or associative network, concepts are represented by nodes with (weighted) links to related nodes. One of the first applications of associative networks in computational semantics was the network representation developed by Quillian [29]. In this model, meanings of words were represented by nodes, with weighted links to other nodes that appeared in the word's dictionary definition. This also enabled each word to be used as the central node in a “spreading activation network”. Since nodes were created for each dictionary definition of a word (i.e., one node for each give word-sense), this could be used to generate lists of words related to each sense of a word, which enabled the network to be applied to the problem of disambiguation, an idea later developed by Lesk [30] and McDonald [31]. Subsequent work in word sense disambiguation derived these associations from a few seed associations subsequently extended by new associates learned from free text [32]. In information retrieval, associative networks have been implemented, whose nodes include documents and authors as well as words [33], and in a larger sense, and there are by now several examples of associative networks or graph models of terms being built from free text (see Widdows 2004, Ch. 2 [34]).

While research and implementation of associative semantic networks has continued to flourish, the description of these models as “neural networks” appears to have been declining. For example, in the mid 1990s, Landauer and Dumais saw fit to give one description of LSA as providing a collection of intervening “latent nodes” between term-nodes and context-nodes [7]. By the first few years of this century, such descriptions had fallen from fashion to some extent: while associative networks and combinatoric models in general have increased in popularity and usefulness (e.g., Google's PageRank [35] model of the Web), the computational processing of these models is in practical terms far from “neural”. Computers (even modern multi-core machines) have far more units of memory than of computation, and attempts to construct massively parallel arrays of processors to behave like human neurones are unwieldy undertakings. Thus it is probably fair to say that research starting in neural networks has contributed considerably to the availability of distributional semantics packages today, but the description and implementation of these systems follows different lines.

In mathematical terms, there are many ways to link combinatoric models and vector space models. The adjacency matrix of a graph is a prominent concept used to apply techniques from linear algebra to combinatorics. Correspondingly, graph models can be built directly from vector space models by considering distributional similarities (e.g., cosine similarities) as network link strengths in a graph, as we later illustrate in one of our visualization examples.

1.5 Rule-based methods for text processing

Distributional approaches to computational semantics are often contrasted with so-called "rule-based methods". The term "rule-based methods" may be used to refer to a great variety of approaches, partly characterized by manual organization of information and automated deduction using this organized information. A typical case might be the annotation of documents with terms from a controlled vocabulary (such as traditional Dewey Decimal or Library of Congress Subject Headings, or Medical Subject Heading (MeSH) terms in MEDLINE), and the use of thesaural relationship between these terms for information retrieval. For example, if a manually created formal knowledge base contains the fact that the "posterior cruciate ligament" is part of the "human knee joint", a query for "knee surgery" could be mapped to an article entitled "arthroscopic surgery of the posterior cruciate ligament" [36].

In natural language processing, rule-based methods work best in systems with small amounts of data, or where large amounts of expert human labor is available. For example, the process of assigning shelf marks to new books is very well established and maintained, and we are not suggesting that automatic distributional methods should replace these in the foreseeable future. However, as document collections have grown, there are many cases where human labor and expertise has been unable to keep up. For example, early generations of internet search engines (e.g., Yahoo and AltaVista in the 1990s) attempted to use human annotation to classify Web pages into a taxonomy, like assigning books to shelf marks, and it proved unscalable and unreliable given the rate at which new Web pages were being created. Nowadays, large scale internet search engines (such as Google, Yahoo and LiveSearch) use automatic methods of indexing almost exclusively.

One of the main difficulties in automating rule-based methods is the accuracy required of semantic annotation. It is not enough to have an ontology of concepts and their relations: it is also necessary to have a reliable way of recognizing when one of these concepts occurs in text, and as is well known, the mapping of words in text to formal concepts is full of ambiguity, contextuality, pragmatics, and many other subtleties. The biggest initiative for adding semantic annotation to Webpages is the Semantic Web, and so far, the amount of data annotated with Semantic Web concepts is tiny compared to the web as a whole. Also, it is increasingly recognized that to improve this disparity, automatic distributional methods may have a significant role to play in bridging the gap (see e.g. [37]).

Rule-based systems lead to the availability of many more hypotheses than when basing hypotheses on plain text. While this is in itself not a bad thing, it can lead to an intractably large search space (both for a formal reasoning agent or a human user), and distributional approaches are important for pruning away unreasonable paths, for example by using statistical methods for disambiguation (see e.g. [38]).

Complementary roles for rule-based and distributional methods may be available in any application domain that has both large amounts of available text data and some reliable process for human involvement. Examples may include:

- Literature review and organization (e.g., in legal or academic research). This process is already highly automated – in many domains, no professionals conduct literature reviews without the aid of automatic information retrieval systems, which we have pointed out are early examples of systems employing some form of distributional semantics. Nonetheless, it is still a time consuming process to winnow through search results, and there are many results that are missed due to differences in terminology. Smarter search engines based upon more carefully analyzed distributional information can help here, for example, by clustering results.

- Terminology acquisition. One of the earliest systems for automatic lexical acquisition used pattern-based distributional information to extract groups of related terms, claiming that the accuracy, though not high enough to put the results automatically into a usable lexicon, was high enough to be of considerable use to a human lexicographer or knowledge engineer [39]. This again falls into the generally useful pattern whereby computers make suggestions, and humans make decisions. In addition, empirical distributional information can be combined with pattern-based and human-curated techniques to reduce false-positives from ambiguous usages [40].
- Knowledge-rich data mining [41]. Recognizing the problem that machine learning can find correlations much more effectively than causes, some researchers in the field are focusing on the challenge of bootstrapping a small amount of hand-curated information to enable more effective automatic analysis of large data sets. There are many ways in which hand-curated data can be used as training data for a distributional system, for example to improve clustering or disambiguation, or to provide relevance feedback to a search engine.

A survey of such examples and techniques also demonstrates that the contrast between “distributional” and “rule-based” methods is hard to define in practice: for example, many pattern based systems for automatic text analysis and extraction analyze the distributional properties of text that matches rule-based patterns. The attempt to define the boundary between different families of techniques sometimes falls back on the distinction between “symbolic” and “spatial”, or “logical” and “geometric” approaches: however, this distinction is challenged by Birkhoff and von Neumann’s description of both geometry and logic in terms of lattice theory [25], and indeed by Aristotle’s introduction of logic itself. For example, Aristotle equates “inclusion” with “predication”, deliberately aligning logical and geometric approaches with the definition:

> “That one term be included in another as in a whole is the same as for the other to be predicated of all of the first.” (Prior Analytics, Ch 1)

A much clearer and more practical distinction can be drawn between manual and automatic methods. The past fifty years has seen the relative economic value of humans and computers change dramatically, and apparently permanently: for almost any system today, the most valuable resource is the human user’s time. The argument for complementary development then becomes very clear: any task that *can* reasonably be done by a computer *should* be done by a computer. Tasks that involve trawling through large numbers of documents to generate hypotheses are obvious candidates for automation, provided the precision of those hypotheses is high enough to make it worthwhile for a human user to review them.

Part of the supposed distinction between rule-based and distributional methods stems from this root: thus far in the history of computational linguistics, rule-based systems have relied on humans to create the rules, whereas distributional systems have relied less on human-generated assumptions, and more on large amounts of data speaking for itself. As with all new and evolving scientific disciplines, the appropriate balance between rationalist and empiricist approaches is being gradually explored in computational linguistics, and we are still very far from hearing the last word in this debate. In general, empiricist/statistical methods are very much in the ascendancy in computational linguistics at the time of writing, partly as resources have become increasingly available and cost-effective, and partly as hand-coded expert systems have shown poor scalability and poor robustness over many years. In many domains, however, hand coded methods have a crucial role to play, and the goal of research and development should be to find an appropriate combination, not to pick one approach and champion it against others. In the applications section later in this paper, we will further describe some of the ways in which distributional methods can enhance traditional ontologies with context aware concept learning and disambiguation.

2. Existing Applications of Distributional Semantics

This section outlines some of the practical applications in which distributional models have performed effectively. The main common theme for most of these applications to date is that they are tasks for which distance between words and other concepts is a useful thing to be able to measure.

2.1 Document Retrieval

The limitations of exact-match keyword-based information retrieval systems was the first motivation for using dimensionality reduction as a means of accessing latent distributional similarities [42]. Though promising, latent semantic indexing for search engines has not proved to be a “killer app”. This is partly because, in spite of some good results, precision and recall were not reliably improved on many test collections. It also became gradually clear that the “semantic search engine” needs to be much more effective than a simple “keyword search engine”, because users have become accustomed to keyword searches and tend to prefer tools that they can manipulate in an increasingly predictable fashion. This is not to imply that keyword search engines do not have many limitations: they do. However, these limitations are relatively easy for users to understand and hence to supplement. A successful deployment of semantic search engines would need not only to improve retrieval results: it would need to improve results in a way that users can understand and control. This may imply that there is not one ideal semantic search engine, but several possibilities, the design of which would involve detailed knowledge of a particular user community. The University of Pittsburgh’s Technology Matching Project, which employs a search engine tailored to support queries related to technology management presents one example of the use of distributional semantics to customize a search engine to the needs of a specific user community [43].

2.2 Synonym Tests

A standard test of language aptitude is to ask a student which of a collection of words is most similar to a given priming word: for example, is optimistic most similar to hopeful, reliable, religious, or sanctimonious [44]. Such exercises are often called synonym tests, and were used as one of the first ways to evaluate LSA. In particular, when the closest LSA-estimated distance between the priming word and the set of alternative answers was taken as a response to the Test of English as a Foreign Language (TOEFL) synonym test, the resulting score (64.4%) approached the average score obtained by a large sample of applicants to U.S. colleges who were second-language English speakers (64.5%) [7]. These results were obtained using a training corpus designed to approximate the quantity and content of the average American college freshman’s lifetime reading, and have since been improved using related methods and the much larger British National Corpus [45].

While synonym tests are often cited as validation of the semantic relations produced by distributional methods, it is worth noting that these tests measure a specific type of semantic relation only. Consequently improved synonym test performance does not necessarily provide motivation for methodological decisions related to applications requiring more general sorts of semantic relations. In addition, as these tests consider term-term similarities only, performance improvements in this test will not necessarily translate to performance enhancements in applications addressing larger units of text than individual terms. In addition to these concerns, Tom Landauer has raised the issue that LSA in particular does not capture word-word similarity explicitly [46]. Rather, LSA is intended to model the meaning of paragraph-length utterances encountered in communication, as well as the way in which humans are able to learn language from such encounters. Consequently the emergence of meaningful similarities between terms is a second-order effect of their respective contributions to the passages they occur in, and the exclusive evaluation of similarities between terms does not adequately address LSA's validity as a model of human language acquisition.

2.3 Word Sense Disambiguation

Many words in natural language have the same written form but different meanings. These meanings may be closely and systematically related (e.g., bank meaning "financial institution" and bank meaning "the building in which a financial institution does business"), or they may be words with quite different meanings that happen to be written in the same way (e.g., bank meaning "financial institution" and bank meaning "the land on the edge of a river"). This case is often used in the literature on word sense disambiguation as an example of accidental ambiguity, though several dictionaries of etymology trace both bank as in "commercial bank" and bank as in "river bank" to an early Germanic origin meaning a bench or table or place where things become piled up: the general point being that the more we research lexical ambiguity, the more systematic it proves to be [47]. However, there are certainly some examples of purely accidental ambiguity, such as proper names (John Smith refers to several completely different individuals) and acronyms (PCA refers to both Principal Component Analysis and Patient Controlled Analgesia).

The task of distinguishing automatically between the different available senses of a term is called word sense disambiguation, and there are several distributional techniques that have been used to address this challenge. The basic notion goes back again to Firth's dictum "you shall know a word by the company it keeps"-if the ambiguous term bank occurs with terms like savings, money, customer it is likely to be a financial bank, whereas if it occurs with terms like erosion, river, mooring it is more likely to be a river bank. In general, distributional methods relying on seed data like fixed collocations and statistical smoothing based on more general similarities have performed quite well [32]. More particularly for our purposes, Wordspace methods based on LSA have been used not only to disambiguate words in content into known senses, but to infer the appropriate word senses to begin with by clustering the appropriate context vectors. This more flexible approach is described as Word Sense Discrimination [13]. The use of clustering initially to obtain the word senses is typical of unsupervised as contrasted with supervised methods in machine learning.

2.4 Bilingual Information Extraction

The Wordspace idea can be adapted to more than one language at once if bilingual parallel corpora can be obtained. A parallel corpus is one that consists of pairs of documents which are known translations of one another. Good examples are proceedings from administrative bodies with more than one official language: for example, the Canadian parliamentary records are available in English and French; there is a similar Hong Kong corpus in English and Chinese; and the largest available multilingual corpus is that of the European Parliament, currently available in Danish, German, Greek, English, Spanish, French, Finnish, Portuguese, Dutch, and Swedish [48].

Once a parallel corpus is obtained, creating vectors for words and documents from both languages in the same Wordspace can be done relatively easily, by saying that any two words that occur in either document A or its translation document A′ cooccur with each other - so the methodologies that work for building a monolingual Wordspace for words in A work bilingually for words in A and A′. Care must be taken to use effective notions of context: for example, if context windows are to be used, then the corpora must be carefully aligned, or if documents are to be used as contextual units, the documents must be reasonably small conceptual units (e.g., for parliamentary proceedings, the transcript of a particular debate is a much more useful contextual until for measuring co-ocurrence than the transcript of all business on a particular day).

A natural way to test bilingual Wordspace models is to select a word from one language, find the nearest vector representing a word from the other language, and compare the results with a bilingual dictionary to see if they are translations of one another. This experiment has been known to give accuracies exceeding 90% for pairs with high cosine similarities [49]. In addition, the cases in which the automatic Wordspace translation differs from the dictionary translation are sometimes interesting in themselves: as well as cases where the dictionary is correct and the Wordspace model is incorrect, there are cases where the Wordspace is giving useful information that the dictionary does not contain. In the experiments described in [49], these included acronyms in the English dictionary that were not yet in the German dictionary but had identical German usages, and parts of many-to-one mappings (both “lung” and “transplant” were mapped to “lungentransplant”, surprisingly correct in context but not a word of its own in English). In recent models build from the Europarl data using the Semantic Vectors software, the English “house” was translated to the French “assemblé”. While dictionaries give the word “maison” as the most common translation of house, in the context of the parliamentary proceedings, the Wordspace translation is exactly right.

2.5 Essay grading

LSA has also been proposed as a means of automatically grading content-based essay examinations [50]. A number of methods that build on the semantic relatedness derived from a LSA matrix have been developed and evaluated for this purpose. These methods generally use a semantic space derived from the content being examined. A single vector is derived for each of the essays to be evaluated, as well as a set of landmark essays of known grade. These essay vectors are derived from the vectors for each term in the entire essay using the normalized vector sum, also occasionally referred to as the vector average. Examples of grading metrics that have been used include the distance between a new essay and an expert response, the distance-weighted average grade of the 10 closest landmark essays, and the vector length of the new essay in the semantic space. In this application, the semantic space contains only domain-relevant concepts. Consequently the vector length metric measures the amount of domain-relevant text included in the essay response. This metric has attracted a degree of criticism, as a longer essay containing redundant but relevant content would score well when this metric alone is employed. Landauer and his colleagues refer to these metrics as “quality” (vector average of nearest known essays) and “quantity” (vector length). The combination of these scores performed best in an extensive evaluation of student essays on the physiology of the heart (n=94) and psychology (n=273), with the correlation between the system’s score and the mean grade assigned by two professional human graders approaching the graders’ agreement with one another. These and other metrics are utilized by the Intelligent Essay Assessor, a commercial product that performs automated essay grading based on LSA [51].

2.6 Visualization

Several visualization methods have been employed in order to evaluate or demonstrate the degree to which a particular semantic space covers a selected content area. As the vectors used in spatial models are high-dimensional (usually $n > 100$), it is necessary to reduce the dimensionality of the data if they are to be visualized in two or three dimensions. Landauer and his colleagues use the GGobi [52] software package, which provides a number of algorithms to support the visualization of high-dimensional data, to create several large-scale complex multidimensional visualizations of subsets of a document collection [53]. The authors conclude that only a limited representation of meaning can be displayed in three dimensions. In contrast to this approach, other authors have employed visualization methods to explore the relationships between small groups of related terms in semantic space. Burgess and Lund used multidimensional scaling (a technique employed in the psychological literature to visualize data on human estimates of pairwise similarity between terms) to project small groups of words into two-dimensional space, in order to demonstrate the clustering of similar concepts [54]. Widdows and Cederberg use a second round of SVD to scale a matrix comprised of the vector representations for a small group of related words down to two dimensions for visualization purposes, revealing clusters of similar concepts, including those drawn from bilingual corpora [55]. Cohen employs Pathfinder network scaling [56] (which has also been used to visualize pairwise similarity rankings) to explore groups of terms in a semantic space derived from Random Indexing of the MEDLINE corpus of abstracts [57]. In this visualization terms are treated as nodes in a network joined by their connectivities to one another. Pathfinder reveals the internal structure of this network by preserving the most significant links only. The figures below illustrate the use of two of these approaches, reduction to two-dimensional space using a second round of SVD (Figure I) and Pathfinder networks (Figure II, graph layout produced using the Prefuse visualization library [58] to visualize the semantic neighborhood of the term “thrombosis” in a Random Indexing space derived from the OHSUMED corpus. Both of these visualizations are reminiscent of those derived from measures of human estimates of semantic distance in empirical psychological research. Despite their limitations, visualizations that can be rendered in two (or three) dimensions present an intuitive and easily interpretable way to explore small semantic neighborhoods.

2.7 Taxonomy Construction and Validation

Structured ontologies and knowledge bases have become regarded as increasingly central to the construction of more sophisticated information systems. Given the cost of building or adapting ontologies manually, ontology learning has naturally become a major research area. Though there are many ontological relationships that may be considered (object part-of object, person owns object, symptom-of condition, drug used-to-treat condition, etc.), most work in ontology learning has to date been devoted to the traditional taxonomic or is-a relation. Many empirical methods have been developed that extract both of the constituents in the relation from particular known patterns or templates. However, a more purely distributional method is to combine some distributional similarity measure with a known set of existing relations to obtain more relations where one of the participants is already known, e.g. classifying unknown words to one of a given small set of “supersenses” [59]. The principle is as follows. Given a seed fact (e.g., “a horse is an animal”) one can find distributionally similar terms to horse (e.g., cow, dog, sheep) and infer that each of these is also an animal. Needless to say, this method used bluntly gives many false positives, e.g., there may be a distributional similarity between horse and cart, but a cart is not an animal. However, it has also been demonstrated that combining distributional similarity measures with pattern-based extraction can improve the accuracy of such extractions, and reduce the errors caused by effects such as ambiguity and over-generalization. In one application, using LSA-based distributional similarity to rank relations extracted using pattern-based methods increased the precision of the top-ranked relations by 18% over a random selection of extracted relations [40].

3. Biomedical Applications of Distributional Semantics

The following section reviews existing applications of distributional semantics in the biomedical domain. A broad range of problems in biomedical informatics have been addressed with these methods, including both bio- and clinical informatics applications. As is the case with other language processing methods, biomedical text presents unique challenges for distributional semantics research. Adequately sized corpora of clinical narrative are difficult to obtain on account of privacy restrictions, and the tendency of clinical reports to lack uniform structure and contain idiosyncratic expressions and acronyms may result in sparse data on certain terms. The biomedical literature contains many meaningful multi-word phrases and terms containing non-alphabet characters, which require the adaptation of existing distributional methods in order to avoid erroneously splitting these terms into fragments. However, the biomedical domain also presents unique opportunities for distributional semantics research. From a theoretical perspective the structural idiosyncrasies of biomedical narrative are typical of a sublanguage as defined by Harris [60]. The relations between words in a sublanguage are more tightly constrained than in general language, which may allow for the categorization of terms into meaningful semantic classes using distributional methods (for example, see table IV below). Rich domain-specific knowledge resources such as MEDLINE present large corpora for analysis. These have in some cases already been mapped to controlled terminologies such as MeSH terms or the UMLS, presenting unique opportunities for distributional models of term-to-terminology relationships. With the advent of scalable methodological alternatives, freely available implementations and the decrease in cost of both storage space and RAM, many of the pre-existing barriers to this sort of research have been removed, creating opportunities for the further exploration of the utility of these methods in the biomedical domain.

3.1 Gene Clustering Using MEDLINE Abstracts

Text-based gene clustering research evaluates the extent to which the semantic relations between terms derived by distributional methods can be used to generate meaningful biological relations between genes. Homayouni and his colleagues present and evaluate Semantic Gene Organizer [61], a system that determines relationships between genes using the following procedure: (1) “Gene-documents” for each gene in a series of 50 are created by concatenating the titles and abstracts of all MEDLINE citations cross-referenced in the mouse, rat and human entries for each gene in LocusLink (which has since been superceded by Entrez gene). (2) A term-”gene-document” matrix is created, and (3) dimension reduction was performed using SVD. The system is shown to successfully retrieve many of the genes associated with the well-established Reelin signalling pathway, including an association between Reelin and fyn kinase, which was not included in LocusLink at the time of the evaluation. Hierarchical clustering is applied to these genes based on their pairwise distance in the reduced-dimensional space, producing functionally cohesive clusters. This clustering included an apparent anomaly, the inclusion of the oncogene SHC1 with a cluster of genes related to Alzheimer’s disease. While these genes shared one co-citation in the document collection, subsequently published research [62] (not included in the document collection) revealed a functional relationship between this oncogene and the APP gene that is implicated in Alzheimer’s Disease. Glenisson et al evaluate a similar approach, and find improvements in statistical measures of cluster quality, as well as the accuracy with which certain genes are categorized when the terms from related MEDLINE citations rather than the keywords from descriptions in a curated database are used as a basis for the vector-space representation of genes [63], indicating that the additional information in unstructured abstract text is of value for text-based categorization of genes.

3.2 Analogous Approaches to Biological Sequence Analysis

The following discussion of statistical approaches to biological sequence analysis is limited to methods that derive a measure of similarity between genes or biological sequences using an approach that falls within the scope of the article as a whole. Consequently, we do not discuss methods that determine similarity between sequences based on gapped alignment between strings as these methods focus on surface similarities between sequences of symbols rather than employing a model analogous to the semantics underlying word choice. The analogy between sequences of words in human language and biological (for example gene or protein) sequences supports the application of many methods traditionally applied to text processing to problems in bioinformatics [64]. Ganapathiraju and her colleagues draw on this analogy, developing an improved method for the prediction of protein sequences likely to be trans-membrane (TM) segments, parts of a protein that pass through the cell membrane [65]. In this application, amino acid sequences are first transformed into sequences of properties (such as polarity or positive charge) that characterize proteins that exist in the cell membrane. These properties are considered as analogous to the terms in the term-document matrix used in LSA, while the protein segments are considered as documents. In this way, protein segments (documents) are represented in terms of the distribution of the properties of their amino acid components (terms). As in LSA, dimension reduction is performed using the SVD, resulting in a reduced-dimensional dataset. Applying a neural network classifier to this dataset results in more accurate prediction of trans-membrane sequences than several other methods. Stuart and Berry take a similar approach to whole-genome analysis of bacterial phylogeny [66], [67]. In this analysis, protein sequences are represented as vectors representing the frequency with which each possible tetra-peptide sequence occurs in this larger protein sequence. Overlapping sequences of tetra-peptides are considered, in a manner reminiscent of the “sliding window” used in the Schütze’s Wordspace model, resulting in a 160,000 dimensional vector for each protein. SVD is performed on this matrix to generate a reduced-dimensional matrix allowing for the measurement of similarities between proteins, some of which are consistent with biologically meaningful categorizations of proteins into families. Species vectors are then constructed as the vector sum of all of the reduced-dimensional protein vectors for a given species, enabling species-species comparison using the cosine metric. These cosine distances are considered as evolutionary distances, allowing for the generation of a phylogenetic tree. Unlike standard phylogenetic methods, this approach is able to scale to the entire genome of each organism considered. As this method does not depend on the alignment of local sequences, it derives similarities between proteins based on the distribution of tetra-peptide sequences regardless of the order in which these sequences appear. Consequently, while many of the clusters in the generated phylogenetic tree did correspond to classical classifications, a number of unconventional associations were also generated. This contrast between methods that rely on local sequence and those that emphasize higher-order distributional similarities is also evident in emerging methods of distributional semantics, some of which have the capacity to capture sequence information (see section 5.2, below) without requiring sequence alignment. It is apparent even from early evaluations of these methods that incorporating word order into a distributional model results in the capture of very different sorts of (nonetheless meaningful) associations between terms [68], suggesting that the application of these methods to protein or gene sequences may reveal as-yet-undetected biologically meaningful associations.

3.3 Literature-based Knowledge Discovery

Gordon and Dumais propose the use of Latent Semantc Indexing (LSI), for literature-based knowledge discovery [69], the identification of knowledge in the research literature that is not explicitly stated. This study evaluates the utility of LSI as a tool to support the knowledge discovery process described by Swanson, in which two disparate literatures, “Literature A” and “Literature C”, are bridged by a third, “Literature B” with some concepts in common with both “A” and “C”. By using this approach, Swanson was able to discover the supplementation of dietary fish oil as a previously unrecognized therapeutic alternative for the treatment of Raynaud’s Syndrome. This seminal discovery, and the framework that yielded it, have influenced much subsequent research into literature-based knowledge discovery, including the work of Gordon and Dumais. These authors demonstrate the ability of LSI to identify “B” literatures, the first step of Swanson’s two-step process. In addition, the potential of LSI as a tool to map between disjoint literatures is explored. Given LSI’s usefulness in information retrieval as a tool to identify relevant documents that do not contain the specific term used in a keyword search, and ability of LSA to find meaningful associations between terms that do not directly co-occur, it seems reasonable to expect that LSA/LSI would also be able to identify indirect connections betweens disparate bodies of literature. However, in this investigation the authors were not able to discover the Raynaud's-to-fish-oil connection directly. Interestingly, eicosopentaenoic acid, the active ingredient of fish oil, is recovered as the 208th ranked indirect neighbor of Raynaud's. In addition, calcium dobesilate and niceritrol are revealed as plausible therapeutic alternatives for this condition. Bruza, Cole and colleagues present a similar approach to literature-based knowledge discovery based on empirical distributional semantics [70],[71]. In this case, the HAL (sliding-window) model, rather than the LSA (term-document) model is used as a basis for the representation of term vectors. This work is further distinguished by its cognitive approach, which highlights the empirical and theoretical support for semantic spaces as models of meaning, and proximity within such spaces as a model of abductive reasoning, the cognitive process through which novel hypotheses are generated. This research confirms Gordan and Dumais' finding that distributional statistics can support the automated discovery of B terms related to Swanson's discoveries. In addition, by weighting those dimensions of the vector for "Raynaud" that correspond to a manually curated set of these B terms, the terms "fish" and "oil" were retrieved within the top 10 near neighbors when particular distance metrics and statistical weighting functions were used, effectively replicating Swanson's discovery using as a corpus a set of MEDLINE titles from core clinical journals between 1980 and 1985, a period preceding the publication of any articles with titles including both "Raynaud's" AND "fish oil". However, these results were the best obtained over a number of runs with different statistical parameters. Many of the other runs ranked these terms far lower in the list than would be reasonable to expect a user to explore, and it is not necessarily the case that the statistical weighting that produced optimal results for this discovery would perform as well in simulating other historical discoveries. While this corpus of titles is small by today's standards, a recent application of Random Indexing illustrates that meaningful associations can be derived from the entire MEDLINE corpus of abstracts using distributional methods [57], suggesting it may be possible to extend these methods to the MEDLINE corpus as a whole, and in doing so incorporate knowledge from basic medical science as well as clinical journals into the discovery model.

3.4 Information Retrieval

As we have discussed previously, some of the distributional methods discussed in this paper are historically related to the vector space and probabilistic approaches to document retrieval. While biomedical information retrieval is not the primary focus of this review (see Hersh for a thorough review of this topic [72]), there are approaches to specific problems in information retrieval that have been addressed using distributional semantics methods. The "related article" feature in PubMED allows users to access articles similar in content to one of the results of a previous search. The feature is supported by the PubMED related articles (pmra) algorithm [73] which is based on a probabilistic model of the likelihood of each document being about a particular topic. In this model, each term is considered as a topic, and the probability that a document focuses on this topic is estimated according to the distributional statistics of each term in the specific document and entire corpus. While aspects of this approach are similar to probabilistic models used in distributional semantics, a notable difference is the modeling of individual terms as independent topics. In contrast, models such as Griffiths and Steyvers' Probabilistic Topics Model [23] model the distribution of terms across topics in addition to the distribution of terms across documents in the corpus. One would anticipate that incorporating this additional detail into the underlying model would find similarities between documents about a similar topic that do not share many terms. However, the parameter estimation process used by these models is computationally demanding, and is unlikely to scale well to a corpus the size of MEDLINE. Random Indexing, however, would present a scalable alternative to address this issue, which was one motivation for the development of LSI. To the best of our knowledge Random Indexing has not been extensively evaluated in an information retrieval context, presenting a research opportunity for its formal evaluation in the context of information retrieval from MEDLINE.

The ability of distributional models to model the meaning underlying terms in a text passage is exploited by SEGOPubMED [74] which uses LSA to map between MEDLINE abstracts and the Gene Ontology (GO) with performance improvements over simple string matching in an information retrieval task. Such mapping between free text in MEDLINE abstracts and an ontology or controlled terminology would allow for the integration of MEDLINE with other databases, improving scientists' ease of access to this information. Distributional approaches to text-to-vocabulary mapping related to the MeSH terminology have also been evaluated. As the MEDLINE database is manually indexed using MeSH terms, an automated method that is able to derive relationships between this controlled terminology and the text of the articles it annotates would be useful to assist with human indexing by recommending MeSH terms and assist users with the refinement of their text-based queries by suggesting suitable MeSH terms for query expansion. Yang and Chute evaluate the Linear Least Squares Fit (LLSF) method [75] as a means to map between text and MeSH terms. In contrast to the methods described previously in this paper, LLSF requires a labeled training set of categorized documents as input. From these, a document-by-term and document-by-category matrix are derived. The LLSF method calculates a mapping between these two matrices, producing term-by-category associations. In this case, the categories assigned are MESH terms, but the method generalizes to any human-indexed set of documents, and a clinical application of this method is discussed in section 3.5, below. LLSF is shown to outperform baseline methods on an information retrieval task in an evaluation on a partitioned set of MEDLINE documents, as well as in the accuracy of assignment of MeSH terms to MEDLINE abstracts as compared to those assigned by human indexers with a fourfold gain in average precision over ranking based term matching between terms in the abstract and those in the category name [76]. As the authors note, the reliance of LLSF on SVD for dimension reduction limits its ability to scale to large test corpora. However, the recent emergence of a scalable alternative to the use of SVD in distributional models (see Random Indexing, section 5.1) suggests that it may be possible to extend this method to much larger annotated corpora. The same authors achieve comparable results using Expert Networks, which take a similar approach to that used in the essay-grading applications of LSA: a MEDLINE abstract to be indexed is projected into a vector space with terms as dimensions, and the k-nearest-neighboring human-indexed abstracts in this space are retrieved [77]. The MeSH terms used to index these neighbors are then weighted by similarity distance and frequency (how many neighboring abstracts they have been used to index), to select the most likely labels for the new abstract. As the similarity scores between documents depend on direct term matching, this algorithm does not have the ability to infer meaningful connections between documents that cover the same topic but do not share any terms. Nonetheless, this method was able to match the performance of LLSF, which one would expect to have similar inferencing capacity to distributional methods such as LSA that also rely on SVD. While the classification accuracy was similar to LLSF, this Expert Networks are considerably more computationally efficient as they do not require SVD. The Pindex system [78] takes a probabilistic approach to this problem. Pindex determines the conditional probability that each MeSH term is assigned to a MEDLINE abstract given the terms occurring in this abstract. These probabilities are then used to assign MeSH terms to clinical texts, correctly assigning MeSH terms that captured approximately half the concepts present in the text. The design of systems to assist human indexers in assigning MeSH terms to MEDLINE documents is the primary concern of the NLM indexing initiative [79] which relies on mappings between document abstracts and UMLS concepts produced by the MetaMap system [80]. Unlike the methods reviewed in this manuscript, MetaMap assigns categories by measuring the extent to which noun phrases in the text and those in the UMLS meta-thesaurus share sequences of terms, rather than the extent to which their distributions across a corpus of documents are similar.

3.5 Automatic Generation of Synonyms

The generation of synonyms for the purpose of terminology expansion or automated thesaurus generation is a task well-suited to distributional methods. Cohen and colleagues combine pattern-based, distributional and graph-theory based approaches to synonym identification in the biomedical literature, using a regular expression designed to have high recall and low precision for terms referring to genes or proteins [81]. Once identified, the terms surrounding each pattern are used to identify potential synonyms as terms that occur in similar local contexts. The co-occurrence between potentially synonymous terms is modeled as a network structure, with terms as nodes and the number of co-occurrences between them as links. Network structures are evaluated using a clustering coefficient, with the assumption that networks of synonyms will contain stronger links than networks of unrelated terms. The method is evaluated against synonym pairs in curated databases, obtaining a maximum F-score of 22.21% using a corpus of 50,000 articles.

3.6 Clinical Applications

In many cases, the language contained in clinical narrative is relatively constrained, both in syntax and semantics. The semantic constraints are predictable in that the conceptual territory spanned by, for example, a radiology report is limited to content relevant to this domain. While less immediately apparent, the syntactic constraints of domain-relevant clinical narrative underlie several influential clinical language processing applications, including MedLEE [82] and the Linguistic String Project [83]. These applications build on the theoretical foundation of sub-language theory developed by Zellig Harris [60], which provides a framework to characterize the additional constraints of the language used in a specialized domain. Of note, the “distributional hypothesis” which is often cited as a theoretical motivation for distributional semantics is also attributed to Harris [84]. These additional constraints make it possible to approach clinical narrative using a semantic grammar, in which the domain-relevant meaning of terms in addition to their part of speech are encoded into the grammar rules used to perform Natural Language Processing. On account of its constrained nature, clinical narrative seems less suited to the distributional semantics approach. However, despite these constraints, existing applications suggest a complementary role for these methods.

In an early application LSA is used to extract semantic content from the UMLS [85] Rather than deriving semantic relations from raw text, a matrix is derived from an existing human-curated data source, the UMLS meta-thesaurus. A range of UMLS concepts is selected to determine the dimensions of an initial term-by-concept matrix. Term vectors are augmented in each concept-dimension if they are a part of the description of the main concept, or are defined as synonyms, lexical variants or associated terms. SVD is performed on this matrix. A preliminary evaluation of this technique suggests that meaningful association between terms can be derived. A similar approach based on a reduced-dimensional matrix of term-by-concept vectors derived from the ICD-9-CM is evaluated in an information retrieval application [86]. In evaluation, this system performed worse than both surface string matching and another document retrieval system. As the authors note, the performance degradation of this system may be related to incomplete coverage of synonymous terms in the ICD-9-CM, suggesting that performance may have been improved by incorporating term similarities derived from other knowledge sources. A more encouraging evaluation of a distributional method is presented in subsequent work by the same authors, which exploits the additional knowledge provided by human annotators of a set of clinical notes [87]. This system uses a linear-least squares fit approach (described previously) to find an optimal mapping between terms used by clinicians in clinical diagnoses and terms in the ICD-9-CM codes assigned to these diagnoses by expert medical coders. The system is evaluated for it’s ability to map between unseen texts (n=3,456 diagnoses, approximately the same size as the training set used to build the model) and the codes assigned to these, which were drawn from a set of 376 possible categories. In this evaluation, LLSF was able to retrieve an expert-assigned canonical term as the nearest neighbor 84% of the time, and as one of the one of the top five nearest neighbors 96% of the time. LLSF was also shown to improve performance over baseline methods in an information retrieval evaluation [75]. The approach used is similar in concept to that used to model bilingual corpora, in that it requires as training data sets of concepts expressed in both clinician's language and in a controlled terminology, just as bilingual vector space models require a matched context in each language. Much effort and expense in the informatics community has been directed at the generation and maintenance of knowledge resources to support biomedical informatics applications. The development of methods to evaluate these resources, assist with their maintenance, and tailor them to particular applications is a perennial focus of attention of biomedical informatics research.

In contrast to the corpus-based distributional approach, other methods exist that derive a measure of semantic relatedness from existing knowledge resources. Unlike the methods employed by Chute and his colleagues, many of these methods bear little resemblance in their implementation to the methods of distributional semantics. Rather, measures of semantic similarity are derived directly from defined relationships between terms. For example, the number of ontological relationships that must be traversed to reach one term from another might be considered as a measure of the semantic distance between these terms. Pedersen and his colleagues adapt a number of ontology driven measures of semantic relatedness to the medical domain [88]. In addition, a context vector is derived in an approach similar to that employed by Schütze in his word-sense disambiguation research [13], using as a corpus specific sections of the Mayo Clinic Corpus of Clinical Notes. A significant aspect of this work is the authors' evaluation method: the pairwise semantic relatedness of a set of 30 terms was ranked by 9 medical coders and 3 physicians, using a 4-point rating scale from synonymous to unrelated. This method of evaluating human determination of semantic relatedness has a long history in psychological and cognitive research [89], and previous evaluation studies have used published datasets of human associations to evaluate ontology-dependent methods of semantic relatedness [90]. This study found that the vector-based method had better correlation with physician's assessment of semantic relatedness than any of the ontology-based measures. Correlation with assessment of semantic relatedness by professional medical coders matched the best of the ontology-derived measures.

In a recent application Fan and Friedman [91] combine distributional semantics and domain knowledge to reduce the granularity with which UMLS concepts are classified, in order to enhance the usefulness of the UMLS as a resource to support Natural Language Processing applications. Distributional semantics implementations derive measures of semantic relatedness from the similarity between the contexts in which terms appear. Most of the implementations we have discussed up to this point use either a document or the neighboring terms within a document as a context. Another approach to the definition of context is to employ syntactic relations between words rather than co-occurrence [92], [93], [94], [15]. In this case, a grammar-based parser is employed to derive syntactic relations between terms in the text, and term vectors are derived from recognized syntactic relations only. For example, "object of the verb treats" might feature as a dimension of a term by syntactic relation matrix. Most reports of syntax-based semantic spaces are to be found in the Natural Language Processing or Computational Linguistics literature. These methods have drawn less attention from the cognitive science community, although Pado and Lapata's work is mentioned in the cognitively-oriented Handbook of Latent Semantic Analysis [95] This division in the literature may be explained by a desire on the part of the cognitive science community to determine to what extent meaningful relations can be derived from text alone without the imposition of a manually constructed grammar. In addition there exists a series of publications in the cognitive science literature showing little or no advantage for methods incorporating syntax [96], [97]. Fan and Friedman draw on additional linguistic and domain knowledge to define a set of rules to extract syntactic dependencies from the output of the MetaMap [80] program, which maps free text to UMLS concepts and includes shallow parsing. These syntactic dependencies are used to derive vector representations for individual UMLS concepts, as well as composite representations for broader categories. Distributional similarity is used to classify a test set of concepts into these broader categories with a low error rate when compared to the classification of a domain expert.

In contrast to the constrained language employed in other medical domains, the language used in psychiatric narrative tends to cover broad conceptual territory, as it includes detailed descriptions of the psychosocial context of psychiatric illness, as well as the perspective of the family and patient. The difficulty of defining a semantic grammar for language of such broad scope motivated an investigation of the ability of LSA to make human-like judgments concerning psychiatric discharge summaries [98]. Once the ability of LSA to draw meaningful associations between psychiatric concepts was established, an LSA-derived score for the evaluation of the potential danger posed by psychiatric patients to themselves or others was evaluated against the ratings of a single human rater. This scoring system was similar in nature to the system employed in the essay-grading studies – a new summary was scored based on a the k-nearest neighboring summaries in semantic space. While the correlation between human- and computer-assigned grades exceeded published estimates of the agreement between expert psychiatrists for similar assessments, the inconsistent nature of human evaluation of potential patient risk limits the usefulness of such a system, even if perfect correlation with a single rater could be attained. However, this work suggested LSA and related methods are able to support the development of computational tools able to address the broad conceptual territory of psychiatric clinical narrative. This potential was further explored in a study of the ability of LSA to extract segments of psychiatric narrative according to their diagnostic relevance to a set of clinically important concepts [99]. This study was motivated in part by the difficulty psychiatric residents were shown to have in selectively extracting text elements that are of relevance to diagnostically meaningful clusters of clinical concepts [100]. The system designed in this study employed vector compositional operators defined by Widdows and Peters [101] and a cognitively motivated learning algorithm to model higher level concepts as regions (or subspaces) of a larger semantic space that could conform their boundaries in response to a labeled set of training data. The system was evaluated on a test set of 100 psychiatric discharge summaries, with mean system-rater agreement approaching that between pairs of expert raters. This result suggests that the associations learned by this system are adequate to support expert-like classification of discharge summary content into clinically useful categories. LSA has also been evaluated as a diagnostic tool for the detection of some aspects of thought disorder in schizophrenia [102]. One diagnostic feature of the thought disorder in schizophrenic patients is incoherent speech, speech that lacks meaningful connections between successive utterances. In this study, a series of experiments are performed in which LSA is used to evaluate the coherence between successive utterances. LSA-derived measures of coherence are shown to distinguish between patients and controls, and correlate with clinical measures of thought disorder.

3.7 Applications in Consumer Health

According to published estimates from 2002, more than 70,000 websites disseminated health information and more than 50 million people sought health information online [103]. While using human-maintained indexes to guide navigation of this information is impractical because of the rate of proliferation of online content, there is some suggestion that distributional models may be useful to healthcare consumers exploring health-related websites. A recent application combining semantic vectors generated using the HAL model and a machine learning algorithm to accurately classify the tone of breast cancer websites as either “supportive” or “medical” illustrates the ability of distributional methods to accurately categorize online health information according to the requirements of different user groups [104]. An innovative recent study suggests that distributional models may be able to derive meaningful information regarding the stage of adaptation of patients to their chronic illness based on their e-mail contributions to an online support community [105]. It is interesting to consider the potential of combining these approaches: for example, patients at an early stage of adaptation to their illness could be directed to “supportive” websites. This proposal is admittedly speculative, and these approaches are both at a early stages of development. However, the potential of distributional methods to categorize both online health information and health care consumers in a psychologically meaningful way seems worthy of further evaluation.

4. Recent developments in Distributional Semantics

In this section we review, with examples, some emerging technologies in distributional semantics research which have not yet been extensively evaluated in a biomedical context.

4.1 Random Indexing

While the computational efficiency of the SVD can be improved through the use of parallel computation and iterative methods optimized for sparse matrices [106] (parallel implementations of LSA have reportedly been used to model corpora of more then 500 million words in size [46) the computational demands of these methods still present a barrier to their widespread dissemination as many researchers do not have access to the computational power required to process very large text corpora. Recently, Random Indexing (RI) has emerged as promising alternative to the use of SVD for the dimension reduction step in the generation of term-by-context vectors [16], [20]. RI avoids constructing the term-by-document matrix, generating reduced-dimensional term vectors directly by:

1. Assigning an index vector of zero values of length k (usually > 1000), the preassigned dimensionality of the reduced-dimensional matrix to be generated, to each document.
2. Assigning the values −1 or 1 to a small number (+−20) of cells in each index vector. These nonzero values are randomly distributed across the index vector.
3. Each time a term occurs in a document, the index vector for that document is added to the term vector for the term, generating a reduced-dimensional approximation of the full term-document matrix.

RI and other similar methods are motivated by the Johnson-Lindenstrauss Lemma [107] which states that the distance between points in a vector space will be approximately preserved if they are projected into a reduced-dimensional subspace of sufficient dimensionality. While this procedure requires a fraction of the RAM and processing power of SVD, it is able to produce term-term associations of similar accuracy to those produced by SVD-based LSA [16]. Given the many possible permutations of a small number +1’s and −1’s in a high-dimensional space, it is likely that most of the assigned index vectors will be close-to-orthogonal (almost perpendicular) to one another. Consequently, rather than constructing a full term-document matrix in which each document is represented as an independent dimension, a reduced-dimensional matrix in which each document is represented as a near orthogonal vector in this space is constructed. Initial investigations of this method are promising: while it may seem counterintuitive that such a simple approach could capture meaningful relations, this method has performed as well as SVD-based methods in the TOEFL synonym test [16]. With further enhancements such as sliding-window based indexing, lemmatization [20] and the incorporation of word-order information [108] this model has performed substantially better than LSA on this test. The primary advantage of RI is its scalability. While most evaluations of LSA and related methods are performed on small document collections, RI scales comfortably to much larger corpora, such as the MEDLINE collection of abstracts (around 9 million abstracts) [57]. Some examples of nearest neighboring terms derived from this space are shown in the table below:

The method is also much faster than previous methods (processing, for example the entire MEDLINE corpus in around 30 minutes), allowing for the rapid prototyping of semantic spaces for experimental purposes. In addition, RI implementations tend to support both term-by-document and sliding-window based indexes, allowing for the comparison between these types of indexing procedures in particular tasks. This is an important research area as prior studies suggest these types of indexing procedures extract different sorts of relations between terms [109].

4.2 Incorporating word order

One common criticism of distributional semantics methods is their failure to acknowledge the sequential structure of language. In medicine, for example, such sequencing is important to capture relations between the elements of multi-word entities such as “blood pressure”, “left ventricular failure” and “pulmonary embolism”. Although the HAL model does distinguish between previous and subsequent neighboring terms in its sliding window, it is true that for the most part the methods discussed in this article tend to ignore term sequence. Several recently proposed methodological innovations seek to address this issue by incorporating word sequence information into either vector space [14], [108] or probabilistic [110], [111] representations. Sahlgren *et al* propose a particularly elegant approach to capturing word order information by permuting the 1’s and −1’s in index vectors used in RI according to the location of neighboring terms within a sliding window [108]. As the index vectors are randomly constructed, this permutation effectively creates a new random vector. By using rotation of the 1’s and −1’s within the index vector, a permutation function that is reversible is obtained, allowing for the construction of order-based queries, as illustrated by the following examples drawn from the TOEFL corpus (for further examples see Jones and Mewhort [14] and Sahlrgren *et al* [108]):

In these examples, the ‘?’ represents the position of the missing term, such that ‘king ?’ is searching for terms likely to occur immediately after the term ‘king’. In addition to supporting order-based retrieval, these spaces appear to capture a more tightly constrained sort of semantic relationship, suggesting they may be useful in finding terms of similar semantic types to support the construction of knowledge resources or semantic grammars. In addition, the RI-based implementation remains scalable to large document collections such as the MEDLINE corpus, as shown in the examples of order-based retrieval below:

The specificity of near-neighbor relationships derived from order-based spaces is evident in these results, which can be for the most part neatly categorized as classes of antidepressant, types of pressure and species of staphylococcus. The incorporation of word-order information has also been shown to improve performance on the TOEFL synonym test, achieving a best score of 80% with a model derived from the TASA corpus, [108] as well as consistency with human performance on certain cognitive tasks [14].

4.3 Further Future Trends

In addition to the incorporation of word-order information, recent implementations have introduced further grammatical information by integrating part-of-speech tags [112] or dependency parsing [15]. Other promising directions for future research include the investigation of different models of vector composition [113] and the combination between vector space models and spatially motivated machine learning methods [114]. Random Indexing is unique amongst the methods discussed in this article in its ability to efficiently integrate new documents into an existing semantic space, allowing for the implementation and study of real-time acquisition of semantic knowledge. In addition, the relative speed with which a semantic space can be derived using Random Indexing allows for rapid prototyping of different models, which is beneficial for research purposes.

5. Software Packages

In this section we introduce some software packages with which semantic distance estimates can be derived from text corpora and/or their distributional statistics. This is not intended to be an exhaustive discussion of the software available for this purpose. Rather, we have focused on a few implementations that either have a relatively broad user base or, in the case of probabilistic models, were produced by the originators of these models. We have tried, where possible, to focus on implementations that perform parsing of documents as well as derivation of semantic distance, as it is our intention to present these methods to as broad an audience as possible. The first three packages below all perform the fundamental tasks required for distributional semantics research: parsing of text to generate distributional statistics, derivation of semantic distance between terms from these statistics, and the comparison between terms to retrieve the nearest neighboring terms or documents to a query term (or document). The availability of such packages enables those without experience in text processing to apply distributional semantics methods to their own research.

5.1 Infomap NLP

The Infomap NLP package [12] was developed at Stanford University. It uses context windows (of configurable length) and a user-defined number of content words to collect an initial co-occurrence matrix, which is then reduced using Michael Berry's SVDPACKC library [115]. It was written largely by Hinrich Schütze and Stefan Kaufmann, and released and supported largely by Scott Cederberg and Dominic Widdows. The software is available through SourceForge, under a free BSD license that permits adaptation and commercial use. While popular, with over 4000 downloads in 4 years, the software has often encountered problems in the areas of ease-of-use and scalability. Being written in C, the software is fast and efficient but is not altogether platform independent. There have been many installation and integration problems, especially in the storage layer, which involves integration with the GNU database libraries. The use of SVD imposes scalability limitations: to our knowledge, Infomap NLP has not been used to create models involving more than a few tens of thousands of word vectors.

5.2. General Text Parser

General Text Parser (GTP) was developed at the University of Tennessee by Howard, Tang, Berry and Martin [116]. It has been used in many LSA applications, and is one of the implementations recommended in the recent Handbook of Latent Semantic Analysis [95]. GTP parses text to generate term-document statistics and, like Infomap, performs SVD using the SVDPACKC library. This package is extremely flexible with a wide range of options. C++ and Java implementations are available on request, and a parallel-processing implementation is also available which reportedly scales well to larger datasets.

5.3 Semantic Vectors

The Semantic Vectors package was developed initially by the University of Pittsburgh in collaboration with MAYA Design, and is now maintained by a small development community including both of the current authors. It is hosted by the Google Code website, and is also available for free download under a BSD license. The Semantic Vectors package was designed to solve some of the problems encountered by Infomap NLP. The software is written entirely in Java which, though somewhat slower than C, has made the package very easily portable, and so far no platform specific problems have been encountered. Random projection is used for dimension reduction, which has enabled the software to build models involving several hundreds of thousands of word vectors on a reasonably standard modern laptop (2.4 GHz processor, 2GB of RAM). The package has been downloaded over 2500 times in its first fifteen months, and is actively maintained with new features and tests, support for bilingual models, clustering, visualization, incorporating word-order information, a variety of logical and compositional operators, and considerable flexibility in tunable parameters, memory models, and I/O formats [43].

5.4. Implementations of pLSA and related models

Tom Griffiths and Mark Steyvers, the developers of the Probabilistic Topic Model variant of pLSA, have produced the Topic Modeling Toolbox [117], a Matlab implementation of various related models using the Gibbs Sampler Algorithm for parameter estimation. David Blei provides a C implementation of LDA using variational Expectation Maximization for parameter estimation under the GNU General Public License [118]. This has been ported to Java by Gregor Heinrich, who also provides an implementation of the Gibbs Sampler as part of the Knowceans project [119]. Unlike the previously discussed packages, these packages do not perform parsing of text, but rather require pre-processed term-document statistics as input. Several other implementations of these algorithms are also available online.

6. Conclusions

With the advent of affordable high-volume storage and processing power, the availability of large corpora of text, and the recent emergence of scalable algorithms, many of the barriers to research in empirical distributional semantics have been removed. These methods have been successful in a number of applications, including the emulation of human performance on cognitive tasks, natural language processing and information retrieval. These successes are reflected in a growing body of literature on distributional semantics methods in cognitive science, application-oriented and biomedical informatics journals. The non-specific associative strengths derived by distributional semantics methods do not provide the formal definitions that are required to support rule-driven computational methods. However, the associations derived by distributional semantics have been shown to correspond to psychological studies of human estimates of similarity between terms, and as such capture additional information that is both cognitively valid and practically useful. Recent research has utilized semantic spaces that incorporate dependency relations derived using knowledge-driven natural language processing tools into their similarity estimates. Conversely, associations derived from semantic spaces can be used to customize existing knowledge resources, and improve the accuracy of pattern-based relationship extraction. We anticipate further realization of the potential of these methods in a range of future biomedical informatics applications.

Acknowledgments

The writing of this manuscript was supported in part by the US National Library of Medicine grant (R01 LM07894). The authors would like to acknowledge Google, Inc. for their support of author DW's ongoing research on the subject. We would also like to thank Tom Landauer for providing the TASA corpus and his comments on the validity of the TOEFL test, as well as two anonymous referees for their recommendations and thorough reviews.

References

1. Zhang J. Representations of health concepts: A cognitive perspective. Journal of Biomedical Informatics 2002:3517–24.

2. Hearst, M. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics; 1992. p. 539-545.

3. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics (Oxford, England) 2001;17(Suppl 1):S74–82.

4. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 2003;36:462–477. [PubMed: 14759819]

5. Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers 1996;28:203–208.

6. Schutze H. Word space. Advances in Neural Information Processing Systems 1993;5:895–902.

7. Landauer TK, Dumais ST. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 1997;104:211–240.

8. Hofmann, T. Probabilistic Latent Semantic Analysis. Proceedings of Uncertainty in Artificial Intelligence, UAI’99; 1999. p. 289-296.

9. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research 2003;3:993–1022.

10. Griffiths, T.; Steyvers, M. A probabilistic approach to semantic representation. Proceedings of the 24th Annual Conference of the Cognitive Science Society; 2002. p. 381-386.

11. Hersh, W.; Buckley, C.; Leone, TJ.; Hickam, D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval; 1994. p. 192-201.

12. Cederberg, S.; Widdows, D.; Peters, S. Infomap NLP software: an open-source package for natural language processing. [December 2008]. Webpage. http://infomap-nlp.sourceforge.net/

13. Schütze H. Automatic Word Sense Discrimination. Computational Linguistics 1998;24:97–123.

14. Jones MN, Mewhort DJK. Representing word meaning and order information in a composite holographic lexicon. Psychological Review 2007;114:1–37. [PubMed: 17227180]

15. Pado S, Lapata M. Dependency-Based Construction of Semantic Space Models. Computational Linguistics 2007;33:161–199.

16. Kanerva, P.; Kristofersson, J.; Holst, A. Random indexing of text samples for latent semantic analysis. Proceedings of the 22nd Annual Conference of the Cognitive Science Society; 2000. p. 10-36.

17. Dumais ST. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers 1991;23:229–236.

18. Gorman, J.; Curran, JR. Random Indexing using Statistical Weight Functions. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP); Sydney, Australia. 2006. p. 457-464.

19. Strang, G. Introduction to Linear Algebra. Wellesley Cambridge Pr: 2003.

20. Karlgren J, Sahlgren M. From words to understanding. Foundations of Real-World Intelligence 2001:294–308.

21. Pereira, F.; Tishby, N.; Lee, L. Distributional clustering of English words. Proceedings of the 31st conference on Association for Computational Linguistics; 1993. p. 183-190.

22. Heinrich, G. Parameter estimation for text analysis. 2005. Web: http://www.arbylon.net/publications/text-est.pdf

23. Steyvers, M.; Griffiths, T. Probabilistic Topic Models. In: Landauer, T.; McNamara, D.; Dennis, S.; Kintsch, W., editors. Handbook of Latent Semantic Analysis. Mahwah, N.J: Lawrence Erlbaum Associates; 2007.

24. Shannon CE. Prediction and entropy of printed English. Bell System Technical Journal 1951;30:50–64.

25. Birkhoff G, von Neumann J. ‘The Logic of Quantum Mechanics’. Annals of Mathematics 1936;37:823–843.

26. Rijsbergen, CJV. The Geometry of Information Retrieval. Cambridge University Press; 2004.

27. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975;18:613–620.

28. Robertson S, Spark-Jones K. Relevance Weighting of Search Terms. J Am Soc Information Sciences 1976;27:129–146.

29. Quillian, MR. Semantic memory. Minsky, M., editor. Semantic Information Processing; 1968. p. 216-270.

30. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation; ACM New York, NY, USA. 1986. p. 24-26.

31. McDonald, JE.; Plate, TA.; Schvaneveldt, RW. Pathfinder associative networks: studies in knowledge organization. Ablex Publishing Corp; 1990. Using pathfinder to extract semantic information from text; p. 149-164.

32. Yarowsky, D. Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Cambridge, Massachusetts: Association for Computational Linguistics; 1995. Unsupervised word sense disambiguation rivaling supervised methods; p. 189-196.

33. Belkin, NJ.; Croft, WB. Annual review of information science and technology. Vol. 22. Elsevier Science Inc; 1987. Retrieval techniques; p. 109-145.

34. Widdows D. Geometry and Meaning. Center for the Study of Language and Information/SRI. 2004

35. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 1998:107–117.

36. Volk M, Ripplinger B, Vintar Š, Buitelaar P, Raileanu D, Sacaleanu B. Semantic annotation for concept-based cross-language medical information retrieval. International Journal of Medical Informatics 2002;67:97–112. [PubMed: 12460635]

37. Maedche A, Staab S. Ontology learning for the Semantic Web. Intelligent Systems, IEEE 2001;16:72–79.

38. Charniak, E. Statistical Language Learning. Bradford Books; 1993.

39. Rilo, E.; Jones, R. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. Proceedings of AAAI-99; 1999. p. 474

40. Cederberg, S.; Widdows, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003; 2003. p. 111-118.

41. Domingos P. Toward knowledge-rich data mining. Data Mining and Knowledge Discovery 2007;15 (1):21–28.

42. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science 1990;41:391–407.

43. Widdows, D.; Ferraro, K. Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. To appear in Sixth International Conference on Language Resources and Evaluation (LREC 2008); 2008.

44. VOCABSYN Frameset1 [Internet]. Available from: http://www.edict.com.hk/vlc/vocabsyn/FramesSyn1.htm

45. Rapp, R. Ninth Machine Translation Summit. 2003. Word sense discovery based on sense descriptor dissimilarity; p. 315-322.

46. Landauer TK. personal communication.

47. Pustejowsky YJ. The Generative Lexicon. Computational Linguistics 1991;17(4):409–441.

48. Koehn, P. MT Summit. 2005. Europarl: A parallel corpus for statistical machine translation.

49. Widdows, D.; Peters, S.; Cederberg, S.; Chan, CK.; Steffen, D.; Buitelaar, P. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. Natural Language Processing in Biomedicine ACL 2003 Workshop; 2003. p. 9-16.

50. Landauer, TK.; Laham, D.; Rehder, B.; Schreiner, ME. How Well Can Passage Meaning be Derived without Using Word Order. A Comparison of Latent Semantic Analysis and Humans. Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society; August 7–10, 1997; Stanford University. 1997.

51. Landauer TK, Laham D, Foltz PW. The Intelligent Essay Assessor. IEEE Intelligent Systems 2000;15:27–31.

52. Swayne DF, Lang DT, Buja A, Cook D. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. Computational Statistics and Data Analysis 2003;43:423–444.

53. Landauer, TK.; Laham, D.; Derr, M. From paragraph to graph: Latent semantic analysis for information visualization. Proceedings of the National Academy of Sciences; 2004 Apr. p. 5214-5219.

54. Burgess C, Lund K. The dynamics of meaning in memory. Cognitive dynamics: Conceptual and representational change in humans and machines 2000:117–156.

55. Widdows, D.; Cederberg, S. Monolingual and bilingual concept visualization from corpora. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations; 2003. p. 31-32.

56. Schvaneveldt, RW. Pathfinder associative networks: studies in knowledge organization. Ablex Publishing Corp; Norwood, NJ, USA: 1990.

57. Cohen TA. Exploring MEDLINE Space with Random Indexing and Pathfinder Networks. AMIA Annu Symp Proc 2008:126–30. [PubMed: 18999236]

58. Heer, J.; Card, SK.; Landay, JA. prefuse: a toolkit for interactive information visualization. Conference on Human Factors in Computing Systems; 2005. p. 421-430.

59. Curran, JR. Supersense tagging of unknown nouns using semantic similarity. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; 2005. p. 26-33.

60. Harris ZS. The structure of science information. Journal of Biomedical Informatics 2002;35:215–221. [PubMed: 12755516]

61. Homayouni R, Heinrich K, Wei L, Berry MW. Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics (Oxford, England) 2005;21(1):104–115.

62. Zambrano N, Gianni D, Bruni P, Passaro F, Telese F, Russo T. Fe65 is not involved in the platelet-derived growth factor-induced processing of Alzheimer’s amyloid precursor protein, which activates its caspase-directed cleavage. The Journal of biological chemistry 2004 Apr;279:16161–16169. [PubMed: 14766758]

63. Glenisson P, Antal P, Mathys J, Moreau Y, De Moor B. Evaluation of the vector space representation in text-based gene clustering. Pac Symp Biocomput 2003:391–402. [PubMed: 12603044]

64. Klein-Seetharaman The Use of Analogies for Interdisciplinary Research in the Convergence of Nano-, Bio- and Information Technology. NSF Report on Societal Implications of Nanoscience and Nanotechnology 2005:128–133.

65. Ganapathiraju M, Balakrishnan N, Reddy R, Klein-Seetharaman J. TMpro: Transmembrane Helix Prediction Using Amino Acid Properties and Latent Semantic Analysis. BMC Bioinformatics 2007;8 (10)

66. Stuart GW, Berry MW. A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. Journal of Bioinformatics and Computational Biology 2003:1475–493.

67. Stuart GW, Berry MW. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. BMC Bioinformatics 2004;5:204–217. [PubMed: 15606920]

68. Widdows, D.; Cohen, T. Semantic Vector Combinations and the Synoptic Gospels. to appear. Proceedings of the Third Quantum Interaction Symposium; March 25–27, 2009; DFKI, Saarbrücken.

69. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science 1998;49:674–685.

70. Cole RJ, Bruza PD. A Bare Bones Approach to Literature-Based Discovery: An Analysis of the Raynaud?s/Fish-Oil and Migraine-Magnesium Discoveries in Semantic Space. Discovery Science 2005:84–98.

71. Bruza, P.; Cole, R.; Song, D.; Bari, Z. Towards Operational Abduction from a Cognitive Perspective. Oxford Univ Press; 2006.

72. Bath, PA.; Hersh, William. Information retrieval: a health and biomedical perspective. New York, NY: Springer; 2003.

73. Lin J, Wilbur WJ. PubMed related articles: A probabilistic topic-based model for content similarity. BMC Bioinformatics 2007;8(1):423. [PubMed: 17971238]

74. Vanteru BC, Shaik JS, Yeasin M. Semantically linking and browsing PubMed abstracts with gene ontology. BMC Genomics 2008;(9 Suppl):1S10.

75. Yang Y, Chute CG. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems (TOIS) 1994;12(3):252–277.

76. Yang Y. An evaluation of statistical approaches to MEDLINE indexing. Proc AMIA Annu Fall Symp 1996:358–362. [PubMed: 8947688]

77. Yang Y, Chute CG. An application of Expert Network to clinical classification and MEDLINE indexing. Proc Annu Symp Comput Appl Med Care 1994:157–61. [PubMed: 7949911]

78. Cooper GF, Miller RA. An Experiment Comparing Lexical and Statistical Methods for Extracting MeSH Terms from Clinical Free Text. Am Med Inform Assoc. 1998

79. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. Proc AMIA Symp 2000:17–21. [PubMed: 11079836]

80. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:1717–21.

81. Cohen A, Hersh W, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics 2005;6:103. [PubMed: 15847682]

82. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp 2000:19270–4.

83. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association: JAMIA 1:142–60. [PubMed: 7719796]

84. Harris, ZS. Mathematical structures of language. Interscience Publishers; New York: 1968.

85. Chute CG, Yang Y, Evans DA. Latent Semantic Indexing of medical diagnoses using UMLS semantic structures. Proc Annu Symp Comput Appl Med Care 1991:1859.

86. Chute CG, Yang Y. An evaluation of concept based latent semantic indexing for clinical information retrieval. Proc Annu Symp Comput Appl Med Care 1992;639:43.

87. Yang, Y.; Chute, CG. Proceedings of the 14th conference on Computational linguistics - Volume 2. Nantes, France: Association for Computational Linguistics; 1992. A Linear Least Squares Fit mapping method for information retrieval from natural language texts; p. 447-453.

88. Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 2007;40:288–299. [PubMed: 16875881]

89. Rubenstein H, Goodenough JB. Contextual correlates of synonymy. Commun ACM 1965;8:627–633.

90. Budanitsky, A.; Hirst, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources; 2001.

91. Fan JW, Friedman C. Semantic Classification of Biomedical Concepts Using Distributional Similarity. Journal of the American Medical Informatics Association 2007;14:467–477. [PubMed: 17460124]

92. Grefenstette, G. Corpus-derived first, second and third-order word affinities. Proceedings of Euralex; 1994. p. 279-290.

93. Lin, D. Automatic retrieval and clustering of similar words. Proceedings of the 17th international conference on Computational linguistics; 1998. p. 768-774.

94. Curran, JR.; Moens, M. Scaling context space. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2001. p. 231-238.

95. Landauer, TK.; McNamara, D.; Dennis, S.; Kintsch, W., editors. Lawrence Handbook of Latent Semantic Analysis. Mahwah, N.J: Lawrence Erlbaum Associates; 2007. Handbook of Latent Semantic Analysis.

96. Wiemer-Hastings, P.; Zipitria, I. Rules for syntax, vectors for semantics. Proceedings of the 23rd Annual Conference of the Cognitive Science Society; 2001. p. 1112-1117.

97. Kanejiya D, Kumar A, Prasad S. Automatic evaluation of students' answers using syntactically enhanced LSA. Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing 2003;2:53–60.

98. Cohen, T.; Blatter, B.; Patel, V. Exploring dangerous neighborhoods: latent semantic analysis and computing beyond the bounds of the familiar. AMIA ?. Annual Symposium proceedings/AMIA Symposium. AMIA Symposium; 2005. p. 151-5.

99. Cohen T, Blatter B, Patel V. Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. J of Biomedical Informatics 2008;41(6):1070–1087.

100. Sharda P, Das AK, Cohen TA, Patel V. Customizing clinical narratives for the electronic medical record interface using cognitive methods. International Journal of Medical Informatics 2006;75:346–368. [PubMed: 16125455]

101. Widdows, D.; Peters, S. Mathematics of Language. Vol. 8. Bloomington, Indiana: 2003 Jun. Word Vectors and Quantum Logic Experiments with negation and disjunction.

102. Elvevaag B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. Schizophrenia Research 2007;93:304–316. [PubMed: 17433866]

103. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671–92. [PubMed: 11780707]

104. Chen, G.; Warren, J.; Evans, J. Automatically generated consumer health metadata using semantic spaces [Internet]; Proceedings of the second Australasian workshop on Health data and knowledge management; Wollongong, NSW. Australia: Australian Computer Society, Inc.; 2008. p. 9-15.

105. McArthur, R.; Bruza, P.; Warren, J.; Kralik, D. Projecting Computational Sense of Self: A Study of Transition in a Chronic Illness Online Community. System Sciences, 2006; HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on; 2006. p. 91c

106. Berry, MW.; Mezher, D.; Philippe, B.; Sameh, A. Parallel computation of the singular value decomposition. In: Kontoghiorghes, E., editor. Handbook on Parallel Computing and Statistics. 2003. p. 117-164.

107. Johnson W, Lindenstrauss J. Extension of lipshitz mapping to hilbert space. Contemporary Math 1984;26:189–206.

108. Sahlgren, M.; Holst, A.; Kanerva, P. Permutations as a Means to Encode Order in Word Space. Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08); July 23–26; Washington D.C., USA.

109. Sahlgren, M. PhD Dissertation. Department of Linguistics, Stockholm University; 2006. The Word-Space model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces.

110. Dennis, S. Handbook of Latent Semantic Analysis. 2007. Introducing Word Order Within the LSA Framework.

111. Griffiths TL, Steyvers M, Blei D, Tenenbaum JB. Integrating topics and syntax. Advances in Neural Information Processing Systems 2005;17:537–544.

112. Widdows D. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology 2003;1:197–204.

113. Widdows, D. Semantic Vector Products: Some Initial Investigations. Proceedings of the Second AAAI Symposium on Quantum Interaction; 2008.

114. Sahlgren, M.; Coster, R. Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. Proceedings of the 20th International Conference on Computational Linguistics; COLING. 2004.

115. Berry, M.; Do, T.; O?Brien, G.; Krishna, V.; Varadhan, S. University of Tennessee Computer Science Department Technical Report. 1993. SVDPACKC (Version 1.0) user?s guide; p. CS-93-194.

116. Giles, JT.; Wo, L.; Berry, MW. Statistical Data Mining and Knowledge Discovery. 2001. GTP (General Text Parser) Software for Text Mining.

117. Topic Modeling Toolbox [Internet]. Available from: http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

118. Latent Dirichlet Allocation in C [Internet]. Available from: http://www.cs.princeton.edu/\blei/lda-c/

119. arbylon projects: knowceans [Internet]. Available from: http://www.arbylon.net/projects/

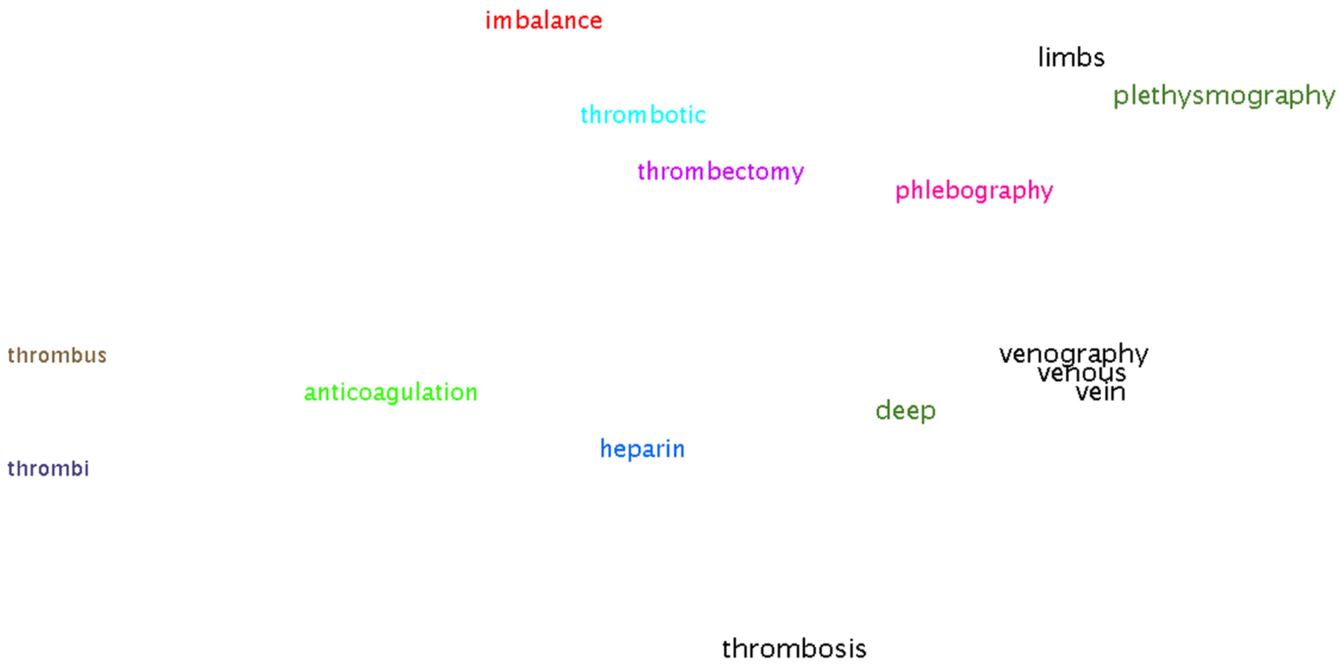

Figure 1.

Visualization of the 15 nearest neighbors of “thrombosis” in the OHSUMED corpus using SVD to reduce the neighborhood to the two dimensions that best capture the variance between the original high-dimensional points. The third most significant dimension is encoded as color and font size. Note the clustering of “thrombus” and “thrombi”, as well as “venography”, “venous” and “vein”.

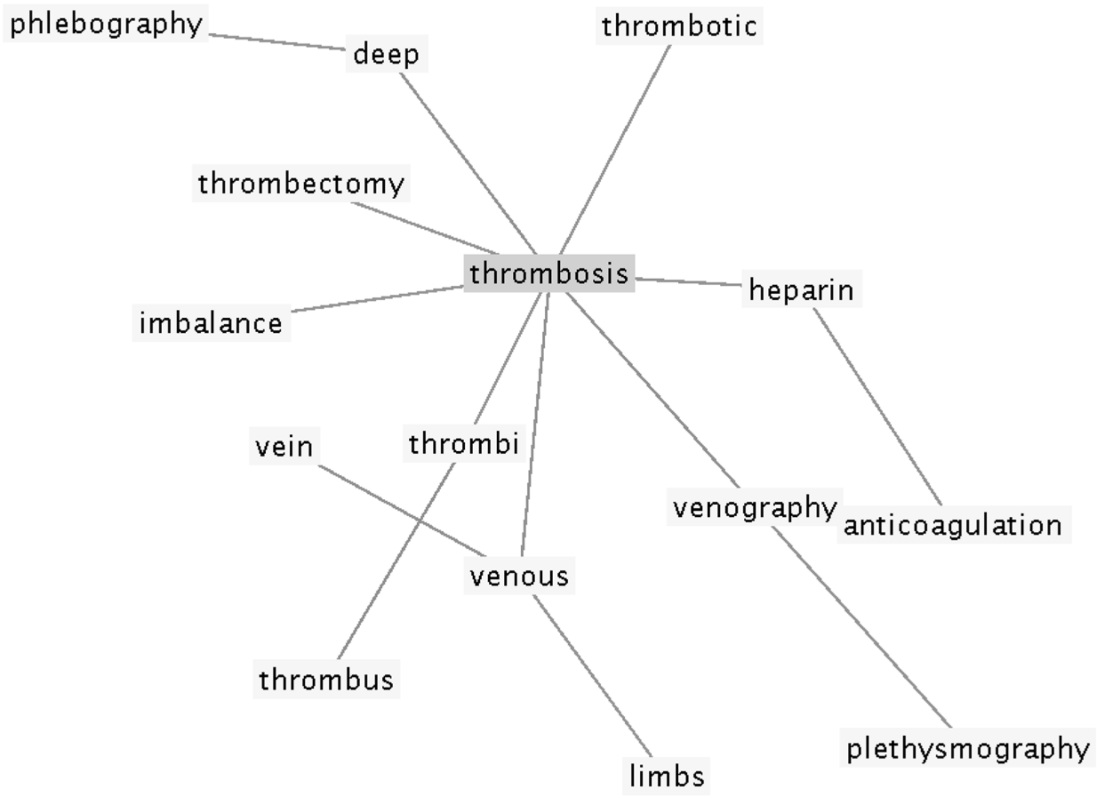

Figure 2.

Visualization of the same 15 nearest neighbors of “thrombosis” using Pathfinder network scaling to preserve the most significant links in the network of terms connected by their semantic similarity. Note the preservation of the connection between “thrombi” and “thrombus”, as well as between “venous” and “vein” and “heparin” (an anticoagulant) and “anticoagulation”.

Table I

nearest neighbors of the term "psychosis" extracted from the OHSUMED corpus

Term	Similarity	Term	Similarity
psychosis	1.00	hallucinations	0.54
psychotic	0.68	headaches	0.54
paranoid	0.58	attacks	0.53
delusions	0.58	mania	0.52
parkinsonism	0.55	antipsychotic	0.52

Table II

nearest neighbors derived from the MEDLINE corpus using RI

staphylococcus	**antidepressants**	**pressure**
1.00:staphylococcus	1.00:antidepressants	1.00:pressure
0.61:aureus	0.41:tricyclic	0.34:blood
0.32:methicillin	0.33:antidepressant	0.33:systolic
0.29:epidermidis	0.18:reuptake	0.28:pressures
0.23:coagulase	0.17:tcas	0.28:mmhg
0.21:mrsa	0.15:tricyclics	0.26:diastolic
0.18:staphylococci	0.14:ssris	0.25:hg
0.15:haemolyticus	0.12:fluoxetine	0.23:arterial
0.15:staphylococcal	0.11:imipramine	0.18:and
0.14:antimicrobial	0.10:depression	0.18:hypertension

Table III

order-based retrieval from the TASA corpus

king ?	king of ?	felt ?	? felt
0.62:aegeus	0.66:course	0.20:resentful	0.25:pamela
0.61:minos	0.52:macedonia	0.18:numb	0.21:tuma
0.53:midas	0.51:humankind	0.18:booted	0.17:madge
0.51:lear	0.50:mankind	0.18:shaky	0.15:benicia
0.49:jr	0.49:sheba	0.17:sorry	0.14:kelvin
0.42:cheng	0.48:chivalry	0.16:strangely	0.14:lena
0.37:tut	0.47:prussia	0.16:joyously	0.13:meribah
0.36:arthur	0.47:humanity	0.16:envious	0.13:khuana
0.34:solomon	0.47:crete	0.15:humiliated	0.12:glenda
0.32:agamemnon	0.44:gibraltar	0.14:constrained	0.12:lindbergh

Table IV

order-based retrieval from the MEDLINE corpus

? antidepressants	**? pressure**	**staphylococcus ?**
0.74:tricyclic	0.90:barometric	0.76:aureas
0.63:tricylic	0.87:intrabolus	0.71:aureus
0.55:trycyclic	0.86:hydrostatic	0.67:epidermidis
0.54:imipraminic	0.84:oncotic	0.64:aureaus
0.49:nontricyclic	0.82:subglottal	0.55:epidemidis
0.26:proserotonergic	0.82:intracompartmental	0.52:auerus
0.16:tetracyclic	0.82:transdiaphragmatic	0.49:saprophyticus
0.14:angiodysplastic	0.81:intracuff	0.47:aures
0.13:nonserotonergic	0.80:subatmospheric	0.46:hyicus
0.13:squamoproliferative	0.79:disjoining	0.45:xylosus