Using distributional similarity to organise biomedical terminology

Julie Weeds, James Dowdall, Gerold Schneider,
Bill Keller and David Weir

We investigate an application of distributional similarity techniques to the problem of structural organisation of biomedical terminology. Our application domain is the relatively small GENIA corpus. Using terms that have been accurately marked-up by hand within the corpus, we consider the problem of automatically determining semantic proximity. Terminological units are defined for our purposes as normalised classes of individual terms. Syntactic analysis of the corpus data is carried out using the Pro3Gres parser and provides the data required to calculate distributional similarity using a variety of measures. Evaluation is performed against a hand-crafted gold standard for this domain in the form of the GENIA ontology. We show that distributional similarity can be used to predict semantic type with a good degree of accuracy, reaching an optimal value of 63.1%.

Keywords: distributional similarity, biomedical terminology, semantic proximity, ontology.

1. Introduction

Lexical resources are commonly organised according to lexico-semantic relations such as synonymy, hyponymy, antonymy and meronymy. For example, the widely-used resource WordNet (Fellbaum 1998) has synonymy and hyponymy as its central organising relations. Word senses are grouped into sets of synonyms, i.e., words that have the same meaning, and then these *synsets* are further organised into a hierarchy, where each child of a node is a type or hyponym of the concept at that node.

Organising a lexical resource according to semantic principles makes it possible for humans and computers to find related words and to derive implicit

Terminology 11:1 (2005), 107–141.
ISSN 0929–9971 / E-ISSN 1569–9994 © John Benjamins Publishing Company

information about words based on the structure of the resource. For example, if one is looking for information about "*amino acid*" and it is known that a "*protein*" is a type of "*amino acid*", then it may be useful to include "*protein*" in a search for information on "*amino acid*".

While much effort has been put into constructing, both manually and automatically, general lexical resources such as WordNet, the need for domain-specific resources is becoming increasingly recognised. This is because specialised domains tend to have large terminological vocabularies, where individual terms are either not used in the general domain, and therefore cannot be found in a general resource, or have technical, domain-dependent meanings.

However, the task of organising a domain vocabulary, such as biomedical terminology, according to semantic relationships is a difficult one, and generally requires expert knowledge about the domain. Further, the process is never finished. There are always new words entering the language and new terms being introduced in a specialised domain. To this end, researchers have begun to investigate a number of ways in which the process might be semi-automated.

The task that we consider in this paper is how new terms might be added to an existing ontology of terminological types. Our approach involves calculating distributional similarity between terms over a domain corpus and hypothesizing that distributionally related terms are also semantically related. We then use the semantic types already assigned to these related terms to predict the semantic type of the unknown or target term. In this way, we make use of the expert knowledge previously supplied in the construction of the hierarchy, but aim to reduce the amount of expert knowledge required in maintaining and updating an existing hierarchy.

The remainder of this paper is organised as follows. In Section 2, we discuss related work on the organisation of terminology. Section 3 then introduces the biomedical domain in which we are working. In particular, we describe the GENIA corpus and the manually constructed GENIA ontology against which our predictions of term similarity are evaluated. In Section 4 we describe the parser (Pro3Gres) used to produce the grammatical dependency relation data that serves as a basis for computing distributional similarity. In Section 5, we discuss distributional similarity itself and consider three alternative measures. In Section 6 we describe a number of experiments in using distributional similarity to determine semantic relatedness of terms. In particular, we investigate whether distributional similarity is correlated with semantic similarity according to the GENIA ontology and whether the distributionally nearest neighbours of a term can be used to predict the semantic type of the term, according to the GENIA ontology. Our results show that distributional similarity techniques can provide a very useful source of information in the semi-automatic placement of new terms in the ontology. Our conclusions and directions for future work are presented in Section 7.

2. Related work

Approaches to the automatic organisation of terminology can be distinguished broadly according to the types of information sources they employ (internal or external) and whether they adopt supervised or unsupervised methods of training. Sources of information internal to the terms include lexical properties such as token sharing and morphological analysis. External sources of information can be statistical, contextual, or ontological. Many successful approaches combine knowledge sources either as a cascade or in parallel.

Techniques exploiting internal sources of information range in sophistication from the analysis of simple lexical inclusion to the terminological variation paradigm. For example, across the entire NLM MEsH thesaurus, simple lexical inclusion between the terms (i.e., where the tokens of one term are included within another) indicates a relation of hyponymy with a precision of 23% (Grabar and Zweigenbaum 2002). Further restricting this relation to ensure that the terms' lexical heads are identical is exploited across the literature as a high precision knowledge source (Mani et al. 2004; Torii et al. 2003; Nenadić et al. 2002b). This is taken as a starting point in clustering terms for the purpose of scientific and technology watch (Ibekwe-SanJuan and SanJuan 2002, 2004), where natural classes of multi-word terms are built around the conceptual head and are further related through the range of syntactic variation. In combination with an external ontology, terminological variation is expanded to include semantic variations, reducing the noise produced by substituting nominal words (SanJuan et al. 2004; Hamon et al. 1998).

Morphological analysis can determine concept families with a precision of 92% within the biomedical domain (Grabar and Zweigenbaum 2000). As shown in (Torii et al. 2003), even the presence of a specific suffix can be used as a feature in the supervised machine learning of semantic types. Dedicated processing of morpho-syntactic variation can determine complex semantic relations between terms such as "antonymy", "result" and "set of" (Daille 2003).

A widely used external source of information is the context within which a term is observed to appear. The notion of term context can be defined as a "bag-of words", with reference to a specific window size around a term (Mani et al. 2004). However, other definitions of context are clearly possible. For example, (Nenadić et al. 2003) demonstrate that using terms rather than words provides better performance at lower recall points within their support-vector machine (SVM) approach to the classification of gene names. Context has also been successfully defined as generalised regular expressions (Nenadić et al. 2002a). The present work adopts a notion of distributional context that is defined in terms of the grammatical relations of *subject* and *object*.

An alternative, complementary external source of information uses shallow parsing around contextual clues (or "cue-phrases") to identify hyponymy and synonymy with some reliability (Hearst 1992; Caraballo 1999; Lin et al. 2003; Morin and Jacquemin 2003; Dowdall et al. 2004). For example, one might expect to see indicators of hyponomy like "*amino acids such as proteins*" occurring in a corpus of biomedical documents. Unfortunately, this approach is likely to have rather low recall in the domain of biomedical research articles because the specified "cue-phrases" appear to be relatively sparse (Nenadić et al. 2002a; Mani et al. 2004). To address this problem, it may be possible to expand the type of corpus to include textbooks (which are naturally more descriptive than discursive and which contain less assumed knowledge) in order to produce a deeper hyponymy hierarchy (Kawasaki et al. 2003).

Of particular relevance to the present work are three studies that use the GENIA corpus and supervised models for determining the semantic type of the terms.

In addition to term identification, in (Chikashi Nobata and ichi Tsujii 1999) terms are classified as belonging to one of four semantic types. The study is based on just 100 abstracts and employs two alternative models of classification. The first model uses supervised learning with external word lists, word frequency and head weighting, and achieves an F-score[1] of 65.8%. The second model uses decision trees based on part-of-speech tags and orthography in addition to the word lists, and pushes the F-score up to 90.1%.

The second study contrasts two models in the combined identification and classification task (Kazama et al. 2002). Word frequency, part-of-speech tags, inflectional morphology and lexical inclusion are used as input to a SVM and Maximum Entropy (ME) model. Over the 670 available abstracts, the SVM is shown to out-perform the ME model. In classifying the terms into one of six semantic types, ME achieves a precision of 53.4% with a recall of 53.0%; the SVM performs slightly better with a precision of 56.2% and a recall of 52.8%.

In a third study that utilises the GENIA corpus at its present size of 2000 abstracts, supervised machine learning is used to classify the terms into one of five semantic types (Torii et al. 2003). Classification is based on a cascade of information sources that includes "f-terms" (where the head of the term is also its classification) the suffix occurring with the head of a term, a measure of term similarity based on a head weighted string matching algorithm and finally the "bag-of-words" context of a term. This approach achieves precision between 84% and 96% with recall between 62% and 90%. These values are high because the classification is attempted over only 5 generic semantic types. Further, the high recall appears to be from the string matching algorithm necessitating a large annotated training set.

Compared to the three studies outlined above, the approach taken here is based solely on the external context of terms. Identifying term similarity does not depend on any annotations in the corpus and the classification task uses all of the GENIA semantic types (see Figure 1). We apply measures of distributional similarity to a parsed corpus and hypothesise that distributionally similar terms are also likely to be semantically related terms. This is in accordance with the *distributional hypothesis* (Harris 1968):

> *The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.*

In recent years, distributional similarity has been applied on this basis to a wide range of problems in natural language processing (Hindle 1990; Grefenstette 1994; Lin 1998a; Curran and Moens 2002; Kilgarriff 2003; Weeds and Weir 2003b; Geffet and Dagan 2004; Linden and Piitulainen 2004)). For such applications, large, general corpora such as the Wall Street Journal or the British National Corpus, are used to discover automatically semantic relationships of the kind found in general, manually-constructed lexical resources such as WordNet (Fellbaum 1998) or Roget 's Thesaurus (Roget 1911).[2]

The use of distributional similarity techniques to predict semantic relationships between terms in a specialised area of knowledge (i.e., biomedicine) has at least two important consequences for the present work. First, it is necessary to employ parsing techniques that can deal reliably with text containing terminological units. Knowledge of multi-word terminology is vital for parsing accuracy in the biomedical domain. Second, in practice, the specialised domain coupled with the need for term annotation results in a much smaller corpus than used in other applications, where the words of interest typically may be assumed to occur over one hundred times. In contrast, the majority of the terms in the domain-specific corpus used in our work occur less than ten times. Consequently, it is necessary to find a technique that will perform well in the presence of very sparse data.

3. The GENIA domain

The GENIA corpus (Kim, J.-D. and Tsujii 2003) consists of 2000 titles and abstracts collected from the MEDLINE repository. The MeSH headings "*human*", "*blood cell*" and "*transcription factor*" were singled out to create a document collection around the topic of biological reactions concerning transcription factors. The resulting documents comprise more than 400,000 words, and have been semi-automatically annotated with part of speech information and manually annotated for terminology. Each instance of a term in the document collection is additionally assigned a single, unambiguous semantic type.

These types are organised into an IS_A hierarchy representing a coarse grained semantic distinction. The resulting hierarchy is known as the GENIA ontology, and is shown here in Figure 1. The ontology can be considered at different levels of specificity. Level 0 is the most specific and corresponds to the leaf nodes of the ontology as shown in the figure. Level 5 is the most general and only involves the three nodes at the top of the ontology, which subsume all other levels.

The huge annotation effort that goes into creating such a resource brings clear advantages for NLP systems. Terminology extraction still remains a semi-automated process involving statistical, linguistic and hybrid algorithms (Castellvi et al. 2001) the results of which always need manual validation. The ability to side step this issue and simulate near perfect terminology extraction allows research effort to be concentrated elsewhere, without the fear that inadequate or inappropriate term extraction methodologies may introduce noise in subsequent processing. The drawback however, is the relatively small size of the corpus.

Language resources used in the development and evaluation of NLP systems typically involve syntactic and/or semantic annotations and have a lower limit of 100,000 words (Marcus et al. 1993; Baker et al. 2003). Whilst the GENIA annotations are invaluable, the considerable effort required to create them keeps the collection at the smaller end of the scale. This is a potential problem for techniques where sparse data is known to adversely effect performance, but

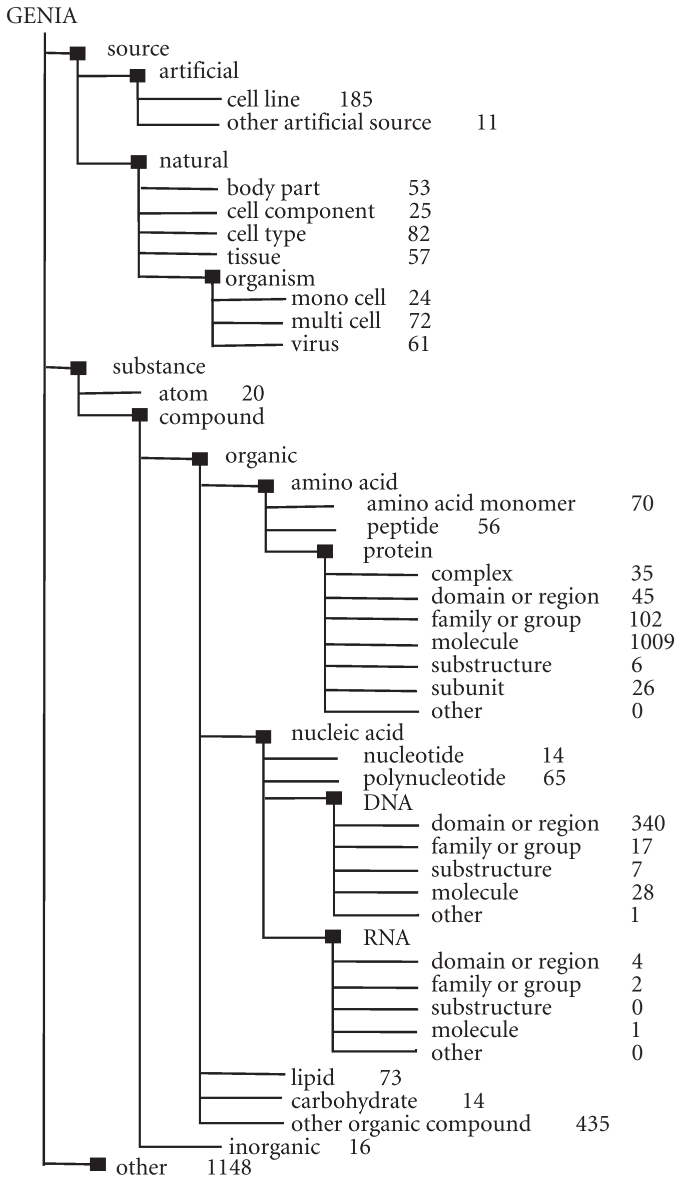


Figure 1. The GENIA ontology

it does reflect the practical problem that technical document collections tend to be smaller than open domain collections for reasons of availability, copyright restrictions and the nature of the subject matter. The GENIA corpus therefore provides a realistic test of performance for a data-driven application such as distributional similarity.

The GENIA corpus is encoded in XML and the ontology is distributed in the DAML+OIL format (Connolly et al. 2001). Terminology is identified using XML tags, with the semantic type of a term as a tag attribute. Syntactically, the terminology takes the form of noun phrases (NPs), the vast majority of which are minimal NPs although coordinated NPs are also represented. In the more complex cases, such as ellipsis in coordinated clauses, the underlying markup disambiguates the terminology as far as possible. The GENIA terminology does not include NPs with attached prepositional phrases as these phrases are considered to consist of distinct terminological units. In total, the corpus identifies 76592 such instances of terms with each assigned one of 36 semantic types. There are two steps in defining the terminological unit for further processing: **term normalisation** and **class identification**.

Term normalisation is designed to identify term instances that refer to the same underlying concept due to arbitrary punctuation use. With larger ontological resources (such as the UMLS (NLM 1998)) term normalisation is aggressive in the sense that terms are lower-cased and stripped of punctuation before the words are sorted alphabetically to produce a normalised representation. Here normalisation is more relaxed, removing punctuation from a word only if the resulting stripped word appears elsewhere in the terminology and the linear order is preserved. This results in 31398 normalised terms.

Next, the normalised terms are gathered into **terminological classes** by exploiting the natural endocentricity of nominal compounds (Barker and Szpakowicz 1998). Following lemmatization using Morpha (Minnen et al. 2001), the head identification algorithm chooses the rightmost non-symbolic word. This excludes words that consist of a sequence of numeric characters, a mixture of alpha-numeric characters or just a single alphabetical character. This ensures that the terms “HMG 88” and “HMG 1” are gathered into the same class. The result of class identification is a set of natural classes of terms that share a common head noun.

This pre-processing of the terminology results in 4797 terminological classes out of which 4104 contain terms with identical semantic types and 558 classes contain terms with 2 or 3 semantic types. A further 135 classes contain terms with more than three semantic types and represent miss-classification due to the highly symbolic nature of the constituent terms and the fact that the head identification algorithm does not take character casing into account. This results, for example, in “75 kD” (of type *protein molecule*), “Kd” (*other name*) and “105 KD” (*peptide*) being grouped together. The number of single typed classes for each level in the ontology is given in Figure 1.

4. The parser

Syntactic analysis of the GENIA corpus is performed by Pro3Gres, a dependency-based linguistic parser that broadly follows the architecture suggested by (Abney 1995). The analysis moves from shallow to deep processing, combining rule-based and statistical decision-making processes to analyse input sentences. The parser makes use of nominal and verbal chunking as a foundation for the dependency rules and a statistical model to build the predicate argument structure between the chunks’ heads. Such hybridisation of chunking and dependency parsing has proven to be practical, fast and robust (Collins 1996; Basili and Zanzotto 2002). By optimising the trade-off between computational efficiency and formal expressivity, Pro3Gres is capable of processing more than 300,000 words per hour.

A hand-written dependency grammar is used to identify possible syntactic structures within each sentence. The grammar contains around 1000 dependency rules, each involving the part-of-speech (POS) tags of a head and its dependent, the dependency relation, lexical information and contextual restrictions. The restrictions express sub-categorisation constraints, such as that only a verb which has an object in its context is allowed to attach a secondary object. The possible syntactic analyses proposed by the dependency grammar are ranked and pruned statistically during parsing, by combining attachment probabilities for the dependency relations used in the grammar. These probabilities were acquired automatically from the Penn Treebank (Marcus et al. 1993). This method of parse selection can be seen as a generalisation of the statistical approach to prepositional phrase attachment developed in (Collins and Brooks 1995). The parser also provides a graceful fallback through partial analysis if no complete parse is available, and uses incrementally aggressive pruning techniques for very long sentences.

Typical examples of the parser output are shown in Figures 2 and 3. The diagrams show the identified GENIA terminology (in boxes), minimal chunks (marked by square braces) and labelled dependency relations between the

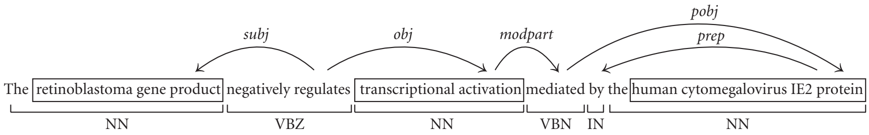

pobj
subj
obj
modpart
prep
The
retinoblastoma gene product
negatively regulates
transcriptional activation
mediated
by
the
human cytomegalovirus IE2 protein
NN
VBZ
NN
VBN
IN
NN

Figure 2. Example sentence parse

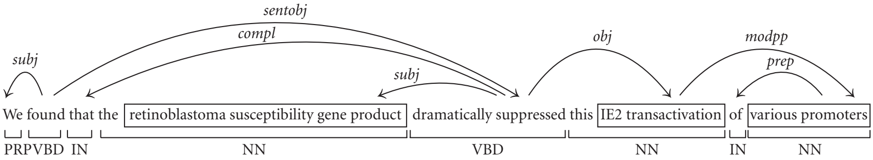


Figure 3. Example sentence parse

heads of chunks (shown as arrows). For example, in the parse of Figure 2, the verb "regulate" has as its subject (*subj*) the chunk "retinoblastoma gene product" and as its object (*obj*) the chunk "transcriptional activation". The latter is modified by a reduced relative clause (*modpart*) with head verb "mediated", which in turn has a prepositional phrase "by ... protein" as dependent. Figure 3 shows an example of a subordinate clause *sentobj* relation introduced by an optional complementizer *compl*. The subordinate object is modified by a prepositional phrase (*modpp*).

Unlike traditional statistical parsers (such as (Collins 1999)) Pro3Gres expresses the majority of long-distance dependencies (Schneider 2003). This is achieved by:

1. relying on Dependency Grammar characteristics
2. expressing Long-Distance Dependencies (LDD) as local dependencies with a dedicated label
3. using statistical post-processing

An example of 4 is the *modpart* (*modification by participle* or *reduced relative*) relation illustrated in the parse of Figure 2, which is assumed to involve a long-distance dependency in the Penn Treebank. The underlying object (past participle) or subject (present participle) relation is recoverable thanks to the dedicated label. Statistical post-processing (4) is used to handle cases involving control relations such as *subject control*. For example, in the sentence "John wants to leave", the proper noun "John" functions not only as the explicit subject of

Table 1. Evaluation of Pro3Gres over 100 random sentences from the GENIA corpus

Parsing	Subject	Object	Noun-PP	Verb-PP	Sub clause
WITH terminology					
Precision	90	94	83	82	71
Recall	86	95	82	84	75
WITHOUT terminology					
Precision	83	70	68	67	63
Recall	75	77	64	68	60

"want", but also as the implicit subject of "leave". A parser that fails to recognize control subjects misses important information (quantitatively, about 3% of all subjects). The lexicalised, statistical post-processing step for control relations selectively converts the dependency tree structure into a graph structure.

The language of the GENIA corpus is very complex and technical, which is attested by the unusually high average sentence length (27 words) and a high token to chunk ratio for NPs (2.3 tokens per chunk). To evaluate the parser performance in this domain, we manually annotated a sample of 100 sentences that had been randomly selected from the GENIA corpus. The manual annotations were the subject, object, PP-attachment and subordinate clause relations. We first ran the parser over the 100 sentences without any consideration of terminology. In this case, the minimal NP and VP chunks used by the parser were solely determined by the LTCHUNK chunker (Finch and Mikheev 1997). Next, we performed the analysis over the same 100 sentences, but using the near-perfect terminology identification provided by the GENIA annotations. A comparison of the results is presented in Table 1.

The results presented in the table show two things. First, despite the complexity of the language represented by the sample sentences, it is clear that the parser is performing very accurately. Second, knowledge of terms has an important and often dramatic impact on parsing performance. Multi-word terminology is known to cause serious problems for NLP systems (Sag et al. 2002; Dowdall et al. 2003) and is a notable characteristic of the biomedical domain represented by the GENIA corpus. The object relation precision is most affected, because many deverbal adjectives such as "reduced" (as in "reduced PMA/Ca2+ activation") may be erroneously interpreted as verb-object relations. The high precision and recall of subject and object relations is of particular importance here as these dependencies provide the contextual features needed to determine distributional similarity between terms.

5. Distributional similarity

In this section, we first introduce the concept of distributional similarity and describe its application to the discovery of semantic relationships. We then discuss three distributional similarity methods used in the literature and in our experimental work.

5.1 Introduction

The intuition underlying distributional similarity is that two words are distributionally similar if they appear in similar contexts. Context, however, can be modelled at a number of different levels. For example, two words might be considered to appear in the same context if they occur in the same document, or the same sentence, or the same grammatical dependency relation (e.g. as the nominal subject or object of a particular verb). In automatic thesaurus generation, it is usual to take grammatical dependency relations as contextual features, since this leads to *tighter* thesauruses (Kilgarriff and Yallop 2000), in which words are related via linguistic relations such as synonymy, hyponymy and antonymy rather than topical relations as might be found in Roget.

Without loss of generality, the similarity between any two words can be defined on a continuous scale between 0 and 1, where 1 represents apparent identity and 0 represents no observed overlap. Thus, one can think of the **neighbours** of a word *w* as being those words that can be ranked in terms of their similarity to *w* (i.e. the set of words which have a non-zero similarity with respect to *w*). In practice however, there may be many neighbours of a word *w* which have very small but non-zero similarity scores. For this reason, it is often more useful to consider only the *k* **nearest neighbours** of *w*, where the parameter *k* may be varied for practical reasons, such as the quantity of text data used to gather word context or the particular application of a thesaurus.

5.2 Measures of distributional similarity

A number of methods have been proposed or adopted for calculating distributional similarity. These measures have been shown to have differing characteristics (Lee 1999; Weeds et al. 2004) which make them useful for different applications or on different data-sets. In this section, we present three distributional similarity methods which have been proposed or adopted in the automatic thesaurus generation literature, and which are used in our experimental work.

These methods are the L_1 Norm, Lin's measure and co-occurrence retrieval (CR). For a more extensive review of measures of distributional similarity, see (Weeds 2003).

In order to increase readability, throughout the following discussion we consider finding similarity between two nouns n_1 and n_2. However, it should be noted that distributional similarity techniques are equally applicable to other parts of speech. We also refer to calculating the similarity between two nouns in terms of their set of **dependency features**, where a dependency feature is a grammatical context in which a noun has occurred within some text corpus. For example, the noun *apple* might have the dependency feature <*apple*, *direct-object-of*, *eat*> (amongst many others), while the noun *girl* may have the distinct dependency feature <*girl*, *subject-of*, *eat*>. The collection of all the contextual features for a given noun defines a point in a multi-dimensional space, and it is the similarity between points in this space which we attempt to measure. Most measures of distributional similarity also take into account the (conditional) probabilities $P(f|n)$ with which each dependency feature f is observed to occur with a given noun n.

5.2.1 *L_1 Norm*

The L_1 Norm is a member of a family of measures, known as the Minkowski Distance, for measuring the distance between two points in space. Distance measures, also referred to as divergence and dissimilarity measures, can be viewed as the inverse of similarity measures; that is, an increase in distance correlates with a decrease in similarity. The L_1 Norm represents the distance travelled between two points given that it is only possible to travel in orthogonal directions and for two nouns, n_1 and n_2 can be written as:

$$dist_{L_1}(n_1, n_2) = \sum_f |P(f|n_1) - P(f|n_2)| \qquad (1)$$

A feature of the L_1 Norm, as shown in (Dagan et al. 1999), is that it can be calculated by considering just the dependency features that occur with both nouns. Consequently, any nouns that do not share any dependency features are at a maximal distance of 2. Conversely, nouns that have identical distributions of dependency features have zero distance between them.

We chose to study the L_1 Norm in this work because it is a popular measure in clustering e.g. (Kaufman and Rousseeuw 1990; Schütze 1993; Dagan et al. 1999) and, whilst being simple to calculate, it has been shown to be as effective as more complicated similarity measures (Lee 1999). Further, recent work (Weeds 2003; Weeds et al. 2004) has shown the L_1 Norm to perform consistently for high and low frequency words, which is likely to be important in this work.

5.2.2 *Lin's measure*

Lin's measure (Lin 1998a) is an information-theoretic measure of similarity which has been shown to perform well in comparison to other measures (Lin 1998a; Weeds 2003) and is becoming a popular choice in applications of distributional similarity (Wiebe 2000; Kilgarriff 2003; McCarthy et al. 2003). It is based on Lin's information-theoretic similarity theorem (Lin 1997, 1998b):

> *The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.*

The information in a description of a word can be measured as the sum of the **pointwise mutual information** (MI) between the word and each dependency feature in the description of the word. The MI between two events measures their relatedness or degree of association (Church and Hanks 1989), and for a noun n and a dependency feature f it can be written as:

$$I(n,f) = \log \frac{P(f,n)}{P(f).P(n)} = \log \frac{P(f|n)}{P(f)} \tag{2}$$

This measures the extent to which the probability of feature f is increased by knowing that the noun is n (or, since it is symmetric, how much the probability of noun is n is increased by knowing that the feature is f). Negative values indicate that the probability of f decreases if we know that the noun is n and a value of zero indicates that the feature and the noun occur together no more or less frequently than one would expect by chance (i.e. assuming independence).

With this definition of MI, the similarity between two nouns n_1 and n_2 can be calculated using Lin's measure as:

$$sim_{lin}(n_1, n_2) = \frac{\sum_{T(n_1)\cap T(n_2)}(I(n_1,f) + I(n_2,f))}{\sum_{T(n_1)} I(n_1,f) + \sum_{T(n_2)} I(n_2,f)} \tag{3}$$

where $T(n) = \{f : I(n,f) > 0\}$. $T(n)$ thus contains the most salient dependency features of a noun n (i.e., those which increase the expectation that the noun is n). Since only these dependency features are considered in the calculation, two nouns n_1 and n_2 will have similarity 0 if there is no overlap in their sets of most salient features (i.e., $T(n_1) \cap T(n_2) = \emptyset$) and they will have similarity 1 when their sets of most salient features are identical (i.e., $T(n_1) = T(n_2)$).

We chose to study Lin's measure in this work because of its wide application and its high performance in previous work. However, Lin's measure has been shown to perform less well at predicting semantically related words for low frequency target words in the general domain (Weeds 2003) and thus we might expect it not to perform as well as other measures in this study.

5.2.3 *Co-occurrence retrieval*

Co-occurrence retrieval (CR), (Weeds and Weir 2003b; Weeds 2003), is based on the idea that similarity between words can be measured by analogy with document retrieval. In document retrieval, there is a set of documents that we would like to retrieve and a set of documents that we actually do retrieve. If we are testing the appropriateness of using one word, n_1, in place of another, n_2, then there is a set of co-occurrences that we would like to retrieve (the dependency features of n_2) and a set of co-occurrences that we do retrieve (the dependency features of n_1). In both document retrieval and co-occurrence retrieval, we can measure the similarity of the two sets in terms of **precision** and **recall**, where precision tells us how much of what was retrieved was correct and recall tells us how much of what we wanted to retrieve was actually retrieved.

An advantage of using co-occurrence retrieval to measure similarity is that it differentiates between two types of dissimilarity (low precision and low recall). When n_1 occurs in contexts that word n_2 does not, the result is a loss of precision, but n_1 may remain a high recall neighbour of n_2. When n_1 does not occur in contexts that n_2 does occur in, the result is a loss of recall, but n_1 may remain a high precision neighbour of n_2. Six different models for calculating precision and recall are proposed in (Weeds 2003). Here we consider only one of these models, the additive, Mutual Information (MI) based CRM, which was shown to consistently outperform the other models (Weeds 2003). In this model the set $T(n)$ of salient dependency features of a word n are first selected using MI:

$$T(n) = \{f : I(n,f) > 0\} \tag{4}$$

The shared features of noun n_1 and noun n_2 are referred to as the set of True Positives (TP):

$$TP = T(n_1) \cap T(n_2) \tag{5}$$

The precision of n_1's retrieval of n_2's features is the proportion of n_1's features that are shared by both nouns, where each feature is weighted by its relative importance according to n_1 (i.e., its MI with n_1):

$$\mathcal{P}(n_1, n_2) = \frac{\sum_{TP} I(n_1, f)}{\sum_{T(n_1)} I(n_1, f)} \tag{6}$$

The recall of n_1's retrieval of n_2's features is the proportion of n_2's features that are shared by both nouns, where each feature is weighted by its relative importance according to n_2 (i.e., its MI with n_2):

$$\mathcal{R}(n_1, n_2) = \frac{\sum_{TP} I(n_2, f)}{\sum_{T(n_2)} I(n_2, f)} \tag{7}$$

Precision and recall both lie in the range [0,1] and are both equal to 1 when each noun has exactly the same features. It should also be noted that the recall of n_1's retrieval of n_2 is equal to the precision of n_2's retrieval of n_1, i.e., $\mathcal{R}(n_1, n_2) = \mathcal{P}(n_2, n_1)$.

(Weeds 2003) investigates a parameterised framework which combines precision and recall with different weights. Here, we consider just one other setting of the framework, which is known as the F-score in Information Retrieval and is the **harmonic mean** of precision and recall:

$$\mathcal{F} = m_h(\mathcal{P}(n_1, n_2), \mathcal{R}(n_1, n_2)) = \frac{2.\mathcal{P}(n_1, n_2).\mathcal{R}(n_1, n_2)}{\mathcal{P}(n_1, n_2) + \mathcal{R}(n_1, n_2)} \tag{8}$$

Note that the harmonic mean of two numbers lies between them, but is always substantially closer to the lower one of the two and attains a maximum when they are equal. In other words, for two words to be considered highly similar by this score, both precision and recall must be high.

We use co-occurrence retrieval in this work as it has been shown to be a useful way of classifying different similarity measures (Weeds et al. 2004). Further, high recall neighbours have been shown to bear more resemblance to sets of neighbours derived from WordNet than high precision or high harmonic mean neighbours in previous work (Weeds 2003). This effect was particularly apparent for low frequency words and thus we would expect high recall neighbours to be more useful here.

6. Evaluating an automatically generated thesaurus

In this section we describe a number of experiments that were conducted in order to evaluate the application of an automatically generated thesaurus to the problem of organising the GENIA terminology. More specifically, our aim was to test the following hypotheses regarding the use of distributional similarity in this domain:

1. distributional similarity predicts semantic similarity for terminology;
2. distributional similarity permits accurate classification of terminology within an existing domain ontology.

A problem that immediately arises in this context is that of data sparseness. The comparatively small size of the GENIA Corpus, coupled with the Zipfian (Zipf 1949) nature of word distribution, means that we have very little co-occurrence data for many of the terms in which we are interested. For example, while there are 31398 terms identified within the GENIA corpus, of these only 1935 (6.2%) occur more than 5 times. It has generally been assumed that the effective application of distributional similarity techniques requires large quantities of data about each word. For example, (Lin 1998a) applies distributional similarity techniques to the problem of automated thesaurus construction, using a 64 million word corpus and only calculating similarity for nouns that occur at least 100 times.[3]

While it would be desirable to substantially extend the corpus before applying distributional similarity techniques, this is not straightforward. Automatic annotation of terminology is not sufficiently accurate for our purposes, and hand-annotation is time-consuming. Instead, we partially address the problem of data-sparseness by applying distributional similarity to the terminological classes rather than the individual terms themselves. This is possible because terms within the same class tend to have the same semantic type. Nevertheless, of 1576 terminological classes, over 50% are represented fewer than five times in the corpus. The number of classes that occur at different frequencies (up to a frequency of 40) is shown in Figure 4. As a consequence, we may expect that the successful application of distributional similarity methods in this domain will still rely on finding a similarity measure that works well for low frequency items. For this reason, in the following experiments we report on the comparative performance of several of the measures described in Section 5.

As a basis for calculating the distributional similarity scores, the GENIA corpus was syntactically analysed using the Pro3Gres parser. The resulting de-

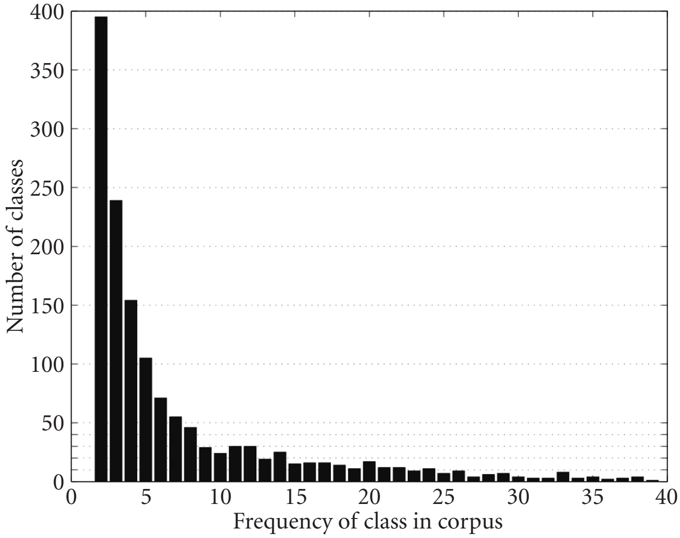


Figure 4. Number of terminological classes with each corpus frequency

pendency parses were then used to extract all those dependency relations of the form ⟨*n*, *subject*, *v*⟩ or ⟨*n*, *object*, *v*⟩, where *n* is a head noun (possibly representing a terminological class), and *v* is a verb. The resulting set of dependency triples provided the raw data required to determine distributional similarity according to the different similarity measures discussed in Section 5: the L_1 Norm (L_1), Lin's measure (`Lin`) and CR (recall ($\mathcal{R}$), precision ($\mathcal{P}$) and harmonic mean ($\mathcal{F}$)). Given a measure of distributional similarity and a set of dependency triples, we found for each terminological class *c* the set of all its neighbours. In general, not every neighbour of a terminological class *c* will itself represent a terminological class. In the following section, where there is a need to restrict attention to just the terminological classes amongst the neighbours, we will refer to these as the **terminological neighbours**.

The neighbours of a class *c* can be ranked according to similarity, so that the neighbour that is most similar to *c* has rank 1, the next most similar rank 2, and so forth. Sets of neighbours were computed twice for each measure: once using all of the available *subject* and *object* dependency triples, and once using just those triples ⟨*n*, *r*, *v*⟩ where the noun *n* had occurred at least five times in the corpus. This was done to allow us to examine the effect of class frequency on the performance of the different distributional similarity measures.

6.1 Distributional similarity and semantic relatedness

One possible way of comparing the ability of the different similarity measures to predict semantic similarity might be to consider the following simple decision task: given three terminological classes, c_1,c_2 and c_3, the goal is to determine whether c_1 is more closely related to c_2 or to c_3. An instance of this task is thus a triple of classes, in which the first and second classes are chosen so as to belong to the same semantic type, while the third belongs to a distinct type. Note that the correct decision is to select class c_2. However, a given measure of distributional similarity will select either c_2 or c_3 depending on which one is distributionally closest to c_1 according to that measure. The measure that is most successful at this task over many trials (i.e., most often chooses c_2 when presented with a large number of different problem instances) may be regarded as the best at predicting semantic relatedness.

While this task is an intuitively appealing way to evaluate the relationship between distributional similarity and semantic similarity, in the present context it turns out to be somewhat problematic. Because of data sparseness, for any given similarity measure, the vast majority of the possible triples c_1, c_2, c_3 will be such that the similarity score of c_1 and c_2 and the similarity score of c_1 and c_3 are extremely low (possibly zero). Unfortunately, such low similarity scores do not provide a reliable basis for choosing between the classes and it turns out to be impossible to make an informed choice about semantic relatedness of terminological classes in a very large number of cases.

One way of attempting to overcome this problem is to perform the evaluation using only those classes where the similarity scores are greater than a given, reasonably large threshold. However, this approach no longer provides a fair comparison of the different similarity measures. This is because different measures may exhibit considerable variation in the rate at which the similarity score drops off as more distant neighbours of a class are considered. For some measures, similarity scores drop off rapidly (yielding a fairly compact set of neighbours) while for others they tail away slowly (yielding a larger and more diffuse set of neighbours). As a result, for triples chosen to evaluate measures where similarity scores drop off rapidly, c_1 and c_3 would typically be closer neighbours than in triples chosen to evaluate measures with similarity scores that tail off slowly, which would presumably favour the latter measures.

In order to avoid these problems, we considered an alternative means of evaluation that is not sensitive to the absolute score that a measure assigns to its neighbours. This is based on the (reasonable) assumption that a distributional similarity measure provides a good basis for determining semantic relatedness of terminological classes if it exhibits a strong, positive correlation between neighbour *rank* and error rate in predicting semantic type. The stronger this correlation, the better the similarity measure at predicting semantic relatedness.

6.1.1 *Neighbour ranking and semantic proximity*

For a given measure of distributional similarity, we calculated the correlation between neighbour rank and error rate. Taking the ranked set of 100 nearest neighbours produced for a given terminological class c, we considered each rank in turn. The ith-ranked neighbour n_i was labeled "correct" if it represented a terminological class with a semantic type matching that of c, and "incorrect" if it represented a terminological class with a semantic type differing from that of c (no label was assigned for neighbours that did not represent terminological classes). Note that in order to avoid equivocation, neighbours with more than one semantic type were also left unlabeled.[4] The error rate at each rank i was then calculated over all of the terminological classes, as the proportion of all the labeled neighbours at rank i that were assigned the label "incorrect".

We might expect the error rate to be affected by the granularity of the classification system used in order to determine the label for each neighbour. The most fine-grained level corresponds to the leaf nodes of the GENIA ontology (level 0) so that a neighbour is labeled as correct or incorrect depending on which of the 36 different leaves it corresponds to. As we "move up" the hierarchy the classification becomes increasingly coarse-grained, until we reach the top of the ontology (level 5) where the labeling decision is made on the basis of which of just 3 different sub-trees of the ontology the neighbour belongs to: *source*, *substance*, or *other*. In order to examine the effect of granularity, we calculated error rates at each of the 6 different levels of the GENIA ontology.

6.1.2 *Results*

The results of the rank correlation experiments are shown in Table 2(a) and Table 2(b). The value of Spearman's rank correlation coefficient[5] calculated between neighbour rank and error rate rank is shown for each level in the ontology and each similarity measure discussed in Section 5, L_1 Norm (L_1), Lin's measure (`Lin`) and CR (recall ($\mathcal{R}$), precision ($\mathcal{P}$) and harmonic mean ($\mathcal{F}$)).

As the figures clearly show, a high positive correlation is demonstrated in all cases. This tells us that neighbour rank reflects the gradient of semantic similarity, with distant neighbours more likely to make an error in matching the semantic type of the target class than close neighbours. The highest positive correlation seen for all frequencies at level 0 in the ontology is for the recall measure (0.934). A scatter plot of neighbour rank against error rate for this case is presented in Figure 5(a). The lowest correlation is seen for the precision measure, which is illustrated in the scatter plot of Figure 5(b).

These results also show that different distributional similarity measures are more effective for different frequencies. For example, over all frequencies, the L_1 Norm outperforms Lin whereas over just the higher frequency terms, Lin outperforms the L_1 Norm. This supports earlier work which suggests that MI and, in particular, Lin's Measure perform poorly for low frequency events (Resnik 1993; Fung and McKeown 1997; Kilgarriff and Tugwell

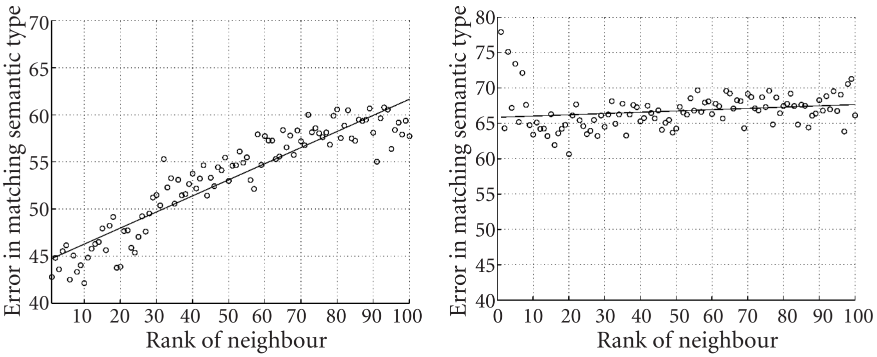
Table 2. Correlation coefficients

	L_1	Lin	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$
5	0.917	0.880	0.920	0.681	0.925
4	0.916	0.887	0.918	0.701	0.923
3	0.917	0.881	0.923	0.707	0.931
2	0.891	0.885	0.933	0.590	0.912
1	0.877	0.886	0.926	0.507	0.909
0	0.862	0.849	0.934	0.380	0.892

(a) Terminological classes of all frequencies

	L_1	Lin	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$
5	0.812	0.842	0.889	0.897	0.887
4	0.802	0.846	0.891	0.898	0.897
3	0.785	0.830	0.894	0.894	0.890
2	0.791	0.864	0.913	0.895	0.887
1	0.799	0.869	0.914	0.900	0.892
0	0.773	0.873	0.911	0.900	0.891

(b) Terminological classes with frequency ≥ 5



(a) *CR Recall Measure* (b) *CR Precision Measure*

Figure 5. Correlation between CR recall measure and CR precision measure and error in semantic type prediction for terminological classes of all frequencies

2001; Weeds and Weir 2003a; Wu and Zhou 2003; Weeds 2003). The high performance of $\mathcal{R}$ and $\mathcal{F}$, which also use MI to select and weight features, supports the claim that MI can be effective for weighting features for low frequency words, provided that only words with high recall of the selected features are considered as neighbours (Weeds 2003; Weeds et al. 2004). However, as can be seen here, frequency of terms only has to increase to a minimum of five for high precision neighbours to also exhibit good correlation with semantic similarity.

It is also possible to read off from these graphs the error with which the first neighbour (and each subsequent neighbour) assigns the correct semantic type to each target terminological class. The error rates for the first neighbour for all measures and all ontological levels are given in Table 3 and Table 4. The tables also contain figures for random classification at each level, as well as a more informed baseline score. The baseline represents the error which would be observed if the first neighbour was always a member of the most populous semantic type (i.e., the semantic type to which most classes belong) at each level in the ontology. For example, at level 0, the most populous semantic type is `other_name`.

Table 3. Error in first neighbour's prediction of semantic type for terminological classes of all frequencies (with one semantic type)

	L_1	`Lin`	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	Random	Baseline
5	29.0	27.8	32.2	41.0	28.2	66.7	49.8
4	30.8	29.5	33.4	42.1	30.0	80.0	50.6
3	31.4	30.6	33.7	42.7	31.0	90.9	50.9
2	46.7	46.7	40.7	60.1	47.5	94.1	57.4
1	48.3	49.0	41.6	61.3	49.0	95.2	57.4
0	52.8	53.8	42.7	77.9	53.2	96.6	57.4

Table 4. Error in first neighbour's prediction of semantic type for terminological classes of frequency ≥ 5 (with one semantic type)

	L_1	`Lin`	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	Random	Baseline
5	16.3	14.6	13.7	39.3	14.5	66.7	35.1
4	16.3	15.1	13.7	39.3	14.9	80.0	35.1
3	16.3	15.1	13.7	39.6	14.9	87.5	35.1
2	21.6	20.4	18.6	47.3	20.9	91.2	35.1
1	21.6	20.4	18.6	47.3	20.9	92.9	35.1
0	22.1	20.8	18.6	47.3	21.3	94.7	35.1

Note that for terminological classes of all frequencies, regardless of similarity measure, the first neighbour is doing far better than chance in predicting the semantic type of the terminological class. With the exception of the precision measure $\mathcal{P}$, the measures are also doing better than the baseline. A very similar picture emerges from Table 4, which also shows that the error rate decreases for higher frequency terminological classes.

With regard to different similarity measures, the results follow the same pattern as for the correlation results. The lowest error rate in prediction of semantic type by the first neighbour is achieved by $\mathcal{R}$ and the highest error rate by $\mathcal{P}$. $\mathcal{F}$, which combines precision and recall, gives intermediate results which are substantially closer to those of $\mathcal{R}$ than those of $\mathcal{P}$. `Lin`, which has been shown (Weeds 2003) to be approximated by $\mathcal{F}$, gives similar results to $\mathcal{F}$ and is the only measure which performs better, relative to other measures ($\mathcal{F}$ and L_1) for high frequency terms.

In summary, the ability of a neighbour to make the correct prediction as to the semantic type of a terminological class tends to decrease as the neighbour becomes more distant (i.e., error is correlated with distributional distance). This supports our first hypothesis that distributional similarity is correlated with semantic similarity. Of the different measures, $\mathcal{R}$ appears to perform the best and $\mathcal{P}$ appears to perform the worst. This means that a useful neighbour needs to have high recall of the most salient features of a terminological class.

While the correlation scores do not vary greatly at different levels in the ontology, the error rate does improve as we move up the ontology. This is to be expected to some extent, as the random assignment of a semantic type will also improve as the number of possible choices decreases. More telling is the observation that the reduction in error rate for the similarity measures generally outstrips that of the baseline.

There is a significant improvement when we only consider terminological classes that have occurred five or more times in the corpus. In part, this could be due to the improvement in the baseline, since the proportion of classes which should be assigned to the most populous semantic type also increases when we consider only the most frequently occurring terminological classes. However, it is also what one would expect given that there is more corpus data for each terminological class for which we are determining neighbours. The overwhelming conclusion here is that even with relatively little corpus data (the majority of terminological classes occurring fewer than 10 times), it is possible to see a clear correlation between distributional similarity and semantic proximity.

6.2 Distributional similarity and classification of terminology

An important potential application of distributional similarity techniques is the organisation of terminology. To determine the extent to which distributional similarity can be used successfully to classify terminology, we considered the problem of assigning an "unknown" terminological class *c* to a semantic type at the most fine-grained level of the GENIA ontology (i.e. leaf nodes at level zero). Our approach makes use of the set of nearest neighbours of a terminological class *c* to select a semantic type for *c* according to a "majority vote" strategy.

6.2.1 *Neighbour selection of semantic type*

Given the observed correlation between neighbour rank and semantic similarity, we might expect the nearest neighbours of a terminological class to be good predictors of its semantic type. To test this, we took each terminological class *c* in turn and found its *k* nearest terminological neighbours. Each of the *k* terminological neighbours of *c* was then used to score the 36 possible semantic types at level 0 of the GENIA ontology. For a neighbour with exactly one semantic type, a score of 1 was assigned to that type; for a neighbour with *N* different semantic types, the score was split equally amongst them, so that each type received a score of $1/N$. The scores obtained in this way were summed over the *k* neighbours of *c*, which was then predicted to belong to the semantic type which received the highest overall score (ties were broken randomly). The type prediction for a terminological class *c* was judged to be correct if *c* belonged to that class according to the GENIA ontology, and otherwise it was judged to be incorrect. Note that in case *c* belonged to several classes, then any one of them would be regarded as correct.

The prediction of semantic type described above is parameterised by the choice of *k*: the number of nearest neighbours that are considered in scoring the different possible types. To investigate the effect that this choice has on prediction accuracy, we ran experiments for different settings, with $k = 10$, 20, 30 and 40. As before, we also considered neighbour sets calculated with reference to all terminological classes, and neighbour sets calculated for those classes represented five or more times.

6.2.2 *Results*

Results showing the percentages assigned correctly for each measure and at each value of *k* are shown in Table 5(a) and Table 5(b). The baseline for

Table 5. Accuracy at assigning semantic types using 10, 20, 30 and 40 nearest neighbours

	L_1	`Lin`	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	Baseline
10	59.4	60.9	54.6	48.0	61.3	42.0
20	60.2	63.0	54.8	54.7	63.0	42.0
30	59.7	62.9	54.8	54.1	63.1	42.0
40	59.3	63.1	53.9	54.8	62.6	42.0

(a) Terminological classes of all frequencies

	L_1	`Lin`	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	Baseline
10	72.3	76.5	69.0	64.1	75.7	58.0
20	71.7	77.0	68.8	70.0	76.8	58.0
30	71.1	75.1	67.0	72.0	75.6	58.0
40	72.4	74.9	66.3	73.0	76.0	58.0

(b) Terminological classes with frequency ≥5

each experiment is calculated as the percentage that would be assigned correctly if every terminological class was assigned to the largest semantic type (`other_name`).

As the results show, all of the measures perform well above the respective baselines in each experiment. The optimal performance over terminological classes of all frequencies is 63.1% and is achieved using the $\mathcal{F}$ measure and 30 nearest neighbours. Examining the results more closely shows that, while the closest neighbours do not always assign the correct semantic type, errors made by these close neighbours can be corrected, to a certain extent, by accumulating evidence from a larger number of more distant neighbours. On the other hand, there comes a point, at around $k = 20$, when the votes of subsequent neighbours begin cancelling each other out, as if these so-called neighbours had been selected at random.

Combining evidence from multiple neighbours produces a different pattern, with respect to similarity measure, from that observed in our earlier experiments. When regarded individually, high recall neighbours showed the highest correlation with semantic similarity. When evidence is combined from multiple neighbours, on the other hand, L_1, `Lin` and $\mathcal{F}$ all outperform $\mathcal{R}$. Best performance over all frequencies is achieved by $\mathcal{F}$ and best performance for higher frequency terms is achieved by `Lin`. Both of these measures require neighbours to have high precision and high recall retrieval of features. This suggests that while precision may introduce some noise into the ranking of neighbours, this noise can be effectively filtered out by considering a cluster of neighbours.

A more detailed analysis of the accuracy of the first ten neighbours[6] at assigning each of the 36 level 0 semantic types in the ontology is presented in Table 6. We report only the analysis for the $\mathcal{F}$ measure as this was the measure that performed best overall, but note that the general pattern observed in the results is typical of all of the measures. The analysis is given in terms of recall

Table 6. Precision and recall in assigning each semantic type using the 10 nearest neighbours of a terminological class

	Class population	All frequencies		Frequency > 5	
		Recall	Precision	Recall	Precision
peptide	56	0	–	0	–
RNA_family_or_group	17	0	–	0	–
amino_acid_monomer	70	0	–	0	–
nucleotide	14	0	–	0	–
cell_component	25	0	–	0	–
cell_type	82	7.79	66.7	8.7	100
protein_N/A	0	0	–	0	–
virus	16.7	61	30.0	14.3	67
polynucleotide	65	0	–	0	–
DNA_domain_or_region	340	9.6	20	8.4	100
DNA_molecule	1009	5.6	100	6.25	100
protein_subunit	26	4.54	33.3	4.76	50
tissue	57	0	–	0	–
mono_cell	24	0	–	0	–
DNA_N/A	1	0	–	0	–
other_artificial_source	11	0	–	0	–
atom	20	0	–	0	–
other_organic_compound	435	68.8	46.0	44.4	44.0
protein_family_or_group	102	1.8	22.2	2.4	40
lipid	73	0	–	0	–
multi_cell	72	11.1	100	10	100
other_name	1148	92.3	75.7	90.1	83.0
RNA_molecule	1	0	–	0	–
cell_line	185	8.3	60.0	2.9	100
DNA_substructure	7	0	–	0	–
body_part	53	0	–	0	–
protein_molecule	1009	92.0	56.6	83.2	71
RNA_domain_or_region	4	0	–	0	–
protein_substructure	6	0	–	0	–
inorganic	16	0	–	0	–
protein_complex	35	0	–	0	–
carbohydrate	14	0	–	0	–
RNA_substructure	0	0	–	0	–
protein_domain_or_region	4	0	–	0	–
DNA_family_or_group	17	9.67	100	13.3	100
Weighted Average	–	48.3	60.0	43.6	72.4

(how many of the terminological classes of that semantic type were assigned to that semantic type by the algorithm) and precision (how many terminological classes assigned to a particular semantic type are correctly assigned to that type).

The analysis shows that the distributional similarity measure tends to exhibit better recall in assigning the most populous semantic types (e.g. `other_name`). This is not surprising given that terminological classes selected randomly as neighbours would exhibit the observed probability distribution of semantic types and thus a majority would tend to vote for the most populous semantic type. However, the distributional similarity measures are not winning simply by always assigning to the most populous type. Other semantic types are also being assigned with high recall. Further, the less populous semantic types, for which recall is typically lower, do tend to be assigned accurately when they are assigned. In other words, if the nearest neighbours of an unknown terminological class indicate that the class is a member of, say, the `multi_cell` semantic type, then we can be very confident that this decision is correct.

When only higher frequency terms are considered, the precision of assignment generally increases whereas the recall of types generally decreases. This is likely to be because by only considering high frequency terms, we are effectively reducing the population of each semantic type.

7. Conclusions and future work

In this paper we have investigated an application of distributional similarity techniques to the problem of organising biomedical terminology drawn from a relatively small, domain-specific corpus: the 400K word GENIA corpus. The work is part of a wider study of techniques that can be used to estimate semantic similarity effectively. Using terms that have been accurately marked up by hand within the corpus, we have considered the problem of automatically determining semantic proximity. Evaluation is performed against a hand-crafted gold standard for this domain in the form of GENIA ontology.

We have demonstrated that, within this domain, distributional similarity is highly correlated with semantic similarity, as defined by the GENIA ontology. Moreover, the distributionally nearest neighbours of any unknown terminological class can be used to predict the semantic type of that class with a reasonably high degree of accuracy. We conclude that such techniques can serve as a rich source of information for the classification of terms, in addition to that provided by terminological variation and contextual parsing methods.

Our work also demonstrates that distributional similarity techniques can be used effectively on relatively sparse data. Indeed, all of the measures we have investigated, with the exception of CR precision have performed comparably. Given just the first neighbour of a terminological class, it has been observed that the CR recall measure $\mathcal{R}$ is best able to predict the semantic type of that class. The CR precision measure $\mathcal{P}$, on the other hand, is least successful amongst the various measure at this prediction task. Previous work (Weeds et al. 2004) shows that high CR precision tends to select low frequency nouns as neighbours. This may explain its particularly poor performance in this application, as the lower frequency terms in the GENIA corpus are very low frequency events and co-occurrence data for such events will tend to exhibit a lower signal-to-noise ratio simply on account of sparseness. However, it appears that combining precision and recall with a measure such as $\mathcal{F}$ or `Lin` achieves better results when evidence is collected from a cluster of neighbours. The optimal performance, achieved using the $\mathcal{F}$ measure and 30 nearest neighbours, over terminological classes of all frequencies is 63.1%. This suggests that while precision can introduce some noise into the neighbour ranking, it does nevertheless provide useful, additional information for determining semantic similarity.

In conclusion, our results demonstrate that the application of distributional similarity techniques is a promising approach to the problem of organising terminology. In future work, we intend to experiment with weighting neighbours' contributions in the semantic type decision task by their distributional ranking. We also believe it may be possible to overcome the biases introduced by having an unequal distribution of terms between semantic types by 1) weighting a neighbour's contribution by our surprise at seeing a neighbour of that semantic type (i.e. smaller semantic classes get higher weights) and/or 2) using an iterative process where the assignment to semantic class gets progressively more fine-grained. Finally, having considered the problem of assigning new terms to an existing set of ontological types, it would also be interesting to determine whether distributional similarity may be used for clustering terminological classes from scratch.

Acknowledgements

This work was funded by the UK Engineering and Physical Sciences Research Council (EPSRC), grant GR/S26408/01 “Natural Language Service Composition” (NatHab).

Notes

1. The F-Score is a standard statistical metric which combines precision and recall into a single measure of overall performance

2. Not all applications of distributional similarity assume the distributional hypothesis. The technique has also been used to identify word-clusters for use in language modelling, where there is no necessary requirement for the clusters to be semantically coherent (Dagan et al. 1994, 1999; Lee 1999).

3. A notable difference between our work and that of Lin is that the corpus used in our experiments is domain-specific and the individual terms are expected to have only a single sense. It is possible that this may reduce the quantity of data required to obtain usable results with measures of distributional similarity.

4. A possible alternative would be to label a neighbour as correct whenever it *shares* a semantic type with *c*, and incorrect otherwise. However, this would result in a more lenient measure of error rate.

5. This is a standard statistical measure that evaluates how well the ranks assigned to a set of objects by two different scoring mechanisms match.

6. We only consider the $k = 10$ results in this analysis since, as more neighbours are considered, it becomes increasingly less likely that the less populous semantic types will be assigned.

References

Abney, S. 1995. “Chunks and dependencies: Bringing processing evidence to bear on syntax.” In Cole, J., G. Green and J. Morgan (eds.), Computational Linguistics and the Foundations of Linguistic Theory, 145–164. CSLI.

Baker, C. F., C. J. Fillmore and B. Cronin. 2003. “The structure of the framenet database.” International Journal of Lexicography 16(3), 281–296.

Barker, K. and S. Szpakowicz. 1998. “Semi-Automatic Recognition of Noun Modifier Relationships.” In Proc. of COLING-ACL98. Montreal, Quebec, Canada.

Basili, R. and F. Zanzotto. 2002. “Parsing engineering and empirical robustness.” Natural Language Engineering 8(1), 21–37.

Caraballo, S. 1999. “Automatic construction of a hypernym-labelled noun hierarchy from text." In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), 120–126.

Castellvi, M. T. C., R. E. Bagot and J. V. Palatresi. 2001. "Automatic term detection: A review of current systems." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.), Recent Advances in Computational Terminology, 53–88. Amsterdam/Philadelphia: John Benjamins.

Chikashi Nobata, N. C. and J. ichi Tsujii. 1999. "Automatic term identification and classification in biology texts." In Proceedings of the fifth Natural Language Processing Pacific Rim Symposium (NLPRS), 369–374. Beijin, China.

Church, K. W. and P. Hanks. 1989. "Word association norms, mutual information and lexicography." In Proceedings of the 27th Annual Conference of the Association for Computational Linguistics (ACL-1989), 76–82.

Collins, M. 1996. "A new statistical parser based on bigram lexical dependencies." In Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, 184–191. Philadelphia.

Collins, M. 1999. "Head-driven statistical models for natural language processing." Ph.D. thesis, University of Pennsylvania.

Collins, M. and J. Brooks. 1995. "Prepositional attachment through a backed-off model." In Proceedings of the Third Workshop on Very Large Corpora. Cambridge, MA.

Connolly, D., F. van Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider and L. A. Stein. 2001. "Daml+oil reference description." W3C Note.

Curran, J. R. and M. Moens. 2002. "Improvements in automatic thesaurus extraction." In ACL-SIGLEX Workshop on Unsupervised Lexical Acquisition. Philadelphia.

Dagan, I., L. Lee and F. Pereira. 1999. "Similarity-based models of word cooccurrence probabilities." Machine Learning Journal 34(1–3).

Dagan, I., F. Pereira and L. Lee. 1994. "Similarity-based estimation of word cooccurrence probabilities." In ACL-94, 272–278.

Daille, B. 2003. "Conceptual structuring through term variations." In Proceedings of the ACL-2003 Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment, 9–16. Saporro, Japan.

Dowdall, J., W. Lowe, J. Ellman, F. Rinaldi and M. Hess. 2004. "The role of multiword terminology in knowledge management." In Proceedings of the International Conference on Language Resources and Evaluation (LREC), 915–918. Lisbon, Portugal.

Dowdall, J., F. Rinaldi, F. Ibekwe-Sanjan and E. Sanjuan. 2003. "Complex structuring of term variants for question answering." In Proceedings of the ACL-2003 Workshop on MultiWord Expressions: Analysis, Acquisition and Treatment, 9–16. Saporro, Japan.

Fellbaum, C. (Ed.). 1998. WordNet: An Electronic Lexical Database. MIT Press.

Finch, S. and A. Mikheev. 1997. "A workbench for finding structure in texts." In Proceedings of Applied Natural Language Processing. Washington.

Fung, P. and K. McKeown. 1997. "A technical word- and term- translation aid using noisy parallel corpora across language groups." Machine Translation 1–2, 53–87.

Geffet, M. and I. Dagan. 2004. "Feature vector quality and distributional similarity." In Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004). Geneva, Switzerland.

Grabar, N. and P. Zweigenbaum. 2000. “Automatic acquisition of domain-specific morphological resources from thesauri.” In Proceedings of RIAO 2000: Content-based Multimedia Information Access, 765–784. Paris, France.

Grabar, N. and P. Zweigenbaum. 2002. “Lexically-based terminology structuring: Some inherent limits.” In Proceedings of the 2nd International Workshop on Computational Terminology (CompuTerm), 36–42. Taipei, Taiwan.

Grefenstette, G. 1994. “Corpus-derived first-, second- and third-order word affinities.” In Proceedings of Euralex, 279–290. Amsterdam.

Hamon, T., A. Nazarenko and C. Gros. 1998. “A step towards the detection of semantic variants of terms in technical documents.” In Proceedings of the 36th conference on Association for Computational Linguistics, 498–504. Montreal, CA.

Harris, Z. S. 1968. Mathematical Structures of Language. New York: Wiley.

Hearst, M. 1992. “Automatic acquisition of hyponyms from large text corpora.” In Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), 539–545. Nantes, France.

Hindle, D. 1990. “Noun classification from predicate-argument structures.” In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL-1990), 268–275.

Ibekwe-SanJuan, F. and E. SanJuan. 2002. “From term variants to research topics.” Journal of Knowledge Organization (ISKO), special issue on Human Language Technology 29(3/4), 181–197.

Ibekwe-SanJuan, F. and E. SanJuan. 2004. “Mining textual data through term variant clustering: The termwatch system.” In Proceedings of Recherche d’Information assisté par ordinateur (RIAO), 487–503. Avignon.

Kaufman, L. and P. J. Rousseeuw. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons.

Kawasaki, Y., J. Kazama and J. Tsujii. 2003. “Extracting biomedical ontology from textbooks and article abstracts.” In Proceedings of the SIGIR’03 Workshop on Text Analysis and Search for Bioinformatics, 44–50.

Kazama, J., T. Makino, Y. Ohta and J. Tsujji. 2002. “Tuning support vector machines for biomedical named entity recognition.” In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, 1–8. Philadelphia, USA.

Kilgarriff, A. 2003. “Thesauruses for natural language processing.” In Proceedings of the Joint Conference on Natural Language Processing and Knowledge Engineering, 5–13. Beijing, China.

Kilgarriff, A. and D. Tugwell. 2001. “WORD SKETCH: Extraction and display of significant collocations for lexicography.” In ACL Workshop on COLLOCATION: Computation extraction, analysis and exploitation. Toulouse.

Kilgarriff, A. and C. Yallop. 2000. “What’s in a thesaurus.” In Second Conference on Language Resources and Evaluation (LREC-00), 1371–1379. Athens.

J.-D. Kim, Y. T., T. Ohta and J. Tsujii. 2003. “Genia corpus a semantically annotated corpus for bio-textmining.” BioInformatics 19(1), i180–i182.

Lee, L. 1999. “Measures of distributional similarity.” In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999), 23–32.

Lin, D. 1997. “Using syntactic dependency as local context to resolve word sense ambiguity.” In Proceedings of ACL/EACL-97, 64–71.

Lin, D. 1998a. “Automatic retrieval and clustering of similar words.” In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL ’98), 768–774. Montreal.

Lin, D. 1998b. “An information-theoretic definition of similarity.” In Proceedings of International Conference on Machine Learning. Madison, Wisconsin.

Lin, D., S. Zhao, L. Qin and M. Zhou. 2003. “Identifying synonyms among distributionally similar words.” In Proceedings of IJCAI-03, 1492–1493.

Linden, K. and J. Piitulainen. 2004. “Discovering synonyms and other related words.” In Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm2004). Geneva, Switzerland.

Mani, I., K. Samuel, K. Concepcion and D. Vogel. 2004. “Automatically inducing ontologies from corpora.” In Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm), 47–54. Geneva, Switzerland.

Marcus, M. P., M. A. Marcinkiewicz and B. Santorini. 1993. “Building a large annotated corpus of English: The Penn Tree Bank.” Computational Linguistics, Special issue on using large corpora 9(2), 313–330.

McCarthy, D., B. Keller and J. Carroll. 2003. “Detecting a continuum of compositionality in phrasal verbs.” In Proceedings of the ACL-2003 Workshop on Multiword Expressions, 73–80. Sapporo, Japan.

Minnen, G., J. Carroll and D. Pearce. 2001. “Applied morphological processing of English.” Natural Language Engineering 7(3), 207–223.

Morin, E. and C. Jacquemin. 2003. “Automatic acquisition and expansion of hypernym links.” Computer and the Humanities.

Nenadić, G., S. Rice, I. Spasić, S. Ananiadou and B. Stapley. 2003. “Selecting text features for gene name classification: From documents to terms.” In Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine, 121–128. Sapporo, Japan.

Nenadić, G., I. Spasić and S. Ananiadou. 2002a. “Automatic discovery of term similarities using pattern mining.” In Proceedings of the 2nd International Workshop on Computational Terminology (CompuTerm), 43–49. Taipei, Taiwan.

Nenadić, G., I. Spasić and S. Ananiadou. 2002b. “Term clustering using a corpus-based similarity measure.” In Proceedings of the 5th International Conference on Text, Speech and Dialogue, 151–154. Springer-Verlag. ISBN 3-540-44129-8.

NLM. 1998. “UMLS Knowledge Sources.” National Library of Medicine, U.S. Dept. of Health and Human Services, 8th edition.

Resnik, P. 1993. “Selection and information: A class-based approach to lexical relationships.” Ph.D. thesis, University of Pennsylvania.

Roget, P. 1911. Thesaurus of English words and phrases. London, UK: Longmans, Green and Co.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. “Multiword expressions: A pain in the neck for NLP.” In Proceedings of the 3rd International Conference on Intelligent text processing and computational linguistics (CICLing-2002), 1–15. Mexico City.

SanJuan, E., J. Dowdall, F. Ibekwe-SanJuan and F. Rinaldi. 2004. "A symbolic approach to automatic multiword term structuring." Submitted to Computer Speech and Language, Special Issue on Multiword Expressions. Elsevier Science. September 2004, 20 pages.

Schneider, G. 2003. "Extracting and using trace-free functional dependencies from the penn treebank to reduce parsing complexity." In Proceedings of Treebanks and Linguistic Theories (TLT). Sweden.

Schütze, H. 1993. "Part-of-speech induction from scratch." In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-1993), 251–258.

Torii, M., S. Kamboj and K. Vijay-Shanker. 2003. "An investigation of various information sources for classifying biological names." In Proceedings of the ACL-03 Workshop: Natural Language Processing in Biomedicine, 113–120. Sapporo, Japan.

Weeds, J. 2003. "Measures and applications of lexical distributional similarity." Ph.D. thesis, Department of Informatics, University of Sussex.

Weeds, J. and D. Weir. 2003a. "Finding and evaluating sets of nearest neighbours." In Proceedings of the 2nd International Conference on Corpus Linguistics. Lancaster, UK.

Weeds, J. and D. Weir. 2003b. "A general framework for distributional similarity." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003). Sapporo, Japan.

Weeds, J., D. Weir and D. McCarthy. 2004. "Characterising measures of lexical distributional similarity." In Proceedings of the 20th International Conference for Computational Linguistics (COLING-2004).

Wiebe, J. 2000. "Learning subjective adjectives from corpora." In Proceedings of AAAI '00.

Wu, H. and M. Zhou. 2003. "Synonymous collocation extraction using translation information." In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), 120–127. Sapporo, Japan.

Zipf, G. 1949. Human Behaviour and the Principle of Least Effort. Addison-Wesley.

8. Authors' addresses

Julie Weeds
Department of Informatics
University of Sussex
Brighton, BN1 9QH
UK
juliewe@sussex.ac.uk

James Dowdall
Department of Informatics
University of Sussex
Brighton, BN1 9QH
UK
j.m.dowdall@sussex.ac.uk

Gerold Schneider
Institute of Computational Linguistics
University of Zurich
Winterthurerstr. 190
CH-8057 Zürich
Switzerland
gschneid@ifi.unizh.ch

Bill Keller
Department of Informatics
University of Sussex
Brighton, BN1 9QH
UK
billk@sussex.ac.uk

David Weir
Department of Informatics
University of Sussex
Brighton, BN1 9QH
UK
Email: davidw@sussex.ac.uk

Julie Weeds graduated from Trinity Hall, Cambridge in 1998 having studied Computer Science. Her MPhil in Computer Speech and Language Processing was awarded at Cambridge University's Engineering Department. Having completed her doctorate on measures and applications of lexical distributional similarity at The University of Sussex in 2003, Julie Weeds is now a research fellow on the Natural Habitats project at Sussex.

James Dowdall graduated from the University of Kent before completing an MPhil in Theoretical Linguistics at Trinity College Dublin. After working as a research assistant at the University of Zürich for three years he is a doctoral student at the University of Sussex. His research interests include computational terminology and ontologies for practical NLP applications.

Gerold Schneider completed his Master Degree at the University of Zürich in 1998. He is currently a research assistant in the European REWERSE project at the Institute of Computational Linguistics, University of Zürich. Besides the development of a low-complexity, broad-coverage probabilistic Dependency Parser for English (Pro3Gres) as part of this doctoral thesis his research interests include Question Answering, Corpus Linguistics and Semantics. He has published over a dozen scientific articles and has been a research assistant in Swiss National Science fund projects at the Universities of Zürich and Geneva.

Bill Keller graduated with a BSc in Computer Science from the University of Warwick in 1983. He was awarded an MA in Cognitive Science from the University of Sussex in 1985 and a PhD in Computational Linguistics, also from Sussex, in 1991. Since 1989 he has lectured at Sussex in Computer Science and Artificial Intelligence, working currently within the department of Informatics. His research interests include corpus-based approaches to language processing, statistical language modelling and machine learning of natural language.

David Weir is a Reader in Computer Science and Artificial Intelligence in the Department of Informatics at the University of Sussex. His research interests are in the area of natural language processing, particularly constrained grammar formalisms, parsing algorithms, lexical knowledge acquisition, natural language generation, and applications of natural language processing to pervasive computing.