Detecting Compositionality of Verb-Object Combinations using Selectional Preferences

Diana McCarthy
University of Sussex
Falmer, East Sussex
BN1 9QH, UK
dianam@sussex.ac.uk

Sriram Venkatapathy
International Institute
of Information Technology
Hyderabad, India
sriram@research.iiit.ac.in

Aravind K. Joshi
University of Pennsylvania,
Philadelphia
PA, USA.
joshi@linc.cis.upenn.edu

Abstract

In this paper we explore the use of selectional preferences for detecting non-compositional verb-object combinations. To characterise the arguments in a given grammatical relationship we experiment with three models of selectional preference. Two use WordNet and one uses the entries from a distributional thesaurus as classes for representation. In previous work on selectional preference acquisition, the classes used for representation are selected according to the coverage of argument tokens rather than being selected according to the coverage of argument types. In our distributional thesaurus models and one of the methods using WordNet we select classes for representing the preferences by virtue of the number of argument types that they cover, and then only tokens under these classes which are representative of the argument head data are used to estimate the probability distribution for the selectional preference model. We demonstrate a highly significant correlation between measures which use these 'type-based' selectional preferences and compositionality judgements from a data set used in previous research. The type-based models perform better than the models which use tokens for selecting the classes. Furthermore, the models which use the automatically acquired thesaurus entries produced the best results. The correlation for the thesaurus models is stronger than any of the individual features used in previous research on the same dataset.

1 Introduction

Characterising the semantic behaviour of phrases in terms of compositionality has particularly attracted attention in recent years (Lin, 1999; Schone and Jurafsky, 2001; Bannard, 2002; Bannard et al., 2003; Baldwin et al., 2003; McCarthy et al., 2003; Bannard, 2005; Venkatapathy and Joshi, 2005). Typically the phrases are putative multiwords and non-compositionality is viewed as an important feature of many such "words with spaces" (Sag et al., 2002). For applications such as paraphrasing, information extraction and translation, it is essential to take the words of non-compositional phrases together as a unit because the meaning of a phrase cannot be obtained straightforwardly from the constituent words. In this work we are investigate methods of determining semantic compositionality of verb-object [1] combinations on a continuum following previous research in this direction (McCarthy et al., 2003; Venkatapathy and Joshi, 2005).

Much previous research has used a combination of statistics and distributional approaches whereby distributional similarity is used to compare the constituents of the multiword with the multiword itself. In this paper, we will investigate the use of selectional preferences of verbs. We will use the preferences to find atypical verb-object combinations as we anticipate that such combinations are more likely to be non-compositional.

[1] We use object to refer to direct objects.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 369–379, Prague, June 2007.

Selectional preferences of predicates have been modelled using the man-made thesaurus WordNet (Fellbaum, 1998), see for example (Resnik, 1993; Li and Abe, 1998; Abney and Light, 1999; Clark and Weir, 2002). There are also distributional approaches which use co-occurrence data to cluster distributionally similar words together. The cluster output can then be used as classes for selectional preferences (Pereira et al., 1993), or one can directly use frequency information from distributionally similar words for smoothing (Grishman and Sterling, 1994).

We used three different types of probabilistic models, which vary in the classes selected for representation over which the probability distribution of the argument heads [2] is estimated. Two use WordNet and the other uses the entries in a thesaurus of distributionally similar words acquired automatically following (Lin, 1998). The first method is due to Li and Abe (1998). The classes over which the probability distribution is calculated are selected according to the minimum description length principle (MDL) which uses the argument head tokens for finding the best classes for representation. This method has previously been tried for modelling compositionality of verb-particle constructions (Bannard, 2002).

The other two methods (we refer to them as 'type-based') also calculate a probability distribution using argument head tokens but they select the classes over which the distribution is calculated using the number of argument head types (of a verb in a corpus) in a given class, rather than the number of argument head tokens in contrast to previous WordNet models (Resnik, 1993; Li and Abe, 1998; Clark and Weir, 2002). For example, if the object slot of the verb *park* contains the argument heads { *car, car, car, car, van, jeep* } then the type-based models use the word type "*car*" only once when determining the classes over which the probability distribution is to be estimated. Classes are selected which maximise the number of types that they cover, rather than the number of tokens. This is done to avoid the selectional preferences being heavily influenced by noise from highly frequent arguments which may be polysemous and some or all of their meanings may not be semantically related to the 'prototypical' arguments of the verb. For example *car* has a **gondola** sense in WordNet.

The third method uses entries in a distributional thesaurus rather than classes from WordNet. The entries used as classes for representation are selected by virtue of the number of argument types they encompass. As with the WordNet models, the tokens are used to estimate a probability distribution over these entries.

In the next section, we discuss related work on identifying compositionality. In section 3, we describe the methods we are using for acquiring our models of selectional preference. In section 4, we test our models on a dataset used in previous research. We compare the three types of models individually and also investigate the best performing model when used in combination with other features used in previous research. We conclude in section 5.

2 Related Work

Most previous work using distributional approaches to compositionality either contrasts distributional information of candidate phrases with constituent words (Schone and Jurafsky, 2001; Bannard et al., 2003; Baldwin et al., 2003; McCarthy et al., 2003) or uses distributionally similar words to detect non-productive phrases (Lin, 1999).

Lin (1999) used his method (Lin, 1998) for automatic thesaurus construction. He identified candidate phrases involving several open-class words output from his parser and filtered these by the log-likelihood statistic. Lin proposed that if there is a phrase obtained by substitution of either the head or modifier in the phrase with a 'nearest neighbour' from the thesaurus then the mutual information of this and the original phrase must be significantly different for the original phrase to be considered non-compositional. He evaluated the output manually.

As well as distributional similarity, researchers have used a variety of statistics as indicators of non-compositionality (Blaheta and Johnson, 2001; Krenn and Evert, 2001). Fazly and Stevenson (2006) use statistical measures of syntactic behaviour to gauge whether a verb and noun combination is likely to be a idiom. Although they are not specifically detecting compositionality, there is a strong corre-

[2] Argument heads are the nouns occurring in the object slot of the target verb.

lation between syntactic rigidity and semantic idiosyncrasy.

Venkatapathy and Joshi (2005) combine different statistical and distributional methods using support vector machines (SVMs) for identifying non-compositional verb-object combinations. They explored seven features as measures of compositionality:

1. frequency
2. pointwise mutual information (Church and Hanks, 1990),
3. least mutual information difference with similar collocations, based on (Lin, 1999) and using Lin's thesaurus (Lin, 1998) for obtaining the similar collocations.
4. The distributed frequency of an object, which takes an average of the frequency of occurrence with an object over all verbs occurring with the object above a threshold.
5. distributed frequency of an object, using the verb, which considers the similarity between the target verb and the verbs occurring with the target object above the specified threshold.
6. a latent semantic approach (LSA) based on (Schütze, 1998; Baldwin et al., 2003) and considering the dissimilarity of the verb-object pair with its constituent verb
7. the same LSA approach, but considering the similarity of the verb-object pair with the verbal form of the object (to capture support verb constructions e.g. *give a smile*

Venkatapathy and Joshi (2005) produced a dataset of verb-object pairs with human judgements of compositionality. We say more about this dataset and Venkatapathy and Joshi's results in section 4 since we use the dataset for our experiments.

In this paper, we investigate the use of selectional preferences to detect compositionality. Bannard (2002) did some pioneering work to try and establish a link between the compositionality of verb particle constructions and the selectional preferences of the multiword and its constituent verb. His results were hampered by models based on (Li and Abe, 1998) which involved rather uninformative models at the roots of WordNet. There are several reasons for this. The classes for the model are selected using MDL by compromising between a simple model with few classes and one which explains the data well. The models are particularly affected by the quantity of data available (Wagner, 2002). Also noise from frequent but idiosyncratic or polysemous arguments weakens the signal. There is scope for experimenting with other approaches such as (Clark and Weir, 2002), however, we feel a type-based approach is worthwhile to avoid the noise introduced from frequent but polysemous arguments and bias from highly frequent arguments which might be part of a multiword rather than a prototypical argument of the predicate in question, for example *eat hat*. In contrast to Bannard, our experiments are with verb-object combinations rather than verb particle constructions. We compare Li and Abe models with WordNet models which use the number of argument types to obtain the classes for representation of the selectional preferences. In addition to experiments with these WordNet models, we propose models using entries in distributional thesauruses for representing preferences.

3 Three Methods for Acquiring Selectional Preferences

All models were acquired from verb-object data extracted using the RASP parser (Briscoe and Carroll, 2002) from the 90 million words of written English from the BNC (Leech, 1992). We extracted verb and common noun tuples where the noun is the argument head of the object relation. The parser was also used to extract the grammatical relation data used for acquisition of the thesaurus described below in section 3.3.

3.1 TCMS

This approach is a reimplementation of Li and Abe (1998). Each selectional preference model (referred to as a tree cut model, or TCM) comprises a set of disjunctive noun classes selected from all the possibilities in the WordNet hyponym hierarchy [3] using MDL (Rissanen, 1978). The TCM covers all the

[3] We use WordNet version 2.1 for the work in this paper.

noun senses in the WordNet hierarchy and is associated with a probability distribution over these noun senses in the hierarchy reflecting the argument head data occurring in the given grammatical relationship with the specified verb. MDL finds the classes in the TCM by considering the cost measured in bits of describing both the model and the argument head data encoded in the model. A compromise is made by having as simple a model as possible using classes further up the hierarchy whilst also providing a good model for the set of argument head tokens (TK).

The classes are selected by recursing from the top of the WordNet hierarchy comparing the cost (or description length) of using the mother class to the cost of using the hyponym daughter classes. In any path, the mother is preferred unless using the daughters would reduce the cost. If using the daughters for the model is less costly than the mother then the recursion continues to compare the cost of the hyponyms beneath.

The cost (or description length) for a set of classes is calculated as the model description length (mdl) and the data description length (ddl) [4] :-

$$\begin{array}{cc} mdl & + \, ddl \\ \frac{k}{2} \times \log |TK| & + - \sum_{tk \in TK} \log p(tk) \end{array} \qquad (1)$$

k, is the number of WordNet classes being currently considered for the TCM minus one. The MDL method uses the size of TK on the assumption that a larger dataset warrants a more detailed model. The cost of describing the argument head data is calculated using the log of the probability estimate from the classes currently being considered for the model. The probability estimate for a class being considered for the model is calculated using the cumulative frequency of all the hyponym nouns under that class that occur in TK, divided by the number of noun senses that these nouns have, to account for their polysemy. This cumulative frequency is also divided by the total number of noun hyponyms under that class in WordNet to obtain a smoothed estimate for all nouns under the class. The probability of the class is obtained by dividing this frequency estimate by the total frequency of the argument heads. The algorithm is described fully by Li and Abe (1998).

[4]See (Li and Abe, 1998) for a full explanation.

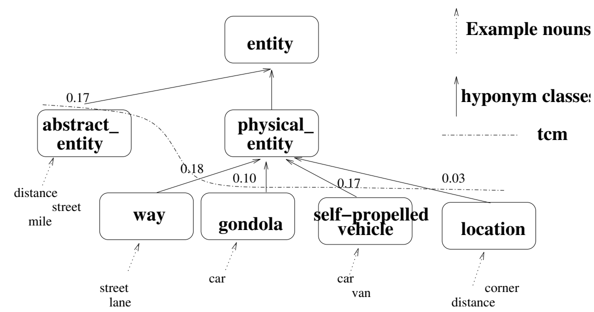

Figure 1: portion of the TCM for the objects of *park*.

A small portion of the TCM for the object slot of *park* is shown in figure 1. WordNet classes are displayed in boxes with a label which best reflects the meaning of the class. The probability estimates are shown for the classes on the TCM. Examples of the argument head data are displayed below the WordNet classes with dotted lines indicating membership at a hyponym class beneath these classes. We cannot show the full TCM due to lack of space, but we show some of the higher probability classes which cover some typical nouns that occur as objects of *park*. Note that probability under the classes **abstract_entity**, **way** and **location** arise because of a systematic parsing error where adverbials such as *distance* in *park illegally some distance from the railway station* are identified by the parser as objects. Systematic noise from the parser has an impact on all the selectional preference models described in this paper.

3.2 WNPROTOs

We propose a method of acquiring selectional preferences which instead of covering all the noun senses in WordNet, just gives a probability distribution over a portion of prototypical classes, we refer to these models as WNPROTOs. A WNPROTO consists of classes within the noun hierarchy which have the highest proportion of word types occurring in the argument head data, rather than using the number of tokens, or frequency, as is used for the TCMs. This allows less frequent, but potentially informative arguments to have some bearing on the models acquired to reduce the impact of highly frequent but polysemous arguments. We then used the frequency data to populate these selected classes.

The classes (C) in the WNPROTO are selected from those which include at least a threshold of 2 argument head types [5] occurring in the training data. Each argument head in the training data is disambiguated according to whichever of the WordNet classes it occurs at or under which has the highest 'type ratio'. Let TY be the set of argument head types in the object slot of the verb for which we are acquiring the preference model. The type ratio for a class (c) is the ratio of noun types ($ty \in TY$) occurring in the training data also listed at or beneath that class in WordNet to the total number of noun types listed at or beneath that particular class in WordNet ($wn_{ty} \in c$). The argument types attested in the training data are divided by the number of WordNet classes that the noun ($classes(ty)$) belongs to, to account for polysemy in the training data.

$$type\ ratio(c) = \frac{\sum_{ty \in TY \in c} \frac{1}{|classes(ty)|}}{|wn_{ty} \in c|} \qquad (2)$$

If more than one class has the same type ratio then the argument is not used for calculating the probability of the preference model. In this way, only arguments that can be disambiguated are used for calculating the probability distribution. The advantage of using the type ratio to determine the classes used to represent the model and to disambiguate the arguments is that it prevents high frequency verb noun combinations from masking the information from prototypical but low frequency arguments. We wish to use classes which are as representative of the argument head types as possible to help detect when an argument head is not related to these classes and is therefore more likely to be non-compositional.

For example, the class **motor_vehicle** is selected for the WNPROTO model of the object slot of *park* even though there are 5 meanings of *car* in WordNet including **elevator_car** and **gondola**. There are 174 occurrences of *car* which overwhelms the frequency of the other objects (e.g. *van* 11, *vehicle* 8) but by looking for classes with a high proportion of types (rather than word tokens) *car* is disambiguated appropriately and the class **motor_vehicle** is selected for representation.

[5]We have experimented with a threshold of 3 and obtained similar results.

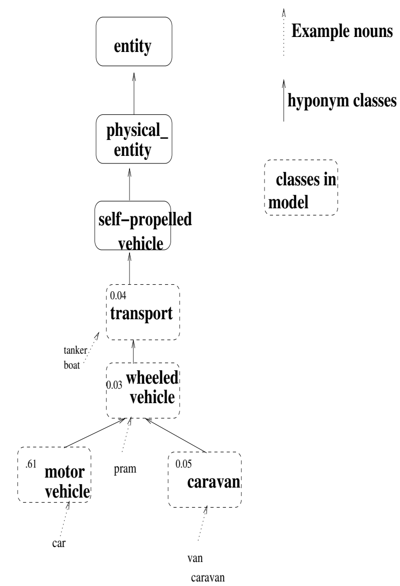

Figure 2: Part of WNPROTO for the object slot of *park*

The relative frequency of each class is obtained from the set of disambiguated argument head tokens and used to provide the probability distribution over this set of classes. Note that in WNPROTO, classes can be subsumed by others in the hyponym hierarchy. The probability assigned to a class is applicable to any descendants in the hyponym hierarchy, except those within any hyponym classes within the WNPROTO. The algorithm for selecting C and calculating the probability distribution is shown as Algorithm 1. Note that we use brackets for comments.

In figure 2 we show a small portion of the WNPROTO for *park*. Again, WordNet classes are displayed in boxes with a label which best reflects the meaning of the class. The probability estimates are shown in the boxes for all the classes included in the WNPROTO. The classes in the WNPROTO model are shown with dashed lines. Examples of the argument head data are displayed below the WordNet classes with dotted lines indicating membership at a hyponym class beneath these classes. We cannot show the full WNPROTO due to lack of space, but we show some of the classes with higher probability which cover some typical nouns that occur as objects of *park*.

Algorithm 1 WNPROTO algorithm

$C = ()$ {classes in WNPROTO}
$D = ()$ {disambiguated $ty \in TY$}
$fD = 0$ {frequency of disambiguated items}
$TY =$ argument head types {nouns occurring as objects of verb, with associated frequencies}
$C1 \in WordNet$
$where\ |ty \in TY occurring\ in\ c \in C1| > 1$
for all $ty \in TY$ **do**
  find $c \in classes(ty) \in C1\ where\ c = \text{argmax}_c\ typeratio(c)$
  if c & $c \notin C$ **then**
    add c to C
    add $ty \leftrightarrow c$ to D {Disambiguated ty with c}
  end if
end for
for all $c \in C$ **do**
  if $|ty \leftrightarrow c \in D| > 1$ **then**
    $fD = fD + frequency(ty)$ {sum frequencies of types under classes to be used in model}
  else
    remove c from C {classes with less than two disambiguated nouns are removed}
  end if
end for
for all $c \in C$ **do**
  $p(c) = \frac{frequency\text{-}of\text{-}all\text{-}tys\text{-}disambiguated\text{-}to\text{-}class(c,D)}{fD}$ {calculating class probabilities}
end for

Algorithm 2 DSPROTO algorithm

$C = ()$ {classes in DSPROTO}
$D = ()$ {disambiguated $ty \in TY$}
$fD = 0$ {frequency of disambiguated items}
$TY =$ argument head types {nouns occurring as objects of verb, with associated frequencies}
$C1 = cty \in TY\ where\ num\text{-}types\text{-}in\text{-}thesaurus(cty, TY) > 1$
order $C1$ by $num\text{-}types\text{-}in\text{-}thesaurus(cty, TY)$ {classes ordered by coverage of argument head types}
for all $cty \in ordered\ C1$ **do**
  $Dcty = ()$ {disambiguated for this class}
  for all $ty \in TY where\ in\text{-}thesaurus\text{-}entry(cty, ty)$ **do**
    if $ty \notin D$ **then**
      add ty to $Dcty$ {types disambiguated to this class only if not disambiguated by a class used already}
    end if
  end for
  if $|Dcty| > 1$ **then**
    add cty to C
    for all $ty \in Dcty$ **do**
      add $ty \leftrightarrow cty$ to D {Disambiguated ty with cty}
      $fD = fD + frequency(ty)$
    end for
  end if
end for
for all $cty \in C$ **do**
  $p(cty) = \frac{frequency\text{-}of\text{-}all\text{-}tys\text{-}disambiguated\text{-}to\text{-}class(cty,D)}{fD}$ {calculating class probabilities}
end for

3.3 DSPROTOs

We use a thesaurus acquired using the method proposed by Lin (1998). For input we used the grammatical relation data from automatic parses of the BNC. For each noun we considered the co-occurring verbs in the object and subject relation, the modifying nouns in noun-noun relations and the modifying adjectives in adjective-noun relations. Each thesaurus entry consists of the target noun and the 50 most similar nouns, according to Lin's measure of distributional similarity, to the target.

The argument head noun types (TY) are used to find the entries in the thesaurus as the 'classes' (C) of the selectional preference for a given verb. As with WNPROTOs, we only cover argument types which form coherent groups with other argument types since we wish i) to remove noise and ii) to be able to identify argument types which are not related with the other types and therefore may be non-compositional. As our starting point we only consider an argument type as a class for C if its entry in the thesaurus covers at least a threshold of 2 types. [6]

To select C we use a best first search. This method processes each argument type in TY in order of the number of the other argument types from TY that it has in its thesaurus entry of 50 similar nouns. An argument head is selected as a class for C ($cty \in C$) [7] if it covers at least 2 of the argument heads that are not in the thesaurus entries of any of the other classes already selected for C. Each argument head is disambiguated by whichever class in C under which it is listed in the thesaurus and which has the largest number of the TY in its thesaurus entry. When the algorithm finishes processing the ordered argument heads to select C, all argument head types are disambiguated by C apart from those which after disambiguation occur in isolation in a class without other argument types. Finally a probability distribution over C is estimated using the frequency (tokens) of argument types that occur in the thesaurus entries for any $cty \in C$. If an argument type occurs in the entry of more than one cty then it is assigned to whichever of these has the largest number of disambiguated argument head types and its token frequency is attributed to that class. We show the algorithm as Algorithm 2.

[6]As with the WNPROTOs, we experimented with a value of 3 for this threshold and obtained similar results.

[7]We use cty for the classes of the DSPROTO. These classes are simply groups of nouns which occur under the entry of a noun (ty) in the thesaurus.

class ($p(c)$)	disambiguated objects (freq)
van (0.86)	car (174) van (11) vehicle (8) ...
mile (0.05)	street (5) distance (4) mile (1) ...
yard (0.03)	corner (4) lane (3) door (1)
backside (0.02)	backside (2) bum (1) butt (1) ...

Figure 3: First four classes of DSPROTO model for *park*

The algorithms for WNPROTO algorithm 1 and DSPROTO (algorithm 2) differ because of the nature of the inventories of candidate classes (WordNet and the distributional thesaurus). There are a great many candidate classes in WordNet. The WNPROTO algorithm selects the classes from all those that the argument heads belong to directly and indirectly by looping over all argument types to find the class that disambiguates each by having the largest type ratio calculated using the undisambiguated argument heads. The DSPROTO only selects classes from the fixed set of argument types. The algorithm loops over the argument types with at least two argument heads in the thesaurus entry and ordered by the number of undisambiguated argument heads in the thesaurus entry. This is a best first search to minimise the number of argument heads used in C but maximise the coverage of argument types.

In figure 3, we show part of a DSPROTO model for the object of *park*. [8] Note again that the class **mile** arises because of a systematic parsing error where adverbials such as *distance* in *park illegally some distance from the railway station* are identified by the parser as objects.

4 Experiments

Venkatapathy and Joshi (2005) produced a dataset of verb-object pairs with human judgements of compositionality. They obtained values of r_s between 0.111 and 0.300 by individually applying the 7 features described above in section 2. The best correlation was given by feature 7 and the second best was feature 3. They combined all 7 features using SVMs and splitting their data into test and training data and achieve a r_s of 0.448, which demonstrates

[8]We cannot show the full model due to lack of space.

significantly better correlation with the human gold-standard than any of the features in isolation

We evaluated our selectional preference models using the verb-object pairs produced by Venkatapathy and Joshi (2005). [9] This dataset has 765 verb-object collocations which have been given a rating between 1 and 6, by two annotators (both fluent speakers of English). Kendall's Tau (Siegel and Castellan, 1988) was used to measure agreement, and a score of 0.61 was obtained which was highly significant. The ranks of the two annotators gave a Spearman's rank-correlation coefficient (r_s) of 0.71.

The Verb-Object pairs included some adjectives (e.g. *happy, difficult, popular*), pronouns and complements e.g. *become director*. We used the subset of 638 verb-object pairs that involved common nouns in the object relationship since our preference models focused on the object relation for common nouns. For each verb-object pair we used the preference models acquired from the RASP parses of the BNC to obtain the probability of the class that this object occurs under. Where the object noun is a member of several classes ($classes(noun) \in C$) in the model, the class with the largest probability is used. Note though that for WNPROTOs we have the added constraint that a hyponym class from C is selected in preference to a hypernym in C. Compositionality of an object noun and verb is computed as:-

$$comp(noun, verb) = max_{c \in classes(noun) \in C}\ p(c|verb) \quad (3)$$

We use the probability of the class, rather than an estimate of the probability of the object, because we want to determine how likely any word belonging to this class might occur with the given verb, rather than the probability of the specific noun which may be infrequent, yet typical, of the objects that occur with this verb. For example, *convertible* may be an infrequent object of *park*, but it is quite likely given its membership of the class **motor_vehicle**. We do not want to assume anything about the frequency of non-compositional verb-object combinations, just that they are unlikely to be members of classes which represent prototypical objects. We will contrast these models with a baseline frequency feature used by Venkatapathy and Joshi.

method	r_s	$p <$ (one tailed)
selectional preferences		
TCM	0.090	0.0119
WNPROTO	0.223	0.00003
DSPROTO	**0.398**	0.00003
features from V&J		
frequency (f1)	0.141	0.00023
MI (f2)	**0.274**	0.00003
Lin99 (f3)	0.139	0.00023
LSA2 (f7)	0.209	0.00003
combination with SVM		
f2,3,7	0.413	0.00003
f1,2,3,7	0.419	0.00003
DSPROTO f1,2,3,7	**0.454**	0.00003

Table 1: Correlation scores for 638 verb object pairs

We use our selectional preference models to provide the probability that a candidate is representative of the typical objects of the verb. That is, if the object might typically occur in such a relationship then this should lessen the chance that this verb-object combination is non-compositional. We used the probability of the classes from our 3 selectional preference models to rank the pairs and then used Spearman's rank-correlation coefficient (r_s) to compare these ranks with the ranks from the gold-standard.

Our results for the three types of preference models are shown in the first section of table 1. [10] All the correlation values are significant, but we note that using the type based selectional preference models achieves a far greater correlation than using the TCMs. The DSPROTO models achieve the best results which is very encouraging given that they only require raw data and an automatic parser to obtain the grammatical relations.

We applied 4 of the features used by Venkatapathy and Joshi (2005) [11] and described in section 2 to our subset of 638 items. These features were ob-

[9] This verb-object dataset is available from http://www.cis.upenn.edu/˜sriramv/mywork.html.

[10] We show absolute values of correlation following (Venkatapathy and Joshi, 2005).

[11] The other 3 features performed less well on this dataset so we do not report the details here. This seems to be because they worked particularly well with the adjective and pronoun data in the full dataset.

tained using the same BNC dataset used by Venkatapathy and Joshi which was obtained using Bikel's parser (Bikel, 2004). We obtained correlation values for these features as shown in table 1 under V&J. These features are feature 1 frequency, feature 2 pointwise mutual information, feature 3 based on (Lin, 1999) and feature 7 LSA feature which considers the similarity of the verb-object pair with the verbal form of the object. Pointwise mutual information did surprisingly well on this 84% subset of the data, however the DSPROTO preferences still outperformed this feature. We combined the DSPROTO and V&J features with an SVM ranking function and used 10 fold cross validation as Venkatapathy and Joshi did. We contrast the result with the V&J features without the preference models. The results in the bottom section of table 1 demonstrate that the preference models can be combined with other features to produce optimal results.

5 Conclusions and Directions for Future Work

We have demonstrated that the selectional preferences of a verbal predicate can be used to indicate if a specific combination with an object is non-compositional. We have shown that selectional preference models which represent prototypical arguments and focus on argument types (rather than tokens) do well at the task. Models produced from distributional thesauruses are the most promising which is encouraging as the technique could be applied to a language without a man-made thesaurus. We find that the probability estimates from our models show a highly significant correlation, and are very promising for detecting non-compositional verb-object pairs, in comparison to individual features used previously.

Further comparison of WNPROTOs and DSPROTOs to other WordNet models are warranted to contrast the effect of our proposal for disambiguation using word types with iterative approaches, particularly those of Clark and Weir (2002). A benefit of the DSPROTOs is that they do not require a hand-crafted inventory. It would also be worthwhile comparing the use of raw data directly, both from the BNC and from google's Web 1T corpus (Brants and Franz, 2006) since web counts have been shown to outperform the Clark and Weir models on a pseudo-disambiguation task (Keller and Lapata, 2003).

We believe that preferences should NOT be used in isolation. Whilst a low preference for a noun may be indicative of peculiar semantics, this may not always be the case, for example *chew the fat*. Certainly it would be worth combining the preferences with other measures, such as syntactic fixedness (Fazly and Stevenson, 2006). We also believe it is worth targeting features to specific types of constructions, for example light verb constructions undoubtedly warrant special treatment (Stevenson et al., 2003)

The selectional preference models we have proposed here might also be applied to other tasks. We hope to use these models in tasks such as diathesis alternation detection (McCarthy, 2000; Tsang and Stevenson, 2004) and contrast with WordNet models previously used for this purpose.

6 Acknowledgements

We acknowledge support from the Royal Society UK for a Dorothy Hodgkin Fellowship to the first author. We thank the anonymous reviewers for their constructive comments on this work.

References

Steven Abney and Marc Light. 1999. Hiding a semantic class hierarchy in a Markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment*, pages 89–96.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment*, pages 65–72.

Colin. Bannard. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Technical Report WP-2002-06, University of Edinburgh, School of Informatics. http://lingo.stanford.edu/pubs/WP-2002-06.pdf.

Colin Bannard. 2005. Learning about the meaning of verb-particle constructions from corpora. *Computer Speech and Language*, 19(4):467–478.

Daniel M. Bikel. 2004. A distributional analysis of a lexicalized statistical parsing model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, July. Association for Computational Linguistics.

Don Blaheta and Mark Johnson. 2001. Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocations*, pages 54–60, Toulouse, France.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1.1. Technical Report.

Edward Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 19(2):263–312.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 337–344, Trento, Italy, April.

Christiane Fellbaum, editor. 1998. *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Ralph Grishman and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International Conference of Computational Linguistics. COLING-94*, volume I, pages 742–747.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Hang Li and Naoki Abe. 1998. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324, Univeristy of Maryland, College Park, Maryland.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 03 Workshop: Multiword expressions: analysis, acquisition and treatment*, pages 73–80.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics. (NAACL)*, pages 256–263, Seattle,WA.

Fernando Pereira, Nattali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Jorma Rissanen. 1978. Modelling by shortest data description. *Automatica*, 14:465–471.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, pages 1–15, Mexico City, Mexico.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Hong Kong.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sidney Siegel and N. John Castellan. 1988. *Non-Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2003. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Vivian Tsang and Suzanne Stevenson. 2004. Using selectional profile distance to detect verb alternations. In *Proceedings of NAACL Workshop on Computational Lexical Semantics (CLS-04)*, pages 30–37, Boston, MA.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 899–906, Vancouver, B.C., Canada.

Andreas Wagner. 2002. Learning thematic role relations for wordnets. In *Proceedings of ESSLLI-2002 Workshop on Machine Learning Approaches in Computational Linguistics*, Trento.