Looking for Prepositional Verbs in Corpus Data

Timothy Baldwin

NICTA Victoria Lab
CSSE
University of Melbourne
Victoria 3010 Australia

CSLI, Stanford University
210 Panama Street, Stanford
CA 94305-4115 USA

tim@csse.unimelb.edu.au

Abstract

We propose a number of unsupervised methods for extracting prepositional verbs (e.g. *refer to, look for*) from corpus data, based on linguistic tests and/or statistical measures. We demonstrate the effectiveness of the individual techniques over a prepositional verb deep lexical acquisition task, and go on to document the successes of an unsupervised classifier combination method.

Keywords: prepositional verb, deep lexical acquisition, multiword expression

1 Introduction

There is growing acknowledgement of the importance of multiword expressions (MWEs) in any holistic language technology solution, particularly in applications which require fine-grained linguistic precision. We define MWEs to be complex lexical items made up of multiple word segments, which are lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic (e.g. *ad hoc, by and large, kick the bucket*, *good morning* and *summer school*, respectively: Sag et al. (2002), Calzolari et al. (2002)).

Any grammar engineering solution to MWEs is made up of two components: (1) the machinery to systematically capture the idiosyncracies of different MWE classes, in the form of a system of lexical types; and (2) the lexical items by which each lexical type is populated. For English verb particle constructions (VPCs), for example, we may encode classes including: (1) intransitive VPCs (e.g. *shoot off*); (2) transitive VPCs which undergo the particle alternation (e.g. *look up* as in *look up the word/look the word up*); and (3) transitive VPCs which strictly occur with split word order (e.g. *have off* as in *have Friday off/*have off Friday*). Each of these lexical types would then be associated with a set of verb–particle pairs which have the predicted syntax. Our interest is in the second of these tasks, that is the population of a lexicon with MWEs classified according to an appropriate set of lexical types.

There are a number of established methods for populating a lexicon. Perhaps the most obvious technique is to mine lexical items from some preexisting machine-readable dictionary. While this may sound trivial, it suffers from two primary shortcomings: (1) for productive MWEs, the coverage of pre-existing dictionaries tends to be patchy (as was shown by Villavicencio (2003) for VPCs); and (2) there are often mismatches in the lexical type systems adopted in different lexicons, such that manual intervention is required to align the lexical type of some or all of the data (Sanfilippo and Poznański, 1992). An alternative strategy is to learn lexical items from corpus data, in a process we term corpus-based **deep lexical acquisition**, that is the acquisition of lexical items from corpus data in a form compatible with some deep lexical resource. Deep lexical acquisition has the advantage over dictionary mining techniques that it is sensitive to the corpus it is applied to, making it possible to tune a lexicon to a particular domain or register. It has the further

In Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, Colchester, UK.

benefit that we are able to fine-tune our extraction method to the peculiarities of the lexical type system in question, and can feed the output directly into the lexicon without worry of misalignment.

The particular MWE type we target for deep lexical acquisition is English prepositional verbs (PVs), that is verbs which select for a specified (transitive) preposition, such as *look for* or *refer to*. Similarly to VPCs, PVs are productive and dictionary coverage is thus variable (see Section 2). This motivates a corpus-based deep lexical acquisition approach to PV learning.

We propose a range of methods for extracting prepositional verbs, each of which operates in an unsupervised fashion, ranking verb–preposition combinations according to a range of corpus-derived statistics of verb, preposition and noun co-occurrence. In evaluation over a range of training corpora, we found the performance of the individual methods to vary considerably, but also that the component methods combine together to produce an extraction technique which is superior to the individual methods.

The remainder of this paper is structured as follows. Section 2 defines PVs and illustrates the difficulty of the extraction task. Section 3 details the full range of methods proposed for extracting PVs. Section 4 evaluates the methods relative to gold-standard dictionary data. Finally, Section 5 concludes the paper with a discussion of past research.

2 The Prepositional Verb Extraction Task

In this section, we provide a linguistic description of PVs as is relevant to our extraction method, and detail the gold-standard lexical resources used in this research.

2.1 The Nature of Prepositional Verbs

We define **prepositional verbs** (PVs) to be verbal MWEs which select for a PP argument made up of a specified preposition head and NP argument. Examples of PVs are *refer to (the book), come across (the letter)* and *skate over (the issue)*. As with many MWEs, PVs cover the full spectrum of semantic compositionality, e.g. *refer to* or *compete in* are fully compositional, whereas *play on (one's fears)* and *grow on (you)* are relatively far removed from the semantics of the simple verbs.

The preposition in a PV can be either fixed or mobile: **fixed prepositions** (e.g. *come <u>across</u> (the letter)*) must occur immediately after the verb, while **mobile prepositions** (e.g. *refer <u>to</u> (the book)*) undergo the same range of variation as non-specified prepositions (e.g. *walk down (the path)*: Huddleston and Pullum (2002)). Basic tests which can be used to distinguish PVs from simple verb–preposition combinations are:

1. the object of the preposition is passivisable in mobile preposition PVs (e.g. *the book was referred to*) but not fixed preposition PVs or simple verb–preposition combinations (e.g. **the letter was come across* and **the path was walked down*, respectively);

2. in fixed preposition PVs, the preposition must follow immediately after the verb (e.g. **the letter across which I came, *across which letter I came, *come suddenly across the letter*), whereas mobile preposition PVs and simple verb–preposition combinations are more flexible (e.g. *the book referred to* and *walk quietly down the path*, respectively);

Were it possible to obtain expert judgements on the syntactic status of verb–preposition combinations, these tests would be sufficient to identify PVs. However, given that we intend to use corpus data as our sole source of evidence in distinguishing PVs from simple verb–preposition combinations, it is doubtful how much leverage the tests are going to give us, particularly for mobile preposition PVs which are separated from simple verb–preposition combinations only by the ability to passivise.

One further construction which is potentially confusable with PVs is transitive VPCs (e.g. *look up (the word)*). Here, however, we can draw on both word order and preposition valence to discriminate the two MWE classes: transitive VPCs generally undergo the particle alternation (e.g. *look up the word/look the word up* vs. *refer to the book/*refer the book to*), and the preposition is intransitive in VPCs but transitive in PVs. Fortunately, part-of-speech taggers and chunk parsers are remarkably effective at disambiguating preposition valence, ameliorating the effects of this potential ambiguity (Baldwin, to appear).

For the purposes of this paper, we will focus exclusively on PVs which select for a single PP. We recognise that there are further lexical types that warrant consideration, including PVs which select for both an NP and a PP (e.g. *intend (the book) for (Kim)*) and PVs which select for two PPs (e.g. *look to (Kim) for (advice)*). However, single-PP PVs are by far the most common type of PV, and thus benefit most from deep lexical acquisition. Also, in developing an extraction technique for single-PP PVs, we can hope to arrive at an extraction technique which can later be extended to other PV lexical types.

2.2 Lexical Resources

In order to carry out deep lexical acquisition, we clearly need a deep lexical resource and associated system of lexical types to tailor our method to. The particular deep lexical resource we select is the LinGO English Resource Grammar (LinGO-ERG: Flickinger et al. (2000), Copestake and Flickinger (2000)), a medium-scale HPSG grammar of English. The PV lexical type that we are interested in is `v_empty_prep_intrans_le` (i.e. single-PP PV), which accounts for around half of the PV lexical entries in the LinGO-ERG lexicon (based on the grammar version of 31 Dec, 2004). Note that the LinGO-ERG does not currently distinguish between fixed and mobile preposition PVs, and we thus do not need to make this distinction in the extraction task.[1]

In order to study patterns of verb–preposition combination over a fixed set of verbs and prepositions, we first identified the 100 most frequent verbs and 10 most frequent transitive prepositions in the written component of the British National Corpus (BNC: Burnard (2000)), as detailed in Appendix A; this sampling was based on the lemmatised output of a custom-built Penn-style tagger. For the 1000 verb–preposition combinations generated by this dataset, we checked for an instance of a `v_empty_prep_intrans_le` lexical item in either the LinGO-ERG or the Longman Phrasal Verbs Dictionary (Dignen et al., 2000). In this way, we identified 135 gold standard PVs for use in evaluation. Note that 75 of the 135 gold-standard PVs were found in the LinGO-ERG and 94 in the Longman dictionary, with an overlap of only 34 PVs. That is, the intersection of the PV content of the two lexical resources accounts for less than half of the PV data found in each, underlining the patchiness of PV coverage in lexical resources.

3 Extraction Methods

We employ a selection of unsupervised corpus-based extraction methods, each of which ranks the set of verb–preposition pairs for relative likelihood of being a PV. Our extraction methods can be categorised as: (a) purely statistical (Simple Verb–Preposition Frequency, the Dice Coefficient, Pointwise Mutual Information, χ^2 and Log-likelihood Ratio); (b) purely linguistic, based on our linguistic tests (Stranded Preposition Frequency, Distance-conditioned Verb–preposition Frequency and Verb–preposition Distance Ratio); and (c) hybrid statistical and linguistic (Skew Divergence). We also present a basic method for system combination.

The corpora we use to derive the feedstock statistics for each method are: the Brown corpus (0.3m words) and the Wall Street Journal (WSJ) corpus (0.7m words), both from the Penn Treebank (Marcus et al., 1993), and also the BNC (90m words). In each case, we chunk parsed the raw text data using a custom-built full text chunk parser based on fnTBL 1.0 (Ngai and Florian, 2001), and lemmatised each word token using morph (Minnen et al., 2001). All verb, preposition and noun token statistics were based on the heads of the respective chunk types in the chunker output.

Simple Verb–preposition Frequency

The most straightforward extraction method is based on the raw frequency of occurrence of each verb–preposition combination in the corpora. In this, we generate a ranking based on the frequency $f(V, P)$ of verb V and preposition P as heads of a verb and preposition chunk, respectively, within 4 chunks of each other; we label this method **V-P Frequency**$_{\text{BASE}}$ in Table 1.

[1] In the current version of the LinGO-ERG, the `v_empty_prep_intrans_le` lexical type allows temporal and locative adverbials but not sentential adverbs to occur between the verb and preposition, thus overgenerating in instances such as **Kim came yesterday across the book* and undergenerating in instances such as *Kim referred frequently to the book*.

Dice Coefficient:

$$\frac{2\,f(V,P)}{f(V,*)+f(*,P)} \tag{1}$$

Pointwise Mutual Information:

$$\log \frac{p(V,P)}{p(V)\,p(P)} \tag{2}$$

χ^2 :

$$\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}} \tag{3}$$

Log-likelihood Ratio:

$$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}} \tag{4}$$

Verb-preposition Distance Ratio:

$$\frac{\sum_{i=0}^{Distance} f(V,P,i)}{\sum_{i=0}^{N} f(V,P,i)} \tag{5}$$

Skew Divergence:

$$s_\alpha(q,r) = D(r \mid\mid \alpha q + (1-\alpha)r) \tag{6}$$

$$D(q \mid\mid r) = \sum_y q(y)(\log q(y) - \log r(y)) \tag{7}$$

Association Measures

We tested a selection of association measures with wide currency in the collocation extraction literature (Schone and Jurafsky, 2001; Pearce, 2002), namely **the Dice Coefficient**, **Pointwise Mutual Information** (Church and Hanks, 1989), χ^2 (**Chi-square**), and **the Log-likelihood Ratio** (Dunning, 1993). Each association measure calculates the deviation between the observed joint frequency of verb V and particle P (i.e. $f(V,P)$) and the expected joint frequency assuming independence between the two lexical items (i.e. $\hat{f}(V,P)$). This takes the form of a direct ratio in the case of the Dice Coefficient and Pointwise Mutual Information (Equations 1 and 2), or alternatively analysis of the verb–preposition contingency table in the case of χ^2 and the Log-likelihood Ratio (Equations 3 and 4).

The manner in which we employ the association measures is, for a given corpus, to calculate the frequency with which each of our 100 most-frequent verbs and 10 most-frequent prepositions occur as the head of verb and preposition chunks, respectively; the joint verb–preposition frequencies are based on strict adjacency. The output of each method takes the form of a descending ranking of preposition–verb pairs, relative to association score.

In our implementation of these measures, we borrowed heavily from the Ngram Statistics Package (Banerjee and Pedersen, 2003).

Stranded Preposition Frequency

In the first linguistic test in Section 2.1, we observed that the object of the preposition is passivisable in mobile preposition PVs, but not simple verb–preposition combinations or fixed preposition PVs. We operationalise this test by calculating the simple frequency of verb–preposition pairs where the transitive preposition immediately follows a verb chunk headed by the verb in question, and the preposition is immediately proceeded by a sentence or clause boundary (i.e. an O chunk headed by any member of the regular expression `[.,:;!?]`), such as in [$_{\text{N}}$ *The book*] [$_{\text{NP}}$ *Kim*] [$_{\text{VP}}$ *referred*] [$_{\text{PP}}$ *to*] [$_{\text{O}}$.]. We term this the “stranded preposition” frequency due to there being no object NP proceeding the preposition.

Distance-conditioned Verb–preposition Frequency

In the second linguistic test in Section 2.1, we observed that the preposition tends to occur immediately after the verb in fixed preposition PVs. We apply this test by calculating the co-occurrence frequency $f(V,P,i)$ of preposition P and the nearest verb V to the left at “chunk distances” $i = 0, 1, .., 4$; the chunk distance is simply the number of chunks intervening between the verb and preposition chunks headed by V and P, respectively. For example, the sentence [$_{\text{NP}}$ *Kim*] [$_{\text{VP}}$ *referred*] [$_{\text{PP}}$ *on*] [$_{\text{NP}}$ *occasion*] [$_{\text{PP}}$ *to*] [$_{\text{NP}}$ *the book*] would constitute an occurrence of *refer to* at distance 2. For each value of i, we generate an independent ranking of verb–preposition pairs, which we then combine by ranking the pairs in ascending order of mean rank. We label this second method **V-P Frequency**$_{\text{DIST}}$ in Table 1.

Verb–preposition Distance Ratio

As a variant on Distance-conditioned Verb–preposition Frequency (above), we calculate the ratio of verb–preposition corpus instances found at a given chunk distance $Distance$ or less as given in Equation 5, where N, the upper bound on chunk distance, was set to 4 throughout evaluation. In accordance with the second linguistic test in Section 2.1, we expect that for smaller values of $Distance$ this ratio will be higher for PVs than for simple verb–preposition combinations. As for the verb–preposition frequency method, we generate the final ranking of verb–preposition pairs in ascending order of rank sum over the individual rankings for $Distance = 0, 1, ..., 4$. Note that while mobile preposition PVs can allow modifiers to occur between the verb and preposition, we predict that actual rates of occurrence relative to simple verb–preposition combinations will be low, such that the ratio will be equally capable of identifying both PV types.

Skew Divergence

Skew divergence provides a means of measuring the distance between two probability distributions. It was proposed by Lee (2001) as an approximation of the Kullback-Leibler (KL) divergence which is robust to unseen events in the probability distributions being compared. It does this in an asymmetric fashion by taking distributions q and r, and for each non-zero event probability $q(y)$ in q, deriving a corresponding event probability in r by interpolating over $q(y)$ and $r(y)$ according to $s_\alpha(q, r)$ as detailed in Equations 6 and 7, where $D(q \mid\mid r)$ is the KL divergence. We follow Lee (2001) in setting α to 0.5 in our experiments.

We apply skew divergence to the task of PV extraction by taking q as the distribution over $P(N|V, P)$ and r as the distribution over $P(N|P)$. That is, we find the distribution of: (a) nouns N governed by the preposition P in verb–preposition combination $V - P$, where V and P are adjacent, and (b) nouns N governed by the preposition P in any context; we then calculate the divergence between these two distributions. Our expectation is that the selectional preferences of a preposition in a given PV are markedly different to those in general, and thus rank the verb–preposition combinations in descending order of skew divergence.

System combination

We carry out system combination by simply summing together the ranks for each verb–preposition pair produced by the individual extraction methods, and reranking the verb–preposition pairs in increasing order of rank sum. This is carried out: (1) across the rankings produced for each of our three corpora, to produce a consolidated ranking for each individual extraction method (the **All** column in Table 1); (2) across the extraction methods for a given corpus (the **Combined** row in Table 1); and (3) across all corpora and all extraction methods, combining a total of 15 base rankings.

4 Evaluation

Evaluation of the extraction methods was performed by taking the gold-standard set of 135 PVs and the ranking of the 1000 preposition–verb combinations generated by each method, and calculating: (a) the Z-score according to the Mann-Whitney test, and (b) the top-N F-score, that is the F-score as calculated over the top-N items in the ranking. With the Mann-Whitney test, the higher the Z-score, the greater the relative proportion of PVs that are contained in the upper reaches of the ranking. The upper bound on the Z-score, generated with the 135 gold-standard PVs ranked 1–135 followed by the remainder of the verb–preposition combinations, is 18.7; the Z-score for a random ranking, averaged over 100 random rankings of the data, is 0.6. The reason for us evaluating according to the top-N F-score is that, in practical applications, we are going to need to select some number of items to skim off the top of the ranking for inclusion in our lexicon as PV lexical items. In our evaluation, we set N to 135 for scaling convenience, such that the upper bound top-N F-score, based on the gold-standard PVs occupying ranks 1–135, is 1.00; the top-N F-score for a random

Extraction Method	Corpus Brown		WSJ		BNC		All	
	Z	F	Z	F	Z	F	Z	F
Random ranking	0.6	0.13	0.6	0.13	0.6	0.13	0.6	0.13
BASELINE:								
V-P Frequency$_{\text{BASE}}$	4.0	0.16	1.8	0.14	0.5	0.13	4.5	0.16
PURELY STATISTICAL:								
Dice Coefficient	**10.8**	0.42	7.4	0.30	10.3	0.39	11.2	0.37
Mutual Information	9.7	0.35	8.5	0.32	9.0	0.37	9.4	0.38
χ^2	9.9	0.36	7.9	0.27	0.2	0.13	8.9	0.36
Log-likelihood	10.0	0.33	7.7	0.24	0.2	0.13	8.8	0.36
LINGUISTIC:								
Stranded P	3.3	0.27	1.3	0.21	10.1	0.32	10.9	0.42
V-P Frequency$_{\text{DIST}}$	9.6	0.39	8.3	0.33	9.2	0.37	11.0	0.44
V-P Distance Ratio	8.7	0.32	8.0	0.30	9.4	0.36	9.2	0.33
HYBRID:								
Skew Divergence	10.3	**0.44**	8.4	0.28	1.7	0.12	8.4	0.37
COMBINED	**10.9**	**0.44**	**8.8**	**0.36**	**10.8**	**0.41**	**11.3**	**0.45**
Perfect Ranking	18.7	1.00	18.7	1.00	18.7	1.00	18.7	1.00

Table 1: Prepositional verb extraction results (Z = Z-score calculated according to the Mann-Whitney test; F = top-135 F-score)

ranking, generated through the same process of random ranking as for the Mann-Whitney test, is 0.13.

The results for each method over the different corpus combinations are presented in Table 1. The baseline for the task is taken to be V-P Frequency$_{\text{BASE}}$, that is the ranking method based on raw verb–preposition frequency which does not take chunk distance into account.

There is an interesting divergence in the performance of the purely statistical methods: the Dice Coefficient and Pointwise Mutual Information performed relatively consistently across all corpus datasets; with V-P Frequency$_{\text{BASE}}$, χ^2 and Log-likelihood, on the other hand, there was a negative correlation between the size of the corpus and performance, with the best performance for a single corpus observed for the Brown corpus and the worst performance (at or below the level of the random baseline) over the BNC. It is interesting to observe that the consistently-performing methods are both based on analysis of joint vs. independent frequencies (i.e. $f(P, V)$ vs. $f(P, *)$ and $f(*, V)$), whereas the remainder of the methods are based exclusively on observed and expected joint frequencies (either simply $f(P, V)$, or $f(P, V)$ vs. $f(P, \bar{V})$, $f(\bar{P}, V)$ and $f(\bar{P}, \bar{V})$)). Further research is required to quantify the impact on the results of the types of statistics utilised in each method. Remarkably, all statistical methods performed best over the Brown corpus despite its modest size, although we have to some degree factored out the effects of data sparseness in focusing exclusively on frequent verbs and prepositions. Corpus combination brought the results for each method up to roughly the highest performance level over a single corpus (namely the Brown corpus). Overall, the Dice Coefficient was the best-performing statistical method.

The purely linguistic extraction methods (Stranded P, V-P Frequency$_{\text{DIST}}$ and V-P Distance Ratio) were all well clear of both the random and V-P Frequency$_{\text{BASE}}$ baselines, and performed well across all corpora in terms of top-N F-score. For Stranded P, we got an appreciable increase in Z-score when using the BNC (and also when combining the three corpora) as the larger data volume dramatically reduced the effects of data spareness, whereas the Z-score was relatively constant for both V-P Frequency$_{\text{DIST}}$ and V-P

Distance Ratio. Contrary to the results for the purely statistical methods, for all three linguistic methods, the results over the BNC were roughly as good or better than results over the other two corpora. The methods also benefitted from corpus combination to a greater degree that the purely statistical methods. Overall, V-P Frequency$_{\mathrm{DIST}}$ was the best-performing linguistic method.

Similarly to a number of the purely statistical methods, Skew Divergence performed best over the small-scale Brown corpus and worst over the large-scale BNC, although the relative drop-off in performance was less pronounced. The top-N F-score for Skew Divergence over the Brown corpus was the best of all methods, equalling the combined method at 0.44, whereas the Z-score tended to be relatively less impressive, ranking 9/10 in corpus combination. This suggests that Skew Divergence is effective in the upper reaches of the ranking, but more erratic towards the tail of the ranking.

There is very little separating the best of the purely statistical (i.e. the Dice Coefficient) and the best of the linguistic extraction methods (i.e. V-P Frequency$_{\mathrm{DIST}}$), and our one hybrid method (Skew Divergence) is at roughly the same level of top-N F-score (but has a lower Z-score – see above).

Corpus combination (i.e. combining the rankings across all three corpora for a given method) led to an equal or higher top-N F-score for 5 out of the 10 extraction methods, and equal or higher Z-score for 7 out of the 10 extraction methods. Method combination (i.e. combining the rankings across the basic extraction methods) led to an equal or higher top-N F-score and Z-score in all cases. The best overall performance was achieved with corpus and method combination in tandem, resulting in a Z-score of 11.3 and top-N F-score of 0.45.

To further explore the impact of corpus combination on the results, and the tailing off of the performance of Skew Divergence observed above, we plotted the precision–recall curves for the different methods over each of the base corpora, and also under corpus combination (see Figures 1–4). That is, for each method, we determined the precision at recall rates of 0.1, 0.2, ... 1.0 in order to analyse the consistency of the generated PV ranking. We focus specifically on the results for V-P Frequency$_{\mathrm{BASE}}$ (**V-P Base**), the Dice Coefficient (**Dice**), the Log-likelihood Ration (**LLR**), V-P Frequency$_{\mathrm{DIST}}$ (**V-P Distance**), Skew Divergence (**Skew**) and method combination (**Combined**). Note that we omit the results for Mutual Information, χ^2, Stranded P and V-P Distance Ratio as they are largely analogous to those for the Dice Coefficient, Log-likelihood Ratio, V-P Frequency$_{\mathrm{DIST}}$ and V-P Frequency$_{\mathrm{DIST}}$, respectively.

In Figures 1–4, we can verify our claims from above that: (1) there is a negative correlation between corpus size and the performance of the Log-likelihood Ratio and Skew Divergence (with the below-baseline results over the BNC self-evident in Figure 3), but that this effect is smoothed under corpus combination (Figure 4); (2) the performance of Skew Divergence under corpus combination is good through the early stages of the ranking, but erratic in the latter half of the ranking; (3) the Dice Coefficient and V-P Frequency$_{\mathrm{DIST}}$ are both consistent performers across all corpora (although the Dice Coefficient is the more consistent throughout the ranking, as borne out in the higher Z-score values); and (4) method combination, particularly when combined with corpus combination, is superior to the individual methods. The smoothing effect of corpus combination is illustrated nicely in the largely linear decreasing curves in Figure 4, as contrasted with the erratic tangle of curves in Figure 1.

5 Discussion

The work of Krenn and Evert (2001) on a German PP-verb extraction task has interesting implications for this research. German PP-verbs are unlike English PVs in that they have fixed lexical form (akin to light verb constructions, e.g. *make a speech*), and Krenn and Evert (2001) made no attempt to learn the valence of the PP-verbs. One intriguing finding of the research is that raw frequency was found to be equivalent in performance to all the statistical "association measure" extraction methods tested. In our case, all tested methods were found to well outperform the raw frequency baseline (V-P Frequency$_{\mathrm{BASE}}$) under corpus combination, except for the selection of purely statistical methods which dropped in performance as the corpus size increased.

Blaheta and Johnson (2001) developed an unsupervised log-linear model for learning English

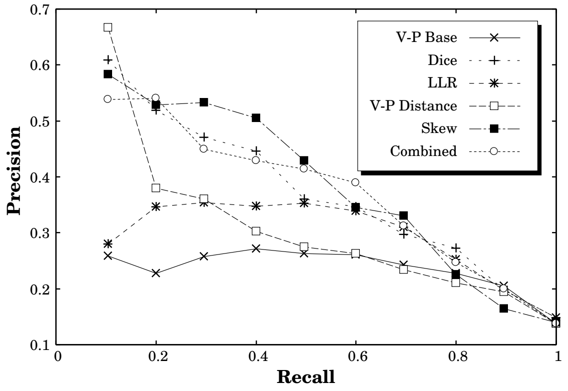

Figure 1: Precision–recall curve over Brown

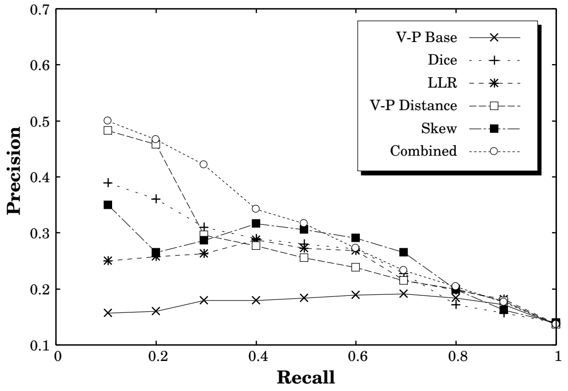

Figure 2: Precision–recall curve over WSJ

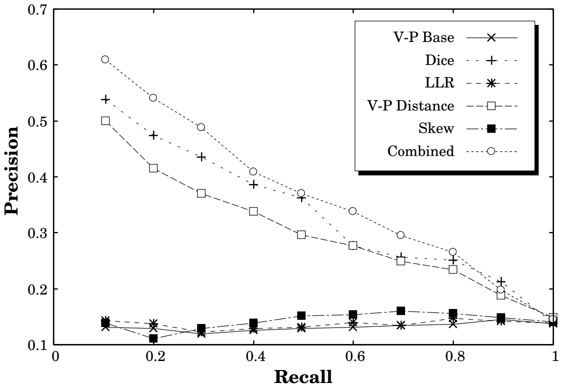

Figure 3: Precision–recall curve over BNC

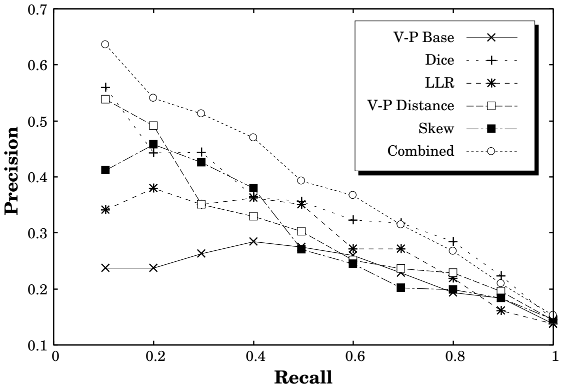

Figure 4: Precision–recall curve over All

"multi-word verbs", by which is meant both VPCs and PVs. As Blaheta and Johnson (2001) blur the distinction between VPCs and PVs, and make no attempt to learn verb valence, direct comparison is difficult. They claim a precision of 0.68 over the top-100 multi-word verbs extracted by their method, whereas our best-performing method produced a precision of 0.48 over the top-100 items, as evaluated in a deep lexical acquisition context with considerably greater syntactic precision.

In related research on the deep lexical acquisition on VPCs, Baldwin (to appear) used taggers, a chunker, a chunk grammar and a full parser to identify VPC instances, and combined the evidence from the individual pre-processors together to produce a supervised method. A significant divergence over our research is the evaluation methodology, in that Baldwin (to appear) took a gold-standard VPC lexicon and pre-annotated three corpora for actual occurrence of the VPC lexical items. In doing so, Baldwin was able to filter out the effects of non-corpus-attested lexical items in the reported results. We have little sense of whether all 135 of our gold-standard PVs occur in the three corpora, and whether low-ranked items are due to a lack of corpus data or some more fundamental shortcoming of our extraction methods. We leave this as an item for future research.

While we have focused on unsupervised extraction methods in this research, there is no doubt that we could benefit from combining our method with supervised techniques for distinguishing between argument and modifier PPs (Buchholz, 1998; Merlo and Leybold, 2001). It would be intriguing to investigate the interface between these two tasks, which we leave for future research.

In conclusion, this paper has proposed a range of unsupervised methods for performing the deep lexical acquisition of English prepositional verbs based on corpus data. The proposed methods draw on a combination of statistical and/or linguistic evidence, and were found to combine together to produce an extraction method with a Z-score of 11.3 and top-N F-score of 0.45.

A Verb and Preposition Data

The 100 most-frequent verbs in the written component of the BNC are:

> *accept, add, agree, allow, appear, apply, ask, be, become, begin, believe, bring, build, buy, call, carry, change, come, consider, continue, create, decide, describe, develop, die, do, draw, establish, exist, expect, fall, feel, find, follow, get, give, go, grow, happen, have, hear, help, hold, include, increase, involve, keep, know, lead, learn, leave, let, like, live, look, lose, make, mean, meet, move, need, offer, pass, pay, play, produce, provide, put, reach, receive, remain, remember, require, return, run, say, see, seem, send, set, show, sit, speak, stand, start, stop, suggest, take, talk, tell, think, try, turn, understand, use, walk, want, win, work, write*

The 10 most-frequent transitive prepositions in the BNC are:

> *as, at, by, for, from, in, of, on, to, with*

Acknowledgements

We would like to thank the anonymous reviewers for their valuable input on this research. Timothy Baldwin is supported by the National Science Foundation under Grant No. BCS-0094638, and the research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and the University of Melbourne.

References

Timothy Baldwin. to appear. The deep lexical acquisition of English verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*.

Satanjeev Banerjee and Ted Pedersen. 2003. The design, implementation, and use of the Ngram Statistics Package. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, pages 17–21, Mexico City, Mexico.

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

Sabine Buchholz. 1998. Distinguishing complements from adjuncts using memory-based learning. In *Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*, pages 41–8, Saarbrücken, Germany.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–40, Las Palmas, Canary Islands.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pages 76–83, Vancouver, Canada.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

Sheila Dignen, Ted Jackson, Jo Leigh, and Evadne Adrian-Vallance, editors. 2000. *Longman Phrasal Verbs Dictionary*. Longman, New York, USA.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Dan Flickinger, Stephan Oepen, Hans Uszkoreit, and Jun'ichi Tsujii. 2000. On building a more efficient grammar by exploiting types. *Journal of Natural Language Engineering* (Special Issue on Efficient Processing with HPSG), 6(1):15–28.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

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/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, pages 39–46, Toulouse, France.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of Artificial Intelligence and Statistics 2001 (AISTATS 2001)*, pages 65–72, Key West, USA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–30.

Paola Merlo and Matthias Leybold. 2001. Automatic distinction of arguments and modifiers: the case of prepositional phrases. In *Proceedings of the ACL/EACL-2001 Workshop on Computational Natural Language Learning (CoNLL-2001)*, pages 121–8, Toulouse, France.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin, Germany.

Antonio Sanfilippo and Victor Poznański. 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP)*, pages 80–7, Trento, Italy.

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108, Pittsburgh, USA.

Aline Villavicencio. 2003. Verb-particle constructions and lexical resources. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64, Sapporo, Japan.