Identifying Phrasal Verbs Using Many Bilingual Corpora

Karl Pichotta*
Department of Computer Science
University of Texas at Austin
pichotta@cs.utexas.edu

John DeNero
Google, Inc.
denero@google.com

Abstract

We address the problem of identifying multiword expressions in a language, focusing on English phrasal verbs. Our *polyglot ranking* approach integrates frequency statistics from translated corpora in 50 different languages. Our experimental evaluation demonstrates that combining statistical evidence from many parallel corpora using a novel ranking-oriented boosting algorithm produces a comprehensive set of English phrasal verbs, achieving performance comparable to a human-curated set.

1 Introduction

A *multiword expression* (MWE), or *noncompositional compound*, is a sequence of words whose meaning cannot be composed directly from the meanings of its constituent words. These idiosyncratic phrases are prevalent in the lexicon of a language; Jackendoff (1993) estimates that their number is on the same order of magnitude as that of single words, and Sag et al. (2002) suggest that they are much more common, though quantifying them is challenging (Church, 2011). The task of identifying MWEs is relevant not only to lexical semantics applications, but also machine translation (Koehn et al., 2003; Ren et al., 2009; Pal et al., 2010), information retrieval (Xu et al., 2010; Acosta et al., 2011), and syntactic parsing (Sag et al., 2002). Awareness of MWEs has empirically proven useful in a number of domains: Finlayson and Kulkarni (2011), for example, use MWEs to attain a significant performance improvement in word sense disambiguation; Venkatapathy and Joshi (2006) use features associated with MWEs to improve word alignment.

We focus on a particular subset of MWEs, English *phrasal verbs*. A phrasal verb consists of a head verb followed by one or more particles, such that the meaning of the phrase cannot be determined by combining the simplex meanings of its constituent words (Baldwin and Villavicencio, 2002; Dixon, 1982; Bannard et al., 2003).[1] Examples of phrasal verbs include *count on* [*rely*], *look after* [*tend*], or *take off* [*remove*], the meanings of which do not involve counting, looking, or taking. In contrast, there are verbs followed by particles that are not phrasal verbs, because their meaning is compositional, such as *walk towards*, *sit behind*, or *paint on*.

We identify phrasal verbs by using frequency statistics calculated from parallel corpora, consisting of bilingual pairs of documents such that one is a translation of the other, with one document in English. We leverage the observation that a verb will translate in an atypical way when occurring as the head of a phrasal verb. For example, the word *look* in the context of *look after* will tend to translate differently from how *look* translates generally. In order to characterize this difference, we calculate a frequency distribution over translations of *look*, then compare it to the distribution of translations of *look* when followed by the word *after*. We expect that idiomatic phrasal verbs will tend to have unexpected translation of their head verbs, measured by the Kullback-Leibler divergence between those distributions.

Our *polyglot ranking* approach is motivated by the hypothesis that using many parallel corpora of different languages will help determine the degree of semantic idiomaticity of a phrase. In order to com-

*Research conducted during an internship at Google.

[1] Nomenclature varies: the term *verb-particle construction* is also used to denote what we call phrasal verbs; further, the term *phrasal verb* is sometimes used to denote a broader class of constructions.

bine evidence from multiple languages, we develop a novel boosting algorithm tailored to the task of ranking multiword expressions by their degree of idiomaticity. We train and evaluate on disjoint subsets of the phrasal verbs in English Wiktionary[2]. In our experiments, the set of phrasal verbs identified automatically by our method achieves held-out recall that nears the performance of the phrasal verbs in WordNet 3.0, a human-curated set. Our approach strongly outperforms a monolingual system, and continues to improve when incrementally adding translation statistics for 50 different languages.

2 Identifying Phrasal Verbs

The task of identifying phrasal verbs using corpus information raises several issues of experimental design. We consider four central issues below in motivating our approach.

Types vs. Tokens. When a phrase is used in context, it takes a particular meaning among its possible senses. Many phrasal verbs admit compositional senses in addition to idiomatic ones—contrast idiomatic "*look down on him for his politics*" with compositional "*look down on him from the balcony*." In this paper, we focus on the task of determining whether a phrase type is a phrasal verb, meaning that it frequently expresses an idiomatic meaning across its many token usages in a corpus. We do not attempt to distinguish which individual phrase tokens in the corpus have idiomatic senses.

Ranking vs. Classification. Identifying phrasal verbs involves relative, rather than categorical, judgments: some phrasal verbs are more compositional than others, but retain a degree of noncompositionality (McCarthy et al., 2003). Moreover, a polysemous phrasal verb may express an idiosyncratic sense more or less often than a compositional sense in a particular corpus. Therefore, we should expect a corpus-driven system not to classify phrases as strictly idiomatic or compositional, but instead assign a ranking or relative scoring to a set of candidates.

Candidate Phrases. We distinguish between the task of identifying candidate multiword expressions and the task of ranking those candidates by their semantic idiosyncracy. With English phrasal verbs, it is straightforward to enumerate all desired verbs followed by one or more particles, and rank the entire set.

[2] http://en.wiktionary.org

Feature	*Description*
φ_L ($\times 50$)	KL Divergence for each language L
μ_1	frequency of phrase given verb
μ_2	PMI of verb and particles
μ_3	μ_1 with interposed pronouns

Table 1: Features used by the polyglot ranking system.

Using Parallel Corpora. There have been a number of approaches proposed for the use of multilingual resources for MWE identification (Melamed, 1997; Villada Moirón and Tiedemann, 2006; Caseli et al., 2010; Tsvetkov and Wintner, 2012; Salehi and Cook, 2013). Our approach differs from previous work in that we identify MWEs using translation distributions of verbs, as opposed to 1–1, 1–m, or m–n word alignments, most-likely translations, bilingual dictionaries, or distributional entropy. To the best of our knowledge, ours is the first approach to use translational distributions to leverage the observation that a verb typically translates differently when it heads a phrasal verb.

3 The Polyglot Ranking Approach

Our approach uses bilingual and monolingual statistics as features, computed over unlabeled corpora. Each statistic characterizes the degree of idiosyncrasy of a candidate phrasal verb, using a single monolingual or bilingual corpus. We combine features for many language pairs using a boosting algorithm that optimizes a ranking objective using a supervised training set of English phrasal verbs. Each of these aspects of our approach is described in detail below; for reference, Table 1 provides a list of the features used.

3.1 Bilingual Statistics

One of the intuitive properties of an MWE is that its individual words likely do not translate literally when the whole expression is translated into another language (Melamed, 1997). We capture this effect

by measuring the divergence between how a verb translates generally and how it translates when heading a candidate phrasal verb.

A *parallel corpus* is a collection of document pairs $\langle D_E, D_F \rangle$, where D_E is in English, D_F is in another language, one document is a translation of the other, and all documents D_F are in the same language. A *phrase-aligned parallel corpus* aligns those documents at a sentence, phrase, and word level. A phrase e aligns to another phrase f if some word in e aligns to some word in f and no word in e or f aligns outside of f or e, respectively. As a result of this definition, the words within an aligned phrase pair are themselves connected by word-level alignments.

Given an English phrase e, define $F(e)$ to be the set of all foreign phrases observed aligned to e in a parallel corpus. For any $f \in F(e)$, let $P(f|e)$ be the conditional probability of the phrase e translating to the phrase f. This probability is estimated as the relative frequency of observing f and e as an aligned phrase pair, conditioned on observing e aligned to any phrase in the corpus:

$$P(f|e) = \frac{N(e,f)}{\sum_{f'} N(e,f')}$$

with $N(e,f)$ the number of times e and f are observed occurring as an aligned phrase pair.

Next, we assign statistics to individual verbs within phrases. The first word of a candidate phrasal verb e is a verb. For a candidate phrasal verb e and a foreign phrase f, let $\pi_1(e,f)$ be the subphrase of f that is most commonly word-aligned to the first word of e. As an example, consider the phrase pair $e =$ *talk down to* and $f =$ *hablar con menosprecio*. Suppose that when e is aligned to f, the word *talk* is most frequently aligned to *hablar*. Then $\pi_1(e,f) =$ *hablar*.

For a phrase e and its set $F(e)$ of aligned translations, we define the *constituent translation probability* of a foreign subphrase x as:

$$P_e(x) = \sum_{f \in F(e)} P(f|e) \cdot \delta\left(\pi_1(e,f), x\right) \quad (1)$$

where δ is the Kronecker delta function, taking value 1 if its arguments are equal and 0 otherwise. Intuitively, P_e assigns the probability mass for every f to its subphrase most commonly aligned to the verb in e. It expresses how this verb is translated in the context of a phrasal verb construction.[3] Equation (1) defines a distribution over all phrases x of a foreign language.

We also assign statistics to verbs as they are translated outside of the context of a phrase. Let $v(e)$ be the verb of a phrasal verb candidate e, which is always its first word. For a single-word verb phrase $v(e)$, we can compute the constituent translation probability $P_{v(e)}(x)$, again using Equation (1). The difference between $P_e(x)$ and $P_{v(e)}(x)$ is that the latter sums over all translations of the verb $v(e)$, regardless of whether it appears in the context of e:

$$P_{v(e)}(x) = \sum_{f \in F(v(e))} P(f|v(e)) \cdot \delta\left(\pi_1(v(e),f), x\right)$$

For a one-word phrase such as $v(e)$, $\pi_1(v(e),f)$ is the subphrase of f that most commonly directly word-aligns to the one word of $v(e)$.

Finally, for a phrase e and its verb $v(e)$, we calculate the Kullback-Leibler (KL) divergence between the translation distribution of $v(e)$ and e:

$$D_{KL}\left(P_{v(e)} \| P_e\right) = \sum_x P_{v(e)}(x) \ln \frac{P_{v(e)}(x)}{P_e(x)} \quad (2)$$

where the sum ranges over all x such that $P_{v(e)}(x) > 0$. This quantifies the difference between the translations of e's verb when it occurs in e, and when it occurs in general. Figure 1 illustrates this computation on a toy corpus.

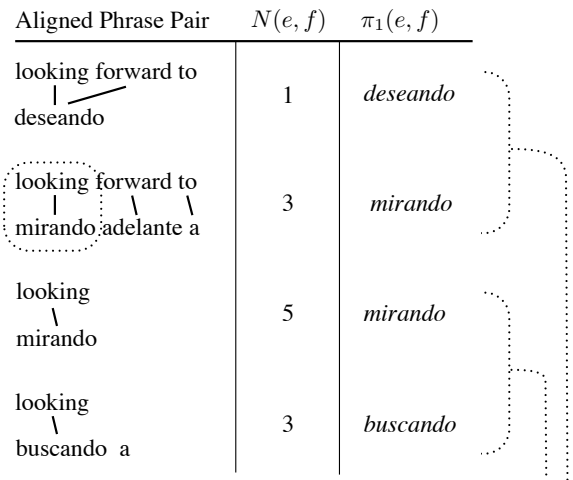
Smoothing. Equation (2) is defined only if, for every x such that $P_{v(e)}(x) > 0$, it is also the case that $P_e(x) > 0$. In order to ensure that this condition holds, we smooth the translation distributions toward uniform. Let D be the set of phrases with non-zero probability under either distribution:

$$D = \{x : P_{v(e)}(x) > 0 \text{ or } P_e(x) > 0\}$$

Then, let U_D be the uniform distribution over D:

$$U_D(x) = \begin{cases} 1/|D| & \text{if } x \in D \\ 0 & \text{if } x \notin D \end{cases}$$

[3] To extend this statistic to other types of multiword expressions, one could compute a similar distribution for other content words in the phrase.



	mirando	*deseando*	*buscando*
$P_{v(e)}(x)$	$\frac{5}{8} = 0.625$	0	$\frac{3}{8} = 0.375$
$P'_{v(e)}(x)$	0.610	0.02	0.373
$P_e(x)$	$\frac{3}{4} = 0.75$	$\frac{1}{4} = 0.25$	0
$P'_e(x)$	0.729	0.254	0.02

$$D_{KL}(P'_{v(e_i)} \| P'_{e_i}) = -0.109 + -0.045 + 1.159 = 1.005$$

Figure 1: The computation of $D_{KL}(P'_{v(e_i)} \| P'_{e_i})$ using a toy corpus, for $e =$ *looking forward to*. Note that the second aligned phrase pair contains the third, so the second's count of 3 must be included in the third's count of 5.

When computing divergence in Equation (2), we use the smoothed distributions P'_e and $P'_{v(e)}$:

$$P'_e(x) = \alpha P_e(x) + (1 - \alpha)U_D(x)$$
$$P'_{v(e)}(x) = \alpha P_{v(e)}(x) + (1 - \alpha)U_D(x).$$

We use $\alpha = 0.95$, which distributes 5% of the total probability mass evenly among all events in D.

Morphology. We calculate statistics for morphological variants of an English phrase. For a candidate English phrasal verb e (for example, *look up*), let E denote the set of inflections of that phrasal verb (for *look up*, this will be [*look*|*looks*|*looked*|*looking*] *up*). We extract the variants in E from the verb entries in English Wiktionary. The final score computed from a phrase-aligned parallel corpus translating English sentences into a language L is the average KL divergence of smoothed constituent translation distributions for any inflected form $e_i \in E$:

$$\varphi_L(e) = \frac{1}{|E|} \sum_{e_i \in E} D_{KL}\left(P'_{v(e_i)} \| P'_{e_i}\right)$$

3.2 Monolingual Statistics

We also collect a number of monolingual statistics for each phrasal verb candidate, motivated by the considerable body of previous work on the topic (Church and Hanks, 1990; Lin, 1999; McCarthy et al., 2003). The monolingual statistics are designed to identify frequent collocations in a language. This set of monolingual features is not comprehensive, as we focus our attention primarily on bilingual features in ths paper.

As above, define E to be the set of morphologically inflected variants of a candidate e, and let V be the set of inflected variants of the head verb $v(e)$ of e. We define three statistics calculated from the phrase counts of a monolingual English corpus. First, we define $\mu_1(e)$ to be the relative frequency of the candidate e, given e's head verb, summed over morphological variants:

$$\begin{aligned}\mu_1(e) &= \ln P(E|V) \\ &= \ln \frac{\sum_{e_i \in E} N(e_i)}{\sum_{v_i \in V} N(v_i)}\end{aligned}$$

where $N(x)$ is the number of times phrase x was observed in the monolingual corpus.

Second, define $\mu_2(e)$ to be the pointwise mutual information (PMI) between V (the event that one of the inflections of the verb in e is observed) and R, the event of observing the rest of the phrase:

$$\begin{aligned}&\mu_2(e) \\ &= \text{PMI}(V, R) \\ &= \lg P(V, R) - \lg\left(P(V)P(R)\right) \\ &= \lg P(E) - \lg\left(P(V)P(R)\right) \\ &= \lg \sum_{e_i \in E} N(e_i) - \lg \sum_{v_i \in V} N(v_i) - \lg N(r) + \lg N\end{aligned}$$

where N is the total number of tokens in the corpus, and logarithms are base-2. This statistic characterizes the degree of association between a verb and its phrasal extension. We only calculate μ_2 for two-word phrases, as it did not prove helpful for longer phrases.

Finally, define $\mu_3(e)$ to be the relative frequency of the phrasal verb e augmented by an accusative pronoun, conditioned on the verb. Let A be the set of phrases in E with an accusative pronoun (*it, them, him, her, me, you*) optionally inserted either at the end of the phrase or directly after the verb. For $e =$ *look up*, $A = \{$*look up, look X up, look up X, looks up, looks X up, looks up X,* $\dots\}$, with X an accusative pronoun. The μ_3 statistic is similar to μ_1, but allows for an intervening or following pronoun:

$$\begin{aligned}\mu_3(e) &= \ln P(A|V) \\ &= \ln \frac{\sum_{e_i \in A} N(e_i)}{\sum_{v_i \in V} N(v_i)}.\end{aligned}$$

This statistic is designed to exploit the intuition that phrasal verbs frequently have accusative pronouns either inserted into the middle (e.g. *look it up*) or at the end (e.g. *look down on him*).

3.3 Ranking Phrasal Verb Candidates

Our goal is to assign a single real-valued score to each candidate e, by which we can rank candidates according to semantic idiosyncrasy. For each language L for which we have a parallel corpus, we defined, in section 3.1, a function $\varphi_L(e)$ assigning real values to candidate phrasal verbs e, which we hypothesize is higher on average for more idiomatic compounds. Further, in section 3.2, we defined real-valued monolingual functions μ_1, μ_2, and μ_3 for which we hypothesize the same trend holds. Because each score individually ranks all candidates, it is natural to view each φ_L and μ_i as a weak ranking function that we can combine with a supervised boosting objective. We use a modified version of AdaBoost (Freund and Schapire, 1995) that optimizes for recall.

For each φ_L and μ_i, we compute a ranked list of candidate phrasal verbs, ordered from highest to lowest value. To simplify learning, we consider only the top 5000 candidate phrasal verbs according to μ_1, μ_2, and μ_3. This pruning procedure excludes candidates that do not appear in our monolingual corpus.

We optimize the ranker using an unranked, incomplete training set of phrasal verbs. We can evaluate the quality of the a ranker by outputting the top N ranked candidates and measuring recall relative to this gold-standard training set. We choose this recall-at-N metric so as to not directly penalize precision errors, as our training set is incomplete.

Algorithm 1 Recall-Oriented Ranking AdaBoost

for $i = 1 : |X|$ **do**
  $w[i] \leftarrow 1/|X|$
end for
for $t = 1 : T$ **do**
  for all $h \in \mathcal{H}$ **do**
    $\epsilon_h \leftarrow 0$
    for $i = 1 : |X|$ **do**
      if $x_i \notin h$ **then**
        $\epsilon_h \leftarrow \epsilon_h + w[i]$
      end if
    end for
  end for
  $h_t \leftarrow \text{argmax}_{h \in \mathcal{H}} |\epsilon_B - \epsilon_h|$
  $\alpha_t \leftarrow \ln(\epsilon_B / \epsilon_{h_t})$
  for $i = 1 : |X|$ **do**
    if $x_i \in h_t$ **then**
      $w[i] \leftarrow \frac{1}{Z} w[i] \exp(-\alpha_t)$
    else
      $w[i] \leftarrow \frac{1}{Z} w[i] \exp(\alpha_t)$
    end if
  end for
end for

Define $\mathcal{H}$ to be the set of N-element sets containing the top proposals for each weak ranker (we use $N = 2000$). That is, each element of $\mathcal{H}$ is a set containing the 2000 highest values for some φ_L or μ_i. We define the baseline error ϵ_B to be $1 - \mathbb{E}[R]$, with R the recall-at-N of a ranker ordering the candidate phrases in the set $\cup\mathcal{H}$ at random. The value $\mathbb{E}[R]$ is estimated by averaging the recall-at-N of 1000 random orderings of $\cup\mathcal{H}$.

Algorithm 1 gives the formulation of the AdaBoost training algorithm that we use to combine weak rankers. The algorithm maintains a weight vector w (summing to 1) which contains a positive real number for each gold standard phrasal verb in the training set X. Initially, w is uniformly set to $1/|X|$. At each iteration of the algorithm, w is modified to take higher values for recently misclassified examples. We repeatedly choose weak rankers $h_t \in \mathcal{H}$ (and corresponding real-valued coefficients α_t) that correctly rank examples with high w values.

Lines 5–12 of Algorithm 1 calculate the weighted error values ϵ_h for every weak ranker set $h \in \mathcal{H}$. The error ϵ_h will be 1 if h contains none of X and 0 if h contains all of X, as w always sums to 1. Line 13 picks the ranker $h_t \in \mathcal{H}$ whose weighted error is as far as possible from the random baseline error ϵ_B. Line 14 calculates a coefficient α_t for h_t, which will be positive if $\epsilon_{h_t} < \epsilon_B$ and negative if $\epsilon_{h_t} > \epsilon_B$. Intuitively, α_t encodes the importance of h_t—it will be high if h_t performs well, and low if it performs poorly. The Z in lines 17 and 19 is the normalizing constant ensuring the vector w sums to 1.

After termination of Algorithm 1, we have weights $\alpha_1, \ldots, \alpha_T$ and lists $h_1, \ldots, h_T$. Define f_t as the function that generated the list h_t (each f_t will be some φ_L or μ_i). Now, we define a final combined function φ, taking a phrase e and returning a real number:

$$\varphi(e) = \sum_{t=1}^{T} \alpha_t f_t(e).$$

We standardize the scores of individual weak rankers to have mean 0 and variance 1, so that their scores are comparable.

The final learned ranker outputs a real value, instead of the class labels frequently found in AdaBoost. This follows previous work using boosting for learning to rank (Freund et al., 2003; Xu and Li, 2007). Our algorithm differs from previous methods because we are seeking to optimize for Recall-at-N, rather than a ranking loss.

4 Experimental Evaluation

4.1 Training and Test Set

In order to train and evaluate our system, we construct a gold-standard list of phrasal verbs from the freely available English Wiktionary. We gather phrasal verbs from three sources within Wiktionary:

1. Entries labeled as *English phrasal verbs*[4],
2. Entries labeled as *English idioms*[5], and
3. The *derived terms*[6] of English verb entries.

[4] http://en.wiktionary.org/wiki/Category:English_phrasal_verbs

[5] http://en.wiktionary.org/wiki/Category:English_idioms

[6] For example, see http://en.wiktionary.org/wiki/take#Derived_terms

about	*across*	*after*	*against*	*along*
among	*around*	*at*	*before*	*behind*
between	*beyond*	*by*	*down*	*for*
from	*in*	*into*	*like*	*off*
on	*onto*	*outside*	*over*	*past*
round	*through*	*to*	*towards*	*under*
up	*upon*	*with*	*within*	*without*

Table 2: Particles and prepositions allowed in phrasal verbs gathered from Wiktionary.

Many of the idioms and derived terms are not phrasal verbs (e.g. *kick the bucket*, *make-or-break*). We filter out any phrases not of the form VP^+, with V a verb, and P^+ denoting one or more occurrences of particles and prepositions from the list in Table 2. We omit prepositions that do not productively form English phrasal verbs, such as *amid* and *as*. This process also omits some compounds that are sometimes called phrasal verbs, such as light verb constructions, e.g. *have a go* (Butt, 2003), and noncompositional verb-adverb collocations, e.g. *look forward*.

There are a number of extant phrasal verb corpora. For example, McCarthy et al. (2003) present graded human compositionality judgments for 116 phrasal verbs, and Baldwin (2008) presents a large set of candidates produced by an automated system, with false positives manually removed. We use Wiktionary instead, in an attempt to construct a maximally comprehensive data set that is free from any possible biases introduced by automatic extraction processes.

4.2 Filtering and Data Partition

The merged list of phrasal verbs extracted from Wiktionary included some common collocations that have compositional semantics (e.g. *know about*), as well as some very rare constructions (e.g. *cheese down*). We removed these spurious results systematically by filtering out very frequent and very infrequent entries. First, we calculated the log probability of each phrase, according to a language model built from a large monolingual corpus of news documents and web documents, smoothed with stupid backoff (Brants et al., 2007). We sorted all Wiktionary phrasal verbs according to this value. Then, we selected the contiguous 75% of the sorted phrases that minimize the variance of this statistic. This method

		Recall-at-1220	
		Dev	Test
Baseline	Frequent Candidates	17.0	19.3
	WordNet 3.0 Frequent	41.6	43.7
	WordNet 3.0 Filtered	49.4	**48.8**
Boosted	Monolingual Only	30.1	30.2
	Bilingual Only	47.1	43.9
	Monolingual+Bilingual	**50.8**	47.9

Table 3: Our boosted ranker combining monolingual and bilingual features (bottom) compared to three baselines (top) gives comparable performance to the human-curated upper bound.

removed a few very frequent phrases and a large number of rare phrases. The remaining phrases were split randomly into a development set of 694 items and a held-out test set of 695 items.

4.3 Corpora

Our monolingual English corpus consists of news articles and documents collected from the web. Our parallel corpora from English to each of 50 languages also consist of documents collected from the web via distributed data mining of parallel documents based on the text content of web pages (Uszkoreit et al., 2010).

The parallel corpora were segmented into aligned sentence pairs and word-aligned using two iterations of IBM Model 1 (Brown et al., 1993) and two iterations of the HMM-based alignment model (Vogel et al., 1996) with posterior symmetrization (Liang et al., 2006). This training recipe is common in large-scale machine translation systems.

4.4 Generating Candidates

To generate the set of candidate phrasal verbs considered during evaluation, we exhaustively enumerated the Cartesian product of all verbs present in the previously described Wiktionary set ($\mathcal{V}$), all particles in Table 2 ($\mathcal{P}$) and a small set of second particles $\mathcal{T} = \{$*with, to, on,* $\epsilon\}$, with ϵ the empty string. The set of candidate phrasal verbs we consider during evaluation is the product $\mathcal{V} \times \mathcal{P} \times \mathcal{T}$, which contains 96,880 items.

4.5 Results

We optimize a ranker using the boosting algorithm described in section 3.3, using the features from Table 1, optimizing performance on the Wiktionary development set described in section 4.2. Monolingual and bilingual statistics are calculated using the corpora described in section 4.3, with candidate phrasal verbs being drawn from the set described in section 4.4.

We evaluate our method of identifying phrasal verbs by computing *recall-at-*N. This statistic is the fraction of the Wiktionary test set that appears in the top N proposed phrasal verbs by the method, where N is an arbitrary number of top-ranked candidates held constant when comparing different approaches (we use $N = 1220$). We do not compute precision, because the test set to which we compare is not an exhaustive list of phrasal verbs, due to the development/test split, frequency filtering, and omissions in the original lexical resource. Proposing a phrasal verb not in the test set is not necessarily an error, but identifying many phrasal verbs from the test set is an indication of an effective method. Recall-at-N is a natural way to evaluate a ranking system where the gold-standard data is an incomplete, unranked set.

Table 3 compares our approach to three baselines using the Recall-at-1220 metric evaluated on both the development and test sets. As a lower bound, we evaluated the 1220 most frequent candidates in our Monolingual corpus (*Frequent Candidates*).

As a competitive baseline, we evaluated the set of phrasal verbs in WordNet 3.0 (Fellbaum, 1998). We selected the most frequent 1220 out of 1781 verb-particle constructions in WordNet (*WordNet 3.0 Frequent*). A stronger baseline resulted from applying the same filtering procedure to WordNet that we did to Wiktionary: sorting all verb-particle entries by their language model score and retaining the 1220 consecutive entries that minimized language model variance (*WordNet 3.0 Filtered*). WordNet is a human-curated resource, and yet its recall-at-N compared to our Wiktionary test set is only 48.8%, indicating substantial divergence between the two resources. Such divergence is typical: lexical resources often disagree about what multiword expressions to include (Lin, 1999).

The three final lines in Table 3 evaluate our

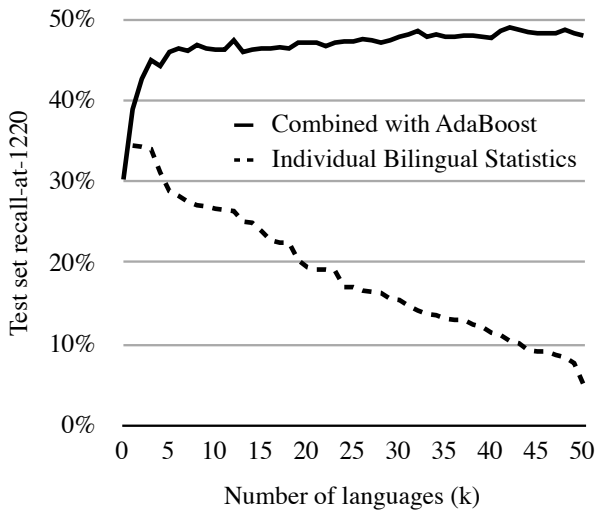


Figure 2: The solid line shows recall-at-1220 when combining the k best-performing bilingual statistics and three monolingual statistics. The dotted line shows the individual performance of the kth best-performing bilingual statistic, when applied in isolation to rank candidates.

boosted ranker. Automatically detecting phrasal verbs using monolingual features alone strongly outperformed the frequency-based lower bound, but underperformed the WordNet baseline. Bilingual features, using features from 50 languages, proved substantially more effective. The combination of both types of features yielded the best performance, outperforming the human-curated WordNet baseline on the development set (on which our ranker was optimized) and approaching its performance on the held-out test set.

4.6 Feature Analysis

The solid line in Figure 2 shows the recall-at-1220 for a boosted ranker using all monolingual statistics and k bilingual statistics, for increasing k. Bilingual statistics are added according to their individual recall, from best-performing to worst. That is, the point at $k = 0$ uses only μ_1, μ_2, and μ_3, the point at $k = 1$ adds the best individually-performing bilingual statistic (Spanish) as a weak ranker, the next point adds the second-best bilingual statistic (German), etc. Boosting maximizes performance on the development set, and evaluation is performed on the test set. We use $T = 53$ (equal to the total number of weak rankers).

	Recall-at-1220	
	Dev	Test
Bilingual only	47.1	43.9
Bilingual+μ_1	48.1	46.9
Bilingual+μ_2	50.1	48.3
Bilingual+μ_3	48.4	46.3
Bilingual+$\mu_1 + \mu_2$	50.2	47.9
Bilingual+$\mu_1 + \mu_3$	49.0	47.4
Bilingual+$\mu_2 + \mu_3$	50.4	**49.4**
Bilingual+$\mu_1 + \mu_2 + \mu_3$	**50.8**	47.9

Table 4: An ablation of monolingual statistics shows that they are useful in addition to the 50 bilingual statistics combined, and no single statistic provides maximal performance.

The dotted line in Figure 2 shows that individual bilingual statistics have recall-at-1220 ranging from 34.4% to 5.0%. This difference reflects the different sizes of parallel corpora and usefulness of different languages in identifying English semantic idiosyncrasy. Combining together the signal of multiple languages is clearly beneficial, and including many low-performing languages still offers overall improvements.

Table 4 shows the effect of adding different subsets of the monolingual statistics to the set of all 50 bilingual statistics. Monolingual statistics give a performance improvement of up to 5.5% recall on the test set, but the comparative behavior of the various combinations of the μ_i is somewhat unpredictable when training on the development set and evaluating on the test set. The pointwise mutual information of a verb and its particles (μ_2) appears to be the most useful feature. In fact, the test set performance of using μ_2 alone outperforms the combination of all three. The best combination even outperforms the WordNet 3.0 baseline on the test set, though optimizing on the development set would not select this model.

4.7 Error Analysis

Table 5 shows the 100 highest ranked phrasal verb candidates by our system that do not appear in either the development or test sets. Most of these candidates are in fact English phrasal verbs that happened to be missing from Wiktionary, some are present in Wiktionary but were removed from the reference

pick up	*pat on*	*tap into*	*fit for*	*charge with*	*suit against*
catch up	*burst into*	*muck up*	*haul up*	*give up*	*get off*
get through	*get up*	*get in*	*tack on*	*buzz about*	*do like*
plump for	*haul in*	*keep up with*	*strap on*	*catch up with*	*suck into*
get round	*chop off*	*slap on*	*pitch into*	*get into*	*inquire into*
drop behind	*get on*	*catch up on*	*pass on*	*cue from*	*carry around*
get around	*get over*	*shoot at*	*pick over*	*shoot by*	*shoot in*
make up to	*get past*	*cast down*	*set up with*	*rule off*	*hand round*
piss on	*hit by*	*break down*	*move for*	*lead off*	*pluck off*
flip through	*edge over*	*strike off*	*plug into*	*keep up*	*go past*
set off	*pull round*	*see about*	*stay on*	*put up*	*sidle up to*
buzz around	*take off*	*set up*	*slap in*	*head towards*	*shoot past*
inquire for	*tuck up*	*lie with*	*well before*	*go on with*	*reel from*
drive along	*snap off*	*barge into*	*whip on*	*put down*	*instance through*
bar from	*cut down on*	*let in*	*tune in to*	*move off*	*suit in*
lean against	*well beyond*	*get down to*	*go across*	*sail into*	*lie over*
hit with	*chow down on*	*look after*	*catch at*		

Table 5: The highest ranked phrasal verb candidates from our full system that do not appear in either Wiktionary set. Candidates are presented in decreasing rank; "pat on" is the second highest ranked candidate.

sets during filtering, and the remainder are in fact not phrasal verbs (true precision errors).

These errors fall largely into two categories. Some candidates are compositional, but contain polysemous verbs, such as *hit by*, *drive along*, and *head towards*. In these cases, prepositions disambiguate the verb, which naturally affects translation distributions. Other candidates are not phrasal verbs, but instead phrases that tend to have a different syntactic role, such as *suit against*, *instance through*, *fit for*, and *lie over* (conjugated as *lay over*). A careful treatment of part-of-speech tags when computing corpus statistics might address this issue.

5 Related Work

The idea of using word-aligned parallel corpora to identify idiomatic expressions has been pursued in a number of different ways. Melamed (1997) tests candidate MWEs by collapsing them into single tokens, training a new translation model with these tokens, and using the performance of the new model to judge candidates' noncompositionality. Villada Moirón and Tiedemann (2006) use word-aligned parallel corpora to identify Dutch MWEs, testing the assumption that the distributions of alignments of MWEs will generally have higher entropies than those of fully compositional compounds. Caseli et al. (2010) generate candidate multiword expressions by picking out sufficiently common phrases that align to single target-side tokens. Tsvetkov and Wintner (2012) generate candidate MWEs by finding one-to-one alignments in parallel corpora which are not in a bilingual dictionary, and ranking them based on monolingual statistics. The system of Salehi and Cook (2013) is perhaps the closest to the current work, judging noncompositionality using string edit distance between a candidate phrase's automatic translation and its components' individual translations. Unlike the current work, their method does not use distributions over translations or combine individual bilingual values with boosting; however, they find, as we do, that incorporating many languages is beneficial to MWE identification.

A large body of work has investigated the identification of noncompositional compounds from monolingual sources (Lin, 1999; Schone and Jurafsky, 2001; Fazly and Stevenson, 2006; McCarthy et al., 2003; Baldwin et al., 2003; Villavicencio, 2003). Many of these monolingual statistics could be viewed as weak rankers and fruitfully incorporated into our framework.

There has also been a substantial amount of work addressing the problem of differentiating between literal and idiomatic instances of phrases in context (Katz and Giesbrecht, 2006; Li et al., 2010;

Sporleder and Li, 2009; Birke and Sarkar, 2006; Diab and Bhutada, 2009). We do not attempt this task; however, techniques for token identification could be used to improve type identification (Baldwin, 2005).

6 Conclusion

We have presented the polyglot ranking approach to phrasal verb identification, using parallel corpora from many languages to identify phrasal verbs. We proposed an evaluation metric that acknowledges the inherent incompleteness of reference sets, but distinguishes among competing systems in a manner aligned to the goals of the task. We developed a recall-oriented learning method that integrates multiple weak ranking signals, and demonstrated experimentally that combining statistical evidence from a large number of bilingual corpora, as well as from monolingual corpora, produces the most effective system overall. We look forward to generalizing our approach to other types of noncompositional phrases.

Acknowledgments

Special thanks to Ivan Sag, who argued for the importance of handling multi-word expressions in natural language processing applications, and who taught the authors about natural language syntax once upon a time. We would also like to thank the anonymous reviewers for their helpful suggestions.

References

Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Natural Language Learning*.

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*.

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech & Language, Special Issue on Multiword Expressions*.

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of english verb-particle constructions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions*.

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*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of European Chapter of the Association for Computational Linguistics*.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*.

Miriam Butt. 2003. The light verb jungle. In *Proceedings of the Workshop on Multi-Verb Constructions*.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1).

Kenneth Church. 2011. How many multiword expressions do people know? In *Proceedings of the ACL Workshop on Multiword Expressions*.

Mona T. Diab and Pravin Bhutada. 2009. Verb noun construction MWE token supervised classification. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Robert Dixon. 1982. The grammar of english phrasal verbs. *Australian Journal of Linguistics*.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Yoav Freund and Robert E. Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Conference on Computational Learning Theory*.

Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*.

Ray Jackendoff. 1993. *The Architecture of the Language Faculty*. MIT Press.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the Association for Computational Linguistics*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the Association for Computational Linguistics*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL Workshop on Multiword Expressions*.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. 2010. Handling named entities and compound verbs in phrase-based statistical machine translation. In *Proceedings of the COLING 2010 Workshop on Multiword Expressions*.

Zhixiang Ren, Yajuan Lu, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the CICLING Conference on Intelligent Text Processing and Computational Linguistics*.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*.

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

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. In *Natural Language Engineering*.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the Conference on Computational Linguistics*.

Sriram Venkatapathy and Aravind K Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL Workshop on Multiword Expressions in a Multilingual Context*.

Aline Villavicencio. 2003. Verb-particle constructions and lexical resources. In *Proceedings of the ACL workshop on Multiword expressions*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the Conference on Computational linguistics*.

Jun Xu and Hang Li. 2007. AdaRank: a boosting algorithm for information retrieval. In *Proceedings of the SIGIR Conference on Research and Development in Information Retrieval*.

Ying Xu, Randy Goebel, Christoph Ringlstetter, and Grzegorz Kondrak. 2010. Application of the tightness continuum measure to chinese information retrieval. In *Proceedings of the COLING Workshop on Multiword Expressions*.