Identifying Phrasal Verbs Using Many Bilingual Corpora

THE UNIVERSITY OF TEXAS AT AUSTIN

Karl Pichotta
University of Texas at Austin

John DeNero
Google

Phrasal Verbs

A **phrasal verb** is a verb followed by one or more particles whose meaning cannot be determined by composing its component words [1].

Phrasal verbs	**Not Phrasal Verbs**
• throw up	• walk towards
• look after	• sit behind
• come up with	• paint on

All of the above are **syntactically similar**. The verb phrases on the right are **compositional**—they mean the sum of their parts; the phrasal verbs on the left are **idiomatic**—their meaning cannot be determined from their components.

Overview

A **bilingual corpus** is a collection of human-translated documents between two languages, whose words have been **automatically aligned** with their translations [2].

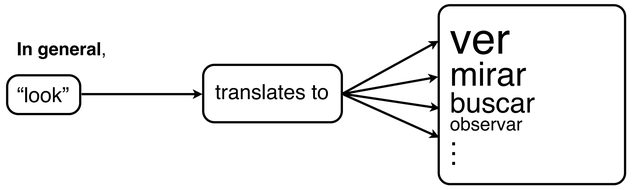
We attempt to **rank** potential phrasal verbs by idiomaticity, using **properties of translations**, measured in many bilingual corpora.

We combine information from bilingual corpora for **many languages**, using a boosting algorithm oriented to the task and evaluation.

We are able to achieve **performance comparable to a human-curated corpus** of phrasal verbs.

Polyglot Ranking

Verbs **translate differently** into foreign languages when heading phrasal verbs.



Head verbs of phrasal verbs **translate differently,** consistent **across many languages**.

Given a bilingual corpus, calculate:

(1) A probability distribution over a verb's translations;

(2) A probability distribution over the verb's translations when heading a candidate phrasal verb.

The **KL Divergence** between these distributions will be higher if the two distributions are very different.

We **combine** the KL divergences for **many different languages** using a modified version of AdaBoost [3] allowing us to rank candidates.

Monolingual statistics are also easily incorporated into the boosting algorithm.

Aligned Phrase Pair	$N(e,f)$	$\pi_1(e,f)$
looking forward to / deseando	1	*deseando*
looking forward to / mirando adelante a	3	*mirando*
looking / mirando	5	*mirando*
looking / buscando a	3	*buscando*

	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

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

Results

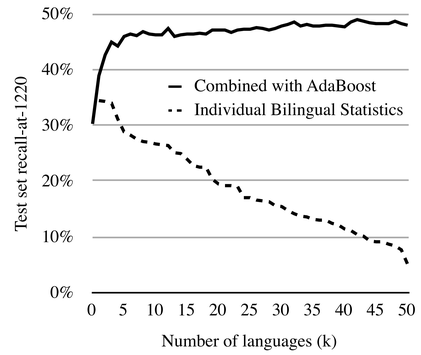
We measure system performance using **Recall at *k***: what percentage of an incomplete gold-standard reference set is in the top *k* guesses?

Our reference set is scraped from English **Wiktionary**.

		Recall at 1220
Baseline	Frequent Candidates	19.3
	WordNet 3.0, Filtered	48.8
Boosted	Monolingual Features	30.2
	Bilingual KL Divergences	43.9
	Monolingual + Bilingual	47.9

- WordNet's human curated collection of phrasal verbs performs only marginally better on this task.
- Bilingual features perform well on their own, and better with monolingual features.

Performance continues to improve when adding individually poorly-performing languages:



[1] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. CICLing 02.

[2] Philipp Koehn. 2010. Statistical Machine Translation. Cambridge University Press, New York, NY, USA.

[3] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1.