Predicting the Semantic Compositionality of Prefix Verbs

Shane Bergsma, Aditya Bhargava, Hua He, Grzegorz Kondrak
Department of Computing Science
University of Alberta
{bergsma,abhargava,hhe,kondrak}@cs.ualberta.ca

Abstract

In many applications, replacing a complex word form by its stem can reduce sparsity, revealing connections in the data that would not otherwise be apparent. In this paper, we focus on prefix verbs: verbs formed by adding a prefix to an existing verb stem. A prefix verb is considered compositional if it can be decomposed into a semantically equivalent expression involving its stem. We develop a classifier to predict compositionality via a range of lexical and distributional features, including novel features derived from web-scale N-gram data. Results on a new annotated corpus show that prefix verb compositionality can be predicted with high accuracy. Our system also performs well when trained and tested on conventional morphological segmentations of prefix verbs.

1 Introduction

Many verbs are formed by adding prefixes to existing verbs. For example, *remarry* is composed of a prefix, *re*-, and a stem, *marry*. We present an approach to predicting the compositionality of prefix verbs. The verb *remarry* is compositional; it means to *marry again*. On the other hand, *retire* is generally non-compositional; it rarely means *to tire again*. There is a continuum of compositionality in prefix verbs, as in other complex word forms and multiword expressions (Bannard et al., 2003; Creutz and Lagus, 2005; Fazly et al., 2009; Xu et al., 2009).

We adopt a definition of compositionality specifically designed to support downstream applications that might benefit from knowledge of verb stems. For example, suppose our corpus contains the following sentence: “Pope Clement VII denied Henry VIII permission to marry again before a decision was given in Rome.” A user might submit the question, “Which pope refused Henry VIII permission to remarry?” If we can determine that the meaning of *remarry* could also be provided via the stem *marry*, we could add *marry* to our search terms. This is known as *morphological query expansion* (Bilotti et al., 2004). Here, such an expansion leads to a better match between question and answer.

Previous work has shown that “full morphological analysis provides at most very modest benefits for retrieval” (Manning et al., 2008). Stemming, lemmatization, and compound-splitting often increase recall at the expense of precision, but the results depend on the morphological complexity of the text’s language (Hollink et al., 2004).

The lack of success in applying morphological analysis in IR is unsurprising given that most previous systems are not designed with applications in mind. For example, the objective of the influential *Linguistica* program is “to produce an output that matches as closely as possible the analysis that would be given by a human morphologist” (Goldsmith, 2001). Unsupervised systems achieve this aim by exploiting learning biases such as minimum description length for lexicons (Goldsmith, 2001; Creutz and Lagus, 2007) and high entropy across morpheme boundaries (Keshava and Pitler, 2006). Supervised approaches learn directly from words annotated by morphologists (Van den Bosch and Daelemans, 1999; Toutanova and Cherry, 2009), often using CELEX, a lexical database that includes morphological information (Baayen et al., 1996).

The conventional approach in morphology is to segment words into separate morphemes even when the words are not entirely compositional combinations of their parts (Creutz and Lagus, 2005). For example, while *co-* is considered a separate morpheme in the verb *cooperate*, the meaning of *cooperate* is not simply *to operate jointly*. These forms are sometimes viewed as *perturbations* of composition (de Marken, 1996). In practice, a user may query, "Which nations do not cooperate with the International Criminal Court?" An expansion of the query to include *operate* may have undesirable consequences.

Rather than relying on conventional standards, we present an algorithm whose objective is to find only those prefix verbs that exhibit semantic compositionality; i.e., prefix verbs that are fully meaning-preserving, sums-of-their-parts. We produce a new corpus, annotated according to this definition. We use these annotated examples to learn a discriminative model of semantic compositionality.

Our classifier relies on a variety of features that exploit the distributional patterns of verbs and stems. We build on previous work that applies semantics to morphology (Yarowsky and Wicentowski, 2000; Schone and Jurafsky, 2001; Baroni et al., 2002), and also on work that exploits web-scale data for semantic analysis (Turney, 2001; Nakov, 2007; Kummerfeld and Curran, 2008). For example, we measure how often a prefix verb appears with a hyphen between the prefix and stem. We also look at the distribution of the stem as a separate word: we calculate the probability of the prefix verb and the separated stem's co-occurrence in a segment of discourse; we also calculate the distributional similarity between the verb and the separated stem. High scores for these measures indicate compositionality. We extract counts from a web-scale N-gram corpus, allowing us to efficiently leverage huge volumes of unlabeled text.

Our system achieves 93.6% accuracy on held-out data, well above several baselines and comparison systems. We also train and test our system on conventional morphological segmentations. Our classifier remains reliable in this setting, making half as many errors as the state-of-the-art unsupervised Morfessor system (Creutz and Lagus, 2007).

2 Problem Definition and Setting

A prefix verb is a derived word with a bound morpheme as prefix. While derivation can change both the meaning and part-of-speech of a word (as opposed to inflection, which does not change "referential or cognitive meaning" (Katamba, 1993)), here the derived form remains a verb.

We define prefix-verb compositionality as a semantic equivalence between a verb and a paraphrase involving the verb's stem. The stem must be used as a verb in the paraphrase. Words can be introduced, if needed, to account for the meaning contributed by the prefix, e.g., *outbuild⇒build more/better/faster than*. A bidirectional entailment between the prefix verb and the paraphrase is required.

Words can have different meanings in different contexts. For example, a nation might "*resort* to force," (non-compositional) while a computer program can "*resort* a linked list" (compositional). We therefore define prefix-verb compositionality as a context-specific property of verb tokens rather than a global property of verb types. However, it is worth noting that we ultimately found the compositionality of types to be very consistent across contexts (Section 5.1.2), and we were unable to leverage contextual information to improve classification accuracy; our final system is essentially type-based. Other recent morphological analyzers have also been type-based (Keshava and Pitler, 2006; Poon et al., 2009).

Our system takes as input a verb token in uninflected form along with its sentence as context. The verb must be divisible into an initial string and a following remainder such that the initial string is on our list of prefixes and the remainder is on our list of stems. Hyphenation is allowed, e.g., both *re-enter* and *reenter* are acceptable inputs. The system determines whether the prefix/stem combination is compositional in the current context. For example, the verb *unionize* in, "The workers must unionize," can be divided into a prefix *un-* and a stem *ionize*. The system should determine that here *unionize* is not a compositional combination of these parts.

The algorithm requires a list of prefixes and stems in a given language. For our experiments, we use both dictionary and corpus-based methods to construct these lists (Section 4).

3 Supervised Compositionality Detection

We use a variety of lexical and statistical information when deciding whether a prefix verb is compositional. We adopt a discriminative approach. We assume some labeled examples are available to train a classifier. Relevant information is encoded in a feature vector, and a learning algorithm determines a set of weights for the features using the training data. As compositionality is a binary decision, we can adopt any standard package for binary classification. In our experiments we use support vector machines.

Our features include both local information that depends only on the verb string (sometimes referred to as lexical features) and also global information that depends on the verb and the stem's distribution in text. Our approach can therefore be regarded as a simple form of semi-supervised learning; we leverage both a small number of labeled examples and a large volume of unlabeled text.

If a frequency or similarity is undefined in our corpus, we indicate this with a separate feature; weights on these features act as a kind of smoothing.

3.1 Features based on Web-Scale N-gram Data

We use web-scale N-gram data to extract distributional features. The most widely-used N-gram corpus is the Google 5-gram Corpus (Brants and Franz, 2006). We use *Google V2*: a new N-gram corpus (also with N-grams of length one-to-five) created from the same one-trillion-word snapshot of the web as the Google 5-gram Corpus, but with enhanced filtering and processing of the source text (Lin et al., 2010). For Google V2, the source text was also part-of-speech tagged, and the resulting part-of-speech tag distribution is included for each N-gram. There are 4.1 billion N-grams in the corpus.

The part-of-speech tag distributions are particularly useful, as they allow us to collect verb-specific counts. For example, while a string like *reuse* occurs 1.1 million times in the web corpus, it is only tagged as a verb 270 thousand times. Conflating the noun/verb senses can lead to misleading scores for certain features. E.g., the hyphenation frequency of *re-use* would appear relatively low, even though *reuse* is semantically compositional.

Lin et al. (2010) also provide a high-coverage, 10-million-phrase set of clusters extracted from the N-grams; we use these for our similarity features (Section 3.1.3). There are 1000 clusters in total. The data does not provide the context vectors for each phrase; rather, each phrase is listed with its 20 most similar clusters, measured by cosine similarity with the cluster centroid. We use these centroid similarities as values in a 1000-dimensional cluster-membership feature space. To calculate the similarity between two verbs, we calculate the cosine similarity between their cluster-membership vectors.

The feature classes in the following four subsections each make use of web-scale N-gram data.

3.1.1 HYPH features

Hyphenated verbs are usually compositional (e.g., *re-elect*). Of course, a particular instance of a compositional verb may or may not occur in hyphenated form. However, across a large corpus, compositional prefix verbs tend to occur in a hyphenated form more often than do non-compositional prefix verbs. We therefore provide real-valued features for how often the verb was hyphenated and unhyphenated on the web. For example, we collect counts for the frequencies of *re-elect* (33K) and *reelect* (9K) in our web corpus, and we convert the frequencies to log-counts. We also give real-valued features for the hyphenated/unhyphenated log-counts using only those occurrences of the verb that were *tagged* as a verb, exploiting the tag distributions in our web corpus as described above.

Nakov and Hearst (2005) previously used hyphenation counts as an indication of a syntactic relationship between nouns. In contrast, we leverage hyphenation counts as an indication of a semantic property of verbs.

3.1.2 COOC features

COOC features, and also the SIM (Section 3.1.3) and YAH (Section 3.2.2) features, concern the association in text between the prefix verb and its stem, where the stem occurs as a separate word. We call this the separated stem.

If a prefix verb is compositional, it is more likely to occur near its separated stem in text. We often see *agree* and *disagree*, *read* and *reread*, etc. occurring in the same segment of discourse. We create features for the association of the prefix verb and its separated stem in a discourse. We include the log-count of how often the verb and stem occur in the same N-gram (of length 2-to-5) in our N-gram corpus. Note that the 2-to-4-gram counts are not strictly a subset of the 5-gram counts, since fewer 5-grams pass the data's minimum frequency threshold.

We also include a real-valued pointwise mutual information (PMI) feature for the verb and separated stem's co-occurrence in an N-gram. For the PMI, we regard occurrence in an N-gram as an event, and calculate the probability that a verb and separated stem jointly occur in an N-gram, divided by the probability of their occurring in an N-gram independently.

3.1.3 SIM features

If a prefix verb is compositional, it should occur in similar contexts to its stem. The idea that a stem and stem+affix should be semantically similar has been exploited previously for morphological analysis (Schone and Jurafsky, 2000). We include a real-valued feature for the distributional similarity of the verb and stem using Lin's thesaurus (Lin, 1998). The coverage of this measure was low: it was non-zero for only 93 of the 1000 prefix verbs in our training set. We therefore also include distributional similarity calculated using the web-scale 10-million-phrase clustering as described above. Using this data, similarity is defined for 615 of the 1000 training verbs. We also explored a variety of WordNet-based similarity measures, but these ultimately did not prove helpful on development data.

3.1.4 FRQ features

We include real-valued features for the raw frequencies of the verb and the stem on the web. If these frequencies are widely different, it may indicate a non-compositional usage. Yarowsky and Wicentowski (2000) use similar statistics to identify words related by inflection, but they gather their counts from a much smaller corpus. In addition, higher-frequency prefix verbs may be *a priori* more likely to be non-compositional. A certain frequency is required for an irregular usage to become familiar to language speakers. The potential correlation between frequency and non-compositionality could thus also be exploited by the classifier via the FRQ features.

3.2 Other Features

3.2.1 LEX features

We provide lexical features for various aspects of a prefix verb. Binary features indicate the occurrence of particular verbs, prefixes, and stems, and whether the prefix verb is hyphenated. While hyphenated prefix verbs are usually compositional, even non-compositional prefix verbs may be hyphenated if the prefix and stem terminate and begin with a vowel, respectively. For example, non-compositional uses of *co-operate* are often hyphenated, whereas the compositional *remarry* is rarely hyphenated. We therefore have indicator features for the conjunction of the prefix and the first letter of the stem (e.g., *co-o*), and also for the prefix conjoined with a flag indicating whether the stem begins with a vowel (e.g., *co+vowel*).

3.2.2 YAH features

While the COOC features capture many cases where the verb and separated stem occur in close proximity (especially, but not limited to, conjunctions), there are many other cases where a longer distance might separate a compositional verb and its separated stem. For example, consider the sentence, "Brush the varnish on, but do not overbrush." Here, the verb and separated stem do not co-occur within a 5-gram window, and their co-occurrence will therefore not be recorded in our N-gram corpus. As an approximation for co-occurrence counts within a longer segment of discourse, we count the number of *pages* on the web where the verb and separated stem co-occur. We use hit-counts returned by the Yahoo search engine API.[1] Similar to our COOC features, we include a real-valued feature for the pointwise mutual information of the prefix verb and separated stem's co-occurrence on a web page, i.e., we use Turney's PMI-IR (Turney, 2001).

Baroni et al. (2002) use similar statistics to help discover morphologically-related words. In contrast to our features, however, their counts are derived from source text that is several orders of magnitude smaller in size.

[1] `http://developer.yahoo.com/search/boss/`

3.2.3 DIC features

One potentially useful resource, when available, is a dictionary of the conventional morphological segmentations of words in the language. Although these segmentations have been created for a different objective than that of our annotations, we hypothesize that knowledge of morphology can help inform our system's predictions. For each prefix verb, we include features for whether or not the prefix and stem are conventionally segmented into separate morphemes, according to a morphological dictionary. Similar to the count-based features, we include a DIC-undefined feature for the verbs that are not in the dictionary; any precompiled dictionary will have imperfect coverage of actual test examples.

Interestingly, DIC features are found to be among our least useful features in the final evaluation.

4 Experiments

4.1 Resources

We use CELEX (Baayen et al., 1996) as our dictionary for the DIC features. We also use CELEX to help extract our lists of prefixes and stems. We take every prefix that is marked in CELEX as forming a new verb by attaching to an existing verb. For stems, we use every verb that occurs in CELEX, but we also extend this list by automatically collecting a large number of words that were automatically tagged as verbs in the NYT section of Gigaword (Graff, 2003). To be included in the extra-verb list, a verb must occur more than ten times and be tagged as a verb more than 70% of the time by a part-of-speech tagger. We thereby obtain 43 prefixes and 6613 stems.[2] We aimed for an automatic, high-precision list for our initial experiments. This procedure is also amenable to human intervention; one could alternatively cast a wider net for possible stems and then manually filter false positives.

4.2 Annotated Data

We carried out a medium-scale annotation to provide training and evaluation data for our experiments.[3] The data for our annotations also comes from the NYT section of Gigaword. We first build a list of possible prefix verbs. We include any verb that a) is composed of a valid prefix and stem; and b) occurs at least twice in the corpus.[4] If the verb occurs less than 50 times in the corpus, we also require that it was tagged as a verb in at least 70% of cases. This results in 2077 possible prefix verbs for annotation.

For each verb type in our list of possible prefix verbs, we randomly select for annotation sentences from Gigaword containing the verb. We take at most three sentences for each verb type so that a few very common types (such as *become*, *understand*, and *improve*) do not comprise the majority of annotated examples. The resulting set of sentences includes a small number of sentences with incorrectly-tagged non-verbs; these are simply marked as non-verbs by our annotators and excluded from our final data sets. A graphical program was created for the annotation; the program automatically links to the online Merriam-Webster dictionary entries for the prefix verb and separated stem. When in doubt about a verb's meaning, our annotators adhere to the dictionary definitions. A single annotator labeled 1718 examples, indicating for each sentence whether the prefix verb was compositional. A second annotator then labeled a random subset of 150 of these examples, and agreement was calculated. The annotators agreed on 137 of the 150 examples. The *Kappa* statistic (Jurafsky and Martin, 2000, page 315), with P(E) computed from the confusion matrices, is 0.82, above the 0.80 level considered to indicate good reliability.

For our experiments, the 1718 annotated examples are randomly divided into 1000 training, 359 development, and 359 held-out test examples.

4.3 Classifier Settings

We train a linear support vector machine classifier using the efficient LIBLINEAR package (Fan et al., 2008). We use L2-loss and L2-regularization. We

[2] The 43 prefixes are: a- ab- ac- ad- as- be- circum- co- col- com- con- cor- counter- cross- de- dis- e- em- en- ex- fore- im- in- inter- ir- mis- out- over- per- photo- post- pre- pro- psycho- re- sub- super- sur- tele- trans- un- under- with-

[3] Our annotated data is publicly available at: `http://www.cs.ualberta.ca/~ab31/verbcomp/`

[4] We found that the majority of single-occurrence verbs in the Gigaword data were typos. We would expect true hapax legomena to be largely compositional, and we could potentially derive better statistics if we include them (Baayen and Sproat, 1996). One possible option, employed in previous work, is to ensure words of interest are "manually corrected for typing errors before further analysis" (Baayen and Renouf, 1996).

optimize the choice of features and regularization hyperparameter on development data, attaining a maximum when $C = 0.1$.

4.4 Evaluation

We compare the following systems:

1. **Base1**: always choose compositional (the majority class).
2. **Base2**: for each prefix, choose the majority class over the verbs having that prefix in training data.
3. **Morf**: the unsupervised Morfessor system (Creutz and Lagus, 2007) (Categories-ML, from 110K-word corpus). If Morfessor splits the prefix and stem into separate morphemes, we take the prediction as compositional. If it does anything else, we take it as non-compositional.
4. **SCD**: **S**upervised **C**ompositionality **D**etection: the system proposed in this paper.

We evaluate using *accuracy*: the percentage of examples classified correctly in held-out test data.

5 Results

We first analyze our annotations, gaining insight into the relation between our definition and conventional segmentations. We also note the consistency of our annotations across contexts. We then provide the main results of our system. Finally, we provide the results of our system when trained and tested on conventional morphological segmentations.

5.1 Analysis of Annotations

5.1.1 Annotation consistency with dictionaries

The majority of our examples are not present in a morphological dictionary, even in one as comprehensive as CELEX. The prefix verbs are in CELEX for only 670 of the 1718 total annotated instances.

For those that are in CELEX, Table 1 provides the confusion matrix that relates the CELEX segmentations to our annotations. The table shows that the major difference between our annotations and CELEX is that our definition of compositionality is more strict than conventional morphological segmentations. When CELEX does not segment the prefix from the stem (case 0), our annotations agree in 183 of 193 cases. When CELEX does split the prefix from the stem (case 1), the meaning is semantically compositional in less than half the cases. This is a key difference between conventional morphology and our semantic definition.

		CELEX segmentation	
		1	0
Compositionality	1	227	10
annotation	0	250	183

Table 1: Confusion matrix on the subset of prefix verb annotations that are also in CELEX. **1** indicates that the prefix and stem are segmented into separate morphemes, **0** indicates otherwise.

It is also instructive to analyze the 10 cases that are semantically compositional but which CELEX did not segment. Most of these are verbs that are conventionally viewed as single morphemes because they entered English as complete words. For example, *await* comes from the Old North French *awaitier*, itself from *waitier*. In practice, it is useful to know that *await* is compositional, i.e. that it can be rephrased as *wait for*. Downstream applications can exploit the compositionality of *await*, but miss the opportunity if using the conventional lack of segmentation.

5.1.2 Annotation consistency across contexts

We next analyze our annotated data to determine the consistency of compositionality across different occurrences of the same prefix-verb type. There are 1248 unique prefix verbs in our 1718 labeled examples: 45 verbs occur three times, 380 occur twice and 823 occur only once. Of the 425 verbs that occur multiple times, only 6 had different annotations in different examples (i.e., six verbs occur in both compositional and non-compositional usages in our dataset). These six instances are subtle, debatable, and largely uninteresting, depending on distinctions like whether the *proclaim* sense of *blazon* can substitute for the *celebrate* sense of *emblazon*, etc.

It is easy to find clearer ambiguities online, such as compositional examples of typically non-compositional verbs (how to *recover* a couch, when to *redress* a wound, etc.). However, in our data verbs like *recover* and *redress* always occur in their more dominant non-compositional sense. People may

Set	#	Base1	Base2	Morf	SCD
Test	359	65.7	87.2	73.8	93.6
$\in$ CELEX	128	30.5	73.4	50.8	89.8
$\notin$ CELEX	231	85.3	94.8	86.6	95.7
$\in$ train	107	69.2	93.5	74.8	97.2
$\notin$ train	252	64.3	84.5	73.4	92.1

Table 2: Number of examples (#) and accuracy (%) on test data, and on in-CELEX vs. not-in-CELEX, and in-training-data vs. not-in-training splits.

consciously or unconsciously recognize the possibility for confusion and systematically hyphenate prefixes from the stem if a less-common compositional usage is employed. For example, our data has "*repress* your feelings" for the non-compositional case but the hyphenated "*re-press* the center" for the compositional usage.[5]

Due to the consistency of compositionality across contexts, context-based *features* may simply not be very useful for classification. All the features we describe in Section 3 depend only on the prefix verb itself and not the verb context. Various context-dependent features did not improve accuracy on our development data and were thus excluded from the final system.

5.2 Main Results

The first row of Table 2 gives the results of all systems on test data. SCD achieves 93.6% accuracy, making one fifth as many errors as the majority-class baseline (Base1) and half as many errors as the more competitive prefix-based predictor (Base2). The substantial difference between SCD and Base2 shows that SCD is exploiting much information beyond the trivial memorization of a decision for each prefix. Morfessor performs better than Base1 but significantly worse than Base2. This indicates that state-of-the-art *unsupervised* morphological segmentation is not yet practical for semantic preprocessing. Of course, Morfessor was also designed with a different objective; in Section 5.3 we compare Morfessor and SCD on conventional morphological segmentations.

Prefix	# Tot	# Comp	SCD
re-	166	147	95.8
over-	26	25	96.2
out-	23	18	91.3
de-	21	**0**	100.0
pre-	19	16	94.7
un-	17	**1**	94.1
dis-	10	**0**	90.0
under-	9	7	77.8
co-	7	6	100.0
en-	5	2	60.0

Table 3: Total number of examples (# Tot), number of examples that are compositional (# Comp), and accuracy (%) of SCD on test data, by prefix.

We further analyzed the systems by splitting the test data two ways.

First, we separate verbs that occur in our morphological dictionary ($\in$ CELEX) from those that do not ($\notin$ CELEX). Despite using the dictionary segmentation itself as a feature, the performance of SCD is worse on the $\in$ CELEX verbs (89.8%). The comparison systems drop even more dramatically on this subset. The $\in$ CELEX verbs comprise the more frequent, irregular verbs in English. Non-compositionality is the majority class on the examples that are in the dictionary.

On the other hand, one would expect verbs that are *not* in a comprehensive dictionary to be largely *compositional*, and indeed most of the $\notin$ CELEX verbs are compositional. However, there is still much to be gained from applying SCD, which makes a third as many errors as the system which always assigns compositional (95.7% for SCD vs. 85.3% for Base1).

Our second way of splitting the data is to divide our test set into prefix verbs that also occurred in training sentences ($\in$ train) and those that did not ($\notin$ train). Over 70% did not occur in training. SCD scores 97.2% accuracy on those that did. The classifier is thus able to exploit the consistency of annotations across different contexts (Section 5.1.2). The 92.1% accuracy on the $\notin$-train portion also shows the features allow the system to generalize well to new, previously-unseen verbs.

Table 3 gives the results of our system on sets of

[5] Note that many examples like *recover*, *repress* and *redress* are only ambiguous in text, not in speech. Pronunciation reduces ambiguity in the same way that hyphens do in text. Conversely, observe that knowledge of compositionality could potentially help speech synthesis.

-LEX	-HYPH	-COOC	-SIM	-YAH	-FRQ	-DIC
85.0	92.8	92.5	93.6	93.6	93.6	93.6
85.5	93.6	92.8	93.0	93.3	93.9	
86.9	90.5	93.3	93.6	93.6		
84.1	90.3	93.3	93.6			
87.5	90.5	93.0				
85.5	89.4					

Table 4: Accuracy (%) of SCD as different feature classes are removed. Performance with all features is 93.6%.

verbs divided according to their prefix. The table includes those prefixes that occurred at least 5 times in the test set. Note that the prefixes have a long tail: these ten prefixes cover only 303 of the 359 test examples. Accuracy is fairly high across all the different prefixes. Note also that the three prefixes *de*-, *un*-, and *dis*- almost always correspond to non-compositional verbs. Each of these prefixes corresponds to a subtle form of negation, and it is usually difficult to paraphrase the negation using the stem. For example, *to demilitarize* does not mean *to not militarize* (or any other simple re-phrasing using the stem as a verb), and so our annotation marks it as non-compositional. Whether such a strict strategy is ultimately best may depend on the target application.

Feature Analysis

We perform experiments to evaluate which features are most useful for this task. Table 4 gives the accuracy of our system as different feature classes are *removed*. A similar table was previously used for feature analysis in Daumé III and Marcu (2005). Each row corresponds to performance with a group of features; each entry is performance with a particular feature class individually removed the group. We remove the least helpful feature class from each group in succession moving group-to-group down the rows.

We first remove the DIC features. These do not impact performance on test data. The last row gives the performance with only HYPH features (85.5, removing LEX), and only LEX features (89.4, removing HYPH). These are found to be the two most effective features for this task, followed by the COOC statistics. The other features, while marginally helpful on development data, are relatively ineffective on the test set. In all cases, removing LEX features hurts the most. Removing LEX not only removes useful stem, prefix, and hyphen information, but it also impairs the ability of the classifier to use the other features to separate the examples.

Base1	Base2	Morf	SCD
76.0	79.6	72.4	86.4

Table 5: Accuracy (%) on CELEX.

5.3 CELEX Experiments and Results

Finally, we train and test our system on prefix verbs where the segmentation decisions are provided by a morphological dictionary. We are interested in whether the strong results of our system could transfer to conventional morphological segmentations. We extract all verbs in CELEX that are valid verbs for our system (divisible into a prefix and verb stem), and take the CELEX segmentation as the label; i.e., whether the prefix and stem are separated into distinct morphemes. We extract 1006 total verbs.

We take 506 verbs for training, 250 verbs as a development set (to tune our classifier's regularization parameter) and 250 verbs as a final held-out test set. We use the same features and classifier as in our main results, except we remove the DIC features which are now the instance labels.

Table 5 shows the performance of our two baseline systems along with Morfessor and SCD. While the majority-class baseline is much higher, the prefix-based baseline is 7% *lower*, indicating that knowledge of prefixes, and lexical features in general, are less helpful for conventional segmentations. In fact, performance only drops 2% when we remove the LEX features, showing that web-scale information alone can enable solid performance on this task. Surprisingly, Morfessor performs worse here, below both baselines and substantially below the supervised system. We confirmed our Morfessor program was generating the same segmentations as the online demo. We also experimented with Linguistica (Goldsmith, 2001), training on a large corpus, but results were worse than with Morfessor.

Accurate segmentation of prefix verbs is clearly part of the mandate of these systems; prefix verb segmentation is simply a very challenging task. Unlike other, less-ambiguous tasks in morphology, a prefix/stem segmentation is plausible for all of our input verbs, since the putative morphemes are by definition valid morphemes in the language.

Overall, the results confirm and extend previous studies that show semantic information is helpful in morphology (Schone and Jurafsky, 2000; Yarowsky and Wicentowski, 2000). However, we reiterate that optimizing systems according to conventional morphology may not be optimal for downstream applications. Furthermore, accuracy is substantially lower in this setting than in our main results. Targeting conventional segmentations may be both more challenging and less useful than focusing on semantic compositionality.

6 Related Work

There is a large body of work on morphological analysis of English, but most of this work does not handle prefixes. Porter's stemmer is a well-known *suffix*-stripping algorithm (Porter, 1980), while publicly-available lemmatizers like *morpha* (Minnen et al., 2001) and PC-KIMMO (Karp et al., 1992) only process inflectional morphology. FreeLing (Atserias et al., 2006) comes with a few simple rules for deterministically stripping prefixes in some languages, but not English (e.g., only *semi*- and *re*- can be stripped when analyzing OOV Spanish verbs).

A number of modern morphological analyzers use supervised machine learning. These systems could all potentially benefit from the novel distributional features used in our model. Van den Bosch and Daelemans (1999) use memory-based learning to analyze Dutch. Wicentowski (2004)'s supervised WordFrame model includes a prefixation component. Results are presented on over 30 languages. Erjavec and Džeroski (2004) present a supervised lemmatizer for Slovene. Dreyer et al. (2008) perform supervised lemmatization on Basque, English, Irish and Tagalog; like us they include results when the set of lemmas is given. Toutanova and Cherry (2009) present a discriminative lemmatizer for English, Bulgarian, Czech and Slovene, but only handle suffix morphology. Poon et al. (2009) present an unsupervised segmenter, but one that is based on a log-linear model that can include arbitrary and interdependent features of the type proposed in our work. We see potential in combining the best elements of both approaches to obtain a system that does not need annotated training data, but can make use of powerful web-scale features.

Our approach follows previous systems for morphological analysis that leverage semantic as well as orthographic information (Yarowsky and Wicentowski, 2000; Schone and Jurafsky, 2001; Baroni et al., 2002). Similar problems also arise in core semantics, such as how to detect the compositionality of multi-word expressions (Lin, 1999; Baldwin et al., 2003; Fazly et al., 2009). Our problem is similar to the analysis of verb-particle constructions or VPCs (e.g., *round up, sell off*, etc.) (Bannard et al., 2003). Web-scale data can be used for a variety of problems in semantics (Lin et al., 2010), including classifying VPCs (Kummerfeld and Curran, 2008).

We motivated our work by describing applications in information retrieval, and here Google is clearly the elephant in the room. It is widely reported that Google has been using stemming since 2003; for example, a search today for *Porter stemming* returns pages describing the *Porter stemmer*, and the returned snippets have words like **stemming**, **stemmer**, and **stem** in bold text. Google can of course develop high-quality lists of morphological variants by paying attention to how users reformulate their queries. User query sessions have previously been used to expand queries using similar terms, such as substituting *feline* for *cat* (Jones et al., 2006). We show that high-quality, IR-friendly stemming is possible even without query data. Furthermore, query data could be combined with our other features for highly discriminative word stemming in context.

Beyond information retrieval, suffix-based stemming and lemmatization have been used in a range of NLP applications, including text categorization, textual entailment, and statistical machine translation. We believe accurate prefix-stripping can also have an impact in these areas.

7 Conclusions and Future Work

We presented a system for predicting the semantic compositionality of prefix verbs. We proposed a new, well-defined and practical definition of compositionality, and we annotated a corpus of sentences according to this definition. We trained a discriminative model to predict compositionality using a range of lexical and web-scale statistical features. Novel features include measures of the frequency of prefix-stem hyphenation, and statistics for the likelihood of the verb and stem co-occurring as separate words in an N-gram. The classifier is highly accurate across a range of prefixes, correctly predicting compositionality for 93.6% of examples.

Our preliminary results provide strong motivation for investigating and applying new distributional features in the prediction of both conventional morphology and in task-directed semantic compositionality. Our techniques could be used on a variety of other complex word forms. In particular, many of our features extend naturally to identifying stem-stem compounds (like *panfry* or *healthcare*). Also, it would be possible for our system to handle inflected forms by first converting them to their lemmas using a morphological analyzer. We could also jointly learn the compositionality of words across their inflections, along the lines of Yarowsky and Wicentowski (2000).

There are also other N-gram-derived features that warrant further investigation. One source of information that has not previously been exploited is the "lexical fixedness" (Fazly et al., 2009) of non-compositional prefix verbs. If prefix verbs are rarely rephrased in another form, they are likely to be non-compositional. For example, in our N-gram data, the count of *quest again* is relatively low compared to the count of *request*, indicating *request* is non-compositional. On the other hand, *marry again* is relatively frequent, indicating that *remarry* is compositional. Incorporation of these and other N-gram counts could further improve classification accuracy.

References

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In *LREC*.

R. Harald Baayen and Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, 72(1).

Harald Baayen and Richard Sproat. 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. *Comput. Linguist.*, 22(2):155–166.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The CELEX2 lexical database. LDC96L14.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *ACL 2003 Workshop on Multiword Expressions*.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *ACL 2003 Workshop on Multiword Expressions*.

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *ACL-02 Workshop on Morphological and Phonological Learning (SIGPHON)*, pages 48–57.

Matthew W. Bilotti, Boris Katz, and Jimmy Lin. 2004. What works better for question answering: Stemming or morphological query expansion? In *Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*.

Thorsten Brants and Alex Franz. 2006. The Google Web 1T 5-gram Corpus Version 1.1. LDC2006T13.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):1–34.

Hal Daumé III and Daniel Marcu. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT-EMNLP*.

Carl de Marken. 1996. Linguistic structure as composition and perturbation. In *ACL*.

Markus Dreyer, Jason Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *EMNLP*.

Tomaž Erjavec and Sašo Džeroski. 2004. Machine learning of morphosyntactic structure: Lemmatising unknown Slovene words. *Applied Artificial Intelligence*, 18:17–41.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *JMLR*, 9:1871–1874.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Comput. Linguist.*, 35(1):61–103.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Comput. Linguist.*, 27(2):153–198.

David Graff. 2003. English Gigaword. LDC2003T05.

Vera Hollink, Jaap Kamps, Christof Monz, and Maarten de Rijke. 2004. Monolingual document retrieval for European languages. *IR*, 7(1):33–52.

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *WWW*.

Daniel Jurafsky and James H. Martin. 2000. *Speech and language processing*. Prentice Hall.

Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. 1992. A freely available wide coverage morphological analyzer for English. In *COLING*.

Francis Katamba. 1993. *Morphology*. MacMillan Press.

Samarth Keshava and Emily Pitler. 2006. A simpler, intuitive approach to morpheme induction. In *2nd Pascal Challenges Workshop*.

Jonathan K. Kummerfeld and James R. Curran. 2008. Classification of verb particle constructions with the Google Web1T Corpus. In *Australasian Language Technology Association Workshop*.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale N-grams. In *LREC*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *ACL*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Nat. Lang. Eng.*, 7(3):207–223.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *CoNLL*.

Preslav Ivanov Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, University of California, Berkeley.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *HLT-NAACL*.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3).

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *LLL/CoNLL*.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *NAACL*.

Kristina Toutanova and Colin Cherry. 2009. A global model for joint lemmatization and part-of-speech prediction. In *ACL-IJCNLP*.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning*.

Antal Van den Bosch and Walter Daelemans. 1999. Memory-based morphological analysis. In *ACL*.

Richard Wicentowski. 2004. Multilingual noise-robust supervised morphological analysis using the word-frame model. In *ACL SIGPHON*.

Ying Xu, Christoph Ringlstetter, and Randy Goebel. 2009. A continuum-based approach for tightness analysis of Chinese semantic units. In *PACLIC*.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *ACL*.