Minimally Supervised Morphological Analysis by Multimodal Alignment

David Yarowsky and **Richard Wicentowski**
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
Email:{yarowsky,richardw}@cs.jhu.edu

Abstract

This paper presents a corpus-based algorithm capable of inducing inflectional morphological analyses of both regular and highly irregular forms (such as brought→bring) from distributional patterns in large monolingual text with no direct supervision. The algorithm combines four original alignment models based on relative corpus frequency, contextual similarity, weighted string similarity and incrementally retrained inflectional transduction probabilities. Starting with no paired <inflection,root> examples for training and no prior seeding of legal morphological transformations, accuracy of the induced analyses of 3888 past-tense test cases in English exceeds 99.2% for the set, with currently over 80% accuracy on the most highly irregular forms and 99.7% accuracy on forms exhibiting non-concatenative suffixation.

1 Task Definition

This paper presents an original and successful algorithm for the nearly unsupervised induction of inflectional morphological analyzers, with a focus on highly irregular forms not typically handled by other morphology induction algorithms. It is useful to consider this task as three separate steps:

1) Estimate a probabilistic alignment between inflected forms and root forms in a given language

2) Train a supervised morphological analysis learner on a weighted subset of these aligned pairs.

3) Use the result of Step 2 as either a stand-alone analyzer or a probabilistic scoring component to iteratively refine the alignment in Step 1.

The target output of Step 1 is an inflection-root mapping such as shown in Table 1, with optional columns giving the hypothesized stem change and suffix analysis as well as part of speech.

ROOT	STEM CHANGE	SUFFIX	INFLECTION	POS
take	ake → ook	+ϵ	took	VBD
take	e → ϵ	+ing	taking	VBG
take	ϵ → ϵ	+s	takes	VBZ
take	e → ϵ	+en	taken	VBN
skip	ϵ → p	+ed	skipped	VBD
defy	y → i	+ed	defied	VBD
defy	y → ie	+s	defies	VBZ
defy	ϵ → ϵ	+ing	defying	VBG
jugar	gar → eg	+a	juega	VPI3S
jugar	gar → eg	+an	juegan	VPI3P
jugar	ar → ϵ	+amos	jugamos	VPI1P
tener	ener → ien	+en	tienen	VPI3P

Table 1: Target output (English and Spanish)

This suffix-focused transformational model is not, as given, sufficient for languages with prefixal, infixal and reduplicative morphologies. But it is remarkably productive across Indo-European languages in its current form and can be extended to other affixational schema when appropriate.

For many applications, once the vocabulary list achieves sufficiently broad coverage, this alignment table effectively *becomes* a morphological analyzer simply by table lookup (independent of necessary contextual ambiguity resolution). While the probabilistic analyzer trained in Step 2 remains useful for previously unseen words, such words are typically quite regular and most of the difficult substance of the lemmatization problem can often be captured by a large root+POS↔inflection mapping table and a simple transducer to handle residual forms. This is not the case for agglutinative languages such as Turkish or Finnish, or for very highly inflected languages such as Czech, where sparse data becomes an issue. But for many languages, and to a quite practical degree, inflectional morphological analysis and generation can be viewed primarily as an *alignment* task on a broad coverage wordlist.

Thus, while this paper will discuss our implementation of a stand-alone probabilistic analyzer and retraining process in Steps 2 and 3, the challenge of large-coverage inflection-root alignment expressed in Step 1 is the core of this work.

1.1 Required and Optional Resources

In further clarification of the task description, the morphological induction described in this paper assumes, and is based on, only the following limited set of (often optional) available resources:

(a) A table (such as Table 2) of the inflectional parts of speech of the given language, along with a list of the canonical suffixes for each part of speech. These suffixes not only serve as mnemonic tags for the POS labels, but they can also be used to obtain a noisy set of candidate examples for each part of speech.[1]

(b) A large unannotated text corpus.

(c) A list of the candidate noun, verb and adjective roots of the language (typically obtainable from a dictionary), and any rough mechanism for identifying the candidate parts of speech of the remaining vocabulary based on aggregate models of context or tag sequence, not morphological analysis. Our concurrent work (Cucerzan and Yarowsky, 2000) focuses on the problem of bootstrapping approximate tag probability distributions by modelling relative word-form occurrence probabilities across indicative lexical contexts (e.g. "*the* <NOUN> *are*" and "*been* <VBG> *the*"), among other predictive variables, with the goal of co-training with the models presented here. It is not necessary to select the part of speech of a word in any given context, only provide an estimate of the candidate tag distributions across a full corpus. The source of these candidate tag estimates is unimportant, however, and the lists can be quite noisy. Their major function is to partially limit the potential alignment space from unrestricted word-to-word alignments across the entire vocabulary.

(d) The current implementation assumes a list of the consonants and vowels of the language.

(e) While not essential to the execution of the algorithm, a list of common function words of the given language is useful to the extraction of context similarity features.

(f) If available, the various distance/similarity tables generated by this algorithm on previously studied languages can be useful as seed information, especially if these languages are closely related (e.g. Spanish and Italian).

[1] The lists need not be exhaustive, and any missing irregular suffixes (e.g. the English past tense $+t$) can be captured via a stem change and null suffix (e.g. send: $d \rightarrow t + \epsilon \Rightarrow$ sent), similar to the representation of take: $ake \rightarrow ook + \epsilon \Rightarrow$ took).

2 Related Work

There is a rich tradition of supervised and unsupervised learning in the domain of morphology. Rumelhart and McClelland (1986), Egedi and Sproat (1988), Ling (1994) and Mooney and Califf (1995) have each investigated the supervised learning of the English past tense from paired training data, the first two using phonologically-based connectionist models and the latter two performing comparative studies with ID3 decision trees and first-order decision lists respectively.

Brent (1993, 1999), de Marcken (1995), Kazakov (1997) and Goldsmith (2000) have each focused on the problem of unsupervised learning of morphological systems as essentially a segmentation task, yielding a morphologically plausible and statistically motivated partition of stems and affixes. Brent and de Marcken both have used a minimum description length framework, with the primary goal of inducing lexemes from boundaryless speech-like streams. Goldsmith specifically sought to induce suffix paradigm classes (e.g. *NULL.ed.ing* vs. *e.ed.ing* vs. *e.ed.es.ing* vs. *ted.tion*) from raw text. However, handling of irregular words was largely excluded from this work, as Goldsmith assumed a strictly concatenative morphology without models for stem changes.

Morphology induction in agglutenative languages such as Turkish and Finnish presents a problem similar to parsing or segmenting a sentence, given the long strings of affixations allowed and the relatively free affix order. Voutilainen (1995) has approached this problem in a finite-state framework, and Hakkani-Tür et al. (2000) have done so using a trigram tagger, with the assumption of a concatenative affixation model.

The two-level model of morphology (Koskenniemi, 1983) has been extremely successful in manually capturing the morphological processes of the world's languages. The context sensitive stem-change models used in this current paper have been partially inspired by this framework. For example, a two-level equivalent capturing *happy + er = happier* is $y{:}i \Leftrightarrow p{:}p\ \_$, quite similar in spirit and function to our probabilistic model $P(y \rightarrow i | ...app, +er)$. Theron and Cloete

English :

Part of Speech	VB	VBD	VBZ	VBG	VBN
Canonical Suffixes	+ϵ	+ed (+t) +ϵ	+s	+ing	+en +ed (+t) +ϵ
Examples (*not* used in training)	jump announce take	jumped announced took	jumps announces takes	jumping announcing taking	jumped announced taken

Spanish:

Part of Speech	VRoot	VPI1s	VPI2s	VPI3s	VPI1p	VPI2p	VPI3p
Canonical Suffixes	+ar +er +ir	+o	+as +es	+a +e	+amos +emos +imos	+áis +éis +ís	+an +en

Table 2: Example parts of speech and their associated canonical suffixes in English and Spanish

(1997) sought to learn a 2-level rule set for English, Xhosa and Afrikaans by supervision from O(4000) aligned inflection-root pairs extracted from dictionaries. Single character insertion and deletions were allowed, and the learned rules supported both prefixation and suffixation. Their supervised learning approach could be applied directly to the aligned pairs induced in this paper.

Finally, Oflazer and Nirenburg (1999) have developed a framework to learn two-level morphological analyzers from interactive supervision in a Elicit-Build-Test loop under the Boas project. Humans provide as-needed feedback regarding errors and omissions. Recently applied to Polish, the model also assumes concatenative morphology and treats non-concatenative irregular forms through table lookup.

Thus there is a notable gap in the research literature for induction of analyzers for irregular morphological processes, including significant stem changing. The algorithm described below directly addresses this gap, while successfully inducing more regular analyses without supervision as well.

3 Lemma Alignment by Frequency Similarity

The motivating dilemma behind our approach to morphological alignment is the question of how one determines that the past tense of *sing* is *sang* and not *singed*. The pairing *sing*→*singed* requires only simple concatenation with the canonical suffix, *+ed*, and *singed* is indeed a legal word in our vocabulary (the past tense of *singe*). And while few irregular verbs have a true word occupying the slot that would be generated by a regular morphological rule, a large corpus is filled with many spelling mistakes or dysfluencies such as *taked* (observed with a frequency of 1), and such errors can wreak havoc in naïve alignment-based methods.

How can we overcome this problem? Relative corpus frequency is one useful evidence source. Observe in Table 3 that in an 80 million word collection of newswire text the relative frequency distribution of *sang/sing* is 1427/1204 (or 1.19/1), which indicates a reasonably close frequency match, while the *singed/sing* ratio is 0.007/1, a substantial disparity.

	VBD:VB	$\frac{VBD}{VB}$	$log(\frac{VBD}{VB})$
sang/sing	1427/1204	1.19	0.17
singed/sing	9/1204	0.007	-4.90
singed/singe	9/2	4.5	1.50
sang/singe	1427/9	158.5	5.06
All VBD/VB		.85	-0.16

Table 3: Example inflection-root frequency ratios

However, simply looking for close relative frequencies between an inflection and its candidate root is inappropriate, given that some inflections are relatively rare and *expected* to occur much less frequently than the root form.

Thus in order to be able to rank the *sang/sing* and *singed/sing* candidates effectively, it is necessary to be able to quantify how well each fits (or deviates from) expected frequency distributions. To do so, we use simple non-parametric statistics to calculate the probability of a particular $\frac{VBD}{VB}$ ratio by examining how frequently other such ratios in a similar range have been seen in the corpus. Figure 1 illustrates such a histogram (based on the log of the ratios to focus more attention on the extrema). The histogram is then smoothed and normalized as an approximation of the probability density function for this estimator ($log(\frac{VBD}{VB})$), which we can then use to quantify to what extent a given candidate $log(\frac{VBD}{VB})$, such as log(sang/sing)=.17, fits our empirically motivated expectations. The relative position of the candidate pairings on the graph suggests that this estimator is indeed informative given the task of ranking potential root-inflection pairings.

However, estimating these distributions presents a problem given that the true alignments (and hence frequency ratios) between inflections

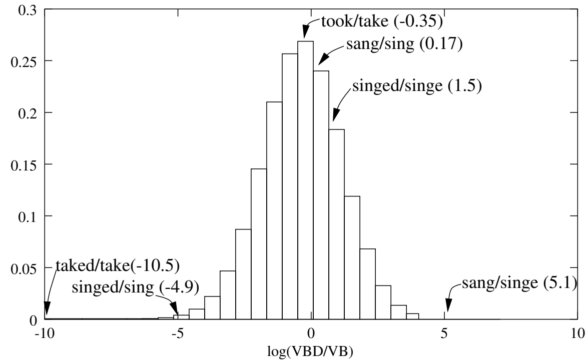

Figure 1: Using the $log(\frac{VBD}{VB})$ estimator to rank potential VBD-VB pairs in English

are not assumed to be known in advance. Thus to approximate this distribution automatically, we make the simplifying assumption that the frequency ratios between inflections and roots (largely an issue of tense and usage) is not significantly different between regular and irregular morphological processes.

Table 4 and Figure 2 illustrate that this simplifying assumption is supported empirically. Despite large lemma frequency differences between regular and irregular English verbs, their relative tense ratios for both $\frac{VBD}{VB}$ and $\frac{VBG}{VB}$ are quite similar in terms of their means and density functions.

VerbType	$\frac{VBD}{VB}$	$\frac{VBG}{VB}$	Avg. Lemma Freq[2]
Regular	.847	.746	861
Irregular	.842	.761	17406

Table 4: Similar regular-irregular frequency ratios

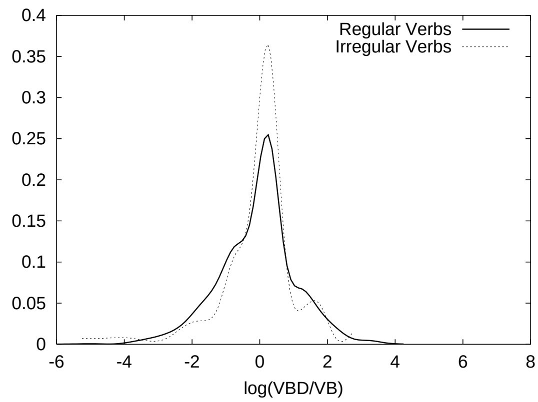

Figure 2: Distributional similarity between regular and irregular forms for VBD/VB

Thus we initially approximate the VBD/VB ratios from an automatically extracted (and noisy) set of verb pairs exhibiting simple and uncontested suffixation with the canonical *+ed* suffix. This distribution is re-estimated as alignments improve, but a single function continues to predict frequency ratios of unaligned (largely irregular) word pairs from the observed frequency of previously aligned (and largely regular) ones.

Furthermore, we are not just limited to using the ratio POS_i/VB to predict the expected frequency of POS_i in the corpus. The expected frequency of a viable past-tense candidate for *sing* should also be estimable from the frequency of any of the other inflectional variants.

Assuming that earlier iterations of the algorithm had filled the SING lemma slots for VBG and VBZ in Table 5 with regular inflections, $\frac{VBD}{VBG}$ and $\frac{VBD}{VZ}$ could also be used as estimators. Figure 3 shows the histogram for the estimator $log(\frac{VBD}{VBG})$.[3]

	Lemma	VB	VBD	VBG	VBZ	VBN
Word	SING	sing	?	singing	sings	?
Freq	?	1204	?	1381	344	?

Table 5: Example lemma frequency profile for *sing*

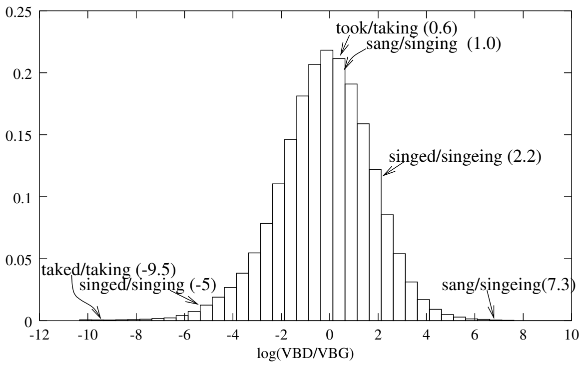

Figure 3: Using the $log(\frac{VBD}{VBG})$ estimator to rank potential VBD-VBG matches in English

We are also not limited to using only a single estimator. In fact, there are considerable robustness advantages to be gained by taking the average of estimators, especially for highly inflected languages where the observed frequency counts may be relatively small. To accomplish this in a general framework, we first estimate the hidden variable of total lemma frequency ($\widehat{LF}$) via a confidence-weighted average of the observed POS_i frequency and a globally estimated $\frac{\widehat{LF}}{POS_i}$ model. Then all subsequent POS_i frequency estimations can be made relative to $\frac{POS_i}{\hat{LF}}$, or a somewhat advantageous variant, $log(\frac{POS_i}{\widehat{LF-POS_i}})$, with this distribution illustrated in Figure 4. Another advantage of this consensus approach is that it only requires T rather than T^2 estimators, which is especially important as the inflectional tagset T grows quite large in some languages.

[3]Using this estimate, we predict a frequency E(VBD)=1567, which is an overestimate relative to the true 1427. In contrast, the distribution for $\frac{VBD}{VBZ}$ is considerably more noisy, given the problems with VBZ forms being confused with plural nouns. This latter measure yields a underestimate of 1184.

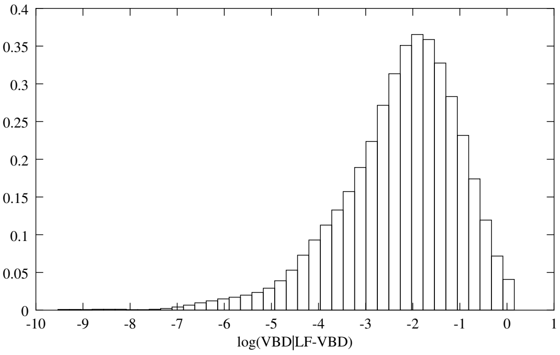

Figure 4: Using the $log(\frac{VBD}{\neg \widehat{VBD}})$ estimator to rank potential VBD-Lemma matches in English

Also, one can alternately conduct the same frequency-distribution-based ranking experiments over suffixes rather than tags. For example, $log(\frac{+ED}{+ING})$ yields a similar estimator to $log(\frac{VBD}{VBG})$, but with somewhat higher variance.[4]

Finally, these frequency-based alignment models can be informative even for more highly inflected languages. Figure 5 illustrates an estimate of the empirical distribution of the $\frac{VPI3P}{VBINF}$ part-of-speech frequency ratios in Spanish, with this estimator strongly favoring the correct but irregular *juegan/jugar* alignment rather than its orthographically similar competitors.

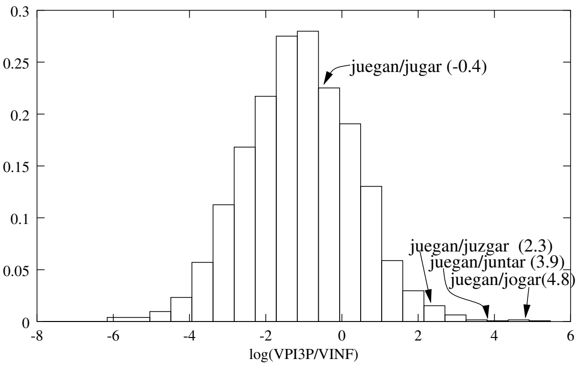

Figure 5: Using the $log(\frac{VPI3P}{VINF})$ estimator to rank potential VBPI3P-VINF pairs in Spanish

[4]This measure also frees one from any need to do part-of-speech distribution estimation. However, when optional variant suffixes (such as *+ed* and *+en*) exist in the canonical suffix set, performance can be improved by modeling this distribution separately for verbs with and without observed distinct $+EN$ forms, as the relative distribution of $log(\frac{+ED}{+ING})$ and $log(\frac{+ED}{ROOT})$ change somewhat substantially in these cases. One does not know in advance, however, whether a given test verb belongs to either set. Thus the initial frequency similarity score should be based on the average of both estimators until the presence or absence of the distinct variant form in the lemma can be ascertained on subsequent iterations.

4 Lemma Alignment by Context Similarity

A second powerful measure for ranking the potential alignments between morphologically related forms is based on the contextual similarity of the candidate forms. For this measure, we computed traditional cosine similarity between vectors of weighted and filtered context features. While this measure also gives relatively high similarity to semantically similar words such as *sip* and *drink*, it is rare even for synonyms to exhibit more similar and idiosyncratic argument distributions and selectional preferences than inflectional variants of the same word (e.g. *sipped*, *sipping* and *sip*). A primary goal in clustering inflectional variants of verbs is to give predominant vector weight to the head-noun objects and subjects of these verbs. However, to minimize needed training resources, we very roughly identified these positions by a set of simple regular expressions over small closed-class parts of speech, with remaining (open-class) content words labeled collectively as CW, e.g.: CW_{subj} (AUX|NEG)* $V_{keyword}$ DET? CW* CW_{obj}.

Such expressions will clearly both extract significant noise and fail to match many legitimate contexts, but because they are applied to a large monolingual corpus, the partial coverage and signal-to-noise ratio are tolerable. Ideally, one would also automatically identify which set of patterns are appropriate for a given language, but this can be accomplished in subsequent iterations of the algorithm by taking previously extracted <inflection,root> pairs and testing which subset of predefined regular expressions is most effective in maximizing the mean context-similarity of the <inflection,root> relative to non-pairs. Similar techniques can be used to weight the relative importance of contextual positions.[5]

For similar reasons, it is useful in subsequent iterations of the algorithm to apply the current analysis modules towards lemmatizing the contextual feature sets. This has the effect of both condensing the contextual signal, and removing potentially distracting correlations with inflectional forms in context.

[5]Another important concept in context similarity measures for morphology that differs from other word clustering measures is the need to downweight or eliminate context words such as subject pronouns that strongly correlate with only one or a few inflectional forms. Giving such words too much weight can cause different verbs of the same person/number to appear more similar to each other than do the different inflections of the same verb. Filtering based on high cross-lemma distributional entropy for a given context word can help eliminate these counter-productive features.

5 Lemma Alignment by Weighted Levenshtein Distance

The third alignment similarity function considers overall stem edit distance using a weighted Levenshtein measure. In morphological systems worldwide, vowels and vowel clusters are relatively mutable through morphological processes, while consonants generally tend to have a lower probability of change during inflection. Rather than treating all string edits as equal, a cost matrix of the form shown in Table 6 is utilized, with initial distance costs δ_1=V-V, δ_2=V^+-V^+, δ_3=C-C and δ_4=C-V^+, initially set to (0.5, 0.6, 1.0, 0.98), a relatively arbitrary assignment reflecting this tendency. However, as subsequent algorithm iterations proceed, this matrix is re-estimated with empirically observed character-to-character stem-change probabilities from the algorithm's current best weighted alignments.

	a	o	ue	m	n	...
a	0	δ_1	δ_2	δ_4	δ_4	...
o	δ_1	0	δ_2	δ_4	δ_4	...
ue	δ_2	δ_2	0	δ_4	δ_4	...
m	δ_4	δ_4	δ_4	0	δ_3	...
n	δ_4	δ_4	δ_4	δ_3	0	...
...	...	...	...	...	...	...

Table 6: Initial Levenshtein cost matrix

More optimally, the initial state of this matrix could be seeded with values partially borrowed from previously trained matrices from other related languages. Alternately, the initial distances could be set partially sensitive to phonological similarities, with dist(/d/,/t/) < dist(/d/,/f/) for example, although this particular distinction emerges readily via iterative re-estimation from the baseline model.

6 Lemma Alignment by Morphological Transformation Probabilities

The goal of this research is not only to extract an accurate table of inflection-root alignments, but also to generalize this mapping function via a generative probabilistic model. The following section describes the creation of this model, as well as how the context-sensitive probability of each morphological transformation can be used as the fourth alignment similarity measure.

At each iteration of the algorithm, this probabilistic mapping function is trained on the table output of the previous iteration, equivalent to the information in Table 1 (e.g. <root,inflection> pairs with optional part-of-speech tags, confidence scores and stemchange+suffix analysis).[6] From this output, we cluster the observed stem changes by the variable-length root context in which they were applied, as illustrated in Table 7.

Root Context	Stem Change	Suffix	Count	Matching Examples
..ray	$\epsilon \rightarrow \epsilon$	+ed	5	spray, stray,...
...ay	$\epsilon \rightarrow \epsilon$	+ed	13	play, spray,...
...oy	$\epsilon \rightarrow \epsilon$	+ed	3	annoy, enjoy,...
...ey	$\epsilon \rightarrow \epsilon$	+ed	5	obey, key,...
...fy	y $\rightarrow$ i	+ed	21	beautify,...
...ry	y $\rightarrow$ i	+ed	7	carry,...
...dy	y $\rightarrow$ i	+ed	4	bloody,...
...y	y $\rightarrow$ i	+ed	43	carry,...
...y	$\epsilon \rightarrow \epsilon$	+ed	21	spray,...
...y	$\epsilon \rightarrow \epsilon$	+ing	83	carry, spray,...
...e	e $\rightarrow \epsilon$	+ed	728	dance,...
...e	e $\rightarrow \epsilon$	+ing	783	dance, take,...
...e	$\epsilon \rightarrow \epsilon$	+ing	1	singe
...ke	ake $\rightarrow$ ook	$+\epsilon$	3	take, shake,...
...ke	ake $\rightarrow$ oke	$+\epsilon$	1	wake
...ke	ke $\rightarrow$ de	$+\epsilon$	1	make
...ay	y $\rightarrow$ id	$+\epsilon$	2	lay, pay
...y	y $\rightarrow$ id	$+\epsilon$	2	lay, pay

Table 7: Stem change data given root context

First note that because the triple of <root> + <stemchange> + <suffix> uniquely determines a resulting inflection, one can effectively compute P(inflection | root, suffix, POS) by P(stemchange | root, suffix, POS), i.e. for any root=$\gamma\alpha$, suffix=$+\sigma$ and inflection=$\gamma\beta\sigma$, $P(\gamma\beta\sigma|\gamma\alpha, +\sigma, POS) = P(\alpha \rightarrow \beta|\gamma\alpha, +\sigma,\text{POS})$.

Using statistics such as shown in Table 7, it is thus possible to compute the generation (or alignment) probability for an inflection given root and suffix using the simple interpolated backoff model in (1) where λ_i is a function of the relative sample size of the conditioning event, and $last_k(root)$ indicates the final k characters of the root.

$$\begin{aligned}
&P(\text{ inflection} \mid \text{root, suffix, POS}) \\
&\quad = P(\ \alpha \rightarrow \beta \mid \text{root, suffix, POS}) \\
&\quad \approx \lambda_1 P(\ \alpha \rightarrow \beta \mid last_3(\text{root}), \text{suffix, POS}) \\
&\qquad + (1-\lambda_1)(\lambda_2 P(\ \alpha \rightarrow \beta \mid last_2(\text{root}), \text{suffix, POS}) \\
&\qquad + (1-\lambda_2)(\lambda_3 P(\ \alpha \rightarrow \beta \mid last_1(\text{root}), \text{suffix, POS}) \\
&\qquad + (1-\lambda_3)(\lambda_4 P(\ \alpha \rightarrow \beta \mid \text{suffix, POS}) \\
&\qquad + (1-\lambda_4) P(\ \alpha \rightarrow \beta\ )
\end{aligned} \tag{1}$$

We only backoff to the extent necessary. Furthermore, note that for English (and most inflections in Spanish), the stem changes observed when adding suffixes are independent of part of speech

[6] If only the pairs are given, with no stemchange+suffix analysis, this analysis can be generated deterministically by removing the longest matching canonical suffix from the inflection and generating the minimal $\alpha \rightarrow \beta + \sigma$ transformation capturing the remaining stem difference.

(i.e. *+s* behaves the same on suffixation for both nouns and verbs), so these probabilities can often be further simplified by deleting the conditioning variable POS, as illustrated in (2).

$$\begin{aligned}
&P(\text{ solidified} \mid \text{solidify, +ed, VBD})\\
&\quad = P(\text{ y}\rightarrow\text{i} \mid \text{solidify, +ed, VBD})\\
&\quad \approx P(\text{ y}\rightarrow\text{i} \mid \text{solidify, +ed})\\
&\quad \approx \lambda_1 P(\text{ y}\rightarrow\text{i} \mid \text{ify, +ed})\\
&\qquad + (1-\lambda_1)(\lambda_2 P(\text{ y}\rightarrow\text{i} \mid \text{fy, +ed})\\
&\qquad + (1-\lambda_2)(\lambda_3 P(\text{ y}\rightarrow\text{i} \mid \text{y, +ed})\\
&\qquad + (1-\lambda_3)(\lambda_4 P(\text{ y}\rightarrow\text{i} \mid \text{+ed})\\
&\qquad + (1-\lambda_4) P(\text{ y}\rightarrow\text{i })
\end{aligned} \tag{2}$$

We have further generalized these variable-length context models via a full hierarchically-smoothed trie architecture, allowing robust specialization to very long root contexts if sample sizes are sufficient.

6.1 Baseline Model for Morphological Transformation Probabilities

On the first iteration, no inflection/root pairs are available for estimating the above models. As prior knowledge is not available regarding $\alpha \rightarrow \beta$ stem-change probabilities, an assumption is made that the cost of each is proportional to the previously described Levenshtein distance between α and β, with the cost of a change increasing geometrically as the distance from the end of the root increases. The rate of this cost increase ultimately depends on the tendency of the language to allow word-internal spelling changes (as in Spanish or Arabic), or strongly favor changes at the point of affixation (as in English).

6.2 Model Improvement by Iterative Re-estimation

The primary goal of iterative retraining is to refine the core morphological transformation model, which not only serves as one of the four similarity models, but is also a primary deliverable of the learning process.

As subsequent iterations proceed, the stem-change probability models are retrained on the output of the prior iteration, weighting each training example with its alignment confidence, and filtering out $\alpha \rightarrow \beta$ changes without a minimum level of support to help reduce noise. The final stem-change probabilities then are an interpolation with the trained model P_j and the initial baseline (P_0) model described in Section 6.1:

$$\begin{aligned}
&P(\ \alpha \rightarrow \beta \mid \text{root, suffix, POS})\\
&\quad = \lambda_j\ P_0(\ \alpha \rightarrow \beta \mid \text{suffix})\\
&\qquad + (1-\lambda_j)\ P_j(\ \alpha \rightarrow \beta \mid \text{root, suffix, POS})
\end{aligned}$$

The Levenshtein distance models are re-estimated as observed in Section 5, while the context similarity model can be improved through better self-learned lemmatization of the modelled context words.

7 Lemma Alignment by Model Combination and the Pigeonhole Principle

As shown empirically below, no single model is sufficiently effective on its own. We applied traditional classifier combination techniques to merge the four models' scores, scaling each to achieve compatible dynamic range. The Frequency, Levenshtein and Context similarity models retain equal relative weight as training proceeds, while the Morphological Transformation (MorphTrans) similarity model increases in relative weight as it becomes better trained.

Table 8 demonstrates the combined measures in action, showing the relative rankings of candidate roots for the inflections *took*, *shook* and *juegan* by the four similarity models after the first iteration (in Columns 2-4). The overall consensus similarity measure at the end of Iteration 1 is shown in Column 1.[7]

Note that even though only one of the four estimators independently ranked *shake* as the most likely root of *shook*, after only the first iteration the consensus choice is correct. The final column of Table 8 shows the retrained MorphTrans similarity measure after convergence. Based on training evidence from the confidently aligned pairs *took/take*, *shook/shake* and *undertook/undertake* from previous iterations, the probability of *ake→ook* has increased significantly, further increasing the confidence in the overall alignments at convergence (not shown), but not changing the previously correct ranking in these cases.

The final alignment constraint that we pursued was based on the pigeonhole principle. This principle suggests that for a given part of speech, a root should not have more than one inflection nor should multiple inflections in the same part of speech share the same root. There are, of course, exceptions to this tendency, such as *travelled/traveled* and *dreamed/dreamt*, which are observed as variant forms of their respected roots.

[7] In addition to the consensus similarity score in subcolumn 2, subcolumn 3 shows the average of the ranks of the candidate root given the inflection *and* the ranks of the candidate inflection given the root. This bidirectional average ranking score favors cases where attraction between root and inflection is mutual, and disfavors cases where higher ranked competition exists for a root's attentions, effectively capturing a weak form of the pigeonhole principle. Thus it was used as the primary ranking criteria (over raw similarity score).

Candidate Roots for the English inflection **TOOK** (1st iteration):

Overall Similarity (Iteration 1)				Context Similarity		Frequency Similarity		Levenshtein Similarity		MorphTrans Similarity (1)		MorphTrans Similarity (C)	
take	.00162	3.8	**1**	**take**	.849	**take**	.072	toot	.333	toot	.002593	**take**	.465578
turn	.00081	8.7	2	turn	.546	tell	.028	tool	.333	tool	.002593	toot	.001296
tell	.00063	15.9	3	tower	.332	turn	.016	toe	.310	tong	.000096	tool	.001296
test	.00041	19.6	4	touch	.324	talk	.014	**take**	.290	tone	.000096	tong	.000048
talk	.00051	21.0	5	tip	.261	test	.001	top	.236	...	...	tone	.000048
tie	.00044	26.7	6	tie	.260	teach	.001	toil	.236	**take**	.000006	tout	.000048

Candidate Roots for the English inflection **SHOOK** (1st iteration):

Overall Similarity (Iteration 1)				Context Similarity		Frequency Similarity		Levenshtein Similarity		MorphTrans Similarity (1)		MorphTrans Similarity (C)	
shake	.00149	5.5	**1**	**shake**	.854	share	.073	shoo	.500	shoot	.002593	**shake**	.465578
shoot	.00126	9.3	2	shave	.323	ship	.068	shoot	.333	shoo	.002593	shoot	.001296
ship	.00104	16.3	3	shape	.210	shift	.062	shoe	.310	shock	.000096	shoo	.001296
shatter	.00061	18.9	4	shore	.194	shop	.060	**shake**	.290	short	.000096	shock	.000048
shop	.00094	19.8	5	shower	.184	**shake**	.058	shop	.236	shout	.000095	short	.000048
shut	.00081	20.6	6	shoot	.162	shut	.052	shout	.236	...	...	shove	.000048
shun	.00039	20.7	7	shock	.154	shoot	.051	show	.236	**shake**	.000003	shore	.000048

Candidate Roots for the Spanish inflection **JUEGAN** (1st iteration):

Overall Similarity (Iteration 1)			Context Similarity		Frequency Similarity		Levenshtein Similarity		MorphTrans Similarity (1)	
jugar	.0024	1	**jugar**	.88	**jugar**	.063	**jugar**	.50	**jugar**	.00129
juzgar	.0006	2	juntar	.38	juzgar	.015	juzgar	.29	jogar	.00129
jurar	.0002	4	jurar	.26	jogar	.009	juntar	.25	juntar	.00004
jogar	.0000	5	justificar	.22	juntar	.004	jurar	.18	juzgar	.00004

Table 8: Example performance of independent and combined similarity measures

The extent to which such overlaps should be penalized depends on the probability of seeing variant inflections in the morphology, but for Spanish and English this is relatively low.

We exploited the pigeonhole principle in two ways simultaneously. The first is a greedy algorithm, in which candidates are aligned in order of decreasing score, and when the the first-choice root for a given inflection has already been taken by another inflection of the same part of speech, the algorithm continues until a free slot is found. The exception is when the highest ranking free form is several orders of magnitude lower than the first choice; here the first-choice alignment is assumed to be correct, but a variant form.

8 Empirical Evaluation

Current empirical evaluation of this work focuses on its accuracy in analyzing the often highly irregular past tense of English verbs. Consistent with prior empirical studies in this field, evaluation was performed on a test set of 3888 inflected words, including 128 highly irregular inflections, 1877 cases where the past tense was formed by simple concatenative suffixation, and 1883 inflections exhibiting a non-concatenative stem change such as gemination or elision.

In execution, for each test inflected form, the analysis algorithm was free to consider alignment to any word in the corpus which had been identified as a potential root verb by the part-of-speech tagging process or occurrence in a dictionary-derived rootlist, *not* just those roots in the test set. It is thus a more challenging evaluation than testing simple alignment accuracy between two clean and extraneous-entry-free wordlists.

Table 9 shows the performance of several of the investigated similarity measures. Frequency similarity (FS), enhanced Levenshtein (LS), and Context similarity (CS) alone achieve only 10%, 31% and 28% overall accuracy respectively. However, the hypothesis that these measures model independent and complementary evidence sources is supported by the roughly additive combined accuracy of 71.6%.[8]

The final performance of the full converged CS+FS+LS+MS model at 99.2% accuracy on the full test set, and 99.7% accuracy on inflections requiring analysis beyond simple concatenative suffixation, is quite remarkable given that the algorithm had absolutely no <inflection,root> examples as training data, and had no prior inventory of stem changes available, with only a slight statistical bias in favor of shorter stem changes with

[8] In fact, in many cases the consensus ranking choice is correct when each independent model's first choice is wrong, actually yielding a small synergistic super additivity.

Combination of Similarity Models	# of Iterations	All Words (3888)	Highly Irregular (128)	Simple Concat. (1877)	Non-Concat. (1883)
FS (*Frequency Sim*)	(Iter 1)	9.8	18.6	8.8	10.1
LS (*Levenshtein Sim*)	(Iter 1)	31.3	19.6	20.0	34.4
CS (*Context Sim*)	(Iter 1)	28.0	32.8	30.0	25.8
CS+FS	(Iter 1)	32.5	64.8	32.0	30.7
CS+FS+LS	(Iter 1)	71.6	76.5	71.1	71.9
CS+FS+LS+MS	(Iter 1)	96.5	74.0	97.3	97.4
CS+FS+LS+MS	(Convg)	**99.2**	**80.4**	**99.9**	**99.7**

Table 9: Performance of combined alignment models on 4 classes of past-tense English verbs

smaller Levenshtein distance, and with the minimal search-simplifying assumption in all the models that candidate alignments must begin with a the same V^*C^* prefix.[9]

Given a starting point where all single character $X \rightarrow Y$ changes at the point of suffixation are equally likely, the processes of elison ($e \rightarrow \epsilon$), gemination (e.g. $\epsilon \rightarrow d$ in the context of d), and $y \rightarrow i$ shift (in the context of a preceding consonant, not vowel) all emerge by the end of the first iteration with high probability in their appropriate contexts, and low probability elsewhere.

Table 10 shows how each of the models perform on a randomly-selected 30% of the highly irregular forms, with correctly selected roots identified in bold. The residual errors are primarily of three types: Two inflections, *went* and *ate*, were not alignable with their correct roots due to different first character. The largest class of errors are due to the pigeonhole principle strongly disfavoring two inflections from sharing the same root.[10]

The remaining errors typically are due to sparse statistics for the lower frequency irregular forms. Mappings such as *slew*↔*slay* are particularly difficult because, with a corpus frequency of only 4, there is too little data to estimate a good context profile or an effectively discriminatory frequency profile. Enlarging the raw corpus size should improve performance in both of these cases.

9 Conclusion

This paper has presented an original algorithm capable of inducing the accurate morphological analysis of even highly irregular verbs, starting with no paired <inflection,root> examples for training and no prior seeding of legal morphological transformations. It does so by treating morphological analysis predominantly as an alignment task in a large corpus, performing the effective collaboration of four original similarity measures based on expected frequency distributions, context, morphologically-weighted Levenshtein similarity and an iteratively bootstrapped model of affixation and stem-change probabilities. This constitutes a significant achievement in that previous approaches to morphology acquisition have either focused on unsupervised induction of quasi-regular concatenative affixation, or handled irregular forms with fully supervised training. In contrast, this paper's essentially unsupervised algorithm achieves over 80% accuracy on the most highly irregular forms, and 99.7% accuracy on analyses requiring some stem change, outperforming Mooney and Califf's fully supervised learning algorithm overall and on both of these measures.

[9]To put the Table 9 results in perspective, Mooney and Califf (1995) achieved 82.5% overall accuracy using a fully supervised decision list learner trained on 250 paired past-tense/root verb pairs (in plain text form). Although they don't breakdown this performance by word type, their included FOILDL program trained from 250 pairs and applied to our evaluation set achieved 100% accuracy on the pairs with simple *+ed* concatenation, 84% accuracy on stem changing (non-concat) pairs and 5% accuracy on the highly irregular pairs, with 89% overall accuracy. Other available supervised learning results (e.g. Ling; Rumelhart and McClelland) are only given for phonological word representations. While not directly comparable with our text-based data, their performance is significantly worse than Mooney and Califf's FOILDL on common phonological paired data, suggesting that FOILDL is a generally competitive reference point for our results.

[10]This was previously noted in the case of *dream* ↔ *dreamed* and *dreamt*, or *burned* ↔ *burned* and *burnt*, with the higher probability analysis typically occupying the root slot and the lower probability form typically forced to seek alignment elsewhere. Indeed, the pigeonhole principle is the most problematic of all the alignment principles used, as it creates nearly as many problems as it fixes. The overall performance advantage is slightly in its favor (with 59 misalignments avoided for 50 problems created), but the cost of this approach is borne heavily by the irregular verbs, and a probabilistic model of when variant forms should be expected/allowed is necessary to fix these cases while preserving the advantages of the principle in downweighting clashing analyses in the more regular verbs.

Test Word	True Root	CS+FS+LS+MS (Convg)	Score	(Itr 1)	CS+FS+LS (Itr 1)	CS+FS (Itr 1)	LS only (Itr 1)
got	get	go	1.30	go	go	go	gut
knew	know	**know**	1.35	**know**	**know**	**know**	**know**
took	take	**take**	1.50	**take**	**take**	**take**	toot
blew	blow	**blow**	1.80	**blow**	**blow**	**blow**	**blow**
became	become	**become**	2.35	**become**	**become**	**become**	**become**
made	make	**make**	2.40	**make**	**make**	**make**	mate
clung	cling	**cling**	2.55	**cling**	**cling**	**cling**	**cling**
drew	draw	**draw**	2.65	**draw**	**draw**	**draw**	**draw**
swore	swear	**swear**	2.80	**swear**	**swear**	**swear**	store
wore	wear	**wear**	3.10	**wear**	**wear**	**wear**	wire
came	come	**come**	3.55	**come**	**come**	**come**	**come**
thought	think	**think**	3.60	**think**	**think**	**think**	thump
flung	fling	**fling**	4.60	**fling**	**fling**	**fling**	**fling**
brought	bring	**bring**	5.35	**bring**	**bring**	**bring**	brighten
strove	strive	**strive**	5.85	**strive**	**strive**	straddle	**strive**
stuck	stick	**stick**	6.00	**stick**	**stick**	stabilize	stock
swept	sweep	**sweep**	6.20	**sweep**	**sweep**	**sweep**	swap
shone	shine	**shine**	6.55	**shine**	**shine**	**shine**	**shine**
woke	wake	**wake**	6.95	**wake**	**wake**	wind	**wake**
clove	cleave	**cleave**	7.35	**cleave**	**cleave**	**cleave**	close
bore	bear	**bear**	7.75	**bear**	bar	**bear**	bare
meant	mean	**mean**	8.20	**mean**	**mean**	manage	mount
lent	lend	**lend**	9.25	**lend**	**lend**	**lend**	**lend**
slew	slay	slit	10.06	slit	slight	slight	slow
struck	strike	**strike**	11.60	**strike**	**strike**	**strike**	strut
bought	buy	**buy**	12.20	**buy**	**buy**	**buy**	bound
bit	bite	**bite**	13.60	**bite**	**bite**	betray	bet
dove	dive	**dive**	17.25	**dive**	**dive**	dash	**dive**
burnt	burn	burp	17.30	burp	burp	burp	**burn**
went	go	want	18.29	want	want	want	want
caught	catch	**catch**	18.35	**catch**	cut	**catch**	cough
dealt	deal	**deal**	21.45	**deal**	**deal**	disagree	**deal**

Table 10: Performance of 4 alignment models on 32 randomly selected highly irregular English verbs

References

M.R. Brent, 1993. Minimal generative models: A middle ground between neurons and triggers. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pages 28–36.

M.R. Brent, 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, pages 71–106.

S. Cucerzan and D. Yarowsky, 2000. Language independent minimally supervised induction of lexical probabilities. *Proceedings of ACL'00*, Hong Kong.

C. de Marcken, 1995. Unsupervised language acquisition. PhD dissertation. MIT.

D. Egedi and R. Sproat, 1988. Connectionist Networks and Natural Language Morphology. UMD Conf on Grammar and Language Processing.

J. Goldsmith, 2000. Unsupervised Learning of the Morphology of a Natural Language. *http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/Paper/paper.html*.

D.Z. Hakkani-Tür, K. Oflazer, G. Tür, 2000. Statistical Morphological Disambiguation for Agglutinative Languages. In *Proceedings of COLING 2000.*

L. Karttunen, 1993. Finite state constraints. In John Goldsmith (ed.) *The Last Phonological Rule*, pages 173–194. Chicago: University of Chicago Press.

D. Kazakov, 1997. Unsupervised learning of naive morphology with genetic algorithms. *ECML/Mlnet Workshop on Empirical Learning of NLP Tasks.*

K. Koskenniemi, 1983. A general computation model for word-form recognition and production. *Pub. 11, Dept. of General Linguistics.* Univ. of Helsinki.

C.X. Ling, 1994. Learning the past tense of English verbs: The symbolic pattern associator vs. connectionist models. *J. Art. Intel. Res.*, 1:209-229.

R. Mooney and M. Califf, 1995. Induction of first-order decision lists: Results on learning the past tense of English verbs. *J. Art. Intel. Res.*, 3:1-24.

K. Oflazer and S. Nirenburg, 1999. Practical bootstrapping of morphological analyzers. *Proceedings of the Conference on Natural Language Learning.*

D. Rumelhart and J. McClelland, 1986. On learning the past tense of English verbs. In J. McClelland, D. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition*, Volume 2. MIT Press.

P. Theron and I. Cloete, 1997. Automatic acquisition of two-level morphological rules. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, pages 103-110.

A. Voutilainen, 1995. Morphological disambiguation. In F. Karlsson, A. Voutilainen, J. Heikkila, and A. Anttila (eds.) *Constraint grammar - A language independent system for parsing unrestricted text*, pages 165–284. The Hague: Mouton de Gruyter.