Universal Morphological Analysis using Structured Nearest Neighbor Prediction

Young-Bum Kim
University of Wisconsin-Madison
ybkim@cs.wisc.edu

João V. Graça
L^2F INESC-ID
Lisboa, Portugal
joao.graca@l2f.inesc-id.pt

Benjamin Snyder
University of Wisconsin-Madison
bsnyder@cs.wisc.edu

Abstract

In this paper, we consider the problem of unsupervised morphological analysis from a new angle. Past work has endeavored to design unsupervised learning methods which explicitly or implicitly encode inductive biases appropriate to the task at hand. We propose instead to treat morphological analysis as a structured prediction problem, where languages with labeled data serve as training examples for unlabeled languages, without the assumption of parallel data. We define a universal morphological feature space in which every language and its morphological analysis reside. We develop a novel structured nearest neighbor prediction method which seeks to find the morphological analysis for each unlabeled language which lies as close as possible in the feature space to a training language. We apply our model to eight inflecting languages, and induce nominal morphology with substantially higher accuracy than a traditional, MDL-based approach. Our analysis indicates that accuracy continues to improve substantially as the number of training languages increases.

1 Introduction

Over the past several decades, researchers in the natural language processing community have focused most of their efforts on developing text processing tools and techniques for English (Bender, 2009), a morphologically simple language. Recently, increasing attention has been paid to the wide variety of other languages of the world. Most of these languages still pose severe difficulties, due to (i) their lack of annotated textual data, and (ii) the fact that they exhibit linguistic structure not found in English, and are thus not immediately susceptible to many traditional NLP techniques.

Consider the example of nominal part-of-speech analysis. The Penn Treebank defines only four English noun tags (Marcus et al., 1994), and as a result, it is easy to treat the words bearing these tags as completely distinct word classes, with no internal morphological structure. In contrast, a comparable tagset for Hungarian includes 154 distinct noun tags (Erjavec, 2004), reflecting Hungarian's rich inflectional morphology. When dealing with such languages, treating words as atoms leads to severe data sparsity problems.

Because annotated resources do not exist for most morphologically rich languages, prior research has focused on unsupervised methods, with a focus on developing appropriate inductive biases. However, inductive biases and declarative knowledge are notoriously difficult to encode in well-founded models. Even putting aside this practical matter, a universally correct inductive bias, if there is one, is unlikely to be be discovered by *a priori* reasoning alone.

In this paper, we argue that languages for which we *have* gold-standard morphological analyses can be used as effective guides for languages *lacking* such resources. In other words, instead of treating each language's morphological analysis as a *de novo* induction problem to be solved with a purely hand-coded bias, we instead learn from our labeled languages what linguistically plausible morphological analyses looks like, and guide our analysis in this direction.

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 322–332,
Edinburgh, Scotland, UK, July 27–31, 2011.

More formally, we recast morphological induction as a new kind of supervised structured prediction problem, where each annotated language serves as a single training example. Each language's noun lexicon serves as a single input x, and the analysis of the nouns into stems and suffixes serves as a complex structured label y.

Our first step is to define a universal morphological feature space, into which each language and its morphological analysis can be mapped. We opt for a simple and intuitive mapping, which measures the sizes of the stem and suffix lexicons, the entropy of these lexicons, and the fraction of word forms which appear without any inflection.

Because languages tend to cluster into well defined morphological groups, we cast our learning and prediction problem in the nearest neighbor framework (Cover and Hart, 1967). In contrast to its typical use in classification problems, where one can simply pick the label of the nearest training example, we are here faced with a structured prediction problem, where locations in feature space depend jointly on the input-label pair (x, y). Finding a nearest neighbor thus consists of *searching* over the space of morphological analyses, until a point in feature space is reached which lies closest to one of the labeled languages. See Figure 1 for an illustration.

To provide a measure of empirical validation, we applied our approach to eight languages with inflectional nominal morphology, ranging in complexity from very simple (English) to very complex (Hungarian). In all but one case, our approach yields substantial improvements over a comparable monolingual baseline (Goldsmith, 2005), which uses the minimum description length principle (MDL) as its inductive bias. On average, our method increases accuracy by 11.8 percentage points, corresponding to a 42% decrease in error relative to a supervised upper bound. Further analysis indicates that accuracy improves as the number of training languages increases.

2 Related Work

In this section, we briefly review prior work on unsupervised morphological induction, as well as multilingual analysis in NLP.

Unsupervised Morphological Induction: Unsupervised morphology remains an active area of research (Schone and Jurafsky, 2001; Goldsmith, 2005; Adler and Elhadad, 2006; Creutz and Lagus, 2005; Dasgupta and Ng, 2007; Creutz and Lagus, 2007; Poon et al., 2009). Many existing algorithms derive morpheme lexicons by identifying recurring patterns in words. The goal is to optimize the compactness of the data representation by finding a small lexicon of highly frequent strings, resulting in a *minimum description length* (MDL) lexicon and corpus (Goldsmith, 2001; Goldsmith, 2005). Later work cast this idea in a probabilistic framework in which the the MDL solution is equivalent to a MAP estimate in a suitable Bayesian model (Creutz and Lagus, 2005). In all these approaches, a locally optimal segmentation is identified using a task-specific greedy search.

Multilingual Analysis: An influential line of prior multilingual work starts with the observation that rich linguistic resources exist for some languages but not others. The idea then is to *project* linguistic information from one language onto others via parallel data. Yarowsky and his collaborators first developed this idea and applied it to the problems of part-of-speech tagging, noun-phrase bracketing, and morphology induction (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2000; Yarowsky and Ngai, 2001), and other researchers have applied the idea to syntactic and semantic analysis (Hwa et al., 2005; Padó and Lapata, 2006) In these cases, the existence of a bilingual parallel text along with highly accurate predictions for one of the languages was assumed.

Another line of work assumes the existence of bilingual parallel texts without the use of any supervision (Dagan et al., 1991; Resnik and Yarowsky, 1997). This idea has been developed and applied to a wide variety tasks, including morphological analysis (Snyder and Barzilay, 2008b; Snyder and Barzilay, 2008a), part-of-speech induction (Snyder et al., 2008; Snyder et al., 2009b; Naseem et al., 2009), and grammar induction (Snyder et al., 2009a; Blunsom et al., 2009; Burkett et al., 2010). An even more recent line of work does away with the assumption of parallel texts and performs joint unsupervised induction for various languages through the use of coupled priors in the context of grammar in-

duction (Cohen and Smith, 2009; Berg-Kirkpatrick and Klein, 2010).

In contrast to these previous approaches, the method proposed in this paper does *not* assume the existence of any parallel text, but *does* assume that labeled data exists for a wide variety of languages, to be used as training examples for our test language.

3 Structured Nearest Neighbor

We reformulate morphological induction as a *supervised* learning task, where each annotated language serves as a single training example for our language-independent model. Each such example consists of an input-label pair (x, y), both of which contain complex internal structure: The input $x \in \mathcal{X}$ consists of a vocabulary list of all words observed in a particular monolingual corpus, and the label $y \in \mathcal{Y}$ consists of the correct morphological analysis of all the vocabulary items in x.[1] Because our goal is to generalize across languages, we define a feature function which maps each (x, y) pair to a universal feature space: $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$.

For each unlabeled input language x, our goal is to predict a complete morphological analysis $y \in \mathcal{Y}$ which maximizes a scoring function on the feature space, $score : \mathbb{R}^d \rightarrow \mathbb{R}$. This scoring function is trained using the n labeled-language examples: $(x, y)_1, \ldots, (x, y)_n$, and the resulting prediction rule for unlabeled input x is given by:

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, score\big(\mathbf{f}(x, y)\big)$$

Languages can be typologically categorized by the type and richness of their morphology. On the assumption that for each test language, at least one typologically similar language will be present in the training set, we employ a *nearest neighbor* scoring function. In the standard nearest neighbor classification setting, one simply predicts the label of the closest training example in the input space.[2] In our structured prediction setting, the mapping to the universal feature space depends crucially on the structure of the proposed label y, not simply the input x. We thus generalize nearest-neighbor prediction to the structured scenario and propose the following prediction rule:

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \min_{\ell} \parallel \mathbf{f}(x, y) - \mathbf{f}(x_\ell, y_\ell) \parallel, \quad (1)$$

where the index ℓ ranges over the training languages. In words, we predict the morphological analysis y for our test language which places it as close as possible in the universal feature space to one of the training languages ℓ.

Morphological Analysis: In this paper we focus on nominal inflectional suffix morphology. Consider the word *utiskom* in Serbian, meaning *impression* with the instrumental case marking. A correct analysis of this word would divide it into a stem (*utisak* = *impression*), a suffix (*-om* = instrumental case), and a phonological deletion rule on the stem's penultimate vowel ($..ak\# \rightarrow ..k\#$).

More generally, as we define it, a morphological analysis of a word type w consists of (i) a stem t, (ii), a suffix f, and (iii) a deletion rule d. Either or both of the suffix and deletion rule can be $NULL$. We allow three types of deletion rules on stems: deletion of final vowels ($..V\# \rightarrow ..\#$), deletion of penultimate vowels ($..VC\# \rightarrow ..C\#$), and removals and additions of final accent marks (e.g. $..\tilde{a}\# \rightarrow ..a\#$). We require that stems be at least three characters long and that suffixes be no more than four. And, of course, we require that after (1) applying deletion rule d to stem t, and (2) adding suffix f to the result, we obtain word w.

Universal Feature Space: We employ a fairly simple and minimal set of features, all of which could plausibly generalize across a wide range of languages. Consider the set of stems T, suffixes F, and deletion rules D, induced by the morphological analyses y of the words x. Our first three features simply count the sizes of these three sets.

These counting features consider only the raw number of unique morphemes (and phonological rules) being used, but not their individual frequency or distribution. Our next set of features considers the empirical *entropy* of these occurrences as distributed across the lexicon of words x by analysis y.

[1] Technically, the label space of each input, $\mathcal{Y}$, should be thought of as a function of the input x. We suppress this dependence for notational clarity.

[2] More generally the majority label of the k-nearest neighbors.

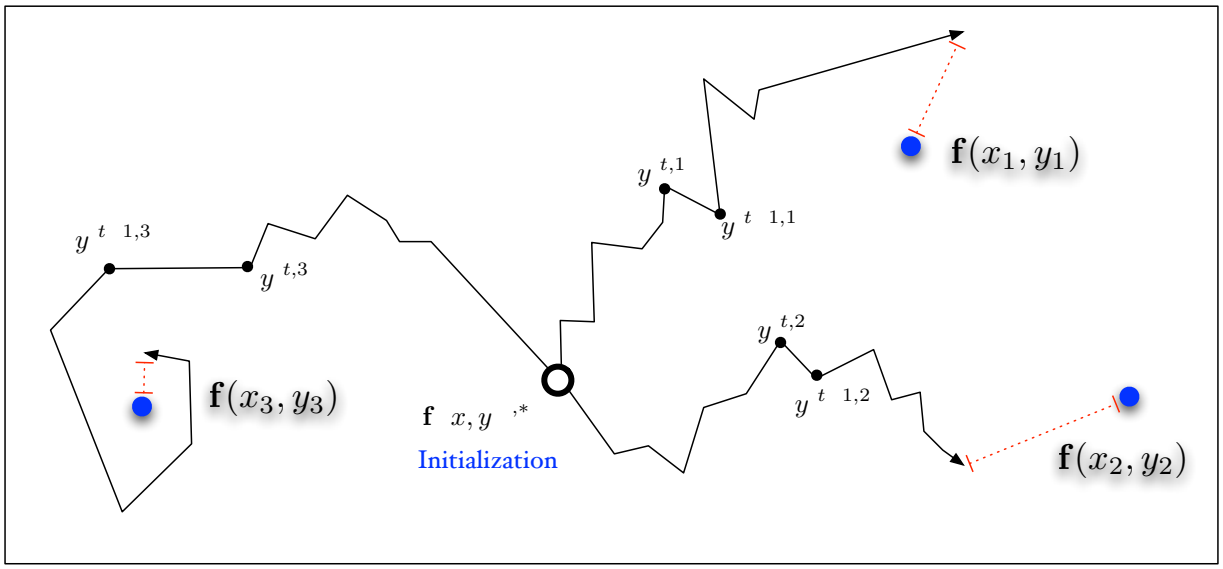

Figure 1: **Structured Nearest Neighbor Search:** The inference procedure for unlabeled test language x, when trained with three labeled languages, $(x_1, y_1), (x_2, y_2), (x_3, y_3)$. Our search procedure iteratively attempts to find labels for x which are as close as possible in feature space to each of the training languages. After convergence, the label which is closest in distance to a training language is predicted, in this case being the label near training language (x_3, y_3).

For example, if the (x, y) pair consists of the analyzed words {*kiss, kiss-es, hug*}, then the empirical distributions over stems, suffixes, and deletion rules would be:

- $P(t = kiss) = 2/3$
- $P(t = hug) = 1/3$
- $P(f = NULL) = 2/3$
- $P(f = -es) = 1/3$
- $P(d = NULL) = 1$

The three entropy features are defined as the shannon entropies of these stem, suffix, and deletion rule probabilities: $H(t), H(f), H(d)$.[3]

Finally, we consider two simple *percentage* features: the percentage of words in x which according to y are left unsegmented (i.e. have the null suffix, 2/3 in the example above), and the percentage of segmented words which employ a deletion rule (0 in the example above). Thus, in total, our model employs 8 universal morphological features. All features are scaled to the unit interval and are assumed to have equal weight.

[3]Note that here and throughout the paper, we operate over word types, ignoring their corpus frequencies.

3.1 Search Algorithm

The main algorithmic challenge for our model lies in efficiently computing the best morphological analysis y for each language-specific word set x, according to Equation 1. Exhaustive search through the set of all possible morphological analyses is impossible, as the number of such analyses grows exponentially in the size of the vocabulary. Instead, we develop a greedy search algorithm in the following fashion (the search procedure is visually depicted in Figure 1).

At each time-step t, we maintain a set of frontier analyses $\left\{y^{(t,\ell)}\right\}_\ell$, where ℓ ranges over the training languages. The goal is to iteratively modify each of these frontier analyses $y^{(t,\ell)} \rightarrow y^{(t+1,\ell)}$ so that the location of the training language in universal feature space — $\mathbf{f}\left(x, y^{(t+1,\ell)}\right)$ — is as close as possible to the location of the training language ℓ: $\mathbf{f}\left(x_\ell, y_\ell\right)$.

After iterating this procedure to convergence, we are left with a set of analyses $\left\{y^{(\ell)}\right\}_\ell$, each of which approximates the analyses which yield minimal distances to a particular training language:

$$y^{(\ell)} \approx \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \parallel \mathbf{f}(x, y) - \mathbf{f}(x_\ell, y_\ell) \parallel .$$

We finally select from amongst these analyses and

make our prediction:

$$\ell^* = \operatorname*{argmin}_{\ell} \parallel \mathbf{f}(x, y^{(\ell)}) - \mathbf{f}(x_\ell, y_\ell) \parallel$$
$$y^* = y^{(\ell^*)}$$

The main outline of our search algorithm is based on the MDL-based greedy search heuristic developed and studied by (Goldsmith, 2005). At a high level, this search procedure alternates between individual analyses of words (keeping the set of stems and suffixes fixed), aggregate discoveries of new stems (keeping the suffixes fixed), and aggregate discoveries of new suffixes (keeping stems fixed). As input, we consider the test words x in our new language, and we run the search in parallel for each training language (x_ℓ, y_ℓ). For each such test-train language pair, the search consists of the following stages:

Stage 0: Initialization

We initially analyze each word $w \in x$ according to peaks in *successor frequency*.[4] If w's n-character prefix $w_{:n}$ has successor frequency > 1 and the surrounding prefixes, $w_{:n-1}$ and $w_{:n+1}$ both have successor frequency $= 1$, then we analyze w as a stem-suffix pair: $(w_{:n}, w_{n+1:})$.[5] Otherwise, we initialize w as an unsuffixed stem. As this procedure tends to produce an overly large set of suffixes F, we further prune F down to the number of suffixes found in the training language, retaining those which appear with the largest number of stems. This initialization stage is carried out once, and afterwards the following three stages are repeated until convergence.

Stage 1: Reanalyze each word

In this stage, we reanalyze each word (in random order). We use the set of stems T and suffixes F obtained from the previous stage, and don't permit the addition of any new items to these lists. Instead, we focus on obtaining better analyses of each word, while also building up a set of phonological deletion rules D. For each word $w \in x$, we consider all possible segmentations of w into a stem-suffix pair (t, f), for which $f \in F$, and where either $t \in T$ or some $t' \in T$ such that t is obtained from t' using a deletion rule d (e.g. by deleting a final or penultimate vowel). For each such possible analysis y', we compute the resulting location in feature space $\mathbf{f}(x, y')$, and select the analysis that brings us closest to our target training language: $y = \operatorname{argmin}_{y'} \parallel \mathbf{f}(x, y') - \mathbf{f}(x_\ell, y_\ell) \parallel$.

Stage 2: Find New Stems

In this stage, we keep our set of suffixes F and deletion rules D from the previous stage fixed, and attempt to find new stems to add to T through an aggregate analysis of unsegmented words. For every string s, we consider the set of words which are currently unsegmented, and can be analyzed as a stem-suffix pair (s, f) for some existing suffix $f \in F$, and some deletion rule $d \in D$. We then consider the joint segmentation of these words into a new stem s, and their respective suffixes. As before, we choose the segmentation if it brings us closer in feature space to our target training language.

Stage 3: Find New Suffixes

This stage is exactly analogous to the previous stage, except we now fix the set of stems T and seek to find new suffixes.

3.2 A Monolingual Supervised Model

In order to provide a plausible upper bound on performance, we also formulate a supervised monolingual morphological model, using the structured perceptron framework (Collins, 2002). Here we assume that we are given some training sequence of inputs and morphological analyses (all within one language): $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We define each input x_i to be a noun w, along with a morphological tag z, which specifies the gender, case, and number of the noun. The goal is to predict the correct segmentation of w into stem, suffix, and phonological deletion rule: $y_i = (t, f, d)$.[6]

To do so, we define a feature function over input-label pairs, (x, y), with the following binary feature templates: (1) According to label y_i, the stem is t

[4] The *successor frequency* of a string prefix s is defined as the number of unique characters that occur immediately after s in the vocabulary.

[5] With the restriction that at this stage we only allow suffixes up to length 5, and stems of at least length 3.

[6] While the assumption of the correct morphological tag as input is somewhat unrealistic, this model still gives us a strong upper bound on how well we can expect our unsupervised model to perform.

	Type Counts				Entropy			Percentage	
	# words	# stems	# suffs	# dels	stem entropy	suff entropy	del entropy	unseg	deleted
BG	4833	3112	21	8	11.4	2.7	0.9	.45	.29
CS	5836	3366	28	12	11.5	3.2	1.6	.38	.53
EN	4178	3453	3	1	11.7	1.0	0.1	.73	.06
ET	6371	3742	141	5	11.5	5.0	0.2	.31	.04
HU	8051	3746	231	7	11.3	5.8	0.5	.23	.11
RO	5578	3297	23	8	11.5	2.9	1.4	.48	.51
SL	6111	3172	32	6	11.3	3.2	1.5	.33	.56
SR	5849	3178	28	5	11.4	2.9	1.4	.33	.53

Table 1: Corpus statistics for the eight languages. The first four columns give the number of unique word, stem, suffix, and phonological deletion rule types. The next three columns give, respectively, the entropies of the distributions of stems, suffixes (including $NULL$), and deletion rules (including $NULL$) over word types. The final two columns give, respectively, the percentage of word types occurring with the $NULL$ suffix, and the number of non-$NULL$ suffix words which use a phonological deletion rule. Note that the final eight columns define the universal feature space used by our model. BG = Bulgarian, CS = Czech, EN = English, ET = Estonian, HU = Hungarian, RO = Romanian, SL = Slovene, SR = Serbian

(one feature for each possible stem). (2) According to label y_i, the suffix and deletion rule are (f, d) (one feature for every possible pair of deletion rules and suffixes). (3) According to label y_i and morphological tag z, the suffix, deletion rule, and gender are respectively (f, d, G). (4) According to label y_i and morphological tag z, the suffix, deletion rule, and case are (f, d, C). (5) According to label y_i and morphological tag z, the suffix, deletion rule, and number are (f, d, N).

We train a set of linear weights on our features using the averaged structured perceptron algorithm (Collins, 2002).

4 Experiments

In this section we turn to experimental findings to provide empirical support for our proposed framework.

Corpus: To test our cross-lingual model, we apply it to a morphologically analyzed corpus of eight languages (Erjavec, 2004). The corpus includes a roughly 100,000 word English text, Orwell's novel "Nineteen Eighty Four," and its translation into seven languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, and Serbian. All the words in the corpus are tagged with morphological stems and a detailed morpho-syntactic analysis. Although the texts are parallel, we note that parallelism is nowhere assumed nor exploited by our model. See Table 1 for a summary of relevant corpus statistics. As indicated in the table, the raw number of nominal word types varies quite a bit across the languages, almost doubling from 4,178 (English) to 8,051 (Hungarian). In contrast, the number of stems appearing within these words is relatively stable across languages, ranging from a minimum of 3,112 (Bulgarian) to a maximum of 3,746 (Hungarian), an increase of just 20%.

In contrast, the number of suffixes across the languages varies quite a bit. Hungarian and Estonian, both Uralic languages with very complex nominal morphology, use 231 and 141 nominal suffixes, respectively. Besides English, the remaining languages employ between 21 and 32 suffixes, and English is the outlier in the other direction, with just three nominal inflectional suffixes.

Baselines and Results: As our unsupervised monolingual baseline, we use the Linguistica program (Goldsmith, 2001; Goldsmith, 2005). We apply Linguistica's default settings, and run the "suffix prediction" option. Our model's search procedure closely mirrors the one used by Linguistica, with the crucial difference that instead of attempting to greedily minimize description length, our algorithm instead tries to find the analysis as close as possible in the universal feature space to that of another language.

To apply our model, we treat each of the eight

	Linguistica	Our Model						Supervised
		Nearest Neighbor		Self (oracle)		Avg.		
		Accuracy	Distance	Accuracy	Distance	Accuracy	Distance	
BG	68.7	84.0 (RO)	0.13	88.7	0.03	68.6	3.90	94.7
CS	60.4	82.8 (BG)	0.40	84.5	0.03	66.3	4.05	93.5
EN	81.1	75.8 (BG)	1.29	89.3	0.10	58.3	4.30	93.4
ET	51.2	66.6 (HU)	0.35	80.9	0.03	52.8	4.57	86.5
HU	64.5	69.3 (ET)	0.81	66.5	1.10	68.0	4.94	94.9
RO	65.6	71.0 (CS)	0.11	71.2	0.15	62.3	3.95	89.1
SL	61.1	82.8 (SR)	0.07	85.5	0.04	61.7	3.69	95.4
SR	64.2	79.1 (SL)	0.06	82.2	0.04	63.0	3.71	94.8
avg.	64.6	76.4	0.40	81.1	0.19	62.6	4.14	92.8

Table 2: **Prediction accuracy** over word types for the Linguistica baseline, our cross-lingual model, and the monolingual supervised perceptron model. For our model, we provide both prediction accuracy and resulting distance to the training language in three different scenarios: (i) **Nearest Neighbor:** The training languages include all seven other languages in our data set, and the predictions with minimal distance to a training language are chosen (the nearest neighbor is indicated in parentheses). (ii) **Self (oracle):** Each language is trained to minimize the distance to its *own* gold-standard analysis. (iii) **Average:** The feature values of all seven training languages are averaged together to create a single objective.

languages in turn as the test language, with the other seven serving as training examples. For each test language, we iterate the search procedure for each training language (performed in parallel), until convergence. The number of required iterations varies from 6 to 36 (depending on the test-training language pair), and each iteration takes no more than 30 seconds of run-time on a 2.4GHz Intel Xeon E5620 processor. We also consider two variants of our method. In the first (**Self (oracle)**), we train each test language to minimize the distance to its *own* gold standard feature values. In the second variant (**Avg.**), we average the feature values of all seven training languages into a single objective. As a plausible upper bound on performance, we implemented the structured perceptron described in Section 3.2. For each language, we train the perceptron on a randomly selected set of 80% of the nouns, and test on the remaining 20%.

The prediction accuracy for all models is calculated as the fraction of word types with correctly predicted suffixes. See Table 2 for the results. For all languages other than English (which is a morphological loner in our group of languages), our model improves over the baseline by a substantial margin, yielding an average increase of 11.8 absolute percentage points, and a reduction in error relative to the supervised upper bound of 42%. Some of the most striking improvements are seen on Serbian and Slovene. These languages are closely related to one another, and indeed our model discovers that they are each others' nearest neighbors. By guiding their morphological analyses towards one another, our model achieves a 21 percentage point increase in the case of Slovene and a 15 percentage point increase in the case of Slovene.

Perhaps unsurprisingly, when each language's gold standard feature values are used as its *own* target (**Self (oracle)** in Table 2), performance increases even further, to an average of 81.1%. By the same token, the resulting distance in universal feature space between training and test analyses is cut in half under this variant, when compared to the non-oracular nearest neighbor method. The remaining errors may be due to limitations of the search procedure (i.e. getting caught in local minima), or to the coarseness of the feature space (i.e. incorrect analyses might map to the same feature values as the correct analysis). Finally, we note that minimizing the distance to the *average* feature values of the seven training languages (**Avg.** in Table 2) yields subpar performance and very large distances between between predicted analyses and target feature values (4.14 compared to 0.40 for nearest neighbor). This

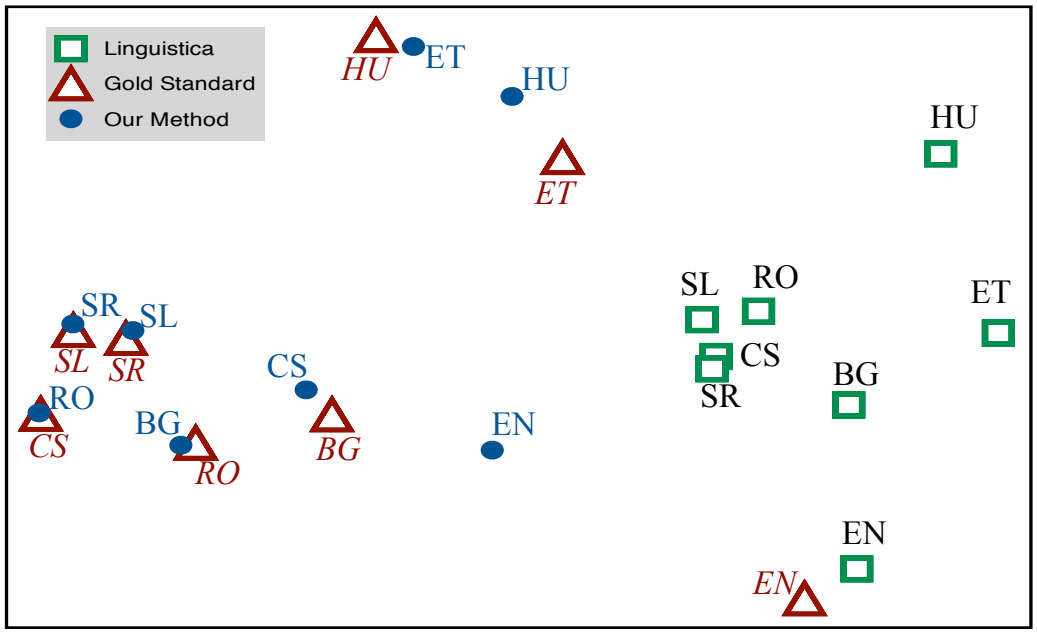

Figure 2: **Locations in Feature Space** of Linguistica predictions (green squares), gold standard analyses (red triangles), and our model's nearest neighbor predictions (blue circles). The original 8-dimensional feature space was reduced to two dimensions using Multidimensional Scaling.

result may indicate that the average feature point between training languages is simply unattainable as an analysis of a real lexicon of nouns.

Visualizing Locations in Feature Space: Besides assessing our method quantitatively, we can also visualize the the eight languages in universal feature space according to (i) their gold standard analyses, (ii) the predictions of our model and (iii) the predictions of Linguistica. To do so, we reduce the 8-dimensional features space down to two dimensions while preserving the distances between the predicted and gold standard feature vectors, using Multidimensional Scaling (MDS). The results of this analysis are shown in Figure 2. With the exception of English, our model's analyses lie closer in feature space to their gold standard counterparts than those of the baseline. It is interesting to note that Serbian and Slovene, which are very similar languages, have essentially swapped places under our model's analysis, as have Estonian and Hungarian (both highly inflected Uralic languages). English has (unfortunately) been pulled towards Bulgarian, the second least inflecting language in our set.

Learning Curves: We also measured the performance of our method as a function of the number of languages in the training set. For each target language, we consider all possible training sets of sizes ranging from 1 to 7 and select the predictions which bring our test language closest in distance to one of the languages in the set. We then average the resulting accuracy over all training sets of each size. Figure 3 shows the resulting learning curves averaged over all test languages (left), as well as broken down by test language (right). The overall trend is clear: as additional languages are added to the training set, test performance improves. In fact, with only one training language, our method performs worse (on average) than the Linguistica baseline. However, with two or more training languages available, our method achieves superior results.

Accuracy vs. Distance: We can gain some insight into these learning curves if we consider the relationship between accuracy (of the test language analysis) and distance to the training language (of the same predicted analysis). The more training languages available, the greater the chance that we can guide our test language into very close proximity to

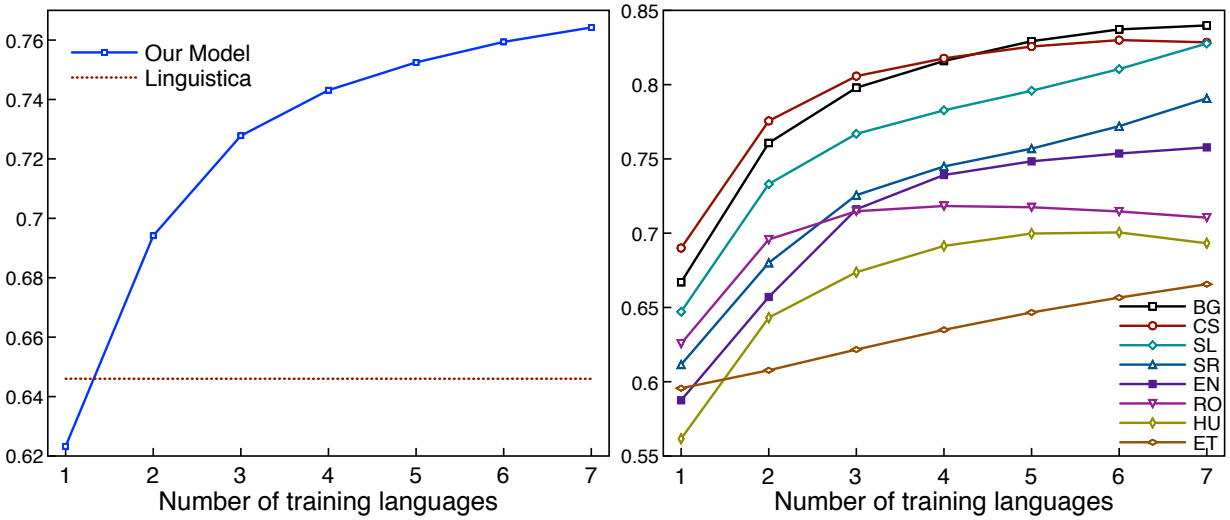

Figure 3: **Learning curves** for our model as the number of training languages increases. The figure on the left shows the average accuracy of all eight languages for increasingly larger training sets (results are averaged over all training sets of size 1,2,3,...). The dotted line indicates the average performance of the baseline. The figure on the right shows similar learning curves, broken down individually for each test language (see Figure 1 for language abbreviations).

one of them. It thus stands to reason that a strong (negative) correlation between distance and accuracy would lead to increased accuracy with larger training sets. In order to assess this correlation, we considered all 56 test-train language pairs and collected the resulting accuracy and distance for each pair. We separately scaled accuracy and distance to the unit interval for each test language (as some test languages are inherently more difficult than others). The resulting plot, shown in Figure 4, shows the expected correlation: When our test language can be guided very closely to the training language, the resulting predictions are likely to be good. If not, the predictions are likely to be bad.

5 Conclusions and Future Work

The approach presented in this paper recasts morphological induction as a structured prediction task. We assume the presence of morphologically labeled languages as *training examples* which guide the induction process for unlabeled test languages. We developed a novel structured nearest neighbor approach for this task, in which all languages and their morphological analyses lie in a universal feature space. The task of the learner is to search through the space of morphological analyses for the test language and return the result which lies closest to one

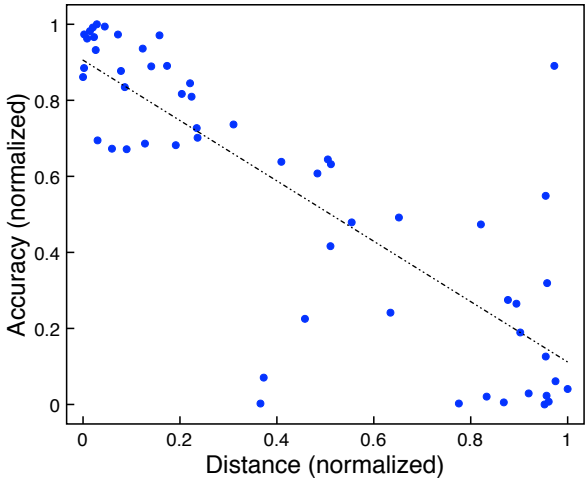

Figure 4: **Accuracy vs. Distance:** For all 56 possible test-train language pairs, we computed test accuracy along with resulting distance in universal feature space to the training language. Distance and accuracy are separately normalized to the unit interval for each test language, and all resulting points are plotted together. A line is fit to the points using least-squares regression.

of the training languages. Our empirical findings validate this approach: On a set of eight different languages, our method yields substantial accuracy gains over a traditional MDL-based approach in the task of nominal morphological induction.

One possible shortcoming of our approach is that it assumes a uniform weighting of the cross-lingual feature space. In fact, some features may be far more relevant than others in guiding our test language to an accurate analysis. In future work, we plan to integrate distance metric learning into our approach, allowing some features to be weighted more heavily than others. Besides potential gains in prediction accuracy, this approach may shed light on deeper relationships between languages than are otherwise apparent.

References

Meni Adler and Michael Elhadad. 2006. An unsupervised morpheme-based hmm for hebrew morphological disambiguation. In *Proceedings of the ACL/CONLL*, pages 665–672.

Emily M. Bender. 2009. Linguistically naïve != language independent: why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pages 26–32, Morristown, NJ, USA. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of the ACL*, pages 1288–1297, Uppsala, Sweden, July. Association for Computational Linguistics.

P. Blunsom, T. Cohn, and M. Osborne. 2009. Bayesian synchronous grammar induction. *Advances in Neural Information Processing Systems*, 21:161–168.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of CoNLL*.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of the NAACL/HLT*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*, pages 1–8.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Publications in Computer and Information Science Report A81, Helsinki University of Technology.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *Proceedings of the ACL*, pages 130–137.

Sajib Dasgupta and Vincent Ng. 2007. Unsupervised part-of-speech acquisition for resource-scarce languages. In *Proceedings of the EMNLP-CoNLL*, pages 218–227.

T. Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC*, volume 4, pages 1535–1538.

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

John Goldsmith. 2005. An algorithm for the unsupervised learning of morphology. Technical report, University of Chicago.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering*, 11(3):311–325.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36(1):341–385.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL*, pages 1161 – 1168.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 209–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX Workshop*

on Tagging Text with Lexical Semantics: Why, What, and How?, pages 79–86.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.

Benjamin Snyder and Regina Barzilay. 2008a. Cross-lingual propagation for morphological analysis. In *Proceedings of the AAAI*, pages 848–854.

Benjamin Snyder and Regina Barzilay. 2008b. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the ACL/HLT*, pages 737–745.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of EMNLP*, pages 1041–1050.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009a. Unsupervised multilingual grammar induction. In *Proceedings of the ACL*, pages 73–81.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2009b. Adding more languages improves unsupervised multilingual part-of-speech tagging: a Bayesian non-parametric approach. In *Proceedings of the NAACL*, pages 83–91.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the NAACL*, pages 1–8.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 207–216, Morristown, NJ, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2000. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, pages 161–168.