Supervised Learning of Complete Morphological Paradigms

Greg Durrett*
Computer Science Division
University of California, Berkeley
gdurrett@cs.berkeley.edu

John DeNero
Google, Inc.
denero@google.com

Abstract

We describe a supervised approach to predicting the set of all inflected forms of a lexical item. Our system automatically acquires the orthographic transformation rules of morphological paradigms from labeled examples, and then learns the contexts in which those transformations apply using a discriminative sequence model. Because our approach is completely data-driven and the model is trained on examples extracted from Wiktionary, our method can extend to new languages without change. Our end-to-end system is able to predict complete paradigms with 86.1% accuracy and individual inflected forms with 94.9% accuracy, averaged across three languages and two parts of speech.

1 Introduction

For natural languages with rich morphology, knowledge of how to inflect base forms is critical for both text generation and analysis. Hand-engineered, rule-based methods for predicting inflections can offer extremely high accuracy, but they are laborious to construct and do not exist with full lexical coverage in all languages. By contrast, a large number of example inflections are freely available in a semi-structured format on the Web. The English Wiktionary[1] is a crowd-sourced lexical resource that includes complete inflection tables for many lexical items in many languages. We present a supervised system that, given only data from Wiktionary, automatically discovers and learns to apply the orthographic transformations governing a language's inflectional morphology.[2]

Our data-driven approach is designed to extend to any language for which we have a substantial number of example inflection tables. The design of our model is guided by three structural assumptions:

1. The inflections of many lexical items are governed by a few repeated morphological paradigms.
2. A morphological paradigm can be decomposed into independent orthographic transformation rules (including prefix, suffix, and stem changes), which are triggered by orthographic context.
3. A base form is transformed in consistent, correlated ways to produce its inflected variants.

Learning proceeds in two stages that both utilize the same training set of labeled inflection tables. First, an inventory of interpretable transformation rules is generated by aligning each base form to all of its inflected forms. Second, a semi-Markov conditional random field (CRF) (Sarawagi and Cohen, 2004) is trained to apply these rules correctly to unseen base forms. As we demonstrate experimentally, the CRF is most effective when jointly predicting all inflected forms of a lexical item together, forcing the system to adopt a single consistent analysis of each base form.

*Research conducted during an internship at Google.

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

[2] See http://eecs.berkeley.edu/~gdurrett for our datasets and code.

Previous work has also described supervised and semi-supervised approaches to predicting inflectional morphology (Yarowsky and Wicentowski, 2000; Wicentowski, 2004; Dreyer and Eisner, 2011). Our approach differs primarily in its use of automatically extracted morphological rules and our discriminative prediction method which jointly models entire inflection tables. These modeling choices are directly inspired by the data setting: Wiktionary contains complete inflection tables for many lexical items in each of a large number of languages, so it is natural to make full use of this information with a joint model of all inflected forms.

We evaluate our predictions on held-out Wiktionary inflection tables for three languages and two parts of speech. Our language-independent method predicts inflections for unseen base forms with accuracies ranging from 88.9% (German nouns) to 99.7% (Spanish verbs). For comparability with previous work, we also evaluate our approach on German verb forms in the CELEX lexical database (Baayen et al., 1995). Our approach outperforms the semi-supervised hierarchical Bayesian model of Dreyer and Eisner (2011), while employing scalable exact inference and interpretable transformation rules.

2 Background: Inflectional Morphology

Among the valid words W and parts of speech P in a language, the base forms $B \subset W \times P$ are the canonical forms of the language's lexical items. A base form relates to an inflected form via an inflectional relation (b, w, a), where $b \in B$ is a base form, $w \in W$ is the inflected form, and a is a vector of morphological attributes. An inflection table $T(b)$ is the set of all such relations for a base form b.

Two partial inflection tables are shown in Table 1, for the base forms (infinitives) of the German verbs *machen* and *schleichen*, containing such inflectional relations as (*machen*, *mache*, [1P,PRES,SING]) and (*machen*, *gemacht*, [PAST PART.]). Only a small sample of the valid attribute combinations are shown; a full inflection table for a German verb in our Wiktionary dataset contains 27 relations.

The goal of this paper is to learn how to map b to $T(b)$. We generate candidate inflection tables by applying compact, interpretable orthographic transformation rules that have been extracted from example tables. As an example of our rule application process, to inflect *machen* appropriately in the forms listed in Table 1, one could apply the following rules:

INFINITIVE	machen	schleichen
1P,PRES,SING	mache	schleiche
2P,PRES,SING	machst	schleichst
3P,PRES,SING	macht	schleicht
PAST PART.	gemacht	geschlichen
...	...	...

Table 1: Two partial inflection tables for the German verbs *machen* (to make) and *schleichen* (to crawl).

1. Replace a suffix *-en* with *-e* for first person, *-st* for second person, *-t* for third person, and *-t* for the past participle.

2. Add a prefix *ge-* for the past participle.

To inflect *schleichen*, one could apply a larger set of three rules:

1. Replace a suffix *-en* with *-e* for first person, *-st* for second person, *-t* for third person, and *-en* for the past participle.

2. Add a prefix *ge-* for the past participle.

3. Delete the first *e* for the past participle.

The inflection tables of other German verbs can be generated using precisely the same rules above, and different inflection patterns may share rules, such as the repeated rule 2. This example illustrates one of our chief assumptions, that the inflections of many base forms can be modeled with a small number of such rules, applied in various combinations.

3 Learning Transformation Rules

From a training set of inflection tables $\{T(b_1), ..., T(b_n)\}$, our system learns a set of orthographic transformation rules. A rule is a function $R : s, a \rightarrow s'$ that takes as input a substring s of a base form and an attribute vector a and outputs a replacement substring s'. The suffix transformation from Section 2 for *machen* can be described using a

Algorithm 1 Learning rules from examples.

Input: n training instances $T(b_1), \ldots, T(b_n)$
Rule set $\mathcal{R} \leftarrow \{\}$
for $i \leftarrow 1$ to n **do**
  Changed source spans $C \leftarrow \{\}$
  for all $a \in A$ **do**
    $C_a \leftarrow \textsc{ProjectSpans}(\textsc{Align}(b_i, T_a(b_i)))$
    $C \leftarrow \textsc{UnionSpans}(C, C_a)$
  end for
  for all $c \in C$ **do**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{\textsc{ExtractRule}(c)\}$
  end for
end for
return $\mathcal{R}$

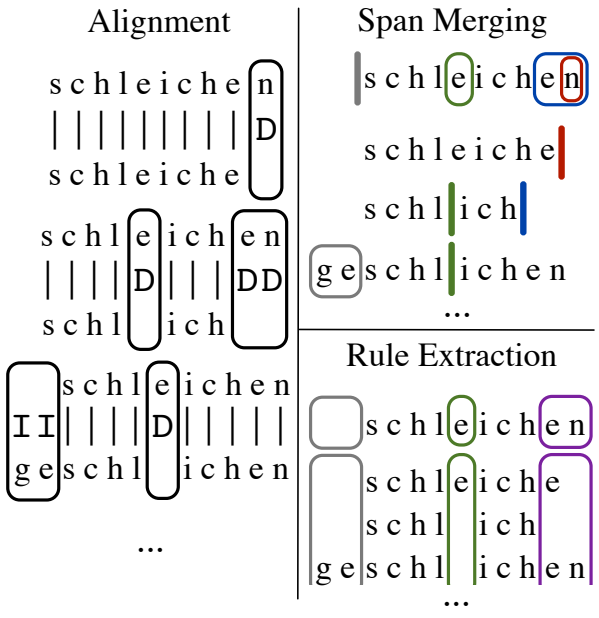

Figure 1: Demonstration of the rule extraction algorithm with the base form *schleichen* and three inflected forms: *schleiche* (first person singular present), *schlich* (first person singular past), and *geschlichen* (past participle). We ideally want to extract appropriate transformation rules like those described in Section 2. In the alignment step, we minimize the edit distance between each inflected form and the base form to identify changed spans. In the span merging step, we project changes onto the base form and take the union of adjacent or overlapping spans. In the rule extraction step, we project these spans back onto the inflected forms to identify transformation rules.

rule with four entries:

$$R(en, [1\text{P,PRES,SING}]) = e$$
$$R(en, [2\text{P,PRES,SING}]) = st$$
$$R(en, [3\text{P,PRES,SING}]) = t$$
$$R(en, [\text{PAST PART.}]) = t$$

Our method for learning rules from examples is described in Algorithm 1 and depicted in Figure 1. We extract rules from each observed inflection table $T(b_i)$ independently, and the final set of rules is simply the union of the sets of rules learned from each example. The procedure for a single inflection table has three steps:

Alignment: Align each inflected form to the base form with an iterated edit-distance algorithm.

Span Merging: Extract the set of spans of the base form that changed to produce the inflected form, and take their union across all attribute vectors to identify maximal changed spans.

Rule Extraction: Extract a rule for each maximal changed span.

Alignment. For each setting of attributes a, we find the lowest-cost transformation of the base form b into the corresponding inflected form $T_a(b)$ using single-character insertions, deletions, and substitutions. This minimum edit distance calculation is computed via the following recurrence, where i is an index into the base form b and j is an index into an inflected form $T_a(b)$:

$$L(i,j) = \min\{L(i, j-1) + I,$$
$$L(i-1, j) + D,$$
$$L(i-1, j-1) + S(i,j)\}$$

I, D, and S are insertion, deletion, and substitution costs, respectively. Tracing the computation of $L(\text{len}(b), \text{len}(T_a(b)))$ yields an optimal sequence of edit operations. The alignments output by this procedure are depicted in the first panel of Figure 1.

The most typical cost scheme sets $I = 1$, $D = 1$, and $S(i,j) = (1 - \mathbb{I}[\text{match}(i,j)])$, i.e. 0 if the ith character of b is the same as the jth character of $T_a(b)$, and 1 otherwise. However, this cost scheme did not yield intuitive alignments for some of our training instances. For example, in the case of the verb *denken* aligning to its past participle *gedacht*,

the initial d and g will be aligned and the following e's will be aligned, preventing the algorithm from recognizing the addition of the prefix *ge-*. To solve this problem, we use a dynamic edit distance cost scheme in which I, D, and unmatched substitutions all have a cost of 0. Matched substitutions have a negative cost $-c_i$, where i is the index in the base form and c_i is the number of other inflected forms for which i is aligned to a matching character. The inflected forms are iteratively realigned with the base form until the c_i converge (Eisner, 2002; Oncina and Sebban, 2006). This cost scheme encourages a single consistent analysis of the base form as it aligns to all of its inflected forms.

Span Merging. From each aligned pair of words, the PROJECTSPANS procedure identifies sequences of character edit operations with contiguous spans of the base form. We construct a set of changed spans C_a of b as follows: include the span (i, j) if and only if no characters between i and j were aligned to matching characters in $T_a(b)$ and no smaller span captures the same set of changes. Projected spans for the inflected forms of *schleichen* are shown in the "Span Merging" panel of Figure 1.

The UNIONSPANS procedure combines two sets of spans by iteratively merging any two spans that are overlapping or adjacent. Repeating this procedure to accumulate spans for each setting of a yields the set C of maximal changed spans for a base form. Any span in C is bordered either by word boundaries or by characters that are match-aligned in every inflected form, meaning that we have isolated a region characterized by a particular orthographic transformation.

Rule Extraction. The final step of Algorithm 1 extracts one rule for each maximal changed span of the base form. The Rule Extraction panel of Figure 1 depicts how maximal changed spans in the base form correspond to transformation rules. Because UNIONSPANS guarantees that match-aligned characters border each maximal changed span, there is no ambiguity about the segmentation of transformations. The EXTRACTRULE procedure produces one rule $R(s, a)$ corresponding to each changed span.

Table 2 contains examples of the transformation rules we extract from German verbs. The extracted rules are interpretable descriptions of common inflection patterns.

Attributes	Suffix				Stem	Pre.
INFINITIVE	en	en	en	n	e	
1P,PRES,SING	e	e	e	e	e	
1P,PAST,SING	te	te		te		
2P,PRES,SING	st	t	st	st	e	
2P,PAST,SING	test	test	st	test		
3P,PRES,SING	t	t	t	t	e	
3P,PAST,SING	te	te		te		
PAST PART.	t	t	en	t		ge
...	...	...	...	...	...	...
Label	$R_{\text{suf},1}$	$R_{\text{suf},2}$	$R_{\text{suf},3}$	$R_{\text{suf},4}$	$R_{\text{st},1}$	$R_{\text{pre},1}$

Table 2: Each column is an example of a morphological transformation rule extracted by our approach. The first four are suffix changes; these apply to, in order, regular verbs such as *machen*, verbs ending in *-zen* or *-sen* such as *setzen*, verbs such as *schleichen* and *beheben*, and verbs ending in *-ern* or *-eln* such as *sprenkeln*. The stem change occurs in strong verbs of the first class such as *schleichen*, *greifen*, and *streiten*. Finally, we learn that *ge-* can be added as a prefix to indicate the past participle.

4 Applying Transformation Rules

For a novel base form b, the inventory of learned transformation rules $\mathcal{R} = \{R(s, a)\}$ can typically generate many candidate inflection tables $T(b)$ for us to choose between. A rule can potentially apply to a base form in a number of places; we define an anchored rule $A = (R, i, j, b)$ to be the application of R to a span (i, j) in b. A is only a valid anchoring if the substring of b between i and j matches the input of rule R.

Given a set $\mathcal{A}$ of non-overlapping anchored rules for b, each entry of $T(b)$ can be deterministically produced by rewriting each anchored rule's span (i, j) using the rule R. Therefore, the task of predicting $T(b)$ is equivalent to selecting a coherent subset $\mathcal{A}$ of anchored rules from the set of all possible anchored rules for this base form. By coherent, we mean that the selected rules are anchored to non-overlapping, non-adjacent[3] spans of b. Figure 2a shows two coherent anchored rule subsets for *schleichen* (the top one being correct). Underlining indi-

[3]During rule extraction, any adjacent changed spans are merged into a single rule. Disallowing adjacent spans here therefore prevents us from synthesizing new rules.

cates length-one spans $S = (i, i+1, b)$ that are not part of any anchored rule in $\mathcal{A}$. We denote the set of such spans by $\mathcal{S}(\mathcal{A})$; this set is uniquely defined for the given base form by the selected anchored rules.

We use a log-linear model to place a conditional distribution over valid anchored rule subsets $\mathcal{A}$ given the base form b:

$$p_w(\mathcal{A}|b) \propto \exp w^T \left(\sum_{A \in \mathcal{A}} \phi(A) + \sum_{S \in \mathcal{S}(\mathcal{A})} \psi(S) \right)$$

where w is a weight vector, $\phi(A)$ computes a feature vector for anchored rule A, and $\psi(S)$ computes a feature vector for preserved spans S. We train this model to maximize the regularized conditional log-likelihood of the training data, which consists of base forms b_i and gold subsets of anchored rules $\mathcal{A}_i^*$ derived using Algorithm 1 on the gold inflection tables.

$$L(w) = \sum_{i=1}^{n} \log p(\mathcal{A}_i^* | b_i) + \frac{\gamma}{2} \|w\|^2.$$

We find $w^* = \arg\max_w L(w)$ using L-BFGS (Liu and Nocedal, 1989), which requires computing $\frac{\partial L}{\partial w}$. This gradient takes the standard form of the difference between gold feature counts and expected feature counts under the model:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{n} \left[\left(\sum_{A \in \mathcal{A}_i^*} \phi(A) + \sum_{S \in \mathcal{S}(\mathcal{A}_i^*)} \psi(S) \right) - \left(\sum_{A \in \mathcal{A}(\mathcal{R}, b)} \mathbb{E}_{p_w} \phi(A) + \sum_{S \in \mathcal{S}(b)} \mathbb{E}_{p_w} \psi(S) \right) \right] - \gamma w$$

where, by a slight abuse of notation, $\mathcal{S}(b)$ is the set of all length-one spans of b.

In general, the normalizer of p_w and the expectation over p_w cannot be computed directly, since there may be exponentially many coherent subsets of anchored rules. However, we note that $\mathcal{A}$ and its corresponding $\mathcal{S}(\mathcal{A})$ form a segmentation of the base form b, with features decomposing over individual segments. Our model can therefore be viewed a semi-Markov model over b (Sarawagi and Cohen, 2004); more precisely, a zeroth-order semi-Markov model, since we do not include features on state transitions. At training time, we can use the forward-backward algorithm for semi-Markov models to compute the gradient of p_w, and at test time, the Viterbi algorithm can exactly find the best rule subset under the model: $\hat{\mathcal{A}} = \arg\max_{\mathcal{A}} p_w(\mathcal{A}|b)$.

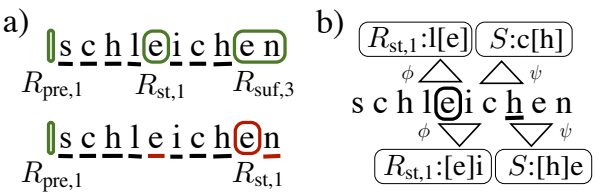

Figure 2: a) Two possible anchored rule sets for *schleichen*. The indicated rules are prefix, stem, and suffix rules as found in Table 2. The top anchoring is correct, while the bottom misplaces the stem change and does not include a suffix change. Underlined letters indicate preserved spans S. b) Bigram context features computed by $\phi(R_{st,1})$, where the stem change is applied to the highlighted *e*, and similar features computed by $\psi(S)$ for the underlined *h*, which is unchanged by the applied rules.

Features. The feature function ϕ captures contextual information in the base form surrounding the site of the anchored rule application. It is well understood that different morphological rules may require examining different amounts of context to apply correctly (Kohonen, 1986; Torkkola, 1993; Shalonova and Golénia, 2010); to this end, we will use local character n-gram features, which have been successfully applied to related problems (Jiampojamarn et al., 2008; Dinu et al., 2012).

A sketch of our feature computation scheme is shown in Figure 2b. Our basic feature template is an indicator on a character n-gram with some offset from the rule application site, conjoined with the identity of the rule R being applied. Our features look at variable amounts of context: we include features on unigrams through 4-grams, starting up to five letters behind the anchored rule span and ending up to five letters past the anchored rule span. These features can model most hand-coded morphological rules, but are in many cases more numerous than necessary. However, we find that regularization is effective at balancing high model capacity with generalization, and reducing the size of the feature set empirically harms overall accuracy.

We also employ factored features that only look at predictions over particular inflected forms; these are

coarser features that are shared between two rules when they predict the same orthographic change for a particular setting of attributes. These features are indicators on R_a (the restriction of R to attributes a), the context n-gram, and its offset from the span.

The feature function ψ is almost identical to ϕ, but instead of indicating a rule appearing in some context, it instead indicates that a particular length-one span is being preserved in its n-gram context. Examples of ψ features are shown in Figure 2b.

Pruning. Thus far, the only requirement on an anchoring A is that the source side of its rule R must match the span it is anchored to in the base form b. We further filter the set of possible A as follows: if every occurrence of R in the training set is preceded by the same character (including a start-of-word character) or followed by the same character (including an end-of-word character), any anchoring A must be preceded or followed accordingly. This stipulation is most useful in restricting prefixing or suffixing insertions, which have an empty source side, to apply only at the beginnings or ends of base forms (rather than at arbitrary points throughout). In doing so, we prune out many erroneous anchored rules and speed up inference substantially without prohibiting correct rule applications.

5 Wiktionary Morphology Data

Our primary source of supervised inflection table data is English Wiktionary. The collective editors of English Wiktionary have created complete, consistent inflection tables for many lexical items in many languages. Previous work has successfully parsed other information from Wiktionary, such as parts of speech, glosses, and etymology (Zesch et al., 2008; Li et al., 2012); however, to our knowledge, inflection tables have not previously been extracted in a format easily amenable to natural language processing applications. These inflection tables are challenging to extract because the layout of tables varies substantially by language (beyond the expected changes due to differing sets of relevant morphological attributes), and some tables contain annotations in addition to word forms.

In order to extract this data, we built a Wiktionary scraper which generates fully structured output by interpreting the templates that generate the rendered inflection tables. Table 3 gives statistics for the number of base forms and inflected forms extracted from Wikitionary. When multiple forms were listed in an inflection table for the same base form and attribute vector, we selected the first in linear order; applying the same principle, we also kept only the first inflection table when more than one was listed for a given base form. Furthermore, base forms and inflected forms separated by spaces, hyphens, or colons were discarded. As a result, we discarded German verb-preposition compounds such as *ablehnen*[4] and Spanish reflexives such as *lavarse*.

Lang/POS	Base forms	Infl. forms per base
DE-NOUNS	2764	8
DE-VERBS	2027	27
ES-VERBS	4055	57
FI-NOUNS	40589	28
FI-VERBS	7249	53

Table 3: Number of full morphology tables extracted from Wiktionary for each language and part of speech pair that we considered, as well as the number of inflected forms associated with each base form.

6 Experiments

We evaluate our model under two experimental conditions. First, we use the German verb lexicon in the CELEX lexical database (Baayen et al., 1995) with the same train/test splits as Dreyer and Eisner (2011). Second, we train on our Wiktionary data described in Section 5 and evaluate on held-out forms from this same dataset.

In each case, we evaluate two variants of our model in order to examine the importance of jointly modeling the production of the entire inflection table. Our JOINT model is exactly as defined in Section 4. For our FACTORED model, the dictionary of rules is extracted separately for each setting of the attributes a; i.e., we run the entire procedure in Section 3 with only one inflected form at a time and forego the UNIONSPANS step. A separate prediction model is trained for each a and so features are not shared across multiple predictions as they are in the JOINT case. Note that this FACTORED approach

[4]This class of verbs was also ignored by Dreyer and Eisner (2011).

	No. of training examples		
	50	100	200
NAÏVE	87.61	87.70	87.70
FACTORED	89.61	91.40	92.64
JOINT	**90.47**	**92.31**	**93.18**
DE11	89.9	91.5	
DE11+CORPUS	90.9	92.2	
ORACLE	*95.47*	*96.09*	*96.77*

Table 4: Accuracies on reconstructing individual inflected forms in CELEX, averaged over the 5415 inflection tables in each of 10 test sets. Three training set sizes are reported. DE11 indicates a reported result from Dreyer and Eisner (2011), with blank results unreported in that work. Our FACTORED model is able to do approximately as well as the DE11 baseline method, and our JOINT model performs better yet, performing comparably to DE11+CORPUS, which uses additional monolingual text. All models substantially outperform the NAÏVE suffixing baseline. The relatively low ORACLE accuracy indicates that some errors arise from failing to apply rules that are not attested in these small training sets.

can produce inflection tables that the JOINT model cannot, due to its ability to "mix and match" orthographic changes in the same inflection table.

We also evaluate a NAÏVE method for applying the joint rules which selects the most common suffix rule available after pruning.[5] Finally, we report the ORACLE accuracy attainable with the morphological rule dictionary of the JOINT model.

For our conditional likelihood objective, we use $\gamma = 0.0002$; this parameter and the feature set were tuned on a small development set and held fixed for all experiments.

6.1 CELEX Experiments

Dreyer and Eisner (2011) construct ten train/test splits of the 5615 German verb forms in the CELEX lexical database, keeping 200 forms for training in each case, which they further subsample. These random splits serve to control for instability due to the small training set sizes. Each infinitive verb form has 22 corresponding inflected forms capturing variation such as person, number, mood, and tense.

Table 4 shows our results compared to those of Dreyer and Eisner (2011). The FACTORED model performs on par with the DE11 baseline model, but the stronger performance of the JOINT model indicates that making joint predictions is important. With 100 training examples, our model is able to equal the performance of DE11+CORPUS, which additionally uses ten million tokens of monolingual German text.

We emphasize that this is not the data condition for which our model was designed. It is unfavorable for two reasons: first, feature-rich models can be learned more stably on larger training sets, and second, the train/test splits are chosen randomly, and therefore the test sets may contain completely irregular verbs using morphological rules that we have never observed. As can be seen from the ORACLE results in Table 4, a substantial fraction of the missed test examples cannot be produced using our extracted rules simply because we have not seen the relevant examples; in many cases, even a human could not generalize correctly from the given examples without exploiting external knowledge of the German language.

6.2 Held-Out Wiktionary Data

Our algorithm was designed with the fundamental assumption that the training set should be a comprehensive description of the morphology of a given language, which is not true for the CELEX data. In order to evaluate on a broader set of languages under these training conditions, we turn to our Wiktionary data. For each language and part of speech, we train on all but 400 inflection tables, holding back 200 examples as a development set and 200 examples as a blind test set.[6] The forms selected for the development and test data were purposely chosen not to be among the 200 most frequently occurring forms in the language, since these common cases can be easily memorized from Wiktionary.

Results are shown in Table 5. As with the CELEX results, we see that the joint prediction improves accuracy over the factored model, obtaining a 9% error reduction on individual forms and a 35% error reduction on exact match. The more pronounced

[5] For example, for German verbs ending in *-en*, this applies the most regular *-en* suffix change, that exhibited by *machen* and many other verbs.

[6] For Finnish nouns, because there were so many inflection tables, we trained only on the first 6000 examples. Using more examples did not significantly change performance.

	Exact table match				Individual form accuracy			
Lang/POS	NAÏVE	FACT.	JOINT	ORACLE	NAÏVE	FACT.	JOINT	ORACLE
DE-VERBS	42.0	74.5	**85.0**	*99.5*	89.13	94.76	**96.19**	*99.98*
DE-NOUNS	12.0	74.0	**79.5**	*98.5*	49.06	88.31	**88.94**	*99.25*
ES-VERBS	81.5	93.5	**95.0**	*99.5*	97.20	99.61	**99.67**	*99.99*
FI-VERBS	33.5	82.0	**87.5**	*99.5*	75.32	**97.23**	96.43	*99.86*
FI-NOUNS	31.0	69.0	**83.5**	*100.0*	61.23	92.14	**93.41**	*100.00*
AVG	40.0	78.6	**86.1**	*99.4*	74.39	94.41	**94.93**	*99.81*

Table 5: Accuracies on reconstructing complete inflection tables and individual inflected forms for held-out base forms in our Wiktionary dataset. Results are shown for our fully JOINT model, a FACTORED model that predicts individual inflected forms independently, a NAÏVE baseline that picks the most common applicable suffix rule, and an ORACLE that selects the best inflection table within our model's capacity. For each language and part of speech, regardless of training set size, evaluation is based on a blind test set of 200 held-out forms.

improvement on exact match is unsurprising, since we expect that the joint predictions should get inflection tables correct in an "all-or-nothing" fashion, whereas factored predictions are more likely to reflect divergent feature weights of the different component models. The NAÏVE baseline performs rather poorly overall, indicating our algorithm is being sophisticated about applying more than just the most common changes. Finally, we note that the ORACLE performance is much higher in this case than on the CELEX data, confirming our intuition that with the appropriate level of supervision our model at least has the capacity to make correct predictions in almost every case.

6.3 Error Analysis

We conducted an error analysis on the output of our JOINT model on German nouns. From 2364 paradigms, we learn 53 different orthographic transformation rules, of which our 200-example development set exhibits 14.[7]

On our development set, 196 inflection tables are within the capacity of our model. Of those 196, 159 are exactly correct. In Table 6, we show the top six rules by frequency in the development set, along with the precision, recall, and F-measure that our model attains for each rule.[8] These rules are mostly interpretable: for example, the first two columns correspond to common suffix rules for feminine and masculine nouns, respectively. Our model's performance is consistently high for each of the rules shown, including a stem change (*a* changing to *ä* in plural forms), providing further evidence that our model is useful for modeling rarer morphological

NOM,SING			a		
NOM,PL	n	e	ä		en
ACC,SING			a		
ACC,PL	n	e	ä		en
DAT,SING			a		
DAT,PL	n	en	ä	n	en
GEN,SING		es	a	s	
GEN,PL	n	e	ä		en
Example	Klasse	Krieg	Haus	Nutzer	Frau
Gold	49	48	26	26	20
Prec	95.7	72.9	88.0	82.8	87.0
Rec	91.8	89.6	84.6	92.3	100.0
F1	93.8	80.4	86.3	87.3	93.0

Table 6: Breakdown of errors by morphological rule being applied by the JOINT model on the DE-NOUNS development set. We show the rule itself, treating the nominative singular as the base form, an example of a German word using that rule, and then the model's accuracy at predicting applications of that rule. Errors are spread out over many rules, but it generally appears that common rules are to blame for the errors that are made, due in large part to gender confusion in this case.

[7] Nineteen of our 53 extracted rules only occur on one example; this suggests a few reasons that fewer rules are applied than are extracted. First, very common base forms with irregular morphology may give rise to completely irregular rules. Second, our edit distance alignment procedure can sometimes merge two adjacent rules if the orthographic context is such that there are multiple minimum-cost analyses. Finally, errors and inconsistencies in Wiktionary can yield nonsense rules that are never applied elsewhere.

[8] Gold rules are obtained by running our rule extraction procedure over the examples in question.

paradigms as well as more common ones.

As a concrete example of an error our model does make, *Löwe* (lion) is incorrectly predicted to have the first suffix, instead of the correct suffix (not shown) which adds an *-n* for accusative, genitive, and dative singular as well. However, making this prediction correctly is essentially beyond the capacity of a model based purely on orthography. Words ending in *-e* are commonly feminine, and none of our other training examples end in *-we*, so guessing that *Löwe* follows a common feminine inflection pattern is reasonable (though *Löwe* is, in fact, masculine). Disambiguating this case requires either features on observed genders, a more complex model of the German language, or observing the word in a large corpus. Generally, when the model fails, as in this case, it is because of a fundamental linguistic information source that it does not have access to.

7 Related Work

Much of the past work on morphology has focused on concatenative morphology using unsupervised methods (Goldsmith, 2001; Creutz and Lagus, 2007; Monson, 2008; Poon et al., 2009; Goldwater et al., 2009) or weak forms of supervision (Snyder and Barzilay, 2008). These methods can handle aspects of derivational morphology that we cannot, such as compounding, but we can handle a much larger subset of inflectional morphology, including more complex prefix and suffix rules, stem changes, and irregular forms. Some unsupervised work has specifically targeted these sorts of phenomena by, for example, learning spelling rules for mildly nonconcatenative cases (Dasgupta and Ng, 2007; Naradowsky and Goldwater, 2009) or mining lemma-base form pairs from a corpus (Schone and Jurafsky, 2001), but it is extremely difficult to make unsupervised methods perform as well as supervised approaches like ours.

Past supervised work on nonconcatenative inflectional morphology has typically targeted individual pairs of base forms and inflected forms for the purposes of inflection (Clark, 2001) or lemmatization (Yarowsky and Wicentowski, 2000; Wicentowski, 2004; Lindén, 2008; Toutanova and Cherry, 2009). Some of these methods may use analysis (Lindén, 2008) or decoding (Toutanova and Cherry, 2009) steps similar to those of our model, but none attempt to jointly predict a complete inflection table based on automatically extracted rules.

Some previous work has addressed the joint analysis (Zajac, 2001; Monson, 2008) or prediction (Lindén and Tuovila, 2009; Dinu et al., 2012) of whole inflection tables, as we do, but rarely are both aspects addressed simultaneously and most approaches are tuned to one particular language or use language-specific, curated resources. In overall setup, our work most closely resembles that of Dreyer and Eisner (2011), but they focus on incorporating large amounts of raw text data rather than using large training sets effectively.

Broadly similar techniques are also employed in systems to filter candidate rules and aid in human annotation of paradigms (Zajac, 2001; Forsberg et al., 2006; Détrez and Ranta, 2012) for resources such as Grammatical Framework (Ranta, 2011).

8 Conclusion

In this work, we presented a method for inflecting base forms in morphologically rich languages: we first extract orthographic transformation rules from observed inflection tables, then learn to apply these rules to new base forms based on orthographic features. Training examples for our supervised method can be collected from Wiktionary for a large number of languages and parts of speech. The changes we extract are interpretable and can be associated with particular classes of words. Moreover, our model can successfully apply these changes to unseen base forms with high accuracy, allowing us to rapidly generate lexicons for new languages of interest.

Our Wiktionary datasets and an open-source version of our code are available at `http://eecs.berkeley.edu/~gdurrett`

Acknowledgments

We are grateful to Klaus Macherey and David Talbot for assistance with the examples and helpful discussions throughout the course of this work. We would also like to thank the three anonymous reviewers for their useful comments.

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX Lexical Database (Release 2). Linguistic Data Consortium, University of Pennsylvania.

Alexander Clark. 2001. Partially Supervised Learning of Morphology with Stochastic Transducers. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, pages 341–348, Tokyo, Japan.

Mathias Creutz and Krista Lagus. 2007. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34, Feb.

Sajib Dasgupta and Vincent Ng. 2007. High Performance, Language-Independent Morphological Segmentation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Grégoire Détrez and Aarne Ranta. 2012. Smart Paradigms and the Predictability and Complexity of Inflectional Morphology. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Liviu P. Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Learning How to Conjugate the Romanian Verb: Rules for Regular and Partially Irregular Verbs. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Markus Dreyer and Jason Eisner. 2011. Discovering Morphological Paradigms from Plain Text Using a Dirichlet Process Mixture Model. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 616–627, Edinburgh, Scotland, UK.

Jason Eisner. 2002. Parameter Estimation for Probabilistic Finite-State Transducers. In *Proceedings of the Association for Computational Linguistics*.

Markus Forsberg, Harald Hammarström, and Aarne Ranta. 2006. Morphological Lexicon Extraction from Raw Text Data. In *Proceedings of Advances in Natural Language Processing*.

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

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1):21–54.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint Processing and Discriminative Training for Letter-to-Phoneme Conversion. In *Proceedings of the Association for Computational Linguistics*.

Teuvo Kohonen. 1986. Dynamically Expanding Context, With Application to the Correction of Symbol Strings in the Recognition of Continuous Speech. In *Proceedings of the International Conference on Pattern Recognition*.

Shen Li, João V. Graça, and Ben Taskar. 2012. Wiki-ly Supervised Part-of-speech Tagging. In *Proceedings of Empirical Methods in Natural Language Processing*.

Krister Lindén and Jussi Tuovila. 2009. Corpus-based Paradigm Selection for Morphological Entries. In *Proceedings of the Nordic Conference of Computational Linguistics*.

Krister Lindén. 2008. A Probabilistic Model for Guessing Base Forms of New Words by Analogy. In *Proceedings of Computational Linguistics and Intelligent Text Processing*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, December.

Christian Monson. 2008. ParaMor: From Paradigm Structure to Natural Language Morphology Induction. *Ph.D. thesis, Carnegie Mellon University*.

Jason Naradowsky and Sharon Goldwater. 2009. Improving Morphology Induction by Learning Spelling Rules. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

Jose Oncina and Marc Sebban. 2006. Learning Stochastic Edit Distance: Application in Handwritten Character Recognition. *Pattern Recognition*, 39(9):1575–1587, September.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised Morphological Segmentation with Log-Linear Models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Sunita Sarawagi and William W. Cohen. 2004. Semi-Markov Conditional Random Fields for Information Extraction. In *Advances in Neural Information Processing Systems 17*.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-Free Induction of Inflectional Morphologies. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Ksenia Shalonova and Bruno Golénia. 2010. Weakly Supervised Morphology Learning for Agglutinating Languages Using Small Training Sets. In *Proceedings of the Conference on Computational Linguistics*.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of the Association for Computational Linguistics*.

Kari Torkkola. 1993. An Efficient Way to Learn English Grapheme-to-Phoneme Rules Automatically. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing: Speech Processing - Volume II*.

Kristina Toutanova and Colin Cherry. 2009. A Global Model for Joint Lemmatization and Part-of-Speech Prediction. In *Proceedings of the Association for Computational Linguistics*.

Richard Wicentowski. 2004. Multilingual Noise-Robust Supervised Morphological Analysis Using the Word-Frame Model. In *Proceedings of the ACL Special Interest Group in Computational Phonology*.

David Yarowsky and Richard Wicentowski. 2000. Minimally Supervised Morphological Analysis by Multimodal Alignment. In *Proceedings of the Association for Computational Linguistics*.

Rémi Zajac. 2001. Morpholog: Constrained and Supervised Learning of Morphology. In *Proceedings of the Conference on Natural Language Learning*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of Language Resources and Evaluation*.