Weighting Finite-State Transductions With Neural Context

Pushpendre Rastogi and Ryan Cotterell and Jason Eisner
Department of Computer Science, Johns Hopkins University
{pushpendre,ryan.cotterell,eisner}@jhu.edu

Abstract
How should one apply deep learning to tasks such as morphological inflection, which stochastically edit one string to get another? A recent approach to such sequence-to-sequence tasks is to compress the input string into a vector that is then used to generate the output string, using recurrent neural networks. In contrast, we propose to keep the traditional architecture, which uses a finite-state transducer to score all possible output strings, but to augment the scoring function with the help of recurrent networks. A stack of bidirectional LSTMs reads the input string from left-to-right and right-to-left, in order to summarize the input context in which a transducer arc is applied. We combine these learned features with the transducer to define a probability distribution over aligned output strings, in the form of a weighted finite-state automaton. This reduces hand-engineering of features, allows learned features to examine unbounded context in the input string, and still permits exact inference through dynamic programming. We illustrate our method on the tasks of morphological inflection and lemmatization.

1 Introduction
Mapping one character sequence to another is a structured prediction problem that arises frequently in NLP and computational linguistics. Common applications include grapheme-to-phoneme (G2P), transliteration, vowelization, normalization, morphology, and phonology. The two sequences may have different lengths.

Traditionally, such settings have been modeled with weighted finite-state transducers (WFSTs) with parametric edge weights (Mohri, 1997; Eisner, 2002). This requires manual design of the transducer states and the features extracted from those states. Alternatively, deep learning has recently been tried for sequence-to-sequence transduction (Sutskever et al., 2014). While training these systems could discover contextual features that a hand-crafted parametric WFST might miss, they dispense with important structure in the problem, namely the monotonic input-output alignment. This paper describes a natural hybrid approach that marries simple FSTs with features extracted by recurrent neural networks.

Our novel architecture allows efficient modeling of globally normalized probability distributions over string-valued output spaces, simultaneously with automatic feature extraction. We evaluate on morphological inflection and lemmatization tasks, showing that our approach strongly outperforms a standard WFST baseline as well as neural sequence-to-sequence models with attention. Our approach also compares reasonably with a state-of-the-art WFST approach that uses task-specific latent variables.

2 Notation and Background
Let $\Sigma_x$ be a discrete input alphabet and $\Sigma_y$ be a discrete output alphabet. Our goal is to define a conditional distribution $p(y \mid x)$ where $x \in \Sigma_x^*$ and $y \in \Sigma_y^*$ and $x$ and $y$ may be of different lengths.

We use italics for characters and boldface for strings. $x_i$ denotes the $i$th character of $x$, and $x_{i:j}$ denotes the substring $x_{i+1} \cdots x_j$ of length $j - i \geq 0$. Note that $x_{i:i} = \varepsilon$, the empty string. Let $n = |x|$.

Our approach begins by hand-specified an unweighted finite-state transducer (FST), $F$, that non-deterministically maps any well-formed input $x$ to all appropriate outputs $y$. An FST is a directed graph whose vertices are called states, and whose arcs are each labeled with some pair $s$ : $t$, representing a possible edit of a source substring $s \in \Sigma_x^*$ into a target substring $t \in \Sigma_y^*$. A path $\pi$ from the FST’s initial state to its final state represents an alignment of $x$ to $y$, where $x$ and $y$ (respectively) are the concatenations of the $s$ and $t$ labels of the arcs along $\pi$. In
general, two strings $x, y$ can be aligned through exponentially many paths, via different edit sequences.

If we represent $x$ as a straight-line finite-state automaton (Figure 1), then composing $x$ with $F$ (Figure 2) yields a new FST, $G$ (Figure 3). The paths in $G$ are in 1-1 correspondence with exactly the paths in $F$ that have input $x$. $G$ can have cycles, allowing outputs of unbounded length.

Each path in $G$ represents an alignment of $x$ to some string in $\Sigma_y^*$. We say $p(y \mid x)$ is the total probability of all paths in $G$ that align $x$ to $y$ (Figure 4).

But how to define the probability of a path? Traditionally (Eisner, 2002), each arc in $F$ would also be equipped with a weight. The weight of a path in $F$, or the corresponding path in $G$, is the sum of its arcs’ weights. We would then define the probability $p(\pi)$ of a path $\pi$ in $G$ as proportional to $\exp w(\pi)$, where $w(\cdot) \in \mathbb{R}$ denotes the weight of an object.

The weight of an arc $\delta \xrightarrow{s,t} \delta'$ in $F$ is traditionally defined as a function of features of the edit $s : t$ and the names $(h, h')$ of the source and target states. In effect, $h$ summarizes the alignment between the prefixes of $x$ and $y$ that precede this edit, and $h'$ summarizes the alignment of the suffixes that follow it.

Thus, while the weight of an edit $s : t$ may depend on context, it traditionally does so only through $h$ and $h'$. So if $F$ has $k$ states, then the edit weight can only distinguish among $k$ different types of preceding or following context.

3 Incorporating More Context

That limitation is what we aim to correct in this paper, by augmenting our representation of context. Our contextual weighting approach will assign weights directly to $G$’s arcs, instead of to $F$’s arcs.

Each arc of $G$ can be regarded as a “token” of an edit arc in $F$: it “applies” that edit to a particular substring of $x$. It has the form $\delta \xrightarrow{i,h} \delta'$, and represents the replacement of $x_i = s$ by $t$. The finite-state composition construction produced this arc of $G$ by combining the arc $\delta \xrightarrow{i \delta t} \delta'$ in $F$ with the path $\delta \xrightarrow{i} \delta'$ in the straight-line automaton representing $x$. The latter automaton uses integers as state names: it is $\delta \xrightarrow{0} x_1, \delta \xrightarrow{1} x_2, \ldots, \delta \xrightarrow{n} \delta'$.

Our top-level idea is to make the weight of this arc in $G$ depend also on $(x, i, j)$, so that it can consider unbounded input context around the edit’s location. Arc weights can now consider arbitrary features of the input $x$ and the position $i, j$—exactly like the potential functions of a linear-chain conditional random field (CRF), which also defines $p(y \mid x)$.

Why not just use a CRF? That would only model a situation that enforced $|y| = |x|$ with each character $y_i$ aligned to $x_i$, since the emissions of a CRF correspond to edits $s : t$ with $|s| = |t| = 1$. An FST can also allow edits with $|s| \neq |t|$, if desired, so it can be fit to $(x, y)$ pairs of different lengths with unknown

---

**Figure 1:** An automaton encoding the English word say.

**Figure 2:** An example transducer $F$, whose state remembers the most recent output character (or $\dollar$ if none). Only a few of the states are shown, with all arcs among them. The $\Sigma$ wildcard matches any symbol in $\Sigma_x$; the “?” wildcard matches the empty string $\epsilon$ or any symbol in $\Sigma_x$.

**Figure 3:** An example of the transducer $G$, which pairs the string $x=say$ with infinitely many possible strings $y$. This $G$ was created as the composition of the straight-line input automaton (Figure 1) and the transducer $F$ (Figure 2). Thus, the state of $G$ tracks the states of those two machines: the position in $x$ and the most recent output character. To avoid a tangled diagram, this figure shows only a few of the states (the start state plus all states of the form $\langle x \rangle$ or $\langle a \rangle$), with all arcs among them.
alignment, summing over their possible alignments.

A standard weighted FST $F$ is similar to a dy-
namic linear-chain CRF. Both are unrolled against
the input $x$ to get a dynamic programming lattice $G$.
But they are not equivalent. By weighting $G$ instead
of $F$, we combine the FST’s advantage (aligning
unequal-length strings $x, y$ via a latent path) with
the CRF’s advantage (arbitrary dependence on $x$).

To accomplish this weighting in practice, sections
4–5 present a trainable neural architecture for
an arc weight function $w = f(s, t, h, h', x, i, j)$.
The goal is to extract continuous features from all
of $x$. While our specific architecture is new, we are
not the first to replace hand-crafted log-linear mod-
els with trainable neural networks (see section 9).\footnote{For CRFs, this sacrifices the convexity of the log-likelihood
training objective. But for FSTs, that objective was usually non-
convex to begin with, because the alignment path is latent.}

Note that as in a CRF, our arc weights cannot
consider arbitrary features of $y$, only of $x$. Still, a
weight’s dependence on states $h, h'$ does let it de-
pend on a finite amount of information about $y$ (also
possible in CRFs/HCRFs) and its alignment to $x$.

In short, our model $p(y \mid x)$ makes the weight of
an $s:t$ edit, applied to substring $x_{i:j}$, depend jointly
on $s:t$ and two summaries of the edit’s context:

- a finite-state summary $(h, h')$ of its context in
  the aligned $(x, y)$ pair, as found by the FST $F$
- a vector-valued summary of the context in $x$
  only, as found by a recurrent neural network

The neural vector is generally a richer summary of
the context, but it considers only the input-side con-
text. We are able to efficiently extract these rich fea-
tures from the single input $x$, but not from each of
the very many possible outputs $y$. The job of the
FST $F$ is to compute additional features that also
depend on the output.\footnote{Each arc of $G$ is used in a known input context, but could
be reused in many output contexts—different paths in $G$. Those contexts are only guaranteed to share $h, h'$. So the arc weight
cannot depend on any other features of the output context.} Thus our model of $p(y \mid x)$
is defined by an FST together with a neural network.

4 Neural Context Features

Our arc weight function $f$ will make use of a vec-
tor $\gamma_{i:j}$ (computed from $x, i, j$) to characterize the
substring $x_{i:j}$ that is being replaced, in context. We
define $\gamma_{i:j}$ as the concatenation of a left vector $\alpha_j$
(describing the prefix $x_{0:j}$) and a right vector $\beta_i$
(describing the suffix $x_{i:n}$), which characterize $x_{i:j}$
jointly with its left or right context. We use $\gamma_i$ to
abbreviate $\gamma_{i-1:i}$, just as $x_i$ abbreviates $x_{i-1:i}$.

To extract $\alpha_j$, we read the string $x$ one character
at a time with an LSTM (Hochreiter and Schmidhu-
er, 1997), a type of trainable recurrent neural net-
work that is good at extracting relevant features from
strings. $\alpha_j$ is the LSTM’s output after $j$ steps (which
read $x_{0:j}$). Appendix A reviews how $\alpha_j \in \mathbb{R}^9$ is
computed for $j = 1, \ldots, n$ using the recursive, dif-
ferentiable update rules of the LSTM architecture.

We also read the string $x$ in reverse with a second
LSTM. $\beta_i \in \mathbb{R}^9$ is the second LSTM’s output after
$n - i$ steps (which read reverse($x_{i:n}$)).

We regard the two LSTMs together as a BiLSTM
function (Graves and Schmidhuber, 2005) that reads
$x$ (Figure 5). For each bounded-length substring
$x_{i:j}$, the BiLSTM produces a characterization $\gamma_{i:j}$
of that substring in context, in $O(n)$ total time.

We now define a “deep BiLSTM,” which stacks
up $K$ BiLSTMs. This deepening is aimed at ex-
tracting the kind of rich features that Sutskever et al.
(2014) and Vinyals et al. (2015) found so effective in a different structured prediction architecture.

The $k$th-level BiLSTM (Figure 6) reads a sequence of input vectors $x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)} \in \mathbb{R}^{d^{(k)}}$, and produces a sequence of vectors $\gamma_1^{(k)}, \gamma_2^{(k)}, \ldots, \gamma_n^{(k)} \in \mathbb{R}^{2q}$. At the initial level $k = 1$, we define $x_i^{(1)} = e_{x_i} \in \mathbb{R}^{d^{(1)}}$, a vector embedding of the character $x_i \in \Sigma$. For $k > 1$, we take $x_i^{(k)}$ to be $\gamma_i^{(k-1)}$, concatenated with $x_i^{(k-1)}$ for good measure. Thus, $d^{(k)} = 2q + d^{(k-1)}$.

After this deep generalization, we define $\gamma_{i:j}$ to be the concatenation of all $\gamma_{i:j}^{(k)}$ (rather than just $\gamma_{i:j}^{(K)}$).

This novel deep BiLSTM architecture has more connections than a pair of deep LSTMs, since $\alpha_i^{(k)}$ depends not only on $\alpha_i^{(k-1)}$ but also on $\beta_i^{(k-1)}$. Thus, while we may informally regard $\alpha_i^{(k)}$ as being a deep summary of the prefix $x_{0:i}$, it actually depends indirectly on all of $x$ (except when $k = 1$).

5 The Arc Weight Function

Given the vector $\gamma_{i:j}$, we can now compute the weight of the edit arc $(i:j) \xrightarrow{s} t$ in $G$, namely $w = f(s, t, h, h', x, i, j)$. Many reasonable functions are possible. Here we use one that is inspired by the log-bilinear language model (Mnih and Hinton, 2007):

$$w = (e_s, \gamma_{i:j}, e_{x_i}, e_{x_{j+1}}) \cdot r_{t,h,h',type(t,s)} \quad (1)$$

The first argument to the inner product is an embedding $e_s \in \mathbb{R}^{d^{(1)}}$ of the source substring $s$, concatenated to the edit’s neural context and also (for good measure) its local context.$^3$ For example, if $|s| = 1$, i.e. $s$ is a single character, then we would use the embedding of that character as $e_s$. Note that the embeddings $e_s$ for $|s| = 1$ are also used to encode the local context characters and the level-1 BiLSTM input. We learn these embeddings, and they form part of our model’s parameter vector $\theta$.

The second argument is a joint embedding of the other properties of the edit: the target substring $t$, the edit arc’s state labels from $F$, and the type of the edit (INS, DEL, or SUB: see section 8). When replacing $s$ in a particular context, which fixes the first argument, we will prefer those replacements whose $r$ embeddings yield a high inner product $w$. We will learn the $r$ embeddings as well; note that their dimensionality must match that of the first argument.

The model’s parameter vector $\theta$ includes the $O((d^{(K)})^2)$ parameters from the $2K$ LSTMs, where $d^{(K)} = O(d^{(1)} + Kq)$. It also $O(d^{(1)}S)$ parameters for the embeddings $e_s$ of the $S$ different input substrings mentioned by $F$, and $O(d^{(K)}T)$ for the embeddings $r_{t,h,h',type(t,s)}$ of the $T$ “actions” in $F$.

6 Training

We train our model by maximizing the conditional log-likelihood objective,

$$\sum_{(x,y^*) \in \text{dataset}} \log p(y^* | x) \quad (2)$$

Recall that $p(y^* | x)$ sums over all alignments. As explained by Eisner (2002), it can be computed as the pathsum of the composition $G \circ y^*$ (Figure 4), divided by the pathsum of $G$ (which gives the normalizing constant for the distribution $p(y | x)$). The pathsum of a weighted FST is the total weight of all paths from the initial state to a final state, and can be computed by the forward algorithm.$^4$

$^3$Our present implementation handles INS edits (for which $j = i$) a bit differently, using $(e_{x_{i+1}}, \gamma_{i:i+1}, e_{x_i}, e_{x_{i+2}})$ rather than $(e_c, \gamma_{i:i+1}, e_{x_i}, e_{x_{i+2}})$. This is conveniently the same vector that is used for all other competing edits at this $i$ position (as they all have $|s| = 1$ in our present implementation); it provides an extra character of lookahead.

$^4$If an FST has cycles, such as the self-loops in the example of Figure 3, then the forward algorithm’s recurrence equations become cyclic, and must be solved as a linear system rather than sequentially. This is true regardless of how the FST’s weights
Eisner (2002) and Li and Eisner (2009) also explain how to compute the partial derivatives of \( p(y^* \mid x) \) with respect to the arc weights, essentially by using the forward-backward algorithm. We backpropagate further from these partials to find the gradient of (2) with respect to all our model parameters. We describe our gradient-based maximization procedure in section 10.3, along with regularization.

Our model does not have to be trained with the conditional log likelihood objective. It could be trained with other objectives such as empirical risk or softmax-margin (Li and Eisner, 2009; Gimpel and Smith, 2010), or with error-driven updates such as in the structured perceptron (Collins, 2002).

7 Inference and Decoding

For a new input \( x \) at test time, we can now construct a weighted FST, \( G \), that defines a probability distribution over all aligned output strings. This can be manipulated to make various predictions about \( y^* \) and its alignment.

In our present experiments, we find the most probable (highest-weighted) path in \( G \) (Dijkstra, 1959), and use its output string \( \hat{y} \) as our prediction. Note that Dijkstra's algorithm is exact; no beam search is required as in some neural sequence models.

On the other hand, \( \hat{y} \) may not be the most probable string—extracting that from a weighted FST is NP-hard (Casacuberta and de la Higuera, 1999). The issue is that the total probability of each \( y \) is split over many paths. Still, this is a well-studied problem in NLP. Instead of the Viterbi approximation, we could have used a better approximation, such as crunching (May and Knight, 2006) or variational decoding (Li et al., 2009). We actually did try crunching the 10000-best outputs but got no significant improvement, so we do not report those results.

8 Transducer Topology

In our experiments, we choose \( F \) to be a simple contextual edit FST as illustrated in Figure 2. Just as in Levenshtein distance (Levenshtein, 1966), it allows all edits \( s \mapsto t \) where \( |s| \leq 1, |t| \leq 1, |s| + |t| \neq 0 \). We consider the edit type to be INS if \( s = \varepsilon \), DEL if \( t = \varepsilon \), and SUB otherwise. Note that copy is a SUB edit with \( s = t \).

For a “memoryless” edit process (Ristad and Yianilos, 1996), the FST would require only a single state. By contrast, we use \( |\Sigma_x| + 1 \) states, where each state records the most recent output character (initially, a special “beginning-of-string” symbol \( \$ \)). That is, the state label \( h \) is the “history” output character immediately before the edit \( s \mapsto t \), so the state label \( h' \) is the history before the next edit, namely the final character of \( h t \). For edits other than DEL, \( h t \) is a bigram of \( y \), which can be evaluated (in context) by the arc weight function \( w = f(s, t, h, h', x, i, j) \).

Naturally, a weighted version of this FST \( F \) is far too simple to do well on real NLP tasks (as we show in our experiments). The magic comes from instead weighting \( G \) so that we can pay attention to the input context \( \gamma_{i,j} \).

The above choice of \( F \) corresponds to the “\((0, 1, 1)\) topology” in the more general scheme of Cotterell et al. (2014). For practical reasons, we actually modify it to limit the number of consecutive INS edits to 3.\(^5\) This trick bounds \( |y| \) to be \( < 4 \cdot (|x| + 1) \), ensuring that the pathsums in section 6 are finite regardless of the model parameters. This simplifies both the pathsum algorithm and the gradient-based training (Dreyer, 2011). Less importantly, since \( G \) becomes acyclic, Dijkstra’s algorithm in section 7 simplifies to the Viterbi algorithm.

9 Related Work

Our model adds to recent work on linguistic sequence transduction using deep learning.

Graves and Schmidhuber (2005) combined BiLSTMs with HMMs. Later, “sequence-to-sequence” models were applied to machine translation by Sutskever et al. (2014) and to parsing by Vinyals et al. (2015). That framework did not model any alignment between \( x \) and \( y \), but adding an “attention” mechanism provides a kind of soft alignment that has improved performance on MT (Bahdanau et al., 2015). Faruqui et al. (2016) apply these methods to morphological reinflection (the only other application to morphology we know of).

---

\(^5\)This multiplies the number of states 4-fold, since each state must also record a count in \([0, 3]\) of immediately preceding INS edits. No INS edit arc is allowed from a state with counter 3.

The counter is not considered by the arc weight function.
et al. (2015) recently augmented the sequence-to-sequence framework with a continuous analog of stack and queue data structures, to better handle long-range dependencies often found in linguistic data.

Some recent papers have used LSTMs or BiLSTMs, as we do, to define probability distributions over action sequences that operate directly on an input sequence. Such actions are aligned to the input. For example, Andor et al. (2016) score edit sequences using a globally normalized model, and Dyer et al. (2015) evaluate the local probability of a parsing action given past actions (and their result) and future words. These architectures are powerful because their LSTMs can examine the output structure; but as a result they do not permit dynamic programming and must fall back on beam search.

Our use of dynamic programming for efficient exact inference has long been common in non-neural architectures for sequence transduction, including FST systems that allow “phrasal” replacements $s : t$ where $|s|,|t| > 1$ (Chen, 2003; Jiaampojamarn et al., 2007; Bisani and Ney, 2008). Our work augments these FSTs with neural networks, much as others have augmented CRFs. In this vein, Durrett and Klein (2015) augment a CRF parser (Finkel et al., 2008) to score constituents with a feedforward neural network. Likewise, FitzGerald et al. (2015) employ feedforward nets as a factor in a graphical model for semantic role labeling. Many CRFs have incorporated feedforward neural networks (Bridle, 1990; Peng et al., 2009; Do and Artieres, 2010; Vinel et al., 2011; Fujii et al., 2012; Chen et al., 2015, and others). Some work augments CRFs with BiLSTMs: Huang et al. (2015) report results on part-of-speech tagging and named entity recognition with a linear-chain CRF-BiLSTM, and Kong et al. (2015) on Chinese word segmentation and handwriting recognition with a semi-CRF-BiLSTM.

## 10 Experiments

We evaluated our approach on two morphological generation tasks of reinflection (section 10.1) and lemmatization (section 10.2). In the reinflection task, the goal is to transduce verbs from one inflected form into another, whereas the lemmatization task requires the model to reduce an inflected verb to its root form.

We compare our WFST-LSTM against two standard baselines, a WFST with hand-engineered features and the Moses phrase-based MT system (Koehn et al., 2007), as well as the more complex latent-variable model of Dreyer et al. (2008). The comparison with Dreyer et al. (2008) is of noted interest since their latent variables are structured particularly for morphological transduction tasks—we are directly testing the ability of the LSTM to structure its hidden layer as effectively as linguistically motivated latent-variables. Additionally, we provide detailed ablation studies and learning curves which show that our neural-WFSA hybrid model can generalize even with very low amounts of training data.

### 10.1 Morphological Reinflection

Following Dreyer (2011), we conducted our experiments on the following transduction tasks from the CELEX (Baayen et al., 1993) morphological database: 13SIA $\mapsto$ 13SKE, 2PIE $\mapsto$ 13PKE, 2PKE $\mapsto$ z and rP $\mapsto$ pA.\(^6\) We refer to these tasks as 13SIA, 2PIE, 2PKE and rP, respectively.

Concretely, each task requires us to map a German inflection into another inflection. Consider the 13SIA task and the German verb abreiben (“to rub off”). We require the model to learn to map a past tense form abreibt to a present tense form abreibe—this involves a combination of stem change and affix generation. Sticking with the same verb abreiben, task 2PIE requires the model to transduce abreibt to abreiben—this requires an insertion and a substitution at the end. The tasks 2PKE and rP are somewhat more challenging since performing well on these tasks requires the model to learn complex transduction: abreiben to abzureiben and abreibt to abgerieben, respectively. These are complex transductions with phenomenon like infixation in specific contexts (abzureiben) and circumfixation (abgerieben) along with additional stem and affix changes. See Dreyer (2011) for more details and examples of these tasks.

We use the datasets of Dreyer (2011). Each experi-

iment sampled a different dataset of 2500 examples from CELEX, dividing this into 500 training + 1000 validation + 1000 test examples. Like them, we report exact-match accuracy on the test examples, averaged over 5 distinct experiments of this form. We also report results when the training and validation data are swapped in each experiment, which doubles the training size.

10.2 Lemmatization

Lemmatization is a special case of morphological re-inflection where we map an inflected form of a word to its lemma (canonical form), i.e., the target inflection is fixed. This task is quite useful for NLP, as dictionaries typically list only the lemma for a given lexical entry, rather than all possible inflected forms. In the case of German verbs, the lemma is taken to be the infinitive form, e.g., we map the past participle abgerieben to the infinitive abreiben.

Following Dreyer (2011), we use a subset of the lemmatization dataset created by Wicentowski (2002) and perform 10-fold experiments on four languages: Basque (5843), English (4915), Irish (1376) and Tagalog (9545), where the numbers in parenthesis indicate the total number of data pairs available. For each experimental fold the total data was divided into train, development and test sets in the proportion of 80:10:10 and we report test accuracy averaged across folds.

10.3 Settings and Training Procedure

We set the hyperparameters of our model to $K = 4$ (stacking depth), $d^1 = 10$ (character embedding dimension), and $q = 15$ (LSTM state dimension). The alphabets $\Sigma_x$ and $\Sigma_y$ are always equal; their size is language-dependent, typically $\approx 26$ but larger in languages like Basque and Irish where our datasets include properly accented characters. With $|\Sigma| = 26$ and the above settings for the hyperparameters, the number of parameters in our models is 352,801.

We optimize these parameters through stochastic gradient descent of the negative log-likelihood objective, and regularize the training procedure through dropout accompanied with gradient clipping and projection of parameters onto L2-balls with small radii, which is equivalent to adding a group-ridge regularization term to the training objective. The learning rate decay schedule, gradient clipping threshold, radii of L2-balls, and dropout frequency were tuned by hand on development data.

In our present experiments, we made one change to the architecture. Treating copy edits like other SUB edits led to poor performance: the system was unable to learn that all SUB edits with $s = t$ were extremely likely. In the experiments reported here, we addressed the problem by simply tying the weights of all copy edits regardless of context, bypassing (1) and instead setting $w = c$ where $c$ is a learned parameter of the model. See section 11 for discussion.

10.4 Results

Table 1 and 2 show our results. We can see that our proposed BiLSTM-WFST model always outperforms all but the most complex latent-variable model of Dreyer (2011); it is competitive with that model, but only beats it once individually. All of Dreyer’s models include output trigram features, while we only use bigrams.

Figure 7 shows learning curves for the 13SIA and 2PKE tasks: test accuracy when we train on less data. Curiously, at 300 data points the performance of our model is tied to Dreyer (2011). We also note

<table>
<thead>
<tr>
<th>Model</th>
<th>13SIA</th>
<th>2PIE</th>
<th>2PKE</th>
<th>rP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses15</td>
<td>85.3</td>
<td>94.0</td>
<td>82.8</td>
<td>70.8</td>
</tr>
<tr>
<td>Dreyer (Backoff)</td>
<td>82.8</td>
<td>88.7</td>
<td>74.7</td>
<td>69.9</td>
</tr>
<tr>
<td>Dreyer (Lat-Class)</td>
<td>84.8</td>
<td>93.6</td>
<td>75.7</td>
<td>81.8</td>
</tr>
<tr>
<td>Dreyer (Lat-Region)</td>
<td>87.5</td>
<td>93.4</td>
<td>87.4</td>
<td>84.9</td>
</tr>
<tr>
<td>BiLSTM-WFST (2 x Data)</td>
<td>85.1</td>
<td>94.4</td>
<td>85.5</td>
<td>83.0</td>
</tr>
<tr>
<td>Model Ensemble</td>
<td>85.8</td>
<td>94.6</td>
<td>86.0</td>
<td>83.8</td>
</tr>
<tr>
<td>BiLSTM-WFST (2 x Data)</td>
<td>87.6</td>
<td>94.8</td>
<td>88.1</td>
<td>85.7</td>
</tr>
</tbody>
</table>

Table 1: Exact match accuracy on the morphological re-inflection task. All the results in the first half of the table are taken from Dreyer (2011), whose experimental setup we copied exactly. The Moses15 result is obtained by applying the SMT toolkit Moses (Koehn et al., 2007) over letter strings with 15-character context windows. Dreyer (Backoff) refers to the $ngrams+x$ model which has access to all the “backoff features.” Dreyer (Lat-Class) is the $ngrams+x+latent\ class$ model, and Dreyer (Lat Region) refers to the $ngrams+x+latent\ class + latent\ change\ region$ model. The “Model Ensemble” row displays the performance of an ensemble including our full model and the 7 models that we performed ablation on. In each column, we boldfaced the highest result and those that were not significantly worse (sign test, $p < 0.05$). Finally, the last row reports the performance of our BiLSTM-WFST when trained on twice the training data.
### Table 2: Lemmatization results on Basque, English, Irish and Tagalog.

Comparison systems marked with (W) are taken from Wicentowski (2002) and systems marked with a (D) are taken from Dreyer (2011). We outperform baselines on all languages and are competitive with the latent-variable approach (ngrams + x + l), beating it in two cases: Irish and Tagalog.

<table>
<thead>
<tr>
<th>Model</th>
<th>Basque</th>
<th>English</th>
<th>Irish</th>
<th>Tagalog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (W)</td>
<td>85.3</td>
<td>91.0</td>
<td>43.3</td>
<td>0.3</td>
</tr>
<tr>
<td>WFAffix (W)</td>
<td>80.1</td>
<td>93.1</td>
<td>70.8</td>
<td>81.7</td>
</tr>
<tr>
<td>ngrams (D)</td>
<td>91.0</td>
<td>92.4</td>
<td>96.8</td>
<td>80.5</td>
</tr>
<tr>
<td>ngrams + x (D)</td>
<td>91.1</td>
<td>93.4</td>
<td>97.0</td>
<td>83.0</td>
</tr>
<tr>
<td>ngrams + x + l (D)</td>
<td>93.6</td>
<td>96.9</td>
<td>97.9</td>
<td>88.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Systems</th>
<th>13SIA</th>
<th>2PIE</th>
<th>2PKE</th>
<th>rP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep BiLSTM w/ Tying</td>
<td>86.8</td>
<td>94.8</td>
<td>87.9</td>
<td>81.1</td>
</tr>
<tr>
<td>Deep BiLSTM (No Context)</td>
<td>86.5</td>
<td>94.3</td>
<td>87.8</td>
<td>78.8</td>
</tr>
<tr>
<td>Deep BiLSTM w/o Copying</td>
<td>86.5</td>
<td>94.6</td>
<td>86.5</td>
<td>80.7</td>
</tr>
<tr>
<td>Shallow BiLSTM</td>
<td>86.4</td>
<td>94.7</td>
<td>86.1</td>
<td>80.6</td>
</tr>
<tr>
<td>Bi-Deep LSTM</td>
<td>86.1</td>
<td>94.2</td>
<td>86.5</td>
<td>78.6</td>
</tr>
<tr>
<td>Deep MonoLSTM</td>
<td>84.0</td>
<td>93.8</td>
<td>85.6</td>
<td>67.3</td>
</tr>
<tr>
<td>Shallow MonoLSTM</td>
<td>84.2</td>
<td>94.5</td>
<td>84.9</td>
<td>68.2</td>
</tr>
<tr>
<td>No LSTM (Local Context)</td>
<td>83.6</td>
<td>88.5</td>
<td>83.2</td>
<td>68.0</td>
</tr>
<tr>
<td>Deep BiLSTM w/o Tying</td>
<td>69.7</td>
<td>78.5</td>
<td>77.9</td>
<td>66.7</td>
</tr>
<tr>
<td>No LSTM (No Context)</td>
<td>70.7</td>
<td>84.9</td>
<td>72.4</td>
<td>64.1</td>
</tr>
</tbody>
</table>

**Figure 7: Learning Curves**

that our model always outperforms the Moses15 baseline on all training set sizes except on the 2PKE task with 50 training samples.

In general, the results from our experiments are a promising indicator that LSTMs are capable of extracting linguistically relevant features for morphology. Our model outperforms all baselines, and is competitive with and sometimes surpasses the latent-variable model of Dreyer et al. (2008) without any of the hand-engineered features or linguistically inspired latent variables.

On morphological reinflection, we outperform all of Dreyer et al.’s models on 2PIE, but fall short of his latent-change region model on the other tasks (outperforming the other models). On lemmatization, we outperform all of Wicentowski’s models on all the languages and all of Dreyer et al.’s models on Irish and Tagalog, but, but not on English and Irish. This suggests that perhaps further gains are possible through using something like Dreyer’s FST as our $F$. Indeed, this would be compatible with much recent work that gets best results from a *combination* of automatically learned neural features and hand-engineered features.

### 10.5 Analysis of Results

We analyzed our lemmatization errors for all the languages on one fold of the datasets. On the English lemmatization task, 7 of our 27 errors simply copied the input word to the output: ate, kept, went, taught, torn, paid, strung. This suggests that our current aggressive parameter tying for copy edits may predict a high probability for a copy edit even in contexts that should not favor it.

Also we found that the FST sometimes produced non-words while lemmatizing the input verbs. For example it mapped *picnicked* $\rightarrow$ *picnick*, *happen* $\rightarrow$ *hapen*, *exceed* $\rightarrow$ *excy* and *lining* $\rightarrow$ *lin*. Since these strings would be rare in a corpus, many such errors could be avoided by a reranking approach that combined the FST’s path score with a string frequency feature.

In order to better understand our architecture and the importance of its various components, we performed an ablation study on the validation portions of the morphological induction datasets, shown in Table 3. We can see in particular that using a BiLSTM instead of an LSTM, increasing the depth of the network, and including local context all helped to improve the final accuracy.

“Deep BiLSTM w/ Tying” refers to our complete model. The other rows are ablation experiments—architectures that are the same as the first row except in the specified way. “Deep BiLSTM (No Context)”
omits local context $e_{x_i}, e_{x_{j+1}}$ from (1). “Deep BiLSTMs w/o Copying” does not concatenate a copy of $x_{i}^{(k-1)}$ into $x_{i}^{(k)}$ and simplifies $\gamma_{i,j}$ to be $\gamma_{i,j}^{(K)}$ only. “Shallow BiLSTM” reduces $K$ from 4 to 1. “Bi-Deep LSTM” replaces our deep BiLSTM with two deep LSTMs that run in opposite directions but do not interact with each other. “Deep MonoLSTM” redefines $\beta_i$ to be the empty vector, i.e., it replaces the deep BiLSTM with a deep left-to-right LSTM. “Shallow MonoLSTM” replaces the deep BiLSTM with a shallow left-to-right LSTM. “No LSTM (Local Context)” omits $x_{i}$ from the weight function altogether. “Deep BiLSTM w/o Tying” does not use the parameter tying heuristic for copy edits. “No LSTM (No Context)" is the simplest model that we consider. It removes $\gamma_{i,j}, e_{x_i}$ and $e_{x_{j+1}}$ and in fact, it is precisely a weighting of the edits in our original FST $F$, without further considering the context in which an edit is applied.

Finally, to compare the performance of our method to baseline neural encoder-decoder models, we trained 1-layer and 4-layer neural sequence-to-sequence models with and without attention, using the publicly available morph-trans toolkit (Faruqui et al., 2016). We show the performance of these models in the lower half of the table. The results consistently show that sequence-to-sequence transduction models that lack the constraints of monotonic alignment perform worse than our proposed models on morphological transduction tasks.

## 11 Future Work

As neither our FST-LSTM model or the latent-variable WFST model of Dreyer et al. (2008) uniformly outperforms the other, a future direction is to improve the FST we use in our model—e.g., by augmenting its states with explicit latent variables. Other improvements to the WFST would be to increase the amount of history $h$ stored in the states, and to allow $s, t$ to be longer than a single character, which would allow the model to segment $x$.

We are not committed to the arc weight function in (1), and we believe that further investigation here could improve performance. The goal is to define the weight of $\overrightarrow{s} \rightarrow \overleftarrow{t}$ in the context summarized by $\alpha_i, \beta_j$ (the context around $x_{i,j} = s$) and/or $\alpha_j, \beta_i$ (which incorporate $s$ as well).

Any parametric function of the variables $(s, t, h, h', \alpha_i, \alpha_j, \beta_i, \beta_j)$ could be used—for example, a neural network or a multilinear function. This function might depend on learned embeddings of the separate objects $s, t, h, h'$, but also on learned joint embeddings of pairs of these objects (which adds finer-grained parameters), or hand-specified properties of the objects such as their phonological features (which adds backoff parameters).

A basic approach along the lines of (1) would use an inner product of some encoding of the arc $(s, t, h, h')$ with some encoding of the context $(s, h, h', \alpha_i, \alpha_j, \beta_i, \beta_j)$. Note that this formulation lets the objects $s, h, h'$ play a dual role—they may appear as part of the arc and/or as part of the context. This is because we must judge whether $s = x_{i,j}$ (with “label” $\overrightarrow{s} \rightarrow \overleftarrow{t}$) is an appropriate input segment given the string context around $x_{i,j}$, but if this segment is chosen, in turn it provides additional context to judge whether $t$ is an appropriate output.

High-probability edits $s : t$ typically have $t \approx s$: a perfect copy, or a modified copy that changes just one or two features such as phonological voicing or orthographic capitalization. Thus, we are interested in learning a shared set of embeddings for $\Sigma_x \cup \Sigma_y$, and making the arc weight depend on features of the “discrepancy vector” $e_t - e_s$, such as this vector’s components\(^7\) and their absolute magnitudes, which would signal discrepancies of various sorts.

## 12 Conclusions

We have presented a hybrid FST-LSTM architecture for string-to-string transduction tasks. This approach combines classical finite-state approaches to transduction and newer neural approaches. We weight the same FST arc differently in different contexts, and use LSTMs to automatically extract features that determine these weights. This reduces the need to engineer a complex topology for the FST or to hand-engineer its weight features. We evaluated one such model on the tasks of morphological reinflection and lemmatization. Our approach outperforms several baselines and is competitive with (and sometimes surpasses) a latent-variable model hand-crafted for morphological transduction tasks.

\(^7\)In various directions—perhaps just the basis directions, which might come to encode distinctive features, via training.
Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency under the Deep Exploration and Filtering of Text (DEFT) Program, agreement number FA8750-13-2-0011; by the National Science Foundation under Grant No. 1423276; and by a DAAD fellowship to the second author. We would like to thank Markus Dreyer for providing the datasets and Manaal Faruqui for providing implementations of the baseline sequence-to-sequence models. Finally we would like to thank Mo Yu, Nanyun Peng, Dingquan Wang and Elan Houri­col­det and Phil Blunsom. 2015. Learning to transduce with unbounded memory. In Proceedings of NIPS, pages 1819–1827.


Alex Graves. 2012. Supervised Sequence Labelling with Recurrent Neural Networks. Springer.


