10-2013

Translating into Morphologically Rich Languages with Synthetic Phrases

Victor Chahuneau
Carnegie Mellon University

Eva Schlinger
Carnegie Mellon University

Noah A. Smith
Carnegie Mellon University, nasmith@cs.cmu.edu

Chris Dyer
Carnegie Mellon University, cdyer@cs.cmu.edu

Published In
Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 1677-1687.

Translating into Morphologically Rich Languages with Synthetic Phrases

Victor Chahuneau Eva Schlinger Noah A. Smith Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{vchahune,eschling,nasmith,cdyer}@cs.cmu.edu

Abstract

Translation into morphologically rich languages is an important but recalcitrant problem in MT. We present a simple and effective approach that deals with the problem in two phases. First, a discriminative model is learned to predict inflections of target words from rich source-side annotations. Then, this model is used to create additional sentence-specific word- and phrase-level translations that are added to a standard translation model as "synthetic" phrases. Our approach relies on morphological analysis of the target language, but we show that an unsupervised Bayesian model of morphology can successfully be used in place of a supervised analyzer. We report significant improvements in translation quality when translating from English to Russian, Hebrew and Swahili.

1 Introduction

Machine translation into morphologically rich languages is challenging, due to lexical sparsity and the large variety of grammatical features expressed with morphology. In this paper, we introduce a method that uses target language morphological grammars (either hand-crafted or learned unsupervisedly) to address this challenge and demonstrate its effectiveness at improving translation from English into several morphologically rich target languages.

Our approach decomposes the process of producing a translation for a word (or phrase) into two steps. First, a meaning-bearing **stem** is chosen and then an appropriate **inflection** is selected using a feature-rich discriminative model that conditions on the source context of the word being translated.

Rather than attempting to directly produce full-sentence translations using such an elementary process, we use our model to generate translations of individual words and short phrases that *augment*—on a sentence-by-sentence basis—the inventory of translation rules obtained using standard translation rule extraction techniques (Chiang, 2007). We call these **synthetic phrases**.

The major advantages of our approach are: (i) synthesized forms are targeted to a specific translation context; (ii) multiple, alternative phrases may be generated with the final choice among rules left to the global translation model; (iii) virtually no language-specific engineering is necessary; (iv) any phrase- or syntax-based decoder can be used without modification; and (v) we can generate forms that were not attested in the bilingual training data.

The paper is structured as follows. We first present our "translate-and-inflect" model for predicting lexical translations into morphologically rich languages given a source word and its context (§2). Our approach requires a morphological grammar to relate surface forms to underlying $\langle \text{stem}, \text{inflection} \rangle$ pairs; we discuss how either a standard morphological analyzer or a simple Bayesian unsupervised analyzer can be used (§3). After describing an efficient parameter estimation procedure for the inflection model (§4), we employ the translate-and-inflect model in an MT system. We describe how we use our model to synthesize translation options (§5) and then evaluate translation quality on English–Russian, English–Hebrew, and English–

Swahili translation tasks, finding significant improvements in all language pairs (§6). We finally review related work (§7) and conclude (§8).

2 Translate-and-Inflect Model

The task of the translate-and-inflect model is illustrated in Fig. 1 for an English–Russian sentence pair. The input will be a sentence $\boldsymbol{e}$ in the source language (in this paper, always English) and any available linguistic analysis of $\boldsymbol{e}$. The output $\boldsymbol{f}$ will be composed of (i) a sequence of stems, each denoted σ and (ii) one morphological inflection pattern for each stem, denoted μ. When the information is available, a stem σ is composed of a lemma and an inflectional class. Throughout, we use Ω_σ to denote the set of possible morphological inflection patterns for a given stem σ. Ω_σ might be defined by a grammar; our models restrict Ω_σ to be the set of inflections observed anywhere in our monolingual or bilingual training data as a realization of σ.[1]

We assume the availability of a deterministic function that maps a stem σ and morphological inflection μ to a target language surface form f. In some cases, such as our unsupervised approach in §3.2, this will be a concatenation operation, though finite-state transducers are traditionally used to define such relations (§3.1). We abstractly denote this operation by $\star$: $f = \sigma \star \mu$.

Our approach consists in defining a probabilistic model over target words $\boldsymbol{f}$. The model assumes independence between each target word f conditioned on the source sentence $\boldsymbol{e}$ and its aligned position i in this sentence.[2] This assumption is further relaxed in §5 when the model is integrated in the translation system.

We decompose the probability of generating each target word f in the following way:

$$p(f \mid \boldsymbol{e}, i) = \sum_{\sigma \star \mu = f} \underbrace{p(\sigma \mid e_i)}_{\text{gen. stem}} \times \underbrace{p(\mu \mid \sigma, \boldsymbol{e}, i)}_{\text{gen. inflection}}$$

Here, each stem is generated independently from a single aligned source word e_i, but in practice we use a standard phrase-based model to generate sequences of stems and only the inflection model operates word-by-word. We turn next to the inflection model.

2.1 Modeling Inflection

In morphologically rich languages, each stem may be combined with one or more inflectional morphemes to express many different grammatical features (e.g., case, definiteness, mood, tense, etc.).

Since the inflectional morphology of a word generally expresses multiple grammatical features, we would like a model that naturally incorporates rich, possibly overlapping features in its representation of both the input (i.e., conditioning context) and output (i.e., the inflection pattern). We therefore use the following parametric form to model inflectional probabilities:

$$u(\mu, \boldsymbol{e}, i) = \exp\Big[\boldsymbol{\varphi}(\boldsymbol{e}, i)^\top \mathbf{W} \boldsymbol{\psi}(\mu) + \boldsymbol{\psi}(\mu)^\top \mathbf{V} \boldsymbol{\psi}(\mu)\Big],$$

$$p(\mu \mid \sigma, \boldsymbol{e}, i) = \frac{u(\mu, \boldsymbol{e}, i)}{\sum_{\mu' \in \Omega_\sigma} u(\mu', \boldsymbol{e}, i)}. \quad (1)$$

Here, $\boldsymbol{\varphi}$ is an m-dimensional *source context* feature vector function, $\boldsymbol{\psi}$ is an n-dimensional *morphology* feature vector function, $\mathbf{W} \in \mathbb{R}^{m \times n}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are parameter matrices. As with the more familiar log-linear parametrization that is written with a single feature vector, single weight vector and single bias vector, this model is linear in its parameters (it can be understood as working with a feature space that is the outer product of the two feature spaces). However, using two feature vectors allows to define overlapping features of both the input *and* the output, which is important for modeling morphology in which output variables are naturally expressed as bundles of features. The second term in the sum in u enables correlations among output features to be modeled independently of input, and as such can be understood as a generalization of the bias terms in multi-class logistic regression (on the diagonal $\mathbf{V}_{ii}$) and interaction terms between output variables in a conditional random field (off the diagonal $\mathbf{V}_{ij}$).

[1] This prevents the model from generating words that would be difficult for the language model to reliably score.

[2] This is the same assumption that Brown et al. (1993) make in, for example, IBM Model 1.

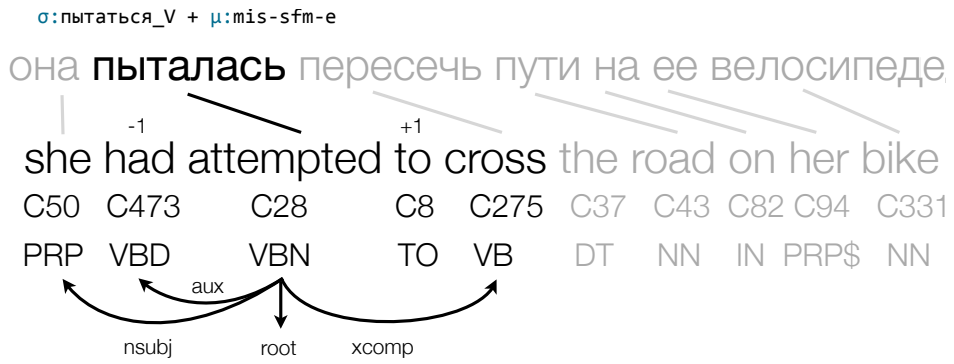

Figure 1: The inflection model predicts a form for the target verb lemma σ =пытаться (*pytat'sya*) based on its source *attempted* and the linear and syntactic source context. The correct inflection string for the observed Russian form in this particular training instance is μ = mis-sfm-e (equivalent to the more traditional morphological string: +MAIN+IND+PAST+SING+FEM+MEDIAL+PERF).

$$\left\{\begin{array}{c}\text{source aligned word } e_i \\ \text{parent word } e_{\pi_i} \text{ with its dependency } \pi_i \rightarrow i \\ \text{all children } e_j \mid \pi_j = i \text{ with their dependency } i \rightarrow j \\ \text{source words } e_{i-1} \text{ and } e_{i+1}\end{array}\right\} \left\{\begin{array}{c}\text{token} \\ \text{part-of-speech tag} \\ \text{word cluster}\end{array}\right\}$$
$$\text{– are } e_i, e_{\pi_i} \text{ at the root of the dependency tree?}$$
$$\text{– number of children, siblings of } e_i$$

Figure 2: Source features $\boldsymbol{\varphi}(\boldsymbol{e}, i)$ extracted from $\boldsymbol{e}$ and its linguistic analysis. π_i denotes the parent of the token in position i in the dependency tree and $\pi_i \rightarrow i$ the typed dependency link.

2.2 Source Context Features: $\boldsymbol{\varphi}(\boldsymbol{e}, i)$

In order to select the best inflection of a target-language word, given the source word it translates and the context of that source word, we seek to exploit as many features of the context as are available. Consider the example shown in Fig. 1, where most of the inflection features of the Russian word (past tense, singular number, and feminine gender) can be inferred from the context of the English word it is aligned to. Indeed, many grammatical functions expressed morphologically in Russian are expressed syntactically in English. Fortunately, high-quality parsers and other linguistic analyzers are available for English.

On the source side, we apply the following processing steps:

- Part-of-speech tagging with a CRF tagger trained on sections 02–21 of the Penn Treebank.
- Dependency parsing with TurboParser (Martins et al., 2010), a non-projective dependency parser trained on the Penn Treebank to produce basic Stanford dependencies.
- Assignment of tokens to one of 600 Brown clusters, trained on 8G words of English text.[3]

We then extract binary features from $\boldsymbol{e}$ using this information, by considering the aligned source word e_i, its preceding and following words, and its syntactic neighbors. These are detailed in Figure 2.

3 Morphological Grammars and Features

We now describe how to obtain morphological analyses and convert them into feature vectors ($\boldsymbol{\psi}$) for our target languages, Russian, Hebrew, and Swahili, using supervised and unsupervised methods.

3.1 Supervised Morphology

The state-of-the-art in morphological analysis uses unweighted morphological transduction rules (usu-

[3]The entire monolingual data available for the translation task of the 8th ACL Workshop on Statistical Machine Translation was used.

ally in the form of an FST) to produce candidate analyses for each word in a sentence and then statistical models to disambiguate among the analyses in context (Hakkani-Tür et al., 2000; Hajič et al., 2001; Smith et al., 2005; Habash and Rambow, 2005, *inter alia*). While this technique is capable of producing high quality linguistic analyses, it is expensive to develop, requiring hand-crafted rule-based analyzers and annotated corpora to train the disambiguation models. As a result, such analyzers are only available for a small number of languages, and, as a practical matter, each analyzer (which resulted from different development efforts) operates differently from the others.

We therefore focus on using supervised analysis for a single target language, Russian. We use the analysis tool of Sharoff et al. (2008) which produces for each word in context a lemma and a fixed-length morphological tag encoding the grammatical features. We process the target side of the parallel data with this tool to obtain the information necessary to extract $\langle \text{lemma}, \text{inflection} \rangle$ pairs, from which we compute σ and morphological feature vectors $\boldsymbol{\psi}(\mu)$.

Supervised morphology features: $\boldsymbol{\psi}(\mu)$. Since a positional tag set is used, it is straightforward to convert each fixed-length tag μ into a feature vector by defining a binary feature for each key-value pair (e.g., Tense=past) composing the tag.

3.2 Unsupervised Morphology

Since many languages into which we might want to translate do not have supervised morphological analyzers, we now turn to the question of how to generate morphological analyses and features using an unsupervised analyzer. We hypothesize that perfect decomposition into rich linguistic structures may not be required for accurate generation of new inflected forms. We will test this hypothesis by experimenting with a simple, *unsupervised* model of morphology that segments words into sequences of morphemes, assuming a (naïve) concatenative generation process and a single analysis per type.

Unsupervised morphological segmentation. We assume that each word can be decomposed into any number of prefixes, a stem, and any number of suffixes. Formally, we let M represent the set of all possible morphemes and define a regular grammar M^*MM^* (i.e., zero or more prefixes, a stem, and zero or more suffixes). To infer the decomposition structure for the words in the target language, we assume that the vocabulary was generated by the following process:

1. Sample morpheme distributions from symmetric Dirichlet distributions: $\theta_p \sim \text{Dir}_{|M|}(\alpha_p)$ for prefixes, $\theta_\sigma \sim \text{Dir}_{|M|}(\alpha_\sigma)$ for stems, and $\theta_s \sim \text{Dir}_{|M|}(\alpha_s)$ for suffixes.

2. Sample length distribution parameters $\lambda_p \sim \text{Beta}(\beta_p, \gamma_p)$ for prefix sequences and $\lambda_s \sim \text{Beta}(\beta_s, \gamma_s)$ for suffix sequences.

3. Sample a vocabulary by creating each word type w using the following steps:

   (a) Sample affix sequence lengths: $l_p \sim \text{Geometric}(\lambda_p)$; $l_s \sim \text{Geometric}(\lambda_s)$.

   (b) Sample l_p prefixes $p_1, \ldots, p_{l_p}$ independently from θ_p; l_s suffixes $s_1, \ldots, s_{l_s}$ independently from θ_s; and a stem $\sigma \sim \theta_\sigma$.

   (c) Concatenate prefixes, the stem, and suffixes: $w = p_1 + \cdots + p_{l_p} + \sigma + s_1 + \cdots + s_{l_s}$.

We use blocked Gibbs sampling to sample segmentations for each word in the training vocabulary. Because of our particular choice of priors, it possible to approximately decompose the posterior over the arcs of a compact finite-state machine. Sampling a segmentation or obtaining the most likely segmentation *a posteriori* then reduces to familiar FST operations. This model is reminiscent of work on learning morphology using adaptor grammars (Johnson et al., 2006; Johnson, 2008).

The inferred morphological grammar is very sensitive to the Dirichlet hyperparameters $(\alpha_p, \alpha_s, \alpha_\sigma)$ and these are, in turn, sensitive to the number of types in the vocabulary. Using $\alpha_p, \alpha_s \ll \alpha_\sigma \ll 1$ tended to recover useful segmentations, but we have not yet been able to find reliable generic priors for these values. Therefore, we selected them empirically to obtain a stem vocabulary size on the parallel data that is one-to-one with English.[4] Future work

[4] Our default starting point was to use $\alpha_p = \alpha_s = 10^{-6}, \alpha_\sigma = 10^{-4}$ and then to adjust all parameters by factors of 10.

Table 1: Corpus statistics.

	Parallel					Parallel+Monolingual		
	Sentences	EN-tokens	TRG-tokens	EN-types	TRG-types	Sentences	TRG-tokens	TRG-types
Russian	150k	3.5M	3.3M	131k	254k	20M	360M	1,971k
Hebrew	134k	2.7M	2.0M	48k	120k	806k	15M	316k
Swahili	15k	0.3M	0.3M	23k	35k	596k	13M	334k

will involve a more direct method for specifying or inferring these values.

Unsupervised morphology features: $\psi(\mu)$. For the unsupervised analyzer, we do not have a mapping from morphemes to structured morphological attributes; however, we can create features from the affix sequences obtained after morphological segmentation. We produce binary features corresponding to the content of each potential affixation position relative to the stem:

For example, the unsupervised analysis μ = wa+ki+wa+STEM of the Swahili word *wakiwapiga* will produce the following features:

$$\psi_{\text{prefix}[-3][\text{wa}]}(\mu) = 1,$$
$$\psi_{\text{prefix}[-2][\text{ki}]}(\mu) = 1,$$
$$\psi_{\text{prefix}[-1][\text{wa}]}(\mu) = 1.$$

4 Inflection Model Parameter Estimation

To set the parameters $\mathbf{W}$ and $\mathbf{V}$ of the inflection prediction model (Eq. 1), we use stochastic gradient descent to maximize the conditional log-likelihood of a training set consisting of pairs of source (English) sentence contextual features ($\boldsymbol{\varphi}$) and target word inflectional features ($\boldsymbol{\psi}$). The training instances are extracted from the word-aligned parallel corpus with the English side preprocessed as discussed in §2.2 and the target side disambiguated as discussed in §3. When morphological category information is available, we train an independent model for each open-class category (in Russian, nouns, verbs, adjectives, numerals, adverbs); otherwise a single model is used for all words (excluding words less than four characters long, which are ignored).

Statistics of the parallel corpora used to train the inflection model are summarized in Table 1. It is important to note here that our richly parameterized model is trained on the full parallel training corpus, not just on a handful of development sentences (which are typically used to tune MT system parameters). Despite this scale, training is simple: the inflection model is trained to discriminate among different inflectional paradigms, not over all possible target language sentences (Blunsom et al., 2008) or learning from all observable rules (Subotin, 2011). This makes the training problem relatively tractable: all experiments in this paper were trained on a single processor using a Cython implementation of the SGD optimizer. For our largest model, trained on 3.3M Russian words, $n = 231K \times m = 336$ features were produced, and 10 SGD iterations were performed in less than 16 hours.

4.1 Intrinsic Evaluation

Before considering the broader problem of integrating the inflection model in a machine translation system, we perform an artificial evaluation to verify that the model learns sensible source sentence-target inflection patterns. To do so, we create an inflection test set as follows. We preprocess the source (English) sentences exactly as during training (§2.2), and using the target language morphological analyzer, we convert each aligned target word to $\langle$stem, inflection$\rangle$ pairs. We perform word alignment on the held-out MT development data for each language pair (cf. Table 1), exactly as if it were going to produce training instances, but instead we use them for testing.

Although the resulting dataset is noisy (e.g., due to alignment errors), this becomes our intrinsic evaluation test set. Using this data, we measure inflection quality using two measurements:[5]

[5]Note that we are not evaluating the stem translation model,

			acc.	ppl.	$\lvert\Omega_\sigma\rvert$
Supervised	Russian	N	64.1%	3.46	9.16
		V	63.7%	3.41	20.12
		A	51.5%	6.24	19.56
		M	73.0%	2.81	9.14
		average	63.1%	3.98	14.49
Unsup.	Russian	all	71.2%	2.15	4.73
	Hebrew	all	85.5%	1.49	2.55
	Swahili	all	78.2%	2.09	11.46

Table 2: Intrinsic evaluation of inflection model (N: nouns, V: verbs, A: adjectives, M: numerals).

- the accuracy of predicting the inflection given the source, source context and target stem, and
- the inflection model perplexity on the same set of test instances.

Additionally, we report the average number of possible inflections for each stem, an upper bound to the perplexity that indicates the inherent difficulty of the task. The results of this evaluation are presented in Table 2 for the three language pairs considered. We remark on two patterns in these results. First, perplexity is substantially lower than the perplexity of a uniform model, indicating our model is overall quite effective at predicting inflections using source context only. Second, in the supervised Russian results, we see that predicting the inflections of adjectives is relatively more difficult than for other parts-of-speech. Since adjectives agree with the nouns they modify in gender and case, and gender is an idiosyncratic feature of Russian nouns (and therefore not directly predictable from the English source), this difficulty is unsurprising.

We can also inspect the weights learned by the model to assess the effectiveness of the features in relating source-context structure with target-side morphology. Such an analysis is presented in Fig. 3.

4.2 Feature Ablation

Our inflection model makes use of numerous feature types. Table 3 explores the effect of removing different kinds of (source) features from the model, evaluated on predicting Russian inflections using supervised morphological grammars.[6] Rows 2–3 show the effect of removing either linear or dependency context. We see that both are necessary for good performance; however removing dependency context substantially degrades performance of the model (we interpret this result as evidence that Russian morphological inflection captures grammatical relationships that would be expressed structurally in English). The bottom four rows explore the effect of source language word representation. The results indicate that lexical features are important for accurate prediction of inflection, and that POS tags and Brown clusters are likewise important, but they seem to capture similar information (removing one has little impact, but removing both substantially degrades performance).

Table 3: Feature ablation experiments using supervised Russian classification experiments.

Features ($\boldsymbol{\varphi}(\boldsymbol{e}, i)$)	acc.
all	54.7%
−linear context	52.7%
−dependency context	44.4%
−POS tags	54.5%
−Brown clusters	54.5%
−POS tags, −Brown cl.	50.9%
−lexical items	51.2%

5 Synthetic Phrases

We turn now to translation; recall that our translate-and-inflect model is used to augment the set of rules available to a conventional statistical machine translation decoder. We refer to the phrases it produces as *synthetic* phrases.

Our baseline system is a standard hierarchical phrase-based translation model (Chiang, 2007). Following Lopez (2007), the training data is compiled into an efficient binary representation which allows extraction of sentence-specific grammars just before decoding. In our case, this also allows the creation of synthetic inflected phrases that are produced conditioning on the sentence to translate.

To generate these synthetic phrases with new inflections possibly unseen in the parallel training

just the inflection prediction model.

[6]The models used in the feature ablation experiment were trained on fewer examples, resulting in overall lower accuracies than seen in Table 2, but the pattern of results is the relevant datapoint here.

Russian supervised	Hebrew	Swahili
Verb: 1st Person `child(nsubj)=I child(nsubj)=we`	Suffix ים (masculine plural) `parent=NNS after=NNS`	Prefix *li* (past) `source=VBD source=VBN`
Verb: Future tense `child(aux)=MD child(aux)=will`	Prefix א (first person sing. + future) `child(nsubj)=I child(aux)='ll`	Prefix *nita* (1st person sing. + future) `child(aux) child(nsubj)=I`
Noun: Animate `source=animals/victims/...`	Prefix כ (preposition like/as) `child(prep)=IN parent=as`	Prefix *ana* (3rd person sing. + present) `source=VBZ`
Noun: Feminine gender `source=obama/economy/...`	Suffix י (possesive mark) `before=my child(poss)=my`	Prefix *wa* (3rd person plural) `before=they child(nsubj)=NNS`
Noun: Dative case `parent(iobj)`	Suffix ה (feminine mark) `child(nsubj)=she before=she`	Suffix *tu* (1st person plural) `child(nsubj)=she before=she`
Adjective: Genitive case `grandparent(poss)`	Prefix כש (when) `before=when before=WRB`	Prefix *ha* (negative tense) `source=no after=not`

Figure 3: Examples of highly weighted features learned by the inflection model. We selected a few frequent morphological features and show their top corresponding source context features.

data, we first construct an additional phrase-based translation model on the parallel corpus preprocessed to replace inflected surface words with their stems. We then extract a set of non-gappy phrases for each sentence (e.g., X → <attempted, пытаться_V>). The target side of each such phrase is re-inflected, conditioned on the source sentence, using the inflection model from §2. Each stem is given its most likely inflection.[7]

The original features extracted for the stemmed phrase are conserved, and the following features are added to help the decoder select good synthetic phrases:

- a binary feature indicating that the phrase is synthetic,
- the log-probability of the inflected forms according to our model,
- the count of words that have been inflected, with a separate feature for each morphological category in the supervised case.

Finally, these synthetic phrases are combined with the original translation rules obtained for the baseline system to produce an extended sentence-specific grammar which is used as input to the decoder. If a phrase already existing in the standard phrase table happens to be recreated, both phrases are kept and will compete with each other with different features in the decoder.

For example, for the large EN→RU system, 6% of all the rules used for translation are synthetic phrases, with 65% of these phrases being entirely new rules.

6 Translation Experiments

We evaluate our approach in the standard discriminative MT framework. We use cdec (Dyer et al., 2010) as our decoder and perform MIRA training to learn feature weights of the sentence translation model (Chiang, 2012). We compare the following configurations:

- A baseline system, using a 4-gram language model trained on the entire monolingual and bilingual data available.
- An enriched system with a class-based n-gram language model[8] trained on the monolingual data mapped to 600 Brown clusters. Class-based language modeling is a strong baseline for scenarios with high out-of-vocabulary rates but in which large amounts of monolingual target-language data are available.
- The enriched system further augmented with our inflected synthetic phrases. We expect the class-based language model to be especially

[7] Several reviewers asked about what happens when k-best inflections are added. The results for $k \in \{2, 4, 8\}$ range from no effect to an improvement over $k = 1$ of about 0.2 BLEU (absolute). We hypothesize that larger values of k could have a greater impact, perhaps in a more "global" model of the target string; however, exploration of this question is beyond the scope of this paper.

[8] For Swahili and Hebrew, $n = 6$; for Russian, $n = 7$.

helpful here and capture some basic agreement patterns that can be learned more easily on dense clusters than from plain word sequences.

Detailed corpus statistics are given in Table 1:

- The Russian data consist of the News Commentary parallel corpus and additional monolingual data crawled from news websites.[9]
- The Hebrew parallel corpus is composed of transcribed TED talks (Cettolo et al., 2012). Additional monolingual news data is also used.
- The Swahili parallel corpus was obtained by crawling the Global Voices project website[10] for parallel articles. Additional monolingual data was taken from the Helsinki Corpus of Swahili.[11]

We evaluate translation quality by translating and measuring the BLEU score of a 2000–3000 sentence-long evaluation corpus, averaging the results over 3 MIRA runs to control for optimizer instability (Clark et al., 2011). Table 4 reports the results. For all languages, using class language models improves over the baseline. When synthetic phrases are added, significant additional improvements are obtained. For the English–Russian language pair, where both supervised and unsupervised analyses can be obtained, we notice that expert-crafted morphological analyzers are more efficient at improving translation quality. Globally, the amount of improvement observed varies depending on the language; this is most likely indicative of the quality of unsupervised morphological segmentations produced and the kinds of grammatical relations expressed morphologically.

Finally, to confirm the effectiveness of our approach as corpus size increases, we use our technique on top of a state-of-the art English–Russian system trained on data from the 8th ACL Workshop on Machine Translation (30M words of bilingual text and 410M words of monolingual text). The setup is identical except for the addition of sparse rule shape indicator features and bigram cluster features. In these large scale conditions, the BLEU score improves from 18.8 to 19.6 with the addition of word clusters and reaches 20.0 with synthetic phrases. Details regarding this system are reported in Ammar et al. (2013).

Table 4: Translation quality (measured by BLEU) averaged over 3 MIRA runs.

	EN→RU	EN→HE	EN→SW
Baseline	14.7±0.1	15.8±0.3	18.3±0.1
+Class LM	15.7±0.1	16.8±0.4	18.7±0.2
+Synthetic			
unsupervised	16.2±0.1	17.6±0.1	19.0±0.1
supervised	16.7±0.1	—	—

7 Related Work

Translation into morphologically rich languages is a widely studied problem and there is a tremendous amount of related work. Our technique of synthesizing translation options to improve generation of inflected forms is closely related to the factored translation approach proposed by Koehn and Hoang (2007); however, an important difference to that work is that we use a discriminative model that conditions on source context to make "local" decisions about what inflections may be used before combining the phrases into a complete sentence translation.

Combination pre-/post-processing solutions are also frequently proposed. In these, the target language is generally transformed from multi-morphemic surface words into smaller units more amenable to direct translation, and then a post-processing step is applied independent of the translation model. For example, Oflazer and El-Kahlout (2007) experiment with partial morpheme groupings to produce novel inflected forms when translating into Turkish; Al-Haj and Lavie (2010) compare different processing schemes for Arabic. A related but different approach is to *enrich* the source language items with grammatical features (e.g., a source sentence like *John saw Mary* is preprocessed into, e.g., John+subj saw+msubj+fobj Mary+obj) so as to make the source and target lexicons have similar morphological contrasts (Avramidis and Koehn, 2008; Yeniterzi and Oflazer, 2010; Chang et al.,

[9] http://www.statmt.org/wmt13/translation-task.html

[10] http://sw.globalvoicesonline.org

[11] http://www.aakkl.helsinki.fi/cameel/corpus/intro.htm

2009). In general, this work suffers from the problem that it is extremely difficult to know *a priori* what the right preprocessing is for a given language pair, data size, and domain.

Several post-processing approaches have relied on supervised classifiers to predict the optimal complete inflection for an incomplete or lemmatized translation. Minkov et al. (2007) present a method for predicting the inflection of Russian and Arabic sentences aligned to English sentences. They train a sequence model to predict target morphological features from the lemmas and the syntactic structures of both aligned sentences and demonstrate its ability to recover accurately inflections on reference translations. Toutanova et al. (2008) apply this method to generate inflections after translation in two different ways: by rescoring inflected n-best outputs or by translating lemmas and re-inflecting them *a posteriori*. El Kholy and Habash (2012) follow a similar method and compare different approaches for generating rich morphology in Arabic after a translation step. Fraser et al. (2012) observe improvements for translation into German with a similar method. As in that work, we model morphological features rather than directly inflected forms. However, that work may be criticized for providing no mechanism to translate surface forms directly, even when evidence for a direct translation is available in the parallel data.

Unsupervised morphology has begun to play a role in translation between morphologically complex languages. Stallard et al. (2012) show that an unsupervised approach to Arabic segmentation performs as well as a supervised segmenter for source-side preprocessing (in terms of English translation quality). For translation *into* morphological rich languages, Clifton and Sarkar (2011) use an unsupervised morphological analyzer to produce morphological affixes in Finnish, injecting some linguistic knowledge in the generation process.

Several authors have proposed using conditional models to predict the probability of phrase translation in context (Gimpel and Smith, 2008; Chan et al., 2007; Carpuat and Wu, 2007; Jeong et al., 2010). Of particular note is the work of Subotin (2011), who use a conditional model to predict morphological features conditioned on rich linguistic features; however, this latter work also conditions on target context, which substantially complicates decoding.

Finally, synthetic phrases have been used for different purposes than generating morphology. Callison-Burch et al. (2006) expanded the coverage of a phrase table by adding synthesized phrases by paraphrasing source language phrases, Chen et al. (2011) produced "fabricated" phrases by paraphrasing both source and target phrases, and Habash (2009) created new rules to handle out-of-vocabulary words. In related work, Tsvetkov et al. (2013) used synthetic phrases to improve generation of (in)definite articles when translating into English from Russian and Czech, two languages which do not lexically mark definiteness.

8 Conclusion

We have presented an efficient technique that exploits morphologically analyzed corpora to produce new inflections possibly unseen in the bilingual training data. Our method decomposes into two simple independent steps involving well-understood discriminative models.

By relying on source-side context to generate additional local translation options and by leaving the choice of the full sentence translation to the decoder, we sidestep the difficulty of computing features on target translations hypotheses. However, many morphological processes (most notably, agreement) are most best modeled using target language context. To capture target context effects, we depend on strong target language models. Therefore, an important extension of our work is to explore the interaction of our approach with more sophisticated language models that more directly model morphology, e.g., the models of Bilmes and Kirchhoff (2003), or, alternatively, ways to incorporate target language context in the inflection model.

We also achieve language independence by exploiting unsupervised morphological segmentations in the absence of linguistically informed morphological analyses.

Code for replicating the experiments is available from `https://github.com/eschling/morphogen`; further details are available in (Schlinger et al., 2013).

Acknowledgments

This work was supported by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533. We would like to thank Kim Spasaro for curating the Swahili development and test sets, Yulia Tsvetkov for assistance with Russian, and the anonymous reviewers for their helpful comments.

References

Hassan Al-Haj and Alon Lavie. 2010. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. In *Proc. of AMTA*.

Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Wang Ling, Austin Matthews, Kenton Murray, Nicola Segall, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. 2013. The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proc. of WMT*.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proc. of ACL*.

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Proc. of NAACL*.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proc. of ACL*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Improved statistical machine translation using paraphrases. In *Proc. of NAACL*.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of EMNLP*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proc. of EAMT*.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*.

Pi-Chuan Chang, Dan Jurafsky, and Christopher D. Manning. 2009. Disambiguating "DE" for Chinese–English machine translation. In *Proc. of WMT*.

Boxing Chen, Roland Kuhn, and George Foster. 2011. Semantic smoothing and fabrication of phrase pairs for SMT. In *Proc. of IWSLT*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *JMLR*, 13:1159–1187.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of ACL*.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proc. of ACL*.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. of ACL*.

Ahmed El Kholy and Nizar Habash. 2012. Translate, predict or generate: Modeling rich morphology in statistical machine translation. In *Proc. of EAMT*.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proc. of EACL*.

Kevin Gimpel and Noah A. Smith. 2008. Rich source-side context for statistical machine translation. In *Proc. of WMT*.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proc. of ACL*.

Nizar Habash. 2009. REMOOV: A tool for online handling of out-of-vocabulary words in machine translation. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: A case study in Czech tagging. In *Proc. of ACL*.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2000. Statistical morphological disambiguation for agglutinative languages. In *Proc. of COLING*.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. 2010. A discriminative lexicon model for complex morphology. In *Proc. of AMTA*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *NIPS*, pages 641–648.

Mark Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proc. SIGMORPHON*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proc. of EMNLP*.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proc. of EMNLP*.

André F.T. Martins, Noah A. Smith, Eric P. Xing, Pedro M.Q. Aguiar, and Mário A.T. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proc. of EMNLP*.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proc. of ACL*.

Kemal Oflazer and İlknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proc. of WMT*.

Eva Schlinger, Victor Chahuneau, and Chris Dyer. 2013. morphogen: Translation into morphologically rich languages with synthetic phrases. *Prague Bulletin of Mathematical Linguistics*, (100).

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *Proc. of LREC*.

Noah A. Smith, David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proc. of EMNLP*.

David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, and Regina Barzilay. 2012. Unsupervised morphology rivals supervised morphology for Arabic MT. In *Proc. of ACL*.

Michael Subotin. 2011. An exponential translation model for target language morphology. In *Proc. ACL*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proc. of ACL*.

Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. 2013. Generating English determiners in phrase-based translation with synthetic translation options. In *Proc. of WMT*.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In *Proc. of ACL*.