Unsupervised Bilingual Morpheme Segmentation and Alignment with Context-rich Hidden Semi-Markov Models

Jason Naradowsky*
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
narad@cs.umass.edu

Kristina Toutanova
Microsoft Research
Redmond, WA 98502
kristout@microsoft.com

Abstract

This paper describes an unsupervised dynamic graphical model for morphological segmentation and bilingual morpheme alignment for statistical machine translation. The model extends Hidden Semi-Markov chain models by using factored output nodes and special structures for its conditional probability distributions. It relies on morpho-syntactic and lexical source-side information (part-of-speech, morphological segmentation) while learning a morpheme segmentation over the target language. Our model outperforms a competitive word alignment system in alignment quality. Used in a monolingual morphological segmentation setting it substantially improves accuracy over previous state-of-the-art models on three Arabic and Hebrew datasets.

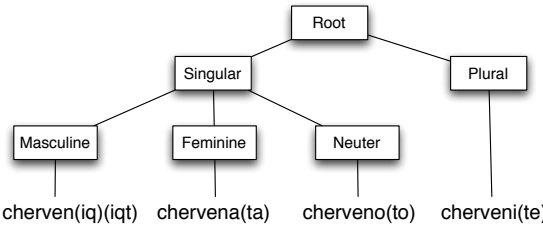
1 Introduction

An enduring problem in statistical machine translation is sparsity. The word alignment models of modern MT systems attempt to capture $p(e_i|f_j)$, the probability that token e_i is a translation of f_j. Underlying these models is the assumption that the word-based tokenization of each sentence is, if not optimal, at least appropriate for specifying a conceptual mapping between the two languages.

However, when translating between unrelated languages – a common task – disparate morphological systems can place an asymmetric conceptual burden on words, making the lexicon of one language much more coarse. This intensifies the problem of sparsity as the large number of word forms created through morphologically productive processes hinders attempts to find concise mappings between concepts.

For instance, Bulgarian adjectives may contain markings for gender, number, and definiteness. The following tree illustrates nine realized forms of the Bulgarian word for *red*, with each leaf listing the definite and indefinite markings.

Table 1: Bulgarian forms of *red*



Contrast this with English, in which this information is marked either on the modified word or by separate function words.

In comparison to a language which isn't morphologically productive on adjectives, the alignment model must observe nine times as much data (assuming uniform distribution of the inflected forms) to yield a comparable statistic. In an area of research where the amount of data available plays a large role in a system's overall performance, this sparsity can be extremely problematic. Further complications are created when lexical sparsity is compounded with the desire to build up alignments over increasingly larger contiguous phrases.

To address this issue we propose an alternative to word alignment: *morpheme alignment*, an alignment that operates over the smallest meaningful subsequences of words. By striving to keep a direct 1-to-1 mapping between corresponding semantic units across languages, we hope to find better estimates

*This research was conducted during the author's internship at Microsoft Research

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 895–904,
Portland, Oregon, June 19-24, 2011.

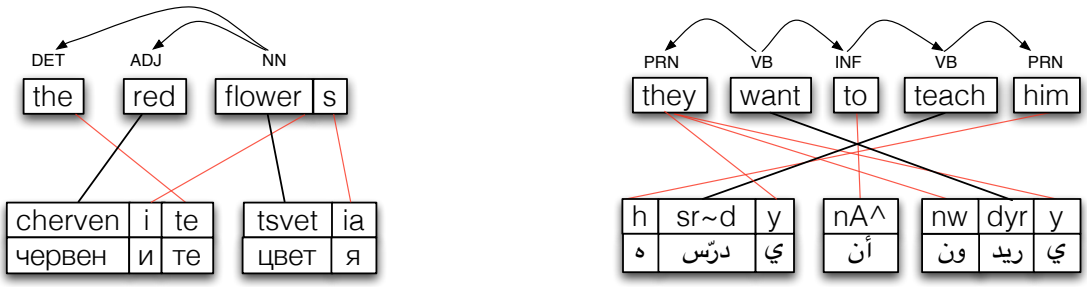


Figure 1: A depiction of morpheme-level alignment. Here dark lines indicate the more stem-focused alignment strategy of a traditional word or phrasal alignment model, while thin lines indicate a more fine-grained alignment across morphemes. In the alignment between English and Bulgarian (a) the morpheme-specific alignment reduces sparsity in the adjective and noun (*red flowers*) by isolating the stems from their inflected forms. Despite Arabic exhibiting templatic morphology, there are still phenomena which can be accounted for with a simpler segmentational approach. The Arabic alignment (b) demonstrates how the plural marker on English *they* would normally create sparsity by being marked in three additional places, two of them inflections in larger wordforms.

for the alignment statistics. Our results show that this improves alignment quality.

In the following sections we describe an unsupervised dynamic graphical model approach to monolingual morphological segmentation and bilingual morpheme alignment using a linguistically motivated statistical model. In a bilingual setting, the model relies on morpho-syntactic and lexical source-side information (part-of-speech, morphological segmentation, dependency analysis) while learning a morpheme segmentation over the target language. In a monolingual setting we introduce effective use of context by feature-rich modeling of the probabilities of morphemes, morpheme-transitions, and word boundaries. These additional sources of information provide powerful bias for unsupervised learning, without increasing the asymptotic running time of the inference algorithm.

Used as a monolingual model, our system significantly improves the state-of-the-art segmentation performance on three Arabic and Hebrew datasets. Used as a bilingual model, our system outperforms the state-of-the-art WDHMM (He, 2007) word alignment model as measured by alignment error rate (AER).

In agreement with some previous work on tokenization/morpheme segmentation for alignment (Chung and Gildea, 2009; Habash and Sadat, 2006), we find that the best segmentation for alignment does not coincide with the gold-standard segmentation and our bilingual model does not outperform our monolingual model in segmentation F-Measure.

2 Model

Our model defines the probability of a target language sequence of words (each consisting of a sequence of morphemes), and alignment from target to source morphemes, given a source language sequence of words (each consisting of a sequence of morphemes).

An example morpheme segmentation and alignment of phrases in English-Arabic and English-Bulgarian is shown in Figure 1. In our task setting, the words of the source and target language as well as the morpheme segmentation of the source (English) language are given. The morpheme segmentation of the target language and the alignments between source and target morphemes are hidden.

The source-side input, which we assume to be English, is processed with a gold morphological segmentation, part-of-speech, and dependency tree analysis. While these tools are unavailable in resource-poor languages, they are often available for at least one of the modeled languages in common translation tasks. This additional information then provides a source of features and conditioning information for the translation model.

Our model is derived from the hidden-markov model for word alignment (Vogel et al., 1996; Och and Ney, 2000). Based on it, we define a dynamic

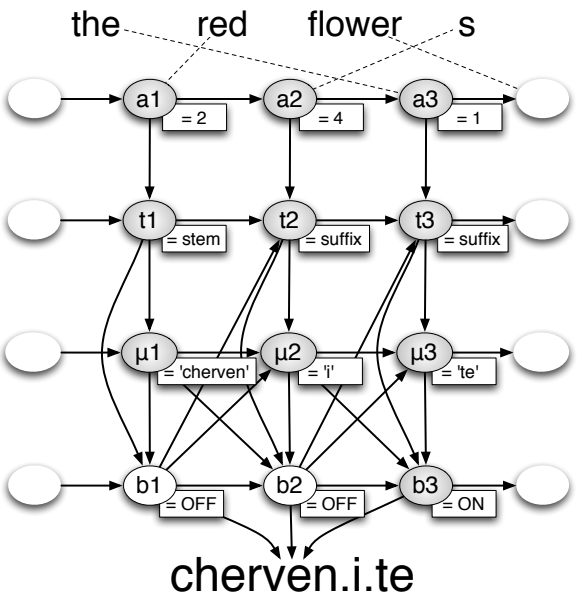


Figure 2: A graphical depiction of the model generating the transliteration of the first Bulgarian word from Figure 1. Trigram dependencies and some incoming/outgoing arcs have been omitted for clarity.

graphical model which lets us encode more linguistic intuition about morpheme segmentation and alignment: (*i*) we extend it to a hidden semi-markov model to account for hidden target morpheme segmentation; (*ii*) we introduce an additional observation layer to model observed word boundaries and thus truly represent target sentences as words composed of morphemes, instead of just a sequence of tokens; (*iii*) we employ hierarchically smoothed models and log-linear models to capture broader context and to better represent the morpho-syntactic mapping between source and target languages. (*iv*) we enrich the hidden state space of the model to encode morpheme types {prefix,suffix,stem}, in addition to morpheme alignment and segmentation information.

Before defining our model formally, we introduce some notation. Each possible morphological segmentation and alignment for a given sentence pair can be described by the following random variables:

Let $\mu_1\mu_2 \ldots \mu_I$ denote I morphemes in the segmentation of the target sentence. For the Example in Figure 1 (a) I=5 and μ_1=*cherven*, μ_2=*i* $\ldots$, and μ_5=*ia*. Let $b_1, b_2, \ldots b_I$ denote Bernoulli variables indicating whether there is a word boundary after morpheme μ_i. For our example, $b_3 = 1$, $b_5 = 1$, and the other b_i are 0. Let $c_1, c_2, \ldots, c_T$ denote the non-space characters in the target string, and $wb_1, \ldots, wb_T$ denote Bernoulli variables indicating whether there is a word boundary after the corresponding target character. For our example, $T = 14$ (for the Cyrillic version) and the only wb variables that are on are wb_9 and wb_{14}. The c and wb variables are observed. Let $s_1 s_2 \ldots s_T$ denote Bernoulli segmentation variables indicating whether there is a morpheme boundary after the corresponding character. The values of the hidden segmentation variables s together with the values of the observed c and wb variables uniquely define the values of the morpheme variables μ_i and the word boundary variables b_i. Naturally we enforce the constraint that a given word boundary $wb_t = 1$ entails a segmentation boundary $s_t = 1$. If we use bold letters to indicate a vector of corresponding variables, we have that $\mathbf{c}, \mathbf{wb}, \mathbf{s}{=}\boldsymbol{\mu}, \mathbf{b}$. We will define the assumed parametric form of the learned distribution using the $\boldsymbol{\mu}, \mathbf{b}$ but the inference algorithms are implemented in terms of the $\mathbf{s}$ and $\mathbf{wb}$ variables.

We denote the observed source language morphemes by $e_1 \ldots e_J$. Our model makes use of additional information from the source which we will mention when necessary.

The last part of the hidden model state represents the alignment between target and source morphemes and the type of target morphemes. Let $ta_i = [a_i, t_i]$, $i = 1 \ldots I$ indicate a factored state where a_i represents one of the J source words (or NULL) and t_i represents one of the three morpheme types {prefix,suffix,stem}. a_i is the source morpheme aligned to μ_i and t_i is the type of μ_i.

We are finally ready to define the desired probability of target morphemes, morpheme types, alignments, and word boundaries given source:

$$P(\boldsymbol{\mu}, \mathbf{ta}, \mathbf{b}|\mathbf{e}) = \prod_{i=1}^{I} P_T(\mu_i|ta_i, b_{i-1}, b_{i-2}, \mu_{i-1}, \mathbf{e}) \cdot P_B(b_i|\mu_i, \mu_{i-1}, ta_i, b_{i-1}, b_{i-2}, \mathbf{e}) \cdot P_D(ta_i|ta_{i-1}, b_{i-1}, \mathbf{e}) \cdot LP(|\mu_i|)$$

We now describe each of the factors used by our model in more detail. The formulation makes explicit the full extent of dependencies we have explored in this work. By simplifying the factors

we can recover several previously used models for monolingual segmentation and bilingual joint segmentation and alignment. We discuss the relationship of this model to prior work and study the impact of the novel components in our experiments.

When the source sentence is assumed to be empty (and thus contains no morphemes to align to) our model turns into a monolingual morpheme segmentation model, which we show exceeds the performance of previous state-of-the-art models. When we remove the word boundary component, reduce the order of the alignment transition, omit the morphological type component of the state space, and retain only minimal dependencies in the morpheme translation model, we recover the joint tokenization and alignment model based on IBM Model-1 proposed by (Chung and Gildea, 2009).

2.1 Morpheme Translation Model

In the model equation, P_T denotes the morpheme translation probability. The standard dependence on the aligned source morpheme is represented as a dependence on the state ta_i and the whole annotated source sentence $\mathbf{e}$. We experimented with multiple options for the amount of conditioning context to be included. When most context is used, there is a bigram dependency of target language morphemes as well as dependence on two previous boundary variables and dependence on the aligned source morpheme e_{a_i} as well as its POS tag.

When multiple conditioning variables are used we assume a special linearly interpolated backoff form of the model, similar to models routinely used in language modeling.

As an example, suppose we estimate the morpheme translation probability as $P_T(\mu_i|e_{a_i}, t_i)$. We estimate this in the M-step, given expected joint counts $c(\mu_i, e_{a_i}, t_i)$ and marginal counts derived from these as follows:

$P_T(\mu_i|e_{a_i}, t_i) = \frac{c(\mu_i, e_{a_i}, t_i) + \alpha_2 P_2(\mu_i|t_i)}{c(e_{a_i}, t_i) + \alpha_2}$

The lower order distributions are estimated recursively in a similar way:

$P_2(\mu_i|t_i) = \frac{c(\mu_i, t_i) + \alpha_1 P_1(\mu_i)}{c(t_i) + \alpha_1}$

$P_1(\mu_i) = \frac{c(\mu_i) + \alpha_0 P_0(\mu_i)}{c(.) + \alpha_0}$

For P_0 we used a unigram character language model. This hierarchical smoothing can be seen as an approximation to hierarchical Dirichlet priors with maximum aposteriori estimation.

Note how our explicit treatment of word boundary variables b_i allows us to use a higher order dependence on these variables. If word boundaries are treated as morphemes on their own, we would need to have a four-gram model on target morphemes to represent this dependency which we are now representing using only a bigram model on hidden morphemes.

2.2 Word Boundary Generation Model

The P_B distribution denotes the probability of generating word boundaries. As a sequence model of sentences the basic hidden semi-markov model completely ignores word boundaries. However, they can be powerful predictors of morpheme segments (by for example, indicating that common prefixes follow word boundaries, or that common suffixes precede them). The log-linear model of (Poon et al., 2009) uses word boundaries as observed left and right context features, and Morfessor (Creutz and Lagus, 2007) includes boundaries as special boundary symbols which can inform about the morpheme state of a morpheme (but not its identity).

Our model includes a special generative process for boundaries which is conditioned not only on the previous morpheme state but also the previous two morphemes and other boundaries. Due to the fact that boundaries are observed their inclusion in the model does not increase the complexity of inference.

The inclusion of this distribution lets us estimate the likelihood of a word consisting of one,two,three, or more morphemes. It also allows the estimation of likelihood that particular morphemes are in the beginning/middle/end of words. Through the included factored state variable ta_i word boundaries can also inform about the likelihood of a morpheme aligned to a source word of a particular pos tag to end a word. We discuss the particular conditioning context for this distribution we found most helpful in our experiments.

Similarly to the P_T distribution, we make use of multiple context vectors by hierarchical smoothing of distributions of different granularities.

2.3 Distortion Model

P_D indicates the distortion modeling distribution we use. [1] Traditional distortion models represent $P(a_j|a_{j-1}, \mathbf{e})$, the probability of an alignment given the previous alignment, to bias the model away from placing large distances between the aligned tokens of consecutively sequenced tokens. In addition to modeling a larger state space to also predict morpheme types, we extend this model by using a special log-linear model form which allows the integration of rich morpho-syntactic context. Log-linear models have been previously used in unsupervised learning for local multinomial distributions like this one in e.g. (Berg-Kirkpatrick et al., 2010), and for global distributions in (Poon et al., 2009).

The special log-linear form allows the inclusion of features targeted at learning the transitions among morpheme types and the transitions between corresponding source morphemes. The set of features with example values for this model is depicted in Table 3. The example is focussed on the features firing for the transition from the Bulgarian suffix *te* aligned to the first English morpheme $\mu_{i-1} = te$, t_{i-1}=suffix, a_{i-1}=1, to the Bulgarian root *tsvet* aligned to the third English morpheme $\mu_i = tsvet$, t_i=root, a_i=3. The first feature is the absolute difference between a_i and $a_{i-1} + 1$ and is similar to information used in other HMM word alignment models (Och and Ney, 2000) as well as phrase-translation models (Koehn, 2004). The alignment positions a_i are defined as indices of the aligned source morphemes. We additionally compute distortion in terms of distance in number of source words that are skipped. This distance corresponds to the feature name WORD DISTANCE. Looking at both kinds of distance is useful to capture the intuition that consecutive morphemes in the same target word should prefer to have a higher proximity of their aligned source words, as compared to consecutive morphemes which are not part of the same target word. The binned distances look at the sign of the distortion and bin the jumps into 5 bins, pooling the distances greater than 2 together. The feature SAME TARGET WORD indicates whether the two consecutive morphemes are part of the same word. In this case, they are not. This feature is not useful on its own because it does not distinguish between different alignment possibilities for ta_i, but is useful in conjunction with other features to differentiate the transition behaviors within and across target words. The DEP RELATION feature indicates the direct dependency relation between the source words containing the aligned source morphemes, if such relationship exists. We also represent alignments to null and have one null for each source word, similarly to (Och and Ney, 2000) and have a feature to indicate null. Additionally, we make use of several feature conjunctions involving the null, same target word, and distance features.

Feature	Value
MORPH DISTANCE	*1*
WORD DISTANCE	*1*
BINNED MORPH DISTANCE	*fore1*
BINNED WORD DISTANCE	*fore1*
MORPH STATE TRANSITION	suffix-root
SAME TARGET WORD	*False*
POS TAG TRANSITION	*DET-NN*
DEP RELATION	*DET←NN*
NULL ALIGNMENT	*False*
conjunctions	...

Figure 3: Features in log-linear distortion model firing for the transition from *te*:suffix:1 to *tsvet*:root:3 in the example sentence pair in Figure 1a.

2.4 Length Penalty

Following (Chung and Gildea, 2009) and (Liang and Klein, 2009) we use an exponential length penalty on morpheme lengths to bias the model away from the maximum likelihood under-segmentation solution. The form of the penalty is:

$LP(|\mu_i|) = \frac{1}{e^{|\mu_i|^{lp}}}$

Here lp is a hyper-parameter indicating the power that the morpheme length is raised to. We fit this parameter using an annotated development set, to optimize morpheme-segmentation F1. The model is extremely sensitive to this value and performs quite poorly if such penalty is not used.

2.5 Inference

We perform inference by EM training on the aligned sentence pairs. In the E-step we compute expected

[1] To reduce complexity of exposition we have omitted the final transition to a special state beyond the source sentence end after the last target morpheme.

counts of all hidden variable configurations that are relevant for our model. In the M-step we re-estimate the model parameters (using LBFGS in the M-step for the distortion model and using count interpolation for the translation and word-boundary models).

The computation of expectations in the E-step is of the same order as an order two semi-markov chain model using hidden state labels of cardinality ($J \times 3$ = number of source morphemes times number of target morpheme types). The running time of the forward and backward dynamic programming passes is $T \times l^2 \times (3J)^2$, where T is the length of the target sentence in characters, J is the number of source morphemes, and l is the maximum morpheme length. Space does not permit the complete listing of the dynamic programming solution but it is not hard to derive by starting from the dynamic program for the IBM-1 like tokenization model of (Chung and Gildea, 2009) and extending it to account for the higher order on morphemes and the factored alignment state space.

Even though the inference algorithm is low polynomial it is still much more expensive than the inference for an HMM model for word-alignment without segmentation. To reduce the running time of the model we limit the space of considered morpheme boundaries as follows:

Given the target side of the corpus, we derive a list of K most frequent prefixes and suffixes using a simple trie-based method proposed by (Schone and Jurafsky, 2000).[2] After we determine a list of allowed prefixes and suffixes we restrict our model to allow only segmentations of the form : ((*p**)*r*(*s**))+ where *p* and *s* belong to the allowed prefixes and suffixes and *r* can match any substring.

We determine the number of prefixes and suffixes to consider using the maximum recall achievable by limiting the segmentation points in this way. Restricting the allowable segmentations in this way not only improves the speed of inference but also leads to improvements in segmentation accuracy.

[2]Words are inserted into a trie with each complete branch naturally identifying a potential suffix, inclusive of its sub-branches. The list comprises of the K most frequent of these complete branches. Inserting the reversed words will then yield potential *prefixes*.

3 Evaluation

For a majority of our testing we borrow the parallel phrases corpus used in previous work (Snyder and Barzilay, 2008), which we refer to as S&B. The corpus consists of 6,139 short phrases drawn from English, Hebrew, and Arabic translations of the Bible. We use an unmodified version of this corpus for the purpose of comparing morphological segmentation accuracy. For evaluating morpheme alignment accuracy, we have also augmented the English/Arabic subset of the corpus with a gold standard alignment between morphemes. Here morphological segmentations were obtained using the previously-annotated gold standard Arabic morphological segmentation, while the English was preprocessed with a morphological analyzer and then further hand annotated with corrections by two native speakers. Morphological alignments were manually annotated. Additionally, we evaluate monolingual segmentation models on the full Arabic Treebank (ATB), also used for unsupervised morpheme segmentation in (Poon et al., 2009).

4 Results

4.1 Morpheme Segmentation

We begin by evaluating a series of models which are simplifications of our complete model, to assess the impact of individual modeling decisions. We focus first on a monolingual setting, where the source sentence aligned to each target sentence is empty.

Unigram Model with Length Penalty

The first model we study is the unigram monolingual segmentation model using an exponential length penalty as proposed by (Liang and Klein, 2009; Chung and Gildea, 2009), which has been shown to be quite accurate. We refer to this model as Model-UP (for unigram with penalty). It defines the probability of a target morpheme sequence as follows: $(\mu_1 \dots \mu_I) = (1-\theta) \prod_{i=1}^{I} \theta P_T(\mu_i) LP(|\mu_i|)$

This model can be (almost) recovered as a special case of our full model, if we drop the transition and word boundary probabilities, do not model morpheme types, and use no conditioning for the morpheme translation model. The only parameter not present in our model is the probability θ of generating a morpheme as opposed to stopping to gener-

ate morphemes (with probability $1 - \theta$). We experimented with this additional parameter, but found it had no significant impact on performance, and so we do not report results including it.

We select the value of the length penalty power by a gird search in the range 1.1 to 2.0, using .1 increments and choosing the values resulting in best performance on a development set containing 500 phrase pairs for each language. We also select the optimal number of prefixes/suffixes to consider by measuring performance on the development set. [3]

Morpheme Type Models

The next model we consider is similar to the unigram model with penalty, but introduces the use of the hidden ta states which indicate only morpheme types in the monolingual setting. We use the ta states and test different configurations to derive the best set of features that can be used in the distortion model utilizing these states, and the morpheme translation model. We consider two variants: (1) Model-HMMP-basic (for HMM model with length penalty), which includes the hidden states but uses them with a simple uniform transition matrix $P(ta_i|ta_{i-1}, b_{i-1})$ (uniform over allowable transitions but forbidding the prefixes from transitioning directly to suffixes, and preventing suffixes from immediately following a word boundary), and (2) a richer model Model-HMMP which is allowed to learn a log-linear distortion model and a feature rich translation model as detailed in the model definition. This model is allowed to use word boundary information for conditioning (because word boundaries are observed), but does not include the P_B predictive word boundary distribution.

Full Model with Word Boundaries

Finally we consider our full monolingual model which also includes the distribution predicting word boundary variables b_i. We term this model Model-FullMono. We detail the best context features for the conditional P_D distribution for each language. We initialize this model with the morpheme translation unigram distribution of ModelHMMP-basic, trained for 5 iterations.

Table 4 details the test set results of the different model configurations, as well as previously reported results on these datasets. For our main results we use the automatically derived list of prefixes and suffixes to limit segmentation points. The names of models that use such limited lists are prefixed by Dict in the Table. For comparison, we also report the results achieved by models that do not limit the segmentation points in this way.

As we can see the unigram model with penalty, Dict-Model-UP, is already very strong, especially on the S&B Arabic dataset. When the segmentation points are not limited, its performance is much worse. The introduction of hidden morpheme states in Dict-HMMP-basic gives substantial improvement on Arabic and does not change results much on the other datasets. A small improvement is observed for the unconstrained models.[4] When our model includes all components except word boundary prediction, Dict-Model-HMMP, the results are substantially improved on all languages. Model-HMMP is also the first unconstrained model in our sequence to approach or surpass previous state-of-the-art segmentation performance.

Finally, when the full model Dict-MonoFull is used, we achieve a substantial improvement over the previous state-of-the-art results on all three corpora, a 6.5 point improvement on Arabic, 6.2 point improvement on Hebrew, and a 9.3 point improvement on ATB. The best configuration of this model uses the same distortion model for all languages: using the morph state transition and boundary features. The translation models used only t_i for Hebrew and ATB and t_i and μ_{i-1} for Arabic. Word boundary was predicted using t_i in Arabic and Hebrew, and additionally using b_{i-1} and b_{i-2} for ATB. The unconstrained models without affix dictionaries are also very strong, outperforming previous state-of-the-art models. For ATB, the unconstrained model slightly outperforms the constrained one.

The segmentation errors made by this system shed light on how it might be improved. We find the dis-

[3] For the S&B Arabic dataset, we selected to use seven prefixes and seven suffixes, which correspond to maximum achievable recall of 95.3. For the S&B Hebrew dataset, we used six prefixes and six suffixes, for a maximum recall of 94.3. The Arabic treebank data required a larger number of affixes: we used seven prefixes and 20 suffixes, for a maximum recall of 98.3.

[4] Note that the inclusion of states in HMMP-basic only serves to provide a different distribution over the number of morphemes in a word, so it is interesting it can have a positive impact.

	Arabic			Hebrew			ATB		
	P	R	F1	P	R	F1	P	R	F1
UP	88.1	55.1	67.8	43.2	87.6	57.9	79.0	54.6	64.6
Dict-UP	85.8	73.1	78.9	57.0	79.4	66.3	61.6	91.0	73.5
HMMP-basic	83.3	58.0	68.4	43.5	87.8	58.2	79.0	54.9	64.8
Dict-HMMP-basic	84.8	76.3	80.3	56.9	78.8	66.1	69.3	76.2	72.6
HMMP	73.6	76.9	75.2	70.2	73.0	71.6	94.0	76.1	84.1
Dict-HMMP	82.4	81.3	81.8	62.7	77.6	69.4	85.2	85.8	85.5
MonoFull	80.5	87.3	83.8	72.2	71.7	72.0	86.2	88.5	**87.4**
Dict-MonoFull	86.1	83.2	**84.6**	73.7	72.5	**73.1**	92.9	81.8	87.0
Poon et. al	76.0	80.2	78.1	67.6	66.1	66.9	88.5	69.2	77.7
S&B-Best	67.8	77.3	72.2	64.9	62.9	63.9	–	–	–
Morfessor	71.1	60.5	65.4	65.4	57.7	61.3	77.4	72.6	74.9

Figure 4: Results on morphological segmentation achieved by monolingual variants of our model (top) with results from prior work are included for comparison (bottom). Results from models with a small, automatically-derived list of possible prefixes and suffixes are labeled as "Dict-" followed by the model name.

tributions over the frequencies of particular errors follow a Zipfian skew across both S&B datasets, with the Arabic being more pronounced (the most frequent error being made 27 times, with 627 errors being made just once) in comparison with the Hebrew (with the most frequent error being made 19 times, and with 856 isolated errors). However, in both the Arabic and Hebrew S&B tasks we find that a tendency to over-segment certain characters off of their correct morphemes and on to other frequently occurring, yet incorrect, particles is actually the cause of many of these isolated errors. In Arabic the system tends to over segment the character *aleph* (totally about 300 errors combined). In Hebrew the source of error is not as overwhelmingly directed at a single character, but *yod* and *he*, the latter functioning quite similarly to the problematic Arabic character and frequently turn up in the corresponding places of cognate words in Biblical texts.

We should note that our models select a large number of hyper-parameters on an annotated development set, including length penalty, hierarchical smoothing parameters α, and the subset of variables to use in each of three component sub-models. This might in part explain their advantage over previous-state-of-the-art models, which might use fewer (e.g. (Poon et al., 2009) and (Snyder and Barzilay, 2008)) or no specifically tuned for these datasets hyper-parameters (Morfessor (Creutz and Lagus, 2007)).

4.2 Alignment

Next we evaluate our full bilingual model and a simpler variant on the task of word alignment. We use the morpheme-level annotation of the S&B English-Arabic dataset and project the morpheme alignments to word alignments. We can thus compare alignment performance of the results of different segmentations. Additionally, we evaluate against a state-of-the-art word alignment system WDHMM (He, 2007), which performs comparably or better than IBM-Model4. The table in Figure 5 presents the results. In addition to reporting alignment error rate for different segmentation models, we report their morphological segmentation F1.

The word-alignment WDHMM model performs best when aligning English words to Arabic words (using Arabic as source). In this direction it is able to capture the many-to-one correspondence between English words and arabic morphemes. When we combine alignments in both directions using the standard grow-diag-final method, the error goes up.

We compare the (Chung and Gildea, 2009) model (termed Model-1) to our full bilingual model. We can recover Model-1 similarly to Model-UP, except now every morpheme is conditioned on an aligned source morpheme. Our full bilingual model outperforms Model-1 in both AER and segmentation F1. The specific form of the full model was selected as in the previous experiments, by choosing the model with best segmentations of the development set.

For Arabic, the best model conditions target mor-

	Arabic						Hebrew		
	Align P	Align R	AER	P	R	F1	P	R	F1
Model-1 (C&G 09)	91.6	81.2	13.9	72.4	76.2	74.3	61.0	71.8	65.9
Bilingual full	91.0	88.3	10.3	90.0	72.0	80.0	63.3	71.2	67.0
WDHMM E-to-A	82.4	96.7	11.1						
WDHMM GDF	82.1	94.6	12.1						

Figure 5: Alignment Error Rate (AER) and morphological segmentation F1 achieved by bilingual variants of our model. AER performance of WDHMM is also reported. Gold standard alignments are not available for the Hebrew data set.

phemes on source morphemes only, uses the boundary model with conditioning on number of morphemes in the word, aligned source part-of-speech, and type of target morpheme. The distortion model uses both morpheme and word-based absolute distortion, binned distortion, morpheme types of states, and aligned source-part-of-speech tags. Our best model for Arabic outperforms WDHMM in word alignment error rate. For Hebrew, the best model uses a similar boundary model configuration but a simpler uniform transition distortion distribution.

Note that the bilingual models perform worse than the monolingual ones in segmentation F1. This finding is in line with previous work showing that the best segmentation for MT does not necessarily agree with a particular linguistic convention about what morphemes should contain (Chung and Gildea, 2009; Habash and Sadat, 2006), but contradicts other results (Snyder and Barzilay, 2008). Further experimentation is required to make a general claim.

We should note that the Arabic dataset used for word-alignment evaluation is unconventionally small and noisy (the sentences are very short phrases, automatically extracted using GIZA++). Thus the phrases might not be really translations, and the sentence length is much smaller than in standard parallel corpora. This warrants further model evaluation in a large-scale alignment setting.

5 Related Work

This work is most closely related to the unsupervised tokenization and alignment models of Chung and Gildea (2009), Xu et al. (2008), Snyder and Barzilay (2008), and Nguyen et al. (2010).

Chung & Gildea (2009) introduce a unigram model of tokenization based on IBM Model-1,which is a special case of our model. Snyder and Barzilay (2008) proposes a hierarchical Bayesian model that combines the learning of monolingual segmentations and a cross-lingual alignment; their model is very different from ours.

Incorporating morphological information into MT has received reasonable attention. For example, Goldwater & McClosky (2005) show improvements when preprocessing Czech input to reflect a morphological decomposition using combinations of lemmatization, pseudowords, and morphemes. Yeniterzi and Oflazer (2010) bridge the morphological disparity between languages in a unique way by effectively aligning English syntactic elements (function words connected by dependency relations) to Turkish morphemes, using rule-based postprocessing of standard word alignment. Our work is partly inspired by that work and attempts to automate both the morpho-syntactic alignment and morphological analysis tasks.

6 Conclusion

We have described an unsupervised model for morpheme segmentation and alignment based on Hidden Semi-Markov Models. Our model makes use of linguistic information to improve alignment quality. On the task of monolingual morphological segmentation it produces a new state-of-the-art level on three datasets. The model shows quantitative improvements in both word segmentation and word alignment, but its true potential lies in its finer-grained interpretation of word alignment, which will hopefully yield improvements in translation quality.

Acknowledgements

We thank the ACL reviewers for their valuable comments on earlier versions of this paper, and Michael J. Burling for his contributions as a corpus annotator and to the Arabic aspects of this paper.

References

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, and Dan Klein. 2010. Unsupervised learning with features. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL)*.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process*.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *North American Chapter of the Association for Computational Linguistics*.

Xiaodong He. 2007. Using word-dependent transition models in HMM based word alignment for statistical machine translation. In *ACL 2nd Statistical MT workshop*, pages 80–87.

Philip Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *AMTA*.

P. Liang and D. Klein. 2009. Online EM for unsupervised models. In *North American Association for Computational Linguistics (NAACL)*.

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the International Conference on Computational Linguistics*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *North American Chapter of the Association for Computation Linguistics - Human Language Technologies 2009 conference (NAACL/HLT-09)*.

Patrick Schone and Daniel Jurafsky. 2000. Knowlege-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2000)*.

Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *ACL*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *In COLING 96: The 16th Int. Conf. on Computational Linguistics*.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised chinese word segmentation for statistical machine translation. In *COLING*.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proceedings of Association of Computational Linguistics*.