Modeling Syntactic Context Improves Morphological Segmentation

Yoong Keok Lee **Aria Haghighi** **Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{yklee, aria42, regina}@csail.mit.edu

Abstract

The connection between part-of-speech (POS) categories and morphological properties is well-documented in linguistics but underutilized in text processing systems. This paper proposes a novel model for morphological segmentation that is driven by this connection. Our model learns that words with common affixes are likely to be in the same syntactic category and uses learned syntactic categories to refine the segmentation boundaries of words. Our results demonstrate that incorporating POS categorization yields substantial performance gains on morphological segmentation of Arabic. [1]

1 Introduction

A tight connection between morphology and syntax is well-documented in linguistic literature. In many languages, morphology plays a central role in marking syntactic structure, while syntactic relations help to reduce morphological ambiguity (Harley and Phillips, 1994). Therefore, in an unsupervised linguistic setting which is rife with ambiguity, modeling this connection can be particularly beneficial.

However, existing unsupervised morphological analyzers take little advantage of this linguistic property. In fact, most of them operate at the vocabulary level, completely ignoring sentence context. This design is not surprising: a typical morphological analyzer does not have access to syntactic information, because morphological segmentation precedes other forms of sentence analysis.

In this paper, we demonstrate that morphological analysis can utilize this connection without assuming access to full-fledged syntactic information. In particular, we focus on two aspects of the morpho-syntactic connection:

- **Morphological consistency within POS categories.** Words within the same syntactic category tend to select similar affixes. This linguistic property significantly reduces the space of possible morphological analyses, ruling out assignments that are incompatible with a syntactic category.

- **Morphological realization of grammatical agreement.** In many morphologically rich languages, agreement between syntactic dependents is expressed via correlated morphological markers. For instance, in Semitic languages, gender and number agreement between nouns and adjectives is expressed using matching suffixes. Enforcing mutually consistent segmentations can greatly reduce ambiguity of word-level analysis.

In both cases, we do not assume that the relevant syntactic information is provided, but instead jointly induce it as part of morphological analysis.

We capture morpho-syntactic relations in a Bayesian model that grounds intra-word decisions in sentence-level context. Like traditional unsupervised models, we generate morphological structure from a latent lexicon of prefixes, stems, and suffixes.

[1] The source code for the work presented in this paper is available at http://groups.csail.mit.edu/rbg/code/morphsyn/.

In addition, morphological analysis is guided by a latent variable that clusters together words with similar affixes, acting as a proxy for POS tags. Moreover, a sequence-level component further refines the analysis by correlating segmentation decisions between adjacent words that exhibit morphological agreement. We encourage this behavior by encoding a transition distribution over adjacent words, using string match cues as a proxy for grammatical agreement.

We evaluate our model on the standard Arabic treebank. Our full model yields 86.2% accuracy, outperforming the best published results (Poon et al., 2009) by 8.5%. We also found that modeling morphological agreement between adjacent words yields greater improvement than modeling syntactic categories. Overall, our results demonstrate that incorporating syntactic information is a promising direction for improving morphological analysis.

2 Related Work

Research in unsupervised morphological segmentation has gained momentum in recent years bringing about significant developments to the area. These advances include novel Bayesian formulations (Goldwater et al., 2006; Creutz and Lagus, 2007; Johnson, 2008), methods for incorporating rich features in unsupervised log-linear models (Poon et al., 2009) and the development of multilingual morphological segmenters (Snyder and Barzilay, 2008a).

Our work most closely relates to approaches that aim to incorporate syntactic information into morphological analysis. Surprisingly, the research in this area is relatively sparse, despite multiple results that demonstrate the connection between morphology and syntax in the context of part-of-speech tagging (Toutanova and Johnson, 2008; Habash and Rambow, 2005; Dasgupta and Ng, 2007; Adler and Elhadad, 2006). Toutanova and Cherry (2009) were the first to systematically study how to incorporate part-of-speech information into lemmatization and empirically demonstrate the benefits of this combination. While our high-level goal is similar, our respective problem formulations are distinct. Toutanova and Cherry (2009) have considered a semi-supervised setting where an initial morphological dictionary and tagging lexicon are provided but the model also has access to unlabeled data. Since a lemmatizer and tagger trained in isolation may produce mutually inconsistent assignments, and their method employs a log-linear reranker to reconcile these decisions. This reranking method is not suitable for the unsupervised scenario considered in our paper.

Our work is most closely related to the approach of Can and Manandhar (2009). Their method also incorporates POS-based clustering into morphological analysis. These clusters, however, are learned as a separate preprocessing step using distributional similarity. For each of the clusters, the model selects a set of affixes, driven by the frequency of their occurrences in the cluster. In contrast, we model morpho-syntactic decisions jointly, thereby enabling tighter integration between the two. This design also enables us to capture additional linguistic phenomena such as agreement. While this technique yields performance improvement in the context of their system, the final results does not exceed state-of-the-art systems that do not exploit this information (for e.g., (Creutz and Lagus, 2007)).

3 Model

Given a corpus of unannotated and unsegmented sentences, our goal is to infer the segmentation boundaries of all words. We represent segmentations and syntactic categories as latent variables with a directed graphical model, and we perform Bayesian inference to recover the latent variables of interest. Apart from learning a compact morpheme lexicon that explains the corpus well, we also model morpho-syntactic relations both within each word and between adjacent words to improve segmentation performance. In the remaining section, we first provide the key linguistic intuitions on which our model is based before describing the complete generative process.

3.1 Linguistic Intuition

While morpho-syntactic interface spans a range of linguistic phenomena, we focus on two facets of this connection. Both of them provide powerful constraints on morphological analysis and can be modeled without explicit access to syntactic annotations.

Morphological consistency within syntactic category. Words that belong to the same syntactic category tend to select similar affixes. In fact, the power of affix-related features has been empirically shown in the task of POS tag prediction (Habash and Rambow, 2005). We hypothesize that this regularity can also benefit morphological analyzers by eliminating assignments with incompatible prefixes and suffixes. For instance, a state-of-the-art segmenter erroneously divides the word "Al{ntxAbAt" into four morphemes "Al-{ntxAb-A-t" instead of three "Al-{ntxAb-At" (translated as "the-election-s".) The affix assignment here is clearly incompatible — determiner "Al" is commonly associated with nouns, while suffix "A" mostly occurs with verbs.

Since POS information is not available to the model, we introduce a latent variable that encodes affix-based clustering. In addition, we consider a variant of the model that captures dependencies between latent variables of adjacent words (analogous to POS transitions).

Morphological realization of grammatical agreement. In morphologically rich languages, agreement is commonly realized using matching suffices. In many cases, members of a dependent pair such as adjective and noun have the exact same suffix. A common example in the Arabic Treebank is the bigram "Al-Df-p Al-grby-p" (which is translated word-for-word as "the-bank the-west") where the last morpheme "p" is a feminine singular noun suffix.

Fully incorporating agreement constraints in the model is difficult, since we do not have access to syntactic dependencies. Therefore, we limit our attention to adjacent words which end with similar strings – for e.g., "p" in the example above. The model encourages consistent segmentation of such pairs. While our string-based cue is a simple proxy for agreement relation, it turns to be highly effective in practice. On the Penn Arabic treebank corpus, our cue has a precision of around 94% at the token-level.

3.2 Generative Process

The high-level generative process proceeds in four phases:

(a) **Lexicon Model**: We begin by generating morpheme lexicons $\boldsymbol{L}$ using parameters $\boldsymbol{\gamma}$. This set of lexicons consists of separate lexicons for prefixes, stems, and suffixes generated in a hierarchical fashion.

(b) **Segmentation Model**: Conditioned on $\boldsymbol{L}$, we draw word types, their segmentations, and also their syntactic categories $(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{T})$.

(c) **Token-POS Model**: Next, we generate the unsegmented tokens in the corpus and their syntactic classes $(\boldsymbol{w}, \boldsymbol{t})$ from a standard first-order HMM which has dependencies between adjacent syntactic categories.

(d) **Token-Seg Model**: Lastly, we generate token segmentations $\boldsymbol{s}$ from a first-order Markov chain that has dependencies between adjacent segmentations.

The complete generative story can be summarized by the following equation:

$$
\begin{aligned}
P(\boldsymbol{w},\boldsymbol{s},\boldsymbol{t},\boldsymbol{W},\boldsymbol{S},\boldsymbol{T},\boldsymbol{L},\boldsymbol{\Theta},\boldsymbol{\theta}|\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta}) = \\
P(\boldsymbol{L}|\boldsymbol{\gamma}) \quad \text{(a)} \\
P(\boldsymbol{W},\boldsymbol{S},\boldsymbol{T},\boldsymbol{\Theta}|\boldsymbol{L},\gamma,\boldsymbol{\alpha}) \quad \text{(b)} \\
P_{\text{pos}}(\boldsymbol{w},\boldsymbol{t},\boldsymbol{\theta}|\boldsymbol{W},\boldsymbol{S},\boldsymbol{T},\boldsymbol{L},\boldsymbol{\alpha}) \quad \text{(c)} \\
P_{\text{seg}}(\boldsymbol{s}|\boldsymbol{W},\boldsymbol{S},\boldsymbol{T},\boldsymbol{L},\boldsymbol{\beta},\boldsymbol{\alpha}) \quad \text{(d)}
\end{aligned}
$$

where $\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{\beta}$ are hyperparameters and parameters whose roles we shall detail shortly.

Our lexicon model captures the desirability of compact lexicon representation proposed by prior work by using parameters $\boldsymbol{\gamma}$ that favors small lexicons. Furthermore, if we set the number of syntactic categories in the segmentation model to one and exclude the token-based models, we recover a segmenter that is very similar to the unigram Dirichlet Process model (Goldwater et al., 2006; Snyder and Barzilay, 2008a; Snyder and Barzilay, 2008b). We shall elaborate on this point in Section 4.

The segmentation model captures morphological consistency within syntactic categories (POS tag), whereas the Token-POS model captures POS tag dependencies between adjacent tokens. Lastly, the Token-Seg model encourages consistent segmentations between adjacent tokens that exhibit morphological agreement.

Lexicon Model The design goal is to encourage morpheme types to be short and the set of affixes (i.e. prefixes and suffixes) to be much smaller than the set of stems. To achieve this, we first draw each morpheme σ in the master lexicon L^* according to a geometric distribution which assigns monotonically smaller probability to longer morpheme lengths:

$$|\sigma| \sim \text{Geometric}(\gamma_l)$$

The parameter γ_l for the geometric distribution is fixed and specified beforehand. We then draw the prefix, the stem, and suffix lexicons (denoted by L_-, L_0, L_+ respectively) from morphemes in L^*. Generating the lexicons in such a hierarchical fashion allows morphemes to be shared among the lower-level lexicons. For instance, once determiner "Al" is generated in the master lexicon, it can be used to generate prefixes or stems later on. To favor compact lexicons, we again make use of a geometric distribution that assigns smaller probability to lexicons that contain more morphemes:

$$\begin{aligned} \text{prefix:} \quad & |L_-| \sim \text{Geometric}(\gamma_-) \\ \text{stem:} \quad & |L_0| \sim \text{Geometric}(\gamma_0) \\ \text{suffix:} \quad & |L_+| \sim \text{Geometric}(\gamma_+) \end{aligned}$$

By separating morphemes into affixes and stems, we can control the relative sizes of their lexicons with different parameters.

Segmentation Model The model independently generates each word type using only morphemes in the affix and stem lexicons, such that each word has exactly one stem and is encouraged to have few morphemes. We fix the number of syntactic categories (tags) to K and begin the process by generating multinomial distribution parameters for the POS tag prior from a Dirichlet prior:

$$\Theta_T \sim \text{Dirichlet}(\alpha_T, \{1, \ldots, K\})$$

Next, for each possible value of the tag $T \in \{1, \ldots, K\}$, we generate parameters for a multinomial distribution (again from a Dirichlet prior) for each of the prefix and the suffix lexicons:

$$\begin{aligned} \Theta_{-|T} &\sim \text{Dirichlet}(\alpha_-, L_-) \\ \Theta_0 &\sim \text{Dirichlet}(\alpha_0, L_0) \\ \Theta_{+|T} &\sim \text{Dirichlet}(\alpha_+, L_+) \end{aligned}$$

By generating parameters in this manner, we allow the multinomial distributions to generate only morphemes that are present in the lexicon. Also, at inference time, only morphemes in the lexicons receive pseudo-counts. Note that the affixes are generated conditioned on the tag; But the stem are not.[2]

Now, we are ready to generate each word type W, its segmentation S, and its syntactic category T. First, we draw the number of morpheme segments $|S|$ from a geometric distribution truncated to generate at most five morphemes:

$$|S| \sim \text{Truncated-Geometric}(\gamma_{|S|})$$

Next, we pick one of the morphemes to be the stem uniformly at random, and thus determine the number of prefixes and suffixes. Then, we draw the syntactic category T for the word. (Note that T is a latent variable which we recover during inference.)

$$T \sim \text{Multinomial}(\Theta_T)$$

After that, we generate each stem σ_0, prefix σ_-, and suffix σ_+ independently:

$$\begin{aligned} \sigma_0 &\sim \text{Multinomial}(\Theta_0) \\ \sigma_-|T &\sim \text{Multinomial}(\Theta_{-|T}) \\ \sigma_+|T &\sim \text{Multinomial}(\Theta_{+|T}) \end{aligned}$$

Token-POS Model This model captures the dependencies between the syntactic categories of adjacent tokens with a first-order HMM. Conditioned on the type-level assignments, we generate (unsegmented) tokens $\boldsymbol{w}$ and their POS tags $\boldsymbol{t}$:

$$\begin{aligned} &P_{\text{pos}}(\boldsymbol{w}, \boldsymbol{t}|\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{\theta}) \\ &= \prod_{w_i, t_i} P(t_{i-1}|t_i, \theta_{t|t}) P(w_i|t_i, \theta_{w|t}) \end{aligned}$$

where the parameters of the multinomial distributions are generated by Dirichlet priors:

$$\begin{aligned} \theta_{t|t} &\sim \text{Dirichlet}(\alpha_{t|t}, \{1, \ldots, K\}) \\ \theta_{w|t} &\sim \text{Dirichlet}(\alpha_{w|t}, \boldsymbol{W_t}) \end{aligned}$$

[2]We design the model as such since the dependencies between affixes and the POS tag are much stronger than those between the stems and tags. In our preliminary experiments, when stems are also generated conditioned on the tag, spurious stems are easily created and associated with garbage-collecting tags.

Here, $\boldsymbol{W_t}$ refers to the set of word types that are generated by tag t. In other words, conditioned on tag t, we can only generate word w from the set of word types in $\boldsymbol{W_t}$ which is generated earlier (Lee et al., 2010).

Token-Seg Model The model captures the morphological agreement between adjacent segmentations using a first-order Markov chain. The probability of drawing a sequence of segmentations $\boldsymbol{s}$ is given by

$$P_{\text{seg}}(\boldsymbol{s}|\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{L}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{(s_{i-1}, s_i)} p(s_i|s_{i-1})$$

For each pair of segmentations s_{i-1} and s_i, we determine: (1) if they should exhibit morpho-syntactic agreement, and (2) if their morphological segmentations are consistent. To answer the first question, we first obtain the final suffix for each of them. Next, we obtain n, the length of the longer suffix. For each segmentation, we define the *ending* to be the last n characters of the word. We then use matching endings as a proxy for morpho-syntactic agreement between the two words. To answer the second question, we use matching final suffixes as a cue for consistent morphological segmentations. To encode the linguistic intuition that words that exhibit morpho-syntactic agreement are likely to be morphological consistent, we define the above probability distribution to be:

$$p(s_i|s_{i-1}) = \begin{cases} \beta_1 & \text{if same endings and same final suffix} \\ \beta_2 & \text{if same endings but different final suffixes} \\ \beta_3 & \text{otherwise (e.g. no suffix)} \end{cases}$$

where $\beta_1 + \beta_2 + \beta_3 = 1$, with $\beta_1 > \beta_3 > \beta_2$. By setting β_1 to a high value, we encourage adjacent tokens that are likely to exhibit morpho-syntactic agreement to have the same final suffix. And by setting $\beta_3 > \beta_2$, we also discourage adjacent tokens with the same endings to be segmented differently. [3]

4 Inference

Given a corpus of unsegmented and unannotated word tokens $\boldsymbol{w}$, the objective is to recover values of all latent variables, including the segmentations $\boldsymbol{s}$.

$$P(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{L}|\boldsymbol{w}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \int P(\boldsymbol{w}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{W}, \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{L}, \boldsymbol{\Theta}, \boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\Theta} d\boldsymbol{\theta}$$

We want to sample from the above distribution using collapsed Gibbs sampling ($\boldsymbol{\Theta}$ and $\boldsymbol{\theta}$ integrated out.) In each iteration, we loop over each word type W_i and sample the following latent variables: its tag T_i, its segmentation S_i, the segmentations and tags for all of its token occurrences $(\boldsymbol{s}_i, \boldsymbol{t}_i)$, and also the morpheme lexicons $\boldsymbol{L}$:

$$P(\boldsymbol{L}, T_i, S_i, \boldsymbol{s}_i, \boldsymbol{t}_i| \boldsymbol{s}_{-i}, \boldsymbol{t}_{-i}, \boldsymbol{S}_{-i}, \boldsymbol{T}_{-i}, \boldsymbol{w}_{-i}, \boldsymbol{W}_{-i}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (1)$$

such that the type and token-level assignments are consistent, i.e. for all $t \in \boldsymbol{t}_i$ we have $t = T_i$, and for all $s \in \boldsymbol{s}_i$ we have $s = S_i$.

4.1 Approximate Inference

Naively sampling the lexicons $\boldsymbol{L}$ is computationally infeasible since their sizes are unbounded. Therefore, we employ an approximation which turns is similar to performing inference with a Dirichlet Process segmentation model. In our approximation scheme, for each possible segmentation and tag hypothesis $(T_i, S_i, \boldsymbol{s}_i, \boldsymbol{t}_i)$, we only consider one possible value for $\boldsymbol{L}$, which we denote the *minimal lexicons*. Hence, the total number of hypothesis that we have to consider is only as large as the number of possibilities for $(T_i, S_i, \boldsymbol{s}_i, \boldsymbol{t}_i)$.

Specifically, we recover the minimal lexicons as follows: for each segmentation and tag hypothesis, we determine the set of distinct affix and stem types in the whole corpus, including the morphemes introduced by segmentation hypothesis under consideration. This set of lexicons, which we call the minimal lexicons, is the most compact ones that are needed to generate all morphemes proposed by the current hypothesis.

Furthermore, we set the number of possible POS tags $K = 5$. [4] For each possible value of the tag, we consider all possible segmentations with at most five segments. We further restrict the stem to have no

[3] Although p sums to one, it makes the model deficient since, conditioned everything already generated, it places some probability mass on invalid segmentation sequences.

[4] We find that increasing K to 10 does not yield improvement.

more than two prefixes or suffixes and also that the stem cannot be shorter than the affixes. This further restricts the space of segmentation and tag hypotheses, and hence makes the inference tractable.

4.2 Sampling equations

Suppose we are considering the hypothesis with segmentation S and POS tag T for word type W_i. Let $\boldsymbol{L} = (L^*, L_-, L_0, L_+)$ be the minimal lexicons for this hypothesis (S, T). We sample the hypothesis $(S, T, s = S, t = T, \boldsymbol{L})$ proportional to the product of the following four equations.

Lexicon Model

$$\begin{aligned} &\prod_{\sigma \in L^*} \gamma_l (1-\gamma_l)^{|\sigma|} && \times \\ &\gamma_- (1-\gamma_-)^{|L_-|} && \times \\ &\gamma_0 (1-\gamma_0)^{|L_0|} && \times \\ &\gamma_+ (1-\gamma_+)^{|L_+|} && (2) \end{aligned}$$

This is a product of geometric distributions involving the length of each morpheme σ and the size of each of the prefix, the stem, and the suffix lexicons (denoted as $|L_-|, |L_0|, |L_+|$ respectively.) Suppose, a new morpheme type σ_0 is introduced as a stem. Relative to a hypothesis that introduces none, this one incurs an additional cost of $(1 - \gamma_0)$ and $\gamma_l (1 - \gamma_l)^{|\sigma_0|}$. In other words, the hypothesis is penalized for increasing the stem lexicon size and generating a new morpheme of length $|\sigma_0|$. In this way, the first and second terms play a role similar to the concentration parameter and base distribution in a DP-based model.

Segmentation Model

$$\begin{aligned} &\frac{\gamma_{|S|}(1-\gamma_{|S|})^{|S|}}{\sum_{j=0}^{5} \gamma_{|S|}(1-\gamma_{|S|})^{j}} && \times \\ &\frac{n_T^{-i} + \alpha}{N^{-i} + \alpha K} && \times \\ &\frac{n_{\sigma_0}^{-i} + \alpha_0}{N_0^{-i} + \alpha_0 |L_0|} && \times \\ &\frac{n_{\sigma_-|T}^{-i} + \alpha_-}{N_{-|T}^{-i} + \alpha_- |L_-|} && \times \\ &\frac{n_{\sigma_+|T}^{-i} + \alpha_+}{N_{+|T}^{-i} + \alpha_+ |L_+|} && (3) \end{aligned}$$

The first factor is the truncated geometric distribution of the number of segmentations $|S|$, and the second factor is the probability of generate the tag T. The rest are the probabilities of generating the stem σ_0, the prefix σ_-, and the suffix σ_+ (where the parameters of the multinomial distribution collapsed out). n_T^{-1} is the number of word types with tag T and N^{-i} is the total number of word types. $n_{\sigma_-|T}^{-i}$ refers to the number of times prefix σ_- is seen in all word types that are tagged with T, and $N_{-|T}^{-i}$ is the total number of prefixes in all word types that has tag T. All counts exclude the word type W_i whose segmentation we are sampling. If there is another prefix, $N_{-|T}^{-i}$ is incremented (and also $n_{\sigma_-|T}^{-i}$ if the second prefix is the same as the first one.) Integrating out the parameters introduces dependencies between prefixes. The rest of the notations read analogously.

Token-POS Model

$$\begin{aligned} &\frac{\alpha_{w|t}{}^{(m^i)}}{(M_t^{-i} + \alpha_{w|t}|\boldsymbol{W_t}|)^{(m^i)}} && \times \\ &\prod_{t=1}^{K} \prod_{t'=1}^{K} \frac{(m_{t'|t}^{-i} + \alpha_{t|t})^{(m_{t'|t}^{i})}}{(M_t^{-i} + \alpha_{t|t})^{(m_{t'|t}^{i})}} && (4) \end{aligned}$$

The two terms are the token-level emission and transition probabilities with parameters integrated out. The integration induces dependences between all token occurrences of word type W which results in ascending factorials defined as $\alpha^{(m)} = \alpha(\alpha + 1) \cdots (\alpha + m - 1)$ (Liang et al., 2010). M_t^{-i} is the number of tokens that have POS tag t, m^i is the number of tokens w_i, and $m_{t'|t}^{-i}$ is the number of tokens t-to-t' transitions. (Both exclude counts contributed by tokens belong to word type W_i.) $|\boldsymbol{W_t}|$ is the number of word types with tag t.

Token-Seg Model

$$\beta_1^{m_{\beta_1}^i} \beta_2^{m_{\beta_2}^i} \beta_3^{m_{\beta_3}^i} \qquad (5)$$

Here, $m_{\beta_1}^i$ refers to the number of transitions involving token occurrences of word type W_i that exhibit morphological agreement. This does not result in ascending factorials since the parameters of transition probabilities are fixed and not generated from Dirichlet priors, and so are not integrated out.

4.3 Staged Training

Although the Gibbs sampler mixes regardless of the initial state in theory, good initialization heuristics often speed up convergence in practice. We therefore train a series of models of increasing complexity (see section 6 for more details), each with 50 iterations of Gibbs sampling, and use the output of the preceding model to initialize the subsequent model. The initial model is initialized such that all words are not segmented. When POS tags are first introduced, they are initialized uniformly at random.

5 Experimental Setup

Performance metrics To enable comparison with previous approaches, we adopt the evaluation set-up of Poon et al. (2009). They evaluate segmentation accuracy on a per token basis, using recall, precision and F1-score computed on segmentation points. We also follow a transductive testing scenario where the same (unlabeled) data is used for both training and testing the model.

Data set We evaluate segmentation performance on the Penn Arabic Treebank (ATB).[5] It consists of about 4,500 sentences of modern Arabic obtained from newswire articles. Following the preprocessing procedures of Poon et al. (2009) that exclude certain word types (such as abbreviations and digits), we obtain a corpus of 120,000 tokens and 20,000 word types. Since our full model operates over sentences, we train the model on the entire ATB, but evaluate on the exact portion used by Poon et al. (2009).

Pre-defined tunable parameters and testing regime In all our experiments, we set $\gamma_l = \frac{1}{2}$ (for length of morpheme types) and $\gamma_{|S|} = \frac{1}{2}$ (for number of morpheme segments of each word.) To encourage a small set of affix types relative to stem types, we set $\gamma_- = \gamma_+ = \frac{1}{1.1}$ (for sizes of the affix lexicons) and $\gamma_0 = \frac{1}{10{,}000}$ (for size of the stem lexicon.) We employ a sparse Dirichlet prior for the type-level models (for morphemes and POS tag) by setting $\alpha = 0.1$. For the token-level models, we set hyperparameters for Dirichlet priors $\alpha_{w|t} = 10^{-5}$ (for unsegmented tokens) and $\alpha_{t|t} = 1.0$ (for POS tags transition.) To encourage adjacent words that exhibit morphological agreement to have the same final suffix, we set $\beta_1 = 0.6, \beta_2 = 0.1, \beta_1 = 0.3$.

In all the experiments, we perform five runs using different random seeds and report the mean score and the standard deviation.

[5] Our evaluation does not include the Hebrew and Arabic Bible datasets (Snyder and Barzilay, 2008a; Poon et al., 2009) since these corpora consists of short phrases that omit sentence context.

Model	R	P	F1	t-test
PCT 09	69.2	88.5	77.7	-
Morfessor-CAT	72.6	77.4	74.9	-
BASIC	71.4	86.7	78.3 (2.9)	-
+POS	75.4	87.4	81.0 (1.5)	+
+TOKEN-POS	75.7	88.5	81.6 (0.7)	$\sim$
+TOKEN-SEG(Full)	**82.1**	**90.8**	**86.2 (0.4)**	++

Table 1: Results on the Arabic Treebank (ATB) data set: We compare our models against Poon et al. (2009) (PCT09) and the Morfessor system (Morfessor-CAT). For our full model (+TOKEN-SEG) and its simplifications (BASIC, +POS, +TOKEN-POS), we perform five random restarts and show the mean scores. The sample standard deviations are shown in brackets. The last column shows results of a paired t-test against the preceding model: ++ (significant at 1%), + (significant at 5%), $\sim$ (not significant), - (test not applicable).

Baselines Our primary comparison is against the morphological segmenter of Poon et al. (2009) which yields the best published results on the ATB corpus. In addition, we compare against the Morfessor Categories-MAP system (Creutz and Lagus, 2007). Similar to our model, their system uses latent variables to induce clustering over morphemes. The difference is in the nature of the clustering: the Morfessor algorithm associates a latent variable for each morpheme, grouping morphemes into four broad categories (prefix, stem, suffix, and non-morpheme) but not introducing dependencies between affixes directly. For both systems, we quote their performance reported by Poon et al. (2009).

6 Results

Comparison with the baselines Table 1 shows that our full model (denoted +TOKEN-SEG) yields a mean F1-score of 86.2, compared to 77.7 and 74.9 obtained by the baselines. This performance gap corresponds to an error reduction of 38.1% over the best published results.

Ablation Analysis To assess relative impact of various components, we consider several simplified variants of the model:

- BASIC is the type-based segmentation model that is solely driven by the lexicon.[6]
- +POS adds latent variables but does not capture transitions and agreement constraints.
- +TOKEN-POS is equivalent to the full model, without agreement constraints.

Our results in Table 1 clearly demonstrate that modeling morpho-syntactic constraints greatly improves the accuracy of morphological segmentation.

We further examine the performance gains arising from improvements due to (1) encouraging morphological consistency within syntactic categories, and (2) morphological realization of grammatical agreement.

We evaluate our models on a subset of words that exhibit morphological consistency. Table 2 shows the accuracies for words that begin with the prefix "Al" (determiner) and end with a suffix "At" (plural noun suffix.) An example is the word "Al-{ntxAb-At" which is translated as "the-election-s". Such words make up about 1% of tokens used for evaluation, and the two affix boundaries constitute about 3% of the all gold segmentation points. By introducing a latent variable to capture dependencies between affixes, +POS is able to improve segmentation performance over BASIC. When dependencies between latent variables are introduced, +TOKEN-POS yields additional improvements.

We also examine the performance gains due to morphological realization of grammatical agreement. We select the set of tokens that share the same final suffix as the preceding token, such as the bigram "Al-Df-p Al-grby-p" (which is translated word-for-word as "the-bank the-west") where the last morpheme "p" is a feminine singular noun suffix. This subset makes up about 4% of the evaluation set, and the boundaries of the final suffixes take up about 5% of the total gold segmentation boundaries.

[6]The resulting model is similar in spirit to the unigram DP-based segmenter (Goldwater et al., 2006; Snyder and Barzilay, 2008a; Snyder and Barzilay, 2008b).

Model	Token		Type	
	F1	Acc.	F1	Acc.
BASIC	68.3	13.9	73.8	24.3
+POS	75.4	26.4	78.5	38.0
+TOKEN-POS	76.5	34.9	82.0	49.6
+TOKEN-SEG	84.0	49.5	85.4	57.7

Table 2: Segmentation performance on words that begin with prefix "Al" (determiner) and end with suffix "At" (plural noun suffix). The mean F1 scores are computed using all boundaries of words in this set. For each word, we also determine if both affixes are recovered while ignoring any other boundaries between them. The other two columns report this accuracy at both the type-level and the token-level.

Model	Token		Type	
	F1	Acc.	F1	Acc.
BASIC	85.6	70.6	79.5	58.6
+POS	87.6	76.4	82.3	66.3
+TOKEN-POS	87.5	75.2	82.2	65.3
+TOKEN-SEG	92.8	91.1	88.9	84.4

Table 3: Segmentation performance on words that have the same final suffix as their preceding words. The F1 scores are computed based on all boundaries within the words, but the accuracies are obtained using only the final suffixes.

Table 3 reveals this category of errors persisted until the final component (+TOKEN-SEG) was introduced.

7 Conclusion

Although the connection between syntactic (POS) categories and morphological structure is well-known, this relation is rarely exploited to improve morphological segmentation performance. The performance gains motivate further investigation into morpho-syntactic models for unsupervised language analysis.

Acknowledgements

This material is based upon work supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-10-1-0533. Thanks to the MIT NLP group and the reviewers for their comments.

References

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

Burcu. Can and Suresh Manandhar. 2009. Unsupervised learning of morphology by using syntactic categories. In *Working Notes, CLEF 2009 Workshop*.

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

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

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the ACL*, pages 673–680.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Heidi Harley and Colin Phillips, editors. 1994. *The Morphology-Syntax Connection*. Number 22 in MIT Working Papers in Linguistics. MIT Press.

Mark Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, June. Association for Computational Linguistics.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2010. Simple type-level unsupervised POS tagging. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 853–861, Cambridge, MA, October. Association for Computational Linguistics.

Percy Liang, Michael I. Jordan, and Dan Klein. 2010. Type-based mcmc. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 573–581, Los Angeles, California, June. Association for Computational Linguistics.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of HLT-NAACL 2009*, pages 209–217, Boulder, Colorado, June. Association for Computational Linguistics.

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

Benjamin Snyder and Regina Barzilay. 2008b. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio, June. Association for Computational Linguistics.

Kristina Toutanova and Colin Cherry. 2009. A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 486–494, Suntec, Singapore, August. Association for Computational Linguistics.

Kristina Toutanova and Mark Johnson. 2008. A bayesian lda-based model for semi-supervised part-of-speech tagging. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1521–1528. MIT Press, Cambridge, MA.