Unsupervised Multilingual Learning for Morphological Segmentation

Benjamin Snyder and Regina Barzilay
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{bsnyder,regina}@csail.mit.edu

Abstract

For centuries, the deep connection between languages has brought about major discoveries about human communication. In this paper we investigate how this powerful source of information can be exploited for unsupervised language learning. In particular, we study the task of morphological segmentation of multiple languages. We present a nonparametric Bayesian model that jointly induces morpheme segmentations of each language under consideration and at the same time identifies cross-lingual morpheme patterns, or *abstract morphemes*. We apply our model to three Semitic languages: Arabic, Hebrew, Aramaic, as well as to English. Our results demonstrate that learning morphological models in tandem reduces error by up to 24% relative to monolingual models. Furthermore, we provide evidence that our joint model achieves better performance when applied to languages from the same family.

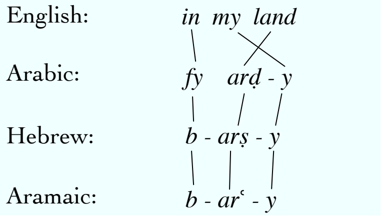
1 Introduction

For centuries, the deep connection between human languages has fascinated linguists, anthropologists and historians (Eco, 1995). The study of this connection has made possible major discoveries about human communication: it has revealed the evolution of languages, facilitated the reconstruction of proto-languages, and led to understanding language universals.

The connection between languages should be a powerful source of information for automatic linguistic analysis as well. In this paper we investigate two questions: *(i)* Can we exploit cross-lingual correspondences to improve unsupervised language learning? *(ii)* Will this joint analysis provide more or less benefit when the languages belong to the same family?

We study these two questions in the context of unsupervised morphological segmentation, the automatic division of a word into morphemes (the basic units of meaning). For example, the English word *misunderstanding* would be segmented into *mis - understand - ing*. This task is an informative testbed for our exploration, as strong correspondences at the morphological level across various languages have been well-documented (Campbell, 2004).

The model presented in this paper automatically induces a segmentation and morpheme alignment from a multilingual corpus of short parallel phrases.[1] For example, given parallel phrases meaning *in my land* in English, Arabic, Hebrew, and Aramaic, we wish to segment and align morphemes as follows:



This example illustrates the potential benefits of unsupervised multilingual learning. The three Semitic languages use cognates (words derived from a common ancestor) to represent the word *land*. They also use an identical suffix (*-y*) to represent the first person possessive pronoun (*my*). These similarities in form should guide the model by constraining

[1] In this paper, we focus on bilingual models. The model can be extended to handle several languages simultaneously as in this example.

the space of joint segmentations. The corresponding English phrase lacks this resemblance to its Semitic counterparts. However, in this as in many cases, no segmentation is required for English as all the morphemes are expressed as individual words. For this reason, English should provide a strong source of disambiguation for highly inflected languages, such as Arabic and Hebrew.

In general, we pose the following question. In which scenario will multilingual learning be most effective? Will it be for related languages, which share a common core of linguistic features, or for distant languages, whose linguistic divergence can provide strong sources of disambiguation?

As a first step towards answering this question, we propose a model which can take advantage of both similarities and differences across languages. This joint bilingual model identifies optimal morphemes for two languages and at the same time finds compact multilingual representations. For each language in the pair, the model favors segmentations which yield high frequency morphemes. Moreover, bilingual morpheme pairs which consistently share a common semantic or syntactic function are treated as *abstract morphemes*, generated by a single language-independent process. These abstract morphemes are induced automatically by the model from recurring bilingual patterns. For example, in the case above, the tuple *(in, fy, b-, b-)* would constitute one of three abstract morphemes in the phrase. When a morpheme occurs in one language without a direct counterpart in the other language, our model can explain away the stray morpheme as arising through a language-specific process.

To achieve this effect in a probabilistic framework, we formulate a hierarchical Bayesian model with Dirichlet Process priors. This framework allows us to define priors over the infinite set of possible morphemes in each language. In addition, we define a prior over abstract morphemes. This prior can incorporate knowledge of the phonetic relationship between the two alphabets, giving potential cognates greater prior likelihood. The resulting posterior distributions concentrate their probability mass on a small group of recurring and stable patterns within and between languages.

We test our model on a multilingual corpus of short parallel phrases drawn from the Hebrew Bible and Arabic, Aramaic, and English translations. The Semitic language family, of which Hebrew, Arabic, and Aramaic are members, is known for a highly productive morphology (Bravmann, 1977). Our results indicate that cross-lingual patterns can indeed be exploited successfully for the task of unsupervised morphological segmentation. When modeled in tandem, gains are observed for all language pairs, reducing relative error by as much as 24%. Furthermore, our experiments show that both related and unrelated language pairs benefit from multilingual learning. However, when common structures such as phonetic correspondences are explicitly modeled, related languages provide the most benefit.

2 Related Work

Multilingual Language Learning Recently, the availability of parallel corpora has spurred research on multilingual analysis for a variety of tasks ranging from morphology to semantic role labeling (Yarowsky et al., 2000; Diab and Resnik, 2002; Xi and Hwa, 2005; Padó and Lapata, 2006). Most of this research assumes that one language has annotations for the task of interest. Given a parallel corpus, the annotations are projected from this source language to its counterpart, and the resulting annotations are used for supervised training in the target language. In fact, Rogati et al., (2003) employ this method to learn arabic morphology assuming annotations provided by an English stemmer.

An alternative approach has been proposed by Feldman, Hana and Brew (2004; 2006). While their approach does not require a parallel corpus it does assume the availability of annotations in one language. Rather than being fully projected, the source annotations provide co-occurrence statistics used by a model in the resource-poor target language. The key assumption here is that certain distributional properties are invariant across languages from the same language families. An example of such a property is the distribution of part-of-speech bigrams. Hana et al., (2004) demonstrate that adding such statistics from an annotated Czech corpus improves the performance of a Russian part-of-speech tagger over a fully unsupervised version.

The approach presented here differs from previous work in two significant ways. First, we do not assume supervised data in any of the languages. Second, we learn a single multilingual model, rather than asymmetrically handling one language at a time. This design allows us to capitalize on structural regularities across languages for the mutual benefit of each language.

Unsupervised Morphological Segmentation Unsupervised morphology is an active area of research (Schone and Jurafsky, 2000; Goldsmith, 2001; Adler and Elhadad, 2006; Creutz and Lagus, 2007; Dasgupta and Ng, 2007).

Most existing algorithms derive morpheme lexicons by identifying recurring patterns in string distribution. The goal is to optimize the compactness of the data representation by finding a small lexicon of highly frequent strings. Our work builds on probabilistic segmentation approaches such as Morfessor (Creutz and Lagus, 2007). In these approaches, models with short description length are preferred. Probabilities are computed for both the morpheme lexicon and the representation of the corpus conditioned on the lexicon. A locally optimal segmentation is identified using a task-specific greedy search.

In contrast to previous approaches, our model induces morphological segmentation for multiple related languages simultaneously. By representing morphemes abstractly through the simultaneous alignment and segmentation of data in two languages, our algorithm capitalizes on deep connections between morpheme usage across different languages.

3 Multilingual Morphological Segmentation

The underlying assumption of our work is that structural commonality across different languages is a powerful source of information for morphological analysis. In this section, we provide several examples that motivate this assumption.

The main benefit of joint multilingual analysis is that morphological structure ambiguous in one language is sometimes explicitly marked in another language. For example, in Hebrew, the preposition meaning “in”, *b-*, is always prefixed to its nominal argument. On the other hand, in Arabic, the most common corresponding particle is *fy*, which appears as a separate word. By modeling cross-lingual morpheme alignments while simultaneously segmenting, the model effectively propagates information between languages and in this case would be encouraged to segment the Hebrew prefix *b-*.

Cognates are another important means of disambiguation in the multilingual setting. Consider translations of the phrase *“...and they wrote it...”*:

- Hebrew: *w-ktb-w ath*
- Arabic: *f-ktb-w-ha*

In both languages, the triliteral root *ktb* is used to express the act of writing. By considering the two phrases simultaneously, the model can be encouraged to split off the respective Hebrew and Arabic prefixes *w-* and *f-* in order to properly align the cognate root *ktb*.

In the following section, we describe a model that can model both generic cross-lingual patterns (*fy* and *b-*), as well as cognates between related languages (*ktb* for Hebrew and Arabic).

4 Model

Overview In order to simultaneously model probabilistic dependencies across languages as well as morpheme distributions within each language, we employ a hierarchical Bayesian model.[2]

Our segmentation model is based on the notion that stable recurring string patterns within words are indicative of morphemes. In addition to learning independent morpheme patterns for each language, the model will prefer, when possible, to join together frequently occurring bilingual morpheme pairs into single *abstract morphemes*. The model is fully unsupervised and is driven by a preference for stable and high frequency cross-lingual morpheme patterns. In addition the model can incorporate character-to-character phonetic correspondences between alphabets as prior information, thus allowing the implicit modeling of cognates.

Our aim is to induce a model which concentrates probability on highly frequent patterns while still allowing for the possibility of those previously unseen. Dirichlet processes are particularly suitable for such conditions. In this framework, we can encode

[2] In (Snyder and Barzilay, 2008) we consider the use of this model in the case where supervised data in one or more languages is available.

prior knowledge over the infinite sets of possible morpheme strings as well as abstract morphemes. Distributions drawn from a Dirichlet process nevertheless produce sparse representations with most probability mass concentrated on a small number of observed and predicted patterns. Our model utilizes a Dirichlet process prior for each language, as well as for the cross-lingual links (*abstract morphemes*). Thus, a distribution over morphemes and morpheme alignments is first drawn from the set of Dirichlet processes and then produces the observed data. In practice, we never deal with such distributions directly, but rather integrate over them during Gibbs sampling.

In the next section we describe our model's "generative story" for producing the data we observe. We formalize our model in the context of two languages $\mathcal{E}$ and $\mathcal{F}$. However, the formulation can be extended to accommodate evidence from multiple languages as well. We provide an example of parallel phrase generation in Figure 1.

High-level Generative Story We have a parallel corpus of several thousand short phrases in the two languages $\mathcal{E}$ and $\mathcal{F}$. Our model provides a generative story explaining how these parallel phrases were probabilistically created. The core of the model consists of three components: a distribution A over bilingual morpheme pairs (*abstract morphemes*), a distribution E over stray morphemes in language $\mathcal{E}$ occurring without a counterpart in language $\mathcal{F}$, and a similar distribution F for stray morphemes in language $\mathcal{F}$.

As usual for hierarchical Bayesian models, the generative story begins by drawing the model parameters themselves – in our case the three distributions A, E, and F. These three distributions are drawn from three separate Dirichlet processes, each with appropriately defined base distributions. The Dirichlet processes ensure that the resulting distributions concentrate their probability mass on a small number of morphemes while holding out reasonable probability for unseen possibilities.

Once A, E, and F have been drawn, we model our parallel corpus of short phrases as a series of independent draws from a phrase-pair generation model. For each new phrase-pair, the model first chooses the number and type of morphemes to be generated. In particular, it must choose how many unaligned stray morphemes from language $\mathcal{E}$, unaligned stray morphemes from language $\mathcal{F}$, and abstract morphemes are to compose the parallel phrases. These three numbers, respectively denoted as m, n, and k, are drawn from a Poisson distribution. This step is illustrated in Figure 1 part (a).

The model then proceeds to independently draw m language $\mathcal{E}$ morphemes from distribution E, n language-$\mathcal{F}$ morphemes from distribution F, and k abstract morphemes from distribution A. This step is illustrated in part (b) of Figure 1.

The $m + k$ resulting language-$\mathcal{E}$ morphemes are then ordered and fused to form a phrase in language $\mathcal{E}$, and likewise for the $n + k$ resulting language-$\mathcal{F}$ morphemes. The ordering and fusing decisions are modeled as draws from a uniform distribution over the set of all possible orderings and fusings for sizes m, n, and k. These final steps are illustrated in parts (c)-(d) of Figure 1. Now we describe the model more formally.

Stray Morpheme Distributions Sometimes a morpheme occurs in a phrase in one language without a corresponding foreign language morpheme in the parallel phrase. We call these "stray morphemes," and we employ language-specific morpheme distributions to model their generation.

For each language, we draw a distribution over all possible morphemes (finite-length strings composed of characters in the appropriate alphabet) from a Dirichlet process with concentration parameter α and base distribution P_e or P_f respectively:

$$\begin{aligned} E|\alpha, P_e &\sim DP(\alpha, P_e) \\ F|\alpha, P_f &\sim DP(\alpha, P_f) \end{aligned}$$

The base distributions P_e and P_f can encode prior knowledge about the properties of morphemes in each of the two languages, such as length and character n-grams. For simplicity, we use a geometric distribution over the length of the string with a final end-morpheme character. The distributions E and F which result from the respective Dirichlet processes place most of their probability mass on a small number of morphemes with the degree of concentration

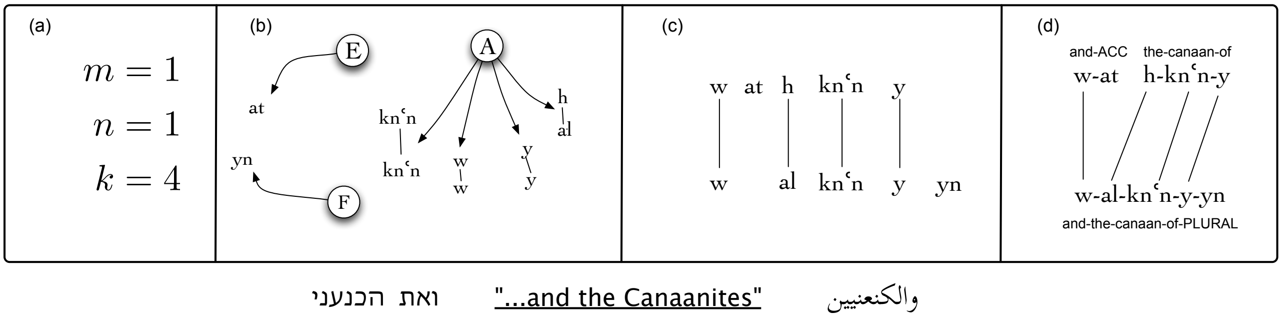


Figure 1: Generation process for a parallel bilingual phrase, with Hebrew shown on top and Arabic on bottom. (a) First the numbers of stray (m and n) and abstract (k) morphemes are drawn from a Poisson distribution. (b) Stray morphemes are then drawn from E and F (language-specific distributions) and abstract morphemes are drawn from A. (c) The resulting morphemes are ordered. (d) Finally, some of the contiguous morphemes are fused into words.

controlled by the prior α. Nevertheless, some non-zero probability is reserved for every possible string.

We note that these single-language morpheme distributions also serve as monolingual segmentation models, and similar models have been successfully applied to the task of word boundary detection (Goldwater et al., 2006).

Abstract Morpheme Distribution To model the connections between morphemes across languages, we further define a model for bilingual morpheme pairs, or *abstract morphemes*. This model assigns probabilities to all pairs of morphemes – that is, all pairs of finite strings from the respective alphabets – (e, f). Intuitively, we wish to assign high probability to pairs of morphemes that play similar syntactic or semantic roles (e.g. *(fy, b-)* for "in" in Arabic and Hebrew). These morpheme pairs can thus be viewed as representing *abstract morphemes*. As with the stray morpheme models, we wish to define a distribution which concentrates probability mass on a small number of highly co-occurring morpheme pairs while still holding out some probability for all other pairs.

We define this abstract morpheme model A as a draw from another Dirichlet process:

$$\begin{aligned} A|\alpha', P' \quad &\sim \quad DP(\alpha', P') \\ (e, f) \quad &\sim \quad A \end{aligned}$$

As before, the resulting distribution A will give non-zero probability to all abstract morphemes (e, f). The base distribution P' acts as a prior on such pairs. To define P', we can simply use a mixture of geometric distributions in the lengths of the component morphemes. However, if the languages $\mathcal{E}$ and $\mathcal{F}$ are related and the regular phonetic correspondences between the letter in the two alphabets are known, then we can use P' to assign higher likelihood to potential cognates. In particular we define the prior $P'(e, f)$ to be the probabilistic string-edit distance (Ristad and Yianilos, 1998) between e and f, using the known phonetic correspondences to parameterize the string-edit model. In particular, insertion and deletion probabilities are held constant for all characters, and substitution probabilities are determined based on the known sound correspondences.

We report results for both the simple geometric prior as well as the string-edit prior.

Phrase Generation To generate a bilingual parallel phrase, we first draw m, n, and k independently from a Poisson distribution. These three integers represent the number and type of the morphemes that compose the parallel phrase, giving the number of stray morphemes in each language $\mathcal{E}$ and $\mathcal{F}$ and the number of coupled bilingual morpheme pairs, respectively.

$$m, n, k \quad \sim \quad Poisson(\lambda)$$

Given these values, we now draw the appropriate number of stray and abstract morphemes from the corresponding distributions:

$$\begin{aligned} e_1, ..., e_m &\sim E \\ f_1, ..., f_n &\sim F \\ (e'_1, f'_1), ..., (e'_k, f'_k) &\sim A \end{aligned}$$

The sets of morphemes drawn for each language are then ordered:

$$\begin{aligned} \tilde{e}_1, ..., \tilde{e}_{m+k} &\sim ORDER|e_1, ..., e_m, e'_1, ..., e'_k \\ \tilde{f}_1, ..., \tilde{f}_{n+k} &\sim ORDER|f_1, ..., f_n, f'_1, ..., f'_k \end{aligned}$$

Finally the ordered morphemes are fused into the words that form the parallel phrases:

$$\begin{aligned} w_1, ..., w_s &\sim FUSE|\tilde{e}_1, ..., \tilde{e}_{m+k} \\ v_1, ..., v_t &\sim FUSE|\tilde{f}_1, ..., \tilde{f}_{n+k} \end{aligned}$$

To keep the model as simple as possible, we employ uniform distributions over the sets of orderings and fusings. In other words, given a set of r morphemes (for each language), we define the distribution over permutations of the morphemes to simply be $ORDER(\cdot|r) = \frac{1}{r!}$. Then, given a fixed morpheme order, we consider fusing each adjacent morpheme into a single word. Again, we simply model the distribution over the $r-1$ fusing decisions uniformly as $FUSE(\cdot|r) = \frac{1}{2^{r-1}}$.

Implicit Alignments Note that nowhere do we explicitly assign probabilities to morpheme alignments between parallel phrases. However, our model allows morphemes to be generated in precisely one of two ways: as a lone stray morpheme or as part of a bilingual abstract morpheme pair. Thus, our model implicitly assumes that each morpheme is either unaligned, or aligned to exactly one morpheme in the opposing language.

If we are given a parallel phrase with already segmented morphemes we can easily induce the distribution over alignments implied by our model. As we will describe in the next section, drawing from these induced alignment distributions plays a crucial role in our inference procedure.

Inference Given our corpus of short parallel bilingual phrases, we wish to make segmentation decisions which yield a set of morphemes with high joint probability. To assess the probability of a potential morpheme set, we need to marginalize over all possible alignments (i.e. possible abstract morpheme pairings and stray morpheme assignments). We also need to marginalize over all possible draws of the distributions A, E, and F from their respective Dirichlet process priors. We achieve these aims by performing Gibbs sampling.

Sampling We follow (Neal, 1998) in the derivation of our blocked and collapsed Gibbs sampler. Gibbs sampling starts by initializing all random variables to arbitrary starting values. At each iteration, the sampler selects a random variable X_i, and draws a new value for X_i from the conditional distribution of X_i given the current value of the other variables: $P(X_i|X_{-i})$. The stationary distribution of variables derived through this procedure is guaranteed to converge to the true joint distribution of the random variables. However, if some variables can be jointly sampled, then it may be beneficial to perform block sampling of these variables to speed convergence. In addition, if a random variable is not of direct interest, we can avoid sampling it directly by marginalizing it out, yielding a collapsed sampler. We utilize variable blocking by jointly sampling multiple segmentation and alignment decisions. We also collapse our Gibbs sampler in the standard way, by using predictive posteriors marginalized over all possible draws from the Dirichlet processes (resulting in Chinese Restaurant Processes).

Resampling For each bilingual phrase, we resample each word in the phrase in turn. For word w in language $\mathcal{E}$, we consider at once all possible segmentations, and for each segmentation all possible alignments. We keep fixed the previously sampled segmentation decisions for all other words in the phrase as well as sampled alignments involving morphemes in other words. We are thus considering at once: all possible segmentations of w along with all possible alignments involving morphemes in w with some subset of previously sampled language-$\mathcal{F}$ morphemes.[3]

[3] We retain morpheme identities during resampling of the morpheme alignments. This procedure is technically justi-

	Arabic			Hebrew		
	precision	recall	F-score	precision	recall	F-score
RANDOM	18.28	19.24	18.75	24.95	24.66	24.80
MORFESSOR	71.10	60.51	65.38	65.38	57.69	61.29
MONOLINGUAL	52.95	78.46	63.22	55.76	64.44	59.78
+ ARABIC/HEBREW	60.40	78.64	68.32	59.08	66.50	62.57
+ ARAMAIC	61.33	77.83	68.60	54.63	65.68	59.64
+ ENGLISH	63.19	74.79	68.49	60.20	64.42	62.23
+ ARAMAIC+PH	66.74	75.46	70.83	60.87	59.73	60.29
+ ARABIC/HEBREW+PH	67.75	77.29	**72.20**	64.90	62.87	**63.87**

Table 1: Precision, recall and F-score evaluated on Arabic and Hebrew. The first three rows provide baselines (random selection, an alternative state-of-the-art system, and the monolingual version of our model). The next three rows show the result of our bilingual model when one of Arabic, Hebrew, Aramaic, or English is added. The final two rows show the result of the bilingual model when character-to-character phonetic correspondences are used in the abstract morpheme prior.

The sampling formulas are easily derived as products of the relevant Chinese Restaurant Processes (with a minor adjustment to take into account the number of stray and abstract morphemes resulting from each decision). See (Neal, 1998) for general formulas for Gibbs sampling from distributions with Dirichlet process priors. All results reported are averaged over five runs using simulated annealing.

5 Experimental Set-Up

Morpheme Definition For the purpose of these experiments, we define *morphemes* to include conjunctions, prepositional and pronominal affixes, plural and dual suffixes, particles, definite articles, and roots. We do not model cases of infixed morpheme transformations, as those cannot be modeled by linear segmentation.

Dataset As a source of parallel data, we use the Hebrew Bible and translations. For the Hebrew version, we use an edition distributed by Westminster Hebrew Institute (Groves and Lowery, 2006). This Bible edition is augmented by gold standard morphological analysis (including segmentation) performed by biblical scholars.

For the Arabic, Aramaic, and English versions, we use the Van Dyke Arabic translation,[4] Targum Onkelos,[5] and the Revised Standard Version (Nelson, 1952), respectively. We obtained gold standard segmentations of the Arabic translation with a hand-crafted Arabic morphological analyzer which utilizes manually constructed word lists and compatibility rules and is further trained on a large corpus of hand-annotated Arabic data (Habash and Rambow, 2005). The accuracy of this analyzer is reported to be 94% for full morphological analyses, and 98%-99% when part-of-speech tag accuracy is not included. We don't have gold standard segmentations for the English and Aramaic portions of the data, and thus restrict our evaluation to Hebrew and Arabic.

To obtain our corpus of short parallel phrases, we preprocessed each language pair using the Giza++ alignment toolkit.[6] Given word alignments for each language pair, we extract a list of phrase pairs that form independent sets in the bipartite alignment graph. This process allows us to group together phrases like *fy ṣbaḥ* in Arabic and *bbqr* in Hebrew while being reasonably certain that all the relevant morphemes are contained in the short extracted phrases. The number of words in such phrases ranges from one to four words in the Semitic languages and up to six words in English. Before performing any experiments, a manual inspection of

fied by augmenting the model with a pair of "morpheme-identity" variables deterministically drawn from each abstract morpheme. Thus the identity of the drawn morphemes can be retained even while resampling their generation mechanism.

[4] http://www.arabicbible.com/bible/vandyke.htm

[5] http://www.mechon-mamre.org/i/t/u/u0.htm

[6] http://www.fjoch.com/GIZA++.html

the generated parallel phrases revealed that many infrequent phrase pairs occurred merely as a result of noisy translation and alignment. Therefore, we eliminated all parallel phrases that occur fewer than five times. As a result of this process, we obtain 6,139 parallel short phrases in Arabic, Hebrew, Aramaic, and English. The average number of morphemes per word in the Hebrew data is 1.8 and is 1.7 in Arabic.

For the bilingual models which employs probabilistic string-edit distance as a prior on abstract morphemes, we parameterize the string-edit model with the chart of Semitic consonant relationships listed on page *xxiv* of (Thackston, 1999). All pairs of corresponding letters are given equal substitution probability, while all other letter pairs are given substitution probability of zero.

Evaluation Methods Following previous work, we evaluate the performance of our automatic segmentation algorithm using F-score. This measure is the harmonic mean of recall and precision, which are calculated on the basis of all possible segmentation points. The evaluation is performed on a random set of 1/5 of the parallel phrases which is unseen during the training phase. During testing, *we do not allow the models to consider any multilingual evidence*. This restriction allows us to simulate future performance on purely monolingual data.

Baselines Our primary purpose is to compare the performance of our bilingual model with its fully monolingual counterpart. However, to demonstrate the competitiveness of this baseline model, we also provide results using MORFESSOR (Creutz and Lagus, 2007), a state-of-the-art unsupervised system for morphological segmentation. While developed originally for Finnish, this system has been successfully applied to a range of languages including German, Turkish and English. The probabilistic formulation of this model is close to our monolingual segmentation model, but it uses a greedy search specifically designed for the segmentation task. We use the publicly available implementation of this system. To provide some idea of the inherent difficulty of this segmentation task, we also provide results from a random baseline which makes segmentation decisions based on a coin weighted with the true segmentation frequency.

6 Results

Table 1 shows the performance of the various automatic segmentation methods. The first three rows provide baselines, as mentioned in the previous section. Our primary baseline is MONOLINGUAL, which is the monolingual counterpart to our model and only uses the language-specific distributions E or F. The next three rows shows the performance of various bilingual models that don't use character-to-character phonetic correspondences to capture cognate information. We find that with the exception of the HEBREW(+ARAMAIC) pair, the bilingual models show marked improvement over MONOLINGUAL. We notice that in general, adding English – which has comparatively little morphological ambiguity – is about as useful as adding a more closely related Semitic language. However, once character-to-character phonetic correspondences are added as an abstract morpheme prior (final two rows), we find the performance of related language pairs outstrips English, reducing relative error over MONOLINGUAL by 10% and 24% for the Hebrew/Arabic pair.

7 Conclusions and Future Work

We started out by posing two questions: *(i)* Can we exploit cross-lingual patterns to improve unsupervised analysis? *(ii)* Will this joint analysis provide more or less benefit when the languages belong to the same family? The model and results presented in this paper answer the first question in the affirmative, at least for the task of morphological segmentation.

We also provided some evidence that considering closely related languages may be more beneficial than distant pairs *if* the model is able to explicitly represent shared language structure (the character-to-character phonetic correspondences in our case). In the future, we hope to apply similar multilingual models to other core unsupervised analysis tasks, including part-of-speech tagging and grammar induction, and to further investigate the role that language relatedness plays in such models. [7]

[7] We acknowledge the support of the National Science Foundation (CAREER grant IIS-0448168 and grant IIS-0415865) and the Microsoft Research Faculty Fellowship. Thanks to members of the MIT NLP group for enlightening discussion.

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.

M. M. Bravmann. 1977. *Studies in Semitic Philology*. Leiden:Brill.

Lyle Campbell. 2004. *Historical Linguistics: An Introduction*. Cambridge: MIT Press.

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.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the ACL*, pages 255–262.

Umberto Eco. 1995. *The Search for the Perfect Language*. Wiley-Blackwell.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*.

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

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

Alan Groves and Kirk Lowery, editors. 2006. *The Westminster Hebrew Bible Morphology Database*. Westminster Hebrew Institute, Philadelphia, PA, USA.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambig uation in one fell swoop. In *Proceedings of the ACL*, pages 573–580.

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A resource-light approach to russian morphology: Tagging russian using czech resources. In *Proceedings of EMNLP*, pages 222–229.

Radford M. Neal. 1998. Markov chain sampling methods for dirichlet process mixture models. Technical Report 9815, Dept. of Statistics and Dept. of Computer Science, University of Toronto, September.

Thomas Nelson, editor. 1952. *The Holy Bible Revised Standard Version*. Thomas Nelson & Sons.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of ACL*, pages 1161 – 1168.

Eric Sven Ristad and Peter N. Yianilos. 1998. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532.

Monica Rogati, J. Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of the ACL*, pages 391–398.

Patrick Schone and Daniel Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the CoNLL*, pages 67–72.

Benjamin Snyder and Regina Barzilay. 2008. Cross-lingual propagation for morphological analysis. In *Proceedings of AAAI*.

Wheeler M. Thackston. 1999. *Introduction to Syriac*. Ibex Publishers.

Chenhai Xi and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-english languages. In *Proceedings of HLT/EMNLP*, pages 851 – 858.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2000. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, pages 161–168.