Cross-lingual Propagation for Morphological Analysis

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

Abstract

Multilingual parallel text corpora provide a powerful means for propagating linguistic knowledge across languages. We present a model which jointly learns linguistic structure for each language while inducing links between them. Our model supports fully symmetrical knowledge transfer, utilizing any combination of supervised and unsupervised data across language barriers. The proposed non-parametric Bayesian model effectively combines cross-lingual alignment with target language predictions. This architecture is a potent alternative to projection methods which decompose these decisions into two separate stages. We apply this approach to the task of morphological segmentation, where the goal is to separate a word into its individual morphemes. When tested on a parallel corpus of Hebrew and Arabic, our joint bilingual model effectively incorporates all available evidence from both languages, yielding significant performance gains.

Introduction

Since the discovery of the Rosetta stone in 1799, parallel text corpora have provided a powerful means for propagating linguistic knowledge across languages. In this paper, we propose a novel statistical framework for such information transfer. We present a model which jointly learns linguistic structure for each language while inducing links between them. This framework can clearly benefit languages with few resources by transferring knowledge from resource-rich languages. However, our model supports fully symmetrical knowledge transfer, allowing us to supplement annotated data in one language with annotated or raw data in another. In essence, our model is able to simultaneously propagate information in both directions between languages.

This framework offers two key advantages. Instead of simply importing annotations from a source to a target language, our method actually models cross-lingual structure. Thus we can utilize any combination of data, both supervised and unsupervised, for any available language. Secondly, our method unifies cross-lingual alignment and prediction, optimizing them jointly. Alignments themselves are then immediately sensitive to the target predictions, obviating the need for any task-specific alignment post-processing. These two properties distinguish our work from the projection framework, commonly used today for cross-lingual knowledge transfer.

In this paper we explore cross-lingual propagation in the context of morphological segmentation. This analysis attempts to automatically segment each word into its basic units of meaning – morphemes. For example, the English word *misunderstanding* would be properly segmented into *mis - understand - ing*. This task is particularly compelling and challenging for cross-lingual learning, as bilingual morpheme-to-morpheme alignment itself depends on predicted segmentation boundaries. For this reason, cross-lingual alignment cannot be treated as a mere prelude to the segmentation task.

The model presented in this paper simultaneously induces a segmentation and bilingual morpheme alignment in a unified probabilistic framework. For example, given parallel phrases meaning “in my land”:

- Arabic: *fy arḍy*
- Hebrew: *barṣy*

we wish to segment and align as follows:

- in: *fy / b-*
- land: *-arḍ- / arṣ-*
- my: *-y / -y*

We define a joint multilingual model which identifies optimal morphemes for each language and at the same time finds compactly representable cross-lingual alignments. For each language, the model favors segmentations which yield high frequency morphemes, incorporating evidence from supervised training data if available.

In addition, we introduce a notion of *abstract morphemes*. Bilingual morpheme pairs which consistently share a common semantic or syntactic function are treated as a unit generated by a single language-independent probabilistic process. These abstract morphemes are induced automatically from recurring bilingual patterns and form the multilingual backbone of the parallel phrases. When a morpheme occurs in one language without a direct counterpart in the other language, the model can explain away the stray morpheme with a language-specific model. The resulting effect of this design is to predict segmentations that yield high frequency

individual-language morphemes as well as common bilingual morpheme pairs.

To achieve this effect in a sound probabilistic framework, we formulate a non-parametric hierarchical Bayesian model with Dirichlet Process priors. This framework allows us to define prior distributions over the infinite set of possible morphemes in each language as well as over all possible aligned pairs of morphemes across languages. Despite assigning positive probability to every possible string, the resulting distributions concentrate their probability mass on a small group of recurring and stable patterns. This feature captures the intuition that hidden structure which most succinctly represents the data should be inferred.

We evaluate our model on a parallel corpus of Hebrew and Arabic short phrases extracted from the Bible. First we show that our model propagates information effectively from a resource-rich source language to a target language with no annotation. Furthermore, we find that even the resulting predictions in the annotated *source* language improve significantly through cross-lingual learning. This result is striking given that the only new source of information is unannotated data in a foreign language. Finally, we find that when the cross-lingual model is given supervised data in both languages, the performance exceeds that of individually trained monolingual models.

Related Work

Cross-lingual Learning Today, most work on cross-lingual learning is performed in the projection framework introduced by (Yarowsky, Ngai, and Wicentowski 2000). By definition projection is asymmetrical: knowledge is transferred from a resource-rich source language to a resource-poor target language. This method follows a two-step architecture. First, annotations are projected from the source language to the target language via alignments. The resulting pseudo-annotations are in turn used for supervised learning in the target language. Since the quality of the projected annotations is crucial for successful learning, the accuracy of the alignment step is always a concern for projection methods. Multiple empirical results have demonstrated that using out-of-the-box alignment methods, such as GIZA++, is insufficient. Therefore, researchers have developed task-specific methods for filtering spurious mappings and post-correcting automatic alignments (Hwa et al. 2005; Rogati, McCarley, and Yang 2003; Padó and Lapata 2005). For instance, in projecting semantic roles, alignments are refined by analyzing syntactic constituents and removing those with low similarity scores. In this and other applications, the alignment correction step is essential for successful learning.

In contrast to the projection framework, our model is fully symmetric in the parallel languages. This feature allows multi-directional knowledge transfer between languages, each with possibly different levels of supervision. Furthermore, our method performs alignment and prediction simultaneously, thus supporting interaction between the two steps. In this way, alignments induced by the model optimize the performance of the task at hand. This design reduces the need for crafting post-correction rules for each application anew, and allows cross-lingual learning when alignment cannot logically be separated from the task.

Projection for Morphological Segmentation Morphological analysis has been extensively studied in the past. We focus our discussion on two algorithms that employ projection for morphological analysis. The first projection algorithm developed by (Yarowsky, Ngai, and Wicentowski 2000) aims to predict the roots of both regular and irregular conjugated words. Root information is projected from English to French using multi-step transitive associations computed from IBM alignments. Since the mapping procedure does not analyze the underlying strings, annotations for words that do not have a corresponding root form in the target corpus cannot be projected. This annotation deficiency could be particularly severe for semitic languages known for their rich morphology.

(Rogati, McCarley, and Yang 2003) address this limitation in the context of Arabic by modeling probabilities of prefixes and suffixes as well as stems. An important source of the model's performance is the availability of an additional large monolingual corpus in the target language. Our segmentation task differs from the task of lemmatization explored in these two papers. Our goal is to fully segment a word into its constituent morphemes, whether those consist of multiple lemmas (as in agglutinative languages) or a rich vocabulary of prepositions and grammatical markers (as in the Semitic languages). In this scenario, alignment and segmentation are logically codependent and thus these two stages cannot be executed in a sequential manner.

Cross-lingual Propagation Scenarios

In this paper we consider three scenarios of cross-lingual knowledge transfer. We assume that during training, a parallel corpus is available. The presence of annotated data varies in each scenario:

1. **Indirect Supervision:** No supervised data is available in the target language. However annotated resources exist for source languages. This is the classical scenario for projection methods (Yarowsky, Ngai, and Wicentowski 2000).
2. **Augmentation:** Supervised data is already available in the language of interest, but can be supplemented with *unsupervised* data in additional languages.
3. **Shared Supervision:** Supervised data is available for both languages.

This range of scenarios mirrors actual multilingual data availability. Even when supervised data is available in both languages, the annotation may not cover the entire parallel corpus. Moreover, each language in the parallel corpus may have a different set of data points annotated. Therefore, in all scenarios with annotations, we assume the presence of additional unannotated parallel data.

Our model effectively generalizes the idea of multilingual knowledge transfer beyond the simple scenario of indirect supervision. While the benefits of indirect supervision seem intuitively plausible (and have been previously demonstrated through the use of annotation projection in other tasks), it is

not at all obvious whether data in a foreign language can be of use when native supervision is already present. We hypothesize that knowledge of shared linguistic structure can in fact yield benefits across the range of these scenarios. The results of this paper provide some evidence for this claim.

Model

Overview

In order to simultaneously model probabilistic dependencies across languages as well as morpheme distributions within each language, we employ a Bayesian graphical model. This setting allows us to easily accommodate supervised as well as unsupervised data in a single framework. This class of models allows inference to be easily performed with a mix of observed and unobserved variables. When the quantity of observed data in a particular language is limited, the model can fall back upon reasonable prior distributions.

Our segmentation model is based on the notion that stable recurring string patterns within words are indicative of morphemes. These patterns are anchored by the presence of supervised segmentations in one or more languages. Our model assigns high probability to morphemes with high actual frequency in supervised portions and high potential frequency in unsupervised portions of the data. In this way, supervised and unsupervised learning are integrated in a single probabilistic framework.

In addition to learning independent morpheme patterns for each language, the model will prefer, when possible, to join together frequently co-occurring bilingual morpheme pairs into single *abstract morphemes*. This aspect of the model is fully unsupervised and is driven by a preference for stable and high frequency cross-lingual morpheme patterns. Our model thus combines the totality of supervised and unsupervised data within each language with unsupervised learning of cross-lingual morpheme alignments.

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 prior information over the infinite set of possible morpheme strings as well as aligned morpheme pairs. 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[1] 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, 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 a generated parallel bilingual phrase in Figure 1.

[1]The use of Dirichlet processes in a Bayesian framework has found recent success across a variety of applications in NLP (Goldwater, Griffiths, and Johnson 2006; Finkel, Grenager, and Manning 2007; Haghighi and Klein 2007).

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 generated probabilistically. 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. This ensures that the resulting distributions concentrate their probability mass on a small number of morphemes while holding out reasonable probability for unseen possibilities.

Once these distributions 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).

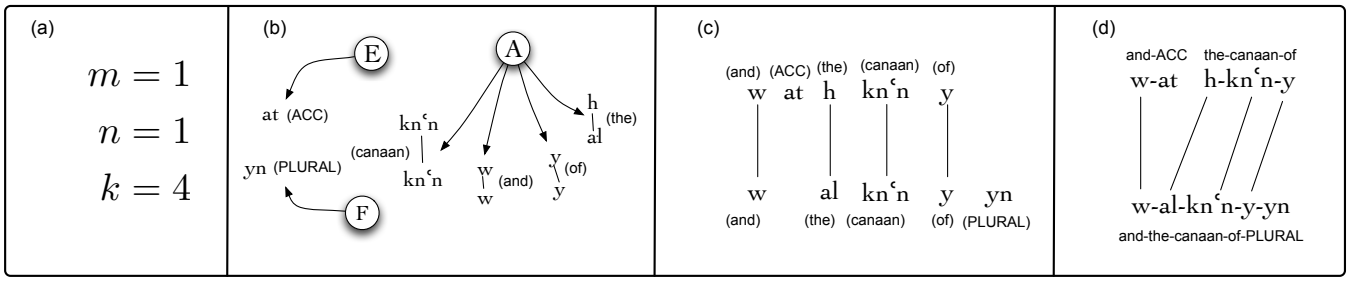
The model then proceeds to independently draw m language-$\mathcal{E}$ morphemes from distribution E, n language-$\mathcal{F}$ morphemes from distribution F, and k abstract morphemes from distribution A. This step is illustrated in part (b) of Figure 1.

The $m + k$ resulting language-$\mathcal{E}$ morphemes are then ordered and fused to form a phrase in language $\mathcal{E}$, and the $n+k$ resulting language-$\mathcal{F}$ morphemes are ordered and fused to form the parallel phrase in language $\mathcal{F}$. 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

To model morphemes which occur in a phrase in one language without a corresponding foreign language morpheme in the parallel phrase ("stray morphemes"), we employ language-specific morpheme distributions.

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:



ואת הכנעני "...and the Canaanites" والكنعنيين

Figure 1: Generation process for a parallel bilingual phrase. (a) First the numbers of stray and abstract 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.

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

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. 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 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, Griffiths, and Johnson 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 single language 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:

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

As before, the resulting distribution A will give non-zero probability to all abstract morphemes (e, f). The base distribution P', which acts as a prior on such pairs, can encode existing cross-lingual knowledge by, for example, taking the string edit distance between e and f into account. However in our experiments for this paper we simply used a mixture of geometric distributions in the lengths of the component morphemes.

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.

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

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 set of morphemes drawn for each language are then ordered:

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

and finally 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 and $FUSE(\cdot|r) = \frac{1}{2^{r-1}}$.

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 in turn each word in the phrase. Consider word w in language $\mathcal{E}$. If the word has been segmented by manual annotation, then that segmentation stays fixed and we need only resample alignment decisions. Otherwise, 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.[2]

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.

Experimental Set-Up

Morpheme Definition Our practical definition of a *morpheme* includes conjunctions, prepositional and pronominal affixes, plural and dual suffixes, particles, and roots. We do not consider 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 an Arabic translation. For the Hebrew version, we use an edition distributed by the Westminster Hebrew Institute (Groves and Lowery 2006). This Bible edition is augmented by gold standard morphological analysis which includes segmentation. This analysis has been performed by linguists and biblical scholars.

For the Arabic version, we use the Van Dyke Arabic translation.[3] We obtain gold standard segmentation 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).

To obtain our corpus of short parallel phrases, we preprocessed the biblical data using the Giza++ alignment toolkit.[4] Given word alignments between the Arabic and Hebrew versions of the Bible, we extract a list of phrase pairs that form

[2] We retain morpheme identities during resampling of the morpheme alignments. This procedure is technically justified 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.

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

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

	key	Arabic			Hebrew		
		precision	recall	f-measure	precision	recall	f-measure
RANDOM	u	18.28	19.24	18.75	24.95	24.66	24.80
MORFESSOR	u	71.10	60.51	65.38	65.38	57.69	61.29
UNSUPERVISED	u	52.95	78.46	63.22	55.76	64.44	59.78
SUPERVISED	s	73.15	92.39	81.65	71.44	79.14	75.09
INDIRECT SUPERVISION	us	64.37	80.81	71.65	63.10	68.02	65.46
AUGMENTATION	su	76.01	92.21	83.33	75.98	78.04	77.00
SHARED SUPERVISION	ss	77.78	92.26	84.39	76.81	79.17	77.98

Table 1: Precision, recall and f-measure for Arabic and Hebrew in various cross-lingual scenarios. The key indicates the combination of supervised/unsupervised for target language (left) and source language (right).

independent sets in the bipartite alignment graph. This process allows us to group together phrases like *fy alṣbaḥ* in Arabic and *bbqr* in Hebrew while being reasonably certain that all the relevant morphemes are contained in the short extracted phrases. The monolingual part of such phrases ranges from one to three words. Before performing any experiments, a manual inspection of the extracted phrases revealed that many infrequent phrase pairs occurred merely as a result of noisy translation and alignment. Therefore, we eliminated all phrase pairs that occur fewer than five times. As a result of this process, we obtain 6,139 phrase pairs. The number of morphemes per word ranges from 1 to 6 for both Hebrew and for Arabic. The average number of morphemes per word in the Hebrew gold data is 1.8, and is 1.7 for Arabic.

Training and Testing We randomly select 1/5 of the data for testing, and 4/5 for training. As discussed above, for the scenarios that involve supervision, we use half of the training set as annotated data and the other half as unannotated data. We emphasize that when testing model performance in each language *no multilingual information on test cases is used*. This allows us to gauge the performance of the model trained in a multilingual setting on future test cases where *only* monolingual data is available.

The training data itself consists of parallel phrases in Hebrew and Arabic. For example, such a training pair would include the Arabic phrase *fy alṣbaḥ* along with the Hebrew *bbqr (in the morning)*. Depending on the exact learning scenario (indirect supervision, augmentation, or shared supervision) some segmentation information may be available. In any case, no alignment between individual morphemes is ever provided.

We give results for all such scenarios in order to test our hypothesis that learning shared linguistic structure across languages can produce significant benefits in a wide range of cases. All results are averaged over five runs of Gibbs sampling with simulated annealing.

Evaluation Methods Following previous work, we evaluate the performance of our segmentation algorithm using F-score. This measure is the harmonic mean of recall and precision, which are obtained from the binary segmentation decisions made between each character in a word.

Baselines Although our main purpose is to compare the various cross-lingual propagation scenarios with our equivalent monolingual model, we also provide two baselines in Table 1. The first, RANDOM, makes segmentation decisions based on a coin weighted with the true segmentation frequency. As one might expect, this results in very poor performance. We also provide results from MORFESSOR(Creutz and Lagus 2007), a state-of-the-art unsupervised morphological segmenter. The results indicate that our own unsupervised monolingual model is quite competitive with MORFESSOR and is thus a reasonable basis for our cross-lingual model.

Results

Table 1 shows the performance of various propagation models on the Arabic-Hebrew dataset. In addition, the table shows the performance of our segmentation model when applied in supervised and unsupervised monolingual settings. In this scenario only the language-specific morpheme distributions E and F are used to model the data (both supervised and unsupervised).

Some of these results are to be expected. The monolingual supervised models achieve relative error reduction of 50% and 38% over their unsupervised monolingual counterparts (for Arabic and Hebrew respectively). Perhaps also not surprising is the performance of the models in the Indirect Supervision scenario. In this setting, the performance is found to lie somewhere between that of the unsupervised and supervised monolingual models. Crudely put, we observe that the benefit of foreign language annotation is somewhere between one third and half that of native annotation. For instance, while the monolingual unsupervised model achieves 63.22% F-measure on Arabic, once supervised Hebrew data is added to the mix, Arabic performance rises to 71.65%, nearly halfway to native Arabic supervised performance of 81.65%.

More surprising are the results in the other propagation scenarios. In the augmentation setting, we find that the addition of unsupervised foreign language data improves performance *even* when native annotated data is available. For example, while the monolingual supervised model achieves 75% F-measure on Hebrew, once unsupervised Arabic data is added, Hebrew performance rises to 77%. Moreover, in the shared supervision scenario, where both native and foreign annotations are available, we see even higher results. Performance rises to 84.39% for Arabic and 77.98% for He-

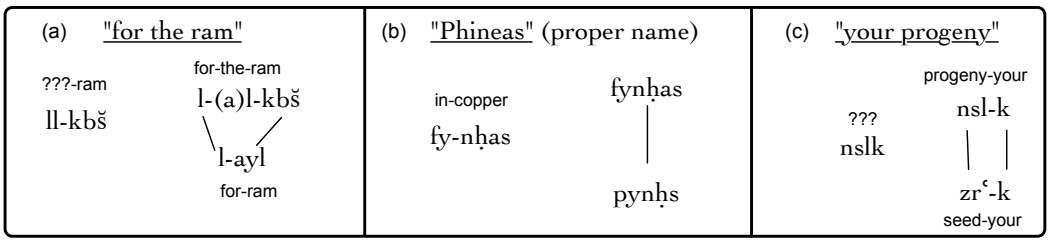


Figure 2: Examples of errors in the monolingual Arabic model, corrected when annotated Hebrew data is provided.

brew. These findings are consistent across both languages and across the three evaluation metrics.

Analysis

First we note that in all the cross-lingual scenarios our model predicts about 0.54 stray morphemes for each language per phrase and about 1.3 abstract morphemes (for a total of 2.6 aligned morphemes). This indicates that as designed, the model prefers to generate the data via cross-lingual patterns when possible.

We also illustrate the benefits of our model with three examples taken from the indirect supervision scenario and shown in Figure 2. Here Hebrew is the annotated source language, and Arabic is the unannotated target language. In example (a), the monolingual segmenter fails to separate the preposition *l-* ("for") from the definite article *-(a)l-* and treats the combination *ll-* as a single morpheme. However, when the parallel Hebrew segmentation is provided, the model recognizes the common abstract morpheme *(l-/l-)* and thus segments the Arabic correctly. In example (b), the monolingual Arabic segmenter made a reasonable guess for a previously unseen word. However, the Hebrew data reveals that this is actually a single morpheme – in fact, a proper name – and the cross-lingual arabic segmenter follows accordingly. Finally, in (c), we have an interesting case where in fact the gold standard (hand-crafted) segmenter mistook the word *nslk* for a single morpheme. However, the cross-lingual Arabic segmenter was able to learn from the parallel Hebrew data that in fact this word is to be analyzed as two morphemes: "your progeny." Unfortunately, in this case our model was penalized for having the correct answer.

Conclusions

In this paper we introduce a novel class of multilingual models. Rather than using parallel data as a means for annotation transfer, we symmetrically model the shared structure between languages using whatever resources are available, both annotated and unannotated. We test this idea on the task of morphological segmentation. Our results demonstrate that our model is able to learn from multiple language sources and to effectively propagate linguistic structure across language barriers.

In the future, we plan to apply our model to more than two languages, and to investigate the effect of genetic distance between the languages analyzed. It remains an open question whether more distant languages can benefit in the same way through joint analysis. Furthermore, we will investigate the application of the cross-lingual propagation method to higher level linguistic structure induction, such as part-of-speech tagging and parsing.[5]

References

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

Finkel, J. R.; Grenager, T.; and Manning, C. D. 2007. The infinite tree. In *Proceedings of the ACL*, 272–279.

Goldwater, S.; Griffiths, T. L.; and Johnson, M. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the ACL*.

Groves, A., and Lowery, K., eds. 2006. *The Westminster Hebrew Bible Morphology Database*. Philadelphia, PA, USA: Westminster Hebrew Institute.

Habash, N., and Rambow, O. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the ACL*, 573–580.

Haghighi, A., and Klein, D. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the ACL*, 848–855.

Hwa, R.; Resnik, P.; Weinberg, A.; Cabezas, C.; and Kolak, O. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering* 11(3):311–325.

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

Padó, S., and Lapata, M. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the HLT*, 859–866.

Rogati, M.; McCarley, J. S.; and Yang, Y. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of the ACL*, 391–398.

Yarowsky, D.; Ngai, G.; and Wicentowski, R. 2000. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT*, 161–168.

[5]Thanks to Branavan, Michael Collins, Yoong Keok Lee, and other members of the MIT NLP group for enlightening discussion. Special thanks to Khaled Al-Masri and John Huehnergard of Harvard for imparting knowledge of Arabic and Semitic Philology.