Climbing the tower of babel: Unsupervised multilingual learning

Climbing the Tower of Babel: Unsupervised Multilingual Learning

Benjamin Snyder BSNYDER@CSAIL.MIT.EDU
Regina Barzilay REGINA@CSAIL.MIT.EDU
MIT Computer Science & Artificial Intelligence Lab, 32 Vassar Street, Cambridge, MA 02139 USA

Abstract

For centuries, scholars have explored the deep links among human languages. In this paper, we present a class of probabilistic models that use these links as a form of naturally occurring supervision. These models allow us to substantially improve performance for core text processing tasks, such as morphological segmentation, part-of-speech tagging, and syntactic parsing. Besides these traditional NLP tasks, we also present a multilingual model for the computational decipherment of lost languages.

1. Overview

Electronic text is currently being produced at a vast and unprecedented scale across the languages of the world. Natural Language Processing (NLP) holds out the promise of automatically analyzing this growing body of text. However, over the last several decades, NLP research efforts have focused on the English language, often neglecting the thousands of other languages of the world (Bender, 2009).

Most of these languages are currently beyond the reach of NLP technology due to several factors. One of these is simply the lack of the kinds of hand-annotated linguistic resources that have helped propel the performance of English language systems. For complex tasks of linguistic analysis, hand-annotated corpora can be prohibitively time-consuming and expensive to produce. For example, the most widely used annotated corpus in the English language, the Penn Treebank (Marcus et al., 1994), took years for a team of professional linguists to produce. It is unrealistic to expect such resources to ever exist for the majority of the world's languages.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Another difficulty for multilingual NLP is that languages exhibit wide variation in their underlying linguistic structure. A model that has been developed for one language may not account for the kinds of structure found in others. In fact, there exists an entire academic discipline devoted to studying and describing systematic cross-lingual variations in language structure, known as linguistic typology (Comrie, 1989).

At first glance, it may seem that linguistic diversity would make developing intelligent text-processing tools for the world's languages a very daunting task. However, we argue that in fact it is possible to harness systematic linguistic diversity and use it to our advantage, utilizing a framework which we call *multilingual learning*. The goal of this enterprise is two-fold:

- To induce more accurate models of individual language structure without any human annotation.
- To induce accurate models of the relationships between languages.

The multilingual learning framework is based on the hypothesis that cross-lingual variations in linguistic structure correspond to *variations in ambiguity*. As an example, consider the syntactically ambiguous English sentence: "I ate pasta with cheese." The prepositional phrase "with cheese" can be interpreted as attaching the noun "pasta" (meaning the pasta had cheese), or could be interpreted as attaching to the verb "ate" (meaning perhaps that the pasta was eaten by means of a cheese-based utensil). As humans, we know that the first of these is the only plausible interpretation, but there is nothing in the sentence itself to indicate the correct parse. In contrast, the parallel sentence in Japanese uses an explicit *genitive marker* to mark the fact that the word for "pasta" is being modified.

This example is an instance of a more general phenomenon: what one language leaves implicit, and thus ambiguous for computers or humans, another will express directly through overt linguistic forms. In the framework of multilingual learning, we treat these vari-

ations in ambiguity as a form of *naturally occurring supervision*: by jointly modeling multiple languages, the idiosyncratic ambiguities of each can be wiped out by information explicit in the others.

The multilingual formulation presents great promise, but also poses novel technical challenges. One such challenge is the discovery of shared cross-lingual structure while allowing significant language-specific idiosyncrasies. To allow an effective balance, our models explain parallel sentences as a combination of multilingual and language specific latent variables in a hierarchical Bayesian framework. Even so, the *scope* of the shared explanatory mechanism is often unknown: some sets of languages exhibit a much larger degree of shared structure than other. For example, parallel phrases in related language pairs like Hebrew and Arabic tend to mirror each other in morphological structure much more than unrelated language pairs (such as English and Hebrew). To account for this variability in shared structure, we employ non-parametric statistical methods which allow for a flexible number of shared variables, as dictated by the languages and data at hand.

Finally, we set scalability in the number of languages as one of our design goals. Massively multilingual data-sets exist (e.g. the Bible, which has been translated into over 1,000 languages) and an ideal multilingual learning technique would scale gracefully in the number of languages. For the task of part-of-speech tagging, we developed a model and learning algorithm that scale *linearly* in the number of languages in terms of both time and space complexity.

We have applied unsupervised multilingual learning to the fundamental NLP tasks of morphological segmentation (Snyder & Barzilay, 2008a;b), part-of-speech tagging (Snyder et al., 2008; 2009b; Naseem et al., 2009), and parsing (Snyder et al., 2009a). We have focused on the use of *parallel corpora* (texts that have been written in one language and translated into other languages). We treat each parallel corpus as a *computational Rosetta Stone* which can help expose the latent structure of each language present. We assume the existence of such a corpus at training time with no human annotations. We do however, assume that reasonably accurate sentence- and word-level alignments have been induced using standard NLP tools (Och & Ney, 2003). At test time, we apply our models to monolingual data in each language. For all three tasks, multilingual learners consistently outperform their monolingual counterparts by a large margin. Remarkably, in the case of part-of-speech tagging, we found that model accuracy continues to increase as more languages are added to the mix. We believe these results point towards a future of ubiquitous and accurate text processing tools for hundreds of the world's languages which lack annotated resources.

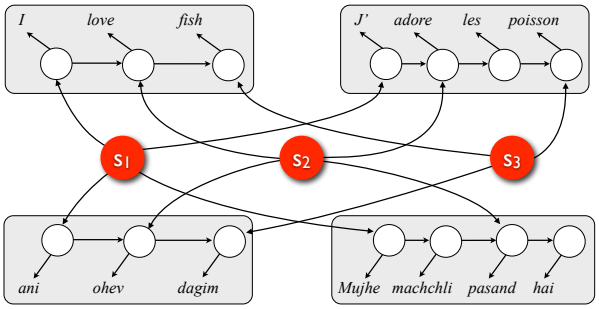


Figure 1. Part-of-speech graphical model structure for example sentence. In this instance, we have three *superlingual tags*: one for the cluster of words corresponding to English "I", one for the cluster of words corresponding to English "love", and one for the cluster of words corresponding to English "fish."

In the sections that follow we sketch the multilingual models that we have developed for three classical NLP tasks: Part-of-speech tagging (Section 2), morphological segmentation (Section 3), and parsing (Section 4). In section 5 we describe a model for the decipherment of lost languages.

2. Part-of-speech Tagging

The goal of part-of-speech tagging is to automatically determine the part-of-speech (noun, verb, adjective, etc) of each word in the context of a given sentence. For example, the word "can" in English may function as an auxiliary verb, a noun, or a regular verb. However, many other languages express these different senses with three distinct lexemes. Thus, at the lexical level, a word with part-of-speech ambiguity in one language may correspond to an unambiguous word in the other language. Languages also differ in their patterns of structural part-of-speech ambiguity. For example, the presence of a definite article (e.g. *the*) in English greatly reduces the ambiguity of the succeeding tag. In languages without definite articles, however, this explicit structural information is absent.

We first describe the structure of our model. We posit a separate Hidden Markov Model (HMM) (Merialdo, 1994) for each language, with an additional layer of latent cross-lingual variables. See Figure 1 for a graphical model depiction. A single cross-lingual variable,

Superlingual value "2"

	Noun	Verb	Determiner
English	0.9	0.1	0.0
French	0.8	0.1	0.1
Hindi	1.0	0.0	0.0

Superlingual value "5"

	Noun	Verb	Determiner
English	0.5	0.4	0.1
French	0.4	0.6	0.0
Hindi	0.5	0.5	0.0

Figure 2. Two stylized examples of superlingual tag values. Each provides a distribution over parts-of-speech for each language.

in our terminology a *superlingual tag*, is present for each cluster of aligned words across languages. These superlingual tags exert influence on the part-of-speech decisions of each word in the associated cluster.

In a standard HMM, we can write the joint probability of a sequence of words $\mathbf{w}$ and part-of-speech tags $\mathbf{y}$ as product of *transition* and *emission* probabilities:

$$P(\mathbf{w}, \mathbf{y}) = \prod_i P(y_i|y_{i-1})P(w_i|y_i)$$

Under our latent variable model, the probability of bilingual parallel sentences $(\mathbf{w}^1, \mathbf{w}^2)$, bilingual part-of-speech sequences $(\mathbf{y}^1, \mathbf{y}^2)$, and superlingual tags $\mathbf{s}$ is given by:

$$\prod_i P(s_i)$$
$$\prod_j P\left(y_j^1|y_{j-1}^1, s_{f(j,1)}\right) P(w_j^1|y_j^1)$$
$$\prod_k P\left(y_k^2|y_{k-1}^2, s_{f(k,2)}\right) P(w_k^2|y_k^2),$$

where $f(m, n)$ gives the index of the superlingual tag associated with word m in language n. Notice that the part-of-speech tagging decisions of each language are independent when conditioning on the superlingual tags $\mathbf{s}$. It is this conditional independence which gives our model some of its crucial properties. Superlingual variables promote cross-lingual regularities (more on this below), yet word order, part-of-speech selection, and even part-of-speech inventory are permitted to vary arbitrarily across languages. In addition, this architecture allows our model to scale linearly in the number of languages: when a language is added to the mix we simply add new directed edges from the existing set of superlingual tags for each sentence.

Intuitively, the value of a superlingual tag represents a particular *multilingual context* that influences each language's part-of-speech selection. Formally, each superlingual value provides a set of multinomial probability distributions — one for each language's part-of-speech inventory. See Figure 2 for two stylized examples. The first shows a superlingual value which predominantly favors nouns across languages, while displaying a slight secondary preference for other parts-of-speech. The second example shows a more complex pattern of part-of-speech preferences, with nouns and verbs almost equally preferred across languages.

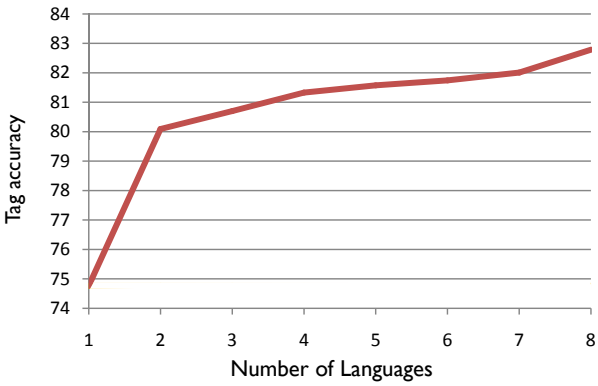


Figure 3. Average part-of-speech prediction accuracy as the number of languages varies (averaged over all subsets of languages for each size).

Give a superlingual tag s and a previous part-of-speech y_{i-1}, we define the generative probability of part-of-speech tag y_i as:

$$P(y_i|y_{i-1}, s) = \frac{P(y_i|y_{i-1}) \cdot P(y_i|s)}{Z},$$

where the first factor is the language-specific transition distribution, the second factor is the part-of-speech distribution provided by the superlingual tag s, and Z is a normalization constant obtained by summing over all possible part-of-speech tags. This parameterization allows a trade-off between language-specific and cross-lingual cues while avoiding the sparsity of a non-factored distribution.

In order to learn repeated cross-lingual patterns, the number of superlingual values must be constrained in some way. Intuitively, we would like to set the number of values to the number of multilingual part-of-speech patterns. However, the number of such patterns is not known *a priori* and may, in fact, depend on the number and properties of the languages under question. Rather than fixing the number of superlingual values to some arbitrary number, we leave it unbounded. To encourage sparse cross-lingual regularities we use a Dirichlet process prior (Ferguson, 1973). Under this non-parametric prior, the distribution over superlingual values must be highly skewed, such that a small finite subset receives a lion's share of the probability mass. The precise number of realized superlingual values will be dictated by the data. In practice we find

that the number of induced values ranges from 11 (for pair of languages) to 17 (for eight languages).

We evaluate our model on a parallel corpus of eight languages: Bulgarian, Czech, English, Estonian, Hungarian, Romanian, Serbian, and Slovene (Erjavec, 2004). We perform inference using Markov Chain Monte Carlo sampling and always test on held out monolingual data for each language. We ran our inference algorithm over all 255 subsets of the eight languages in our corpus, so we could examine the average change in performance as the number of languages increases. In the monolingual scenario, our model reduces to the Bayesian HMM of Goldwater & Griffiths (2007). When a complete part-of-speech dictionary[1] is available and our model is trained using eight languages, average tag prediction accuracy increases from 91.1% for monolingual models to 95%. In more realistic cases, where the tag dictionary is restricted to only frequently occurring words, we see even larger gaps between monolingual and multilingual performance. In one such scenario, where dictionary entries are only available for words occurring more than five times in the corpus, average multilingual performance increases to 82.8% from the monolingual baseline of 74.8%. As seen in Figure 3, accuracy gains steadily as languages are added to the mix.

3. Morphological Segmentation

In the task of morphological analysis, the goal is to segment words into *morphemes*, the smallest units of meaning (e.g. "misunderstanding" segments into three morphemes: "mis understand ing"). While the morphology of English is fairly simple, many languages exhibit a richer and more productive set of morphological patterns. In the unsupervised setting, morphological segmentation consists of finding recurrent prefix and suffix patterns which allow a more compact representation of the many possible derived word forms. Our multilingual model for this task automatically induces a segmentation and morpheme alignment from a multilingual (unannotated) corpus of short parallel phrases. For example, given parallel phrases meaning *in my land* in English, Arabic, Hebrew, and Aramaic, we wish to segment and align morphemes as shown in Figure 4.

This example illustrates the potential benefits of unsupervised multilingual morphological analysis. The three Semitic languages use cognates (words derived from a common ancestor) to represent the word *land*.

[1] i.e. entries indicating the set of potential parts-of-speech for each word

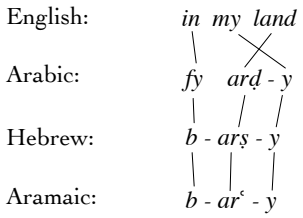


Figure 4. Morphological segmentation and alignment.

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 the space of joint segmentations and alignments. 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. More generally speaking, our model exploits the fact that each language distributes morphemes across words in a unique pattern. Note that morphemes expressed in one language often have no counterpart at all in some other languages, so morphemes must be allowed to remain unaligned.

The technical difficulty when compared to the part-of-speech model of Section 2 is that the *units of alignment* now depend on the results of the model's segmentation predictions. Whereas before we could treat word-level alignments as fixed and observed (as the result of preprocessing with standard NLP word-alignment tools), we must now fold alignment uncertainty into the morphology model itself.

We start with a sketch of the probabilistic process posited by our model for the generation of short bilingual phrases (see Figure 5 for an accompanying example). First, the numbers of unaligned language-specific morphemes (m and n), and the number of aligned morpheme pairs (k) are drawn from a Poisson distribution. These are the number of morphemes that will ultimately compose the bilingual parallel phrase. Next, the morphemes are drawn from the appropriate distributions: m and n morphemes are respectively drawn from language-specific morpheme distributions E and F, and k bilingual morpheme pairs are drawn from A. The resulting morphemes for each language are finally ordered and fused into words.

As in the previous section, the scope of cross-lingual connections (now in the form of aligned morpheme pairs) is not known *a priori*. Indeed, even the number of morphemes in each language is not known in

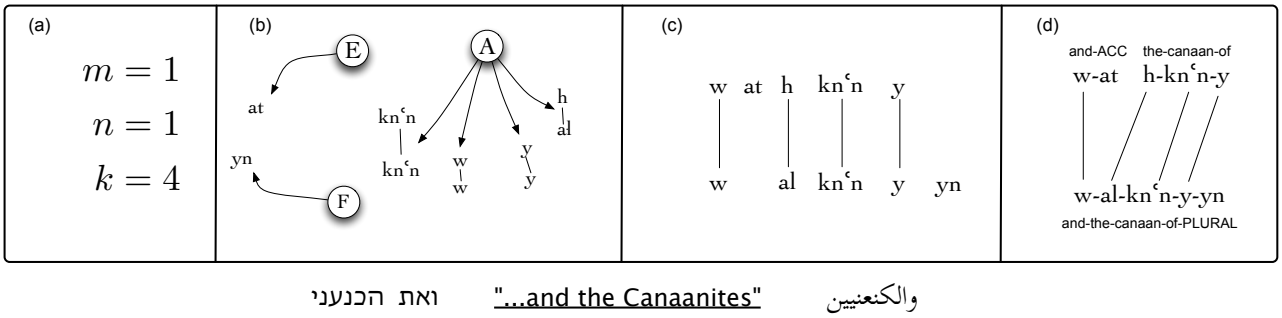


Figure 5. Morphological generation process for a parallel bilingual phrase, with Hebrew shown on top and Arabic on bottom. (a) First the numbers of isolated morphemes (m and n) and aligned morphemes pairs (k) are drawn from a Poisson distribution. (b) Isolated morphemes are then drawn from E and F (language-specific distributions) and aligned morpheme pairs are drawn from A. (c) The resulting morphemes are ordered. (d) Finally, some of the contiguous morphemes are fused into words.

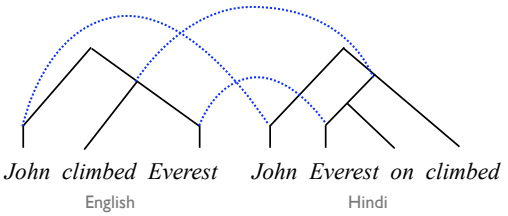
the unsupervised setting. We therefore employ non-parametric Dirichlet process priors on the three morpheme distributions E, F (language-specific), and A (bilingual). In this manner, the data itself can dictate the number of induced cross-lingual morpheme tuples as well as the number of isolated, language-specific morphemes which remain unaligned.

In addition, this model formulation allows us to consider various prior *base distributions* over aligned morpheme tuples (distribution A). In the case of unrelated languages, the base distributions can simply model morpheme length, encoding the fact that shorter morphemes tend to be more frequent than longer morphemes. However, if the languages are related and the phonetic relationship between the writing systems is known, we can employ a *probabilistic string-edit distance* over string tuples. In this way, a segmentation and alignment which align phonetically similar strings will be encouraged.

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. 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 error by as much as 24%. Furthermore, our experiments show that both related and unrelated language pairs benefit from multilingual learning. However, when the phonetic correspondences between related languages are explicitly modeled using the string-edit base distribution, related languages provide the most benefit.

4. Syntactic Parsing

Now we turn to the task of syntactic parsing. The goal of this task is to induce the underlying grammatical structure of each sentence in the form of a parse tree. In the monolingual setting, learning accurate parsing models without human-annotated texts has proven quite difficult (Charniak & Carroll, 1992; Klein & Manning, 2002). Here we consider the unsupervised multilingual scenario, where parsing models are induced simultaneously for pairs of languages using parallel texts. Again, our key premise is that ambiguous syntactic structures in languages may correspond to less uncertain structures in another language, due to systematic differences in word order and manner of expression. Thus, even in the absence of human-annotated trees, we hope to induce an accurate parsing model. Consider the following pair of parsed sentences in English and Hindi/Urdu:



If we know the correspondence between the words of the sentences (but nothing else about the languages in question), we can immediately pick up some important parsing cues. For example, we can rule as unlikely the possibility of parsing “John climbed” as a constituent subtree in English by the fact that the corresponding words in Hindi appear far apart. Likewise, we can avoid parsing “John Everest” in Hindi as a constituent,

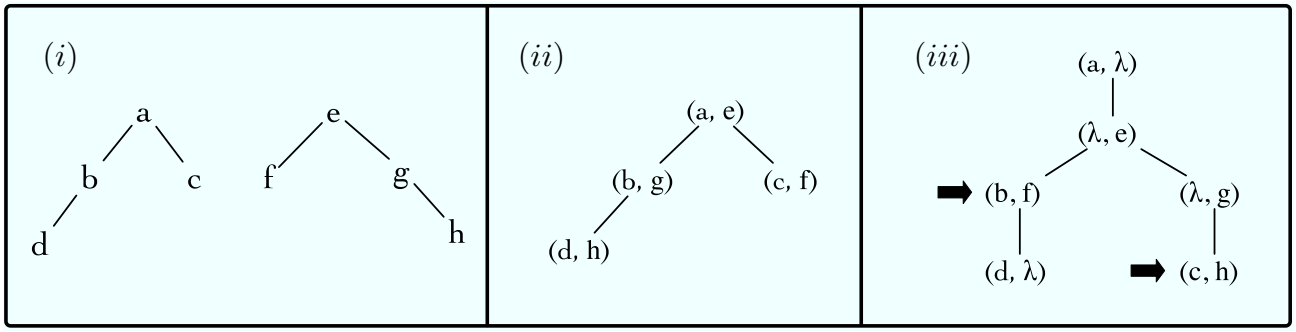


Figure 6. A pair of trees (i) and two possible alignment trees. In (ii), no empty spaces are inserted, but the order of one of the original tree's siblings has been reversed. In (iii), only two pairs of nodes have been aligned (indicated by arrows) and many empty spaces inserted.

through a similar comparison to the English sentence.

However, even in this simplest of sentence pairs, we notice syntactic divergence. While the English sentence uses the simple transitive verb "climbed" to express the fact that John completed his climb of Everest, the verb in the Hindi/Urdu sentence takes the post-positional argument "Everest on." The syntactic divergence in real-life examples becomes only more severe. The key challenge then is representational. We need to parse both sentences with possibly quite divergent trees, while recognizing shared syntactic structure. In effect, we seek to produce two loosely bound trees: node-to-node alignments need only be used where repeated bilingual patterns can be discerned in the data.

We achieve this loose binding of trees by adapting *unordered tree alignment* (Jiang et al., 1995) to a probabilistic setting. Under this formalism, any two trees can be aligned using an *alignment tree*. The alignment tree embeds the original two trees within it: each node is labeled by a pair (x, y), (λ, y), or (x, λ) where x is a node from the first tree, y is a node from the second tree, and λ is an empty space. The individual structure of each tree must be preserved under the embedding with the exception of sibling order (to allow variations in phrase and word order).

The flexibility of this formalism can be demonstrated by two extreme cases: (1) an alignment between two trees may actually align *none* of their individual nodes, instead inserting an empty space λ for each of the original two trees' nodes. (2) if the original trees are isomorphic to one another, the alignment may match their nodes exactly, without inserting any empty spaces. See Figure 6 for an example. An additional benefit of this formalism is computational: The marginalized probability over all possible alignments for any two trees can be efficiently computed with a dynamic program in bi-linear time in the size of the two trees.

We formulated a generative Bayesian model which seeks to explain sentence- and word-aligned parallel sentences through a combination of bilingual and monolingual syntactic parameters. Our model views each bilingual pair of sentences as having been probabilistically generated as follows: First an *alignment tree* is drawn uniformly from the set of all such trees. This alignment tree specifies the structure of each of the two individual trees, as well as the pairs of nodes which are aligned and those which are not aligned (i.e. paired with a λ). For each pair of aligned nodes, a corresponding pair of sentence constituents are jointly drawn from a bilingual distribution. For unaligned nodes (i.e. nodes paired with a λ in the alignment tree), a single sentence constituent is drawn, in this case from a language-specific distributions. Finally word-level alignments are drawn based on the structure of the alignment tree.

To perform inference under this model, we use a Metropolis-Hastings within-Gibbs sampler. We sample pairs of trees and then compute marginalized probabilities over all possible alignments using dynamic programming.

We tested the effectiveness of our bilingual grammar induction model on three corpora of parallel text: English-Korean, English-Urdu and English-Chinese. The model is trained using bilingual data with automatically induced word-level alignments, but is tested on purely monolingual data for each language. In all cases, our model outperforms a state-of-the-art baseline: the Constituent Context Model (CCM) (Klein & Manning, 2002), sometimes by substantial margins. On average, over all the testing scenarios that we studied, our model achieves an absolute increase in F-measure of 8.8 points, and a 19% reduction in error

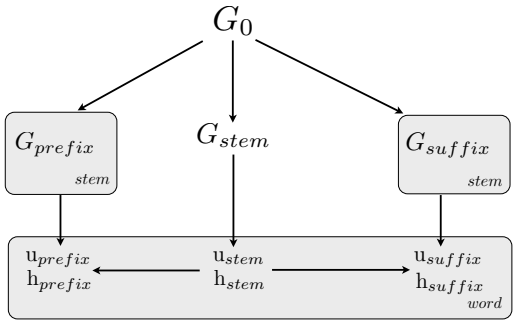


Figure 7. Plate diagram of the decipherment model. The base distribution G_0 defines probabilities over string-pairs based solely on character-level correspondences. The morpheme-pair distributions $G_{stem}, G_{prefix}, G_{suffix}$ directly assign probabilities to highly frequent morpheme pairs. Each stem pair provides a separate distribution over prefix and suffix pairs.

relative to a theoretical upper bound.

5. Lost Language Decipherment

The models discussed in the previous three sections all assumed the existence of multilingual parallel text. For traditional NLP tasks this is a reasonable assumption, as parallel texts are readily available for many of the world's languages. In contrast, our present work focuses on the decipherment of *lost languages*, where parallel texts are not available.

Several lost languages have been manually deciphered by humans in the last two centuries. In each case, the decipherment has been considered a major intellectual breakthrough, often the culmination of decades of scholarly efforts. So far, computers have played little role in this enterprise, even for recently deciphered languages. Skeptics argue that computers do not have the "logic and intuition" required to unravel the mysteries of ancient scripts.[2] We aim to demonstrate that at least some of this logic and intuition *can* be successfully captured by computational models.

Our definition of the computational decipherment task closely follows the setup typically faced by human decipherers (Robinson, 2002). Our input consists of texts in a lost language and a corpus of non-parallel data in a known related language. The decipherment itself involves two related sub-tasks: (i) finding the mapping between alphabets of the known and lost languages, and (ii) translating words in the lost language into corresponding cognates of the known language.

While there is no single formula that human decipherers have employed, manual efforts have focused on several guiding principles. A common starting point is to compare letter and word frequencies between the lost and known languages. In the presence of cognates the correct mapping between the languages will reveal similarities in frequency, both at the character and lexical level. In addition, morphological analysis plays a crucial role here, as highly frequent prefix and suffix correspondences can be particularly revealing. In fact, these three strands of analysis (character frequency, morphology, and lexical frequency) are intertwined throughout the human decipherment process. Partial knowledge of each drives discovery in the others.

We capture these intuitions in a generative Bayesian model. This model assumes that each word in the lost language is composed of morphemes which were generated with latent counterparts in the known language. We model bilingual morpheme pairs as arising through a series of Dirichlet processes. This allows us to assign probabilities based both on character-level correspondences (using a character-edit base distribution) as well as higher-level morpheme correspondences. In addition, our model carries out an implicit morphological analysis of the lost language, utilizing the known morphological structure of the related language. This model structure allows us to capture the interplay between the character- and morpheme-level correspondences that humans have used in the manual decipherment process. See figure 7 for a graphical overview of the model.

We have applied our decipherment model to a corpus of Ugaritic, an ancient Semitic language discovered in 1928 and manually deciphered four years later, using knowledge of Hebrew, a related language. As input to our model, we use the corpus of Ugaritic texts (consisting of 7,386 unique word forms) along with a Hebrew lexicon extracted from the Hebrew Bible. Our model yield an almost perfect decipherment of the Ugaritic alphabetic symbols. In addition, over half of the Ugaritic word forms with Hebrew cognates are correctly deciphered into their Hebrew counterparts.

6. Conclusions and Future Work

In Sections 2, 3, and 4, we described our application of multilingual learning to three traditional NLP tasks. In all cases, we assumed unannotated parallel text at training time and applied the resulting models to

[2] *"Successful archaeological decipherment has turned out to require a synthesis of logic and intuition . . . that computers do not (and presumably cannot) possess."* A. Robinson, "Lost Languages: The Enigma of the World's Undeciphered Scripts" (2002)

monolingual test data. We believe this to be a realistic scenario for a large number of the world's languages, as parallel texts are widely available. Finally, in Section 5, we considered the special case of lost language decipherment, where parallel text is not present, but information about a closely related language is available.

For future work, we pose the following two questions: (i) Can multilingual learning be used to triangulate the *information content* of sentences in multiple languages? (ii) Can knowledge of linguistic typology (and universal features of language) be used to induce more accurate unsupervised models, even without the use of parallel text?

References

Bender, Emily M. Linguistically naïve != language independent: why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pp. 26–32, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Charniak, Eugene and Carroll, Glen. Two experiments on learning probabilistic dependency grammars from corpora. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, pp. 1–13, 1992.

Comrie, Bernard. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell, 1989.

Erjavec, T. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC*, volume 4, pp. 1535–1538, 2004.

Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 1:209–230, 1973.

Goldwater, Sharon and Griffiths, Thomas L. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the ACL*, pp. 744–751, 2007.

Jiang, T., Wang, L., and Zhang, K. Alignment of trees – an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148, 1995.

Klein, Dan and Manning, Christopher D. A generative constituent-context model for improved grammar induction. In *Proceedings of the ACL*, pp. 128–135, 2002.

Marcus, M.P., Santorini, B., and Marcinkiewicz, M.A. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2): 313–330, 1994.

Merialdo, Bernard. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2): 155–171, 1994.

Naseem, Tahira, Snyder, Benjamin, Eisenstein, Jacob, and Barzilay, Regina. Multilingual part-of-speech tagging: two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36(1):341–385, 2009. ISSN 1076-9757.

Och, Franz Josef and Ney, Hermann. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Robinson, Andrew. *Lost Languages: The Enigma of the World's Undeciphered Scripts*. McGraw-Hill, 2002.

Snyder, Benjamin and Barzilay, Regina. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the ACL/HLT*, pp. 737–745, 2008a.

Snyder, Benjamin and Barzilay, Regina. Cross-lingual propagation for morphological analysis. In *Proceedings of the AAAI*, pp. 848–854, 2008b.

Snyder, Benjamin, Naseem, Tahira, Eisenstein, Jacob, and Barzilay, Regina. Unsupervised multilingual learning for POS tagging. In *Proceedings of EMNLP*, pp. 1041–1050, 2008.

Snyder, Benjamin, Naseem, Tahira, and Barzilay, Regina. Unsupervised multilingual grammar induction. In *Proceedings of the ACL*, pp. 73–81, 2009a.

Snyder, Benjamin, Naseem, Tahira, Eisenstein, Jacob, and Barzilay, Regina. Adding more languages improves unsupervised multilingual part-of-speech tagging: a bayesian non-parametric approach. In *Proceedings of the NAACL*, pp. 83–91, 2009b.