Unsupervised Bilingual POS Tagging with Markov Random Fields

Desai Chen **Chris Dyer** **Shay B. Cohen** **Noah A. Smith**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
desaic@andrew.cmu.edu, {cdyer,scohen,nasmith}@cs.cmu.edu

Abstract

In this paper, we give a treatment to the problem of bilingual part-of-speech induction with parallel data. We demonstrate that naïve optimization of log-likelihood with joint MRFs suffers from a severe problem of local maxima, and suggest an alternative – using contrastive estimation for estimation of the parameters. Our experiments show that estimating the parameters this way, using overlapping features with joint MRFs performs better than previous work on the *1984* dataset.

1 Introduction

This paper considers unsupervised learning of linguistic structure—specifically, parts of speech—in parallel text data. This setting, and more generally the multilingual learning scenario, has been found advantageous for a variety of unsupervised NLP tasks (Snyder et al., 2008; Cohen and Smith, 2010; Berg-Kirkpatrick et al., 2010; Das and Petrov, 2011).

We consider globally normalized Markov random fields (MRFs) as an alternative to directed models based on multinomial distributions or locally normalized log-linear distributions. This alternate parameterization allows us to introduce correlated features that, at least in principle, depend on any parts of the hidden structure. Such models, sometimes called "undirected," are widespread in *supervised* NLP; the most notable instances are conditional random fields (Lafferty et al., 2001), which have enabled rich feature engineering to incorporate knowledge and improve performance. We conjecture that the "features view" of NLP problems is also more appropriate in unsupervised settings than the contrived, acyclic causal stories required by directed models. Indeed, as we will discuss below, previous work on multilingual POS induction has had to resort to objectionable independence assumptions to avoid introducing cyclic dependencies in the causal network.

While undirected models are formally attractive, they are computationally demanding, particularly when they are used *generatively*, i.e., as joint distributions over input and output spaces. Inference and learning algorithms for these models are usually intractable on realistic datasets, so we must resort to approximations. Our emphasis here is primarily on the machinery required to support overlapping features, not on weakening independence assumptions, although we weaken them slightly. Specifically, our parameterization permits us to model the relationship between aligned words in any configuration, rather than just those that conform to an acyclic generative process, as previous work in this area has done (§2). We incorporate word prefix and suffix features (up to four characters) in an undirected version of a model designed by Snyder et al. (2008). Our experiments suggest that feature-based MRFs offer advantages over the previous approach.

2 Related Work

The task of unsupervised bilingual POS induction was originally suggested and explored by Snyder et al. (2008). Their work proposes a joint model over pairs of tag sequences and words that can be understood as a pair of hidden Markov models (HMMs)

in which aligned words share states (a fixed and observable word alignment is assumed). Figure 1 gives an example for a French-English sentence pair. Following Goldwater and Griffiths (2007), the transition, emission and coupling parameters are governed by Dirichlet priors, and a token-level collapsed Gibbs sampler is used for inference. The hyperparameters of the prior distributions are inferred from data in an empirical Bayesian fashion.

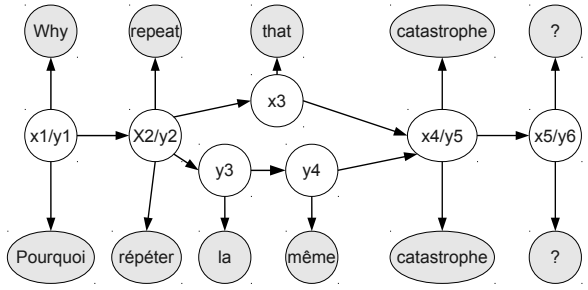


Figure 1: Bilingual Directed POS induction model

When word alignments are monotonic (i.e., there are no crossing links in the alignment graph), the model of Snyder et al. is straightforward to construct. However, crossing alignment links pose a problem: they induce cycles in the tag sequence graph, which corresponds to an ill-defined probability model. Their solution is to eliminate such alignment pairs (their algorithm for doing so is discussed below). Unfortunately, this is a potentially a serious loss of information. Crossing alignments often correspond to systematic word order differences between languages (e.g., SVO vs. SOV languages). As such, leaving them out prevents useful information about entire subsets of POS types from exploiting of bilingual context.

In the monolingual setting, Smith and Eisner (2005) showed similarly that a POS induction model can be improved with spelling features (prefixes and suffixes of words), and Haghighi and Klein (2006) describe an MRF-based monolingual POS induction model that uses features. An example of such a monolingual model is shown in Figure 2. Both papers developed different approximations of the computationally expensive partition function. Haghighi and Klein (2006) approximated by ignoring all sentences of length greater than some maximum, and the "contrastive estimation" of Smith and Eisner (2005) approximates the partition function with a set of automatically distorted training examples which are compactly represented in WFSTs.

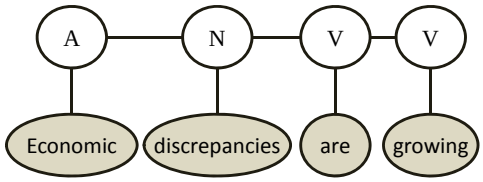


Figure 2: Monolingual MRF tag model (Haghighi and Klein, 2006)

Das and Petrov (2011) also consider the problem of unsupervised bilingual POS induction. They make use of independent conventional HMM monolingual tagging models that are parameterized with feature-rich log-linear models (Berg-Kirkpatrick et al., 2010). However, training is constrained with tag dictionaries inferred using bilingual contexts derived from aligned parallel data. In this way, the complex inference and modeling challenges associated with a bilingual tagging model are avoided.

Finally, multilingual POS induction has also been considered without using parallel data. Cohen et al. (2011) present a multilingual estimation technique for part-of-speech tagging (and grammar induction), where the lack of parallel data is compensated by the use of labeled data for some languages and unlabeled data for other languages.

3 Model

Our model is a Markov random field whose random variables correspond to words in two parallel sentences and POS tags for those words. Let $\mathbf{s} = \langle s_1, \ldots, s_{N_s} \rangle$ and $\mathbf{t} = \langle t_1, \ldots, t_{N_t} \rangle$ denote the two word sequences; these correspond to $N_s + N_t$ observed random variables.[1] Let $\mathbf{x}$ and $\mathbf{y}$ denote the sequences of POS tags for $\mathbf{s}$ and $\mathbf{t}$, respectively. These are the hidden variables whose values we seek to infer. We assume that a word alignment is provided for the sentences. Let $A \subseteq \{1, \ldots, N_s\} \times \{1, \ldots N_t\}$ denote the word correspondences specified by the alignment. The MRF's unnormalized probability S

[1] We use "source" and "target" but the two are completely symmetric in our undirected framework.

assigns:

$$S(\mathbf{s}, \mathbf{t}, \mathbf{x}, \mathbf{y} \mid A, \mathbf{w}) = \exp \mathbf{w}^\top \left(\sum_{i=1}^{N_s} \mathbf{f}_{\text{s-emit}}(s_i, x_i) + \sum_{i=2}^{N_s} \mathbf{f}_{\text{s-tran}}(x_{i-1}, x_i) + \sum_{i=1}^{N_t} \mathbf{f}_{\text{t-emit}}(t_i, y_i) + \sum_{i=2}^{N_t} \mathbf{f}_{\text{t-tran}}(y_{i-1}, y_i) + \sum_{(i,j) \in A} \mathbf{f}_{\text{align-POS}}(x_i, y_j) \right)$$

where $\mathbf{w}$ is a numerical vector of feature weights that parameterizes the model. Each $\mathbf{f}_\bullet$ corresponds to features on pairs of random variables; a source POS tag and word, two adjacent source POS tags, similarly for the target side, and aligned source/target POS pairs. For simplicity, we let $\mathbf{f}$ denote the sum of these five feature vectors. (In most settings, each feature/coordinate will be specific to one of the five addends.) In this paper, the features are indicators for each possible value of the pair of random variables, plus prefix and suffix features for words (up to four characters). These features encode information similar to the Bayesian bilingual HMM discussed in §2. Future work might explore extensions to this basic feature set.

The marginal probability of the words is given by:

$$p(\mathbf{s}, \mathbf{t} \mid A, \mathbf{w}) = \frac{\sum_{\mathbf{x},\mathbf{y}} S(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{t} \mid A, \mathbf{w})}{\sum_{\mathbf{s}',\mathbf{t}'} \sum_{\mathbf{x},\mathbf{y}} S(\mathbf{s}', \mathbf{t}', \mathbf{x}, \mathbf{y} \mid A, \mathbf{w})}.$$

Maximum likelihood estimation would choose weights $\mathbf{w}$ to optimize a product of quantities like the above, across the training data.

A key advantage of this representation is that any alignments may be present. In directed models, crossing links create forbidden cycles in the graphical model. For example, Figure 3 shows a crossing link between "Economic discrepancies" and "divergences economiques." Snyder et al. (2008) dealt with this problem by deleting word correspondences that created cycles. The authors deleted crossing links by considering each alignment link in the order of the source sentence, deleting it if it crossed previous links. Deleting crossing links removes some information about word correspondence.

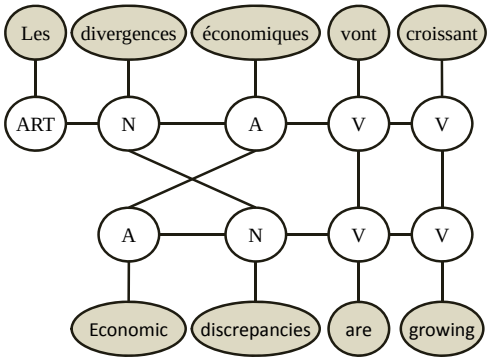


Figure 3: Bilingual tag model.

4 Inference and Parameter Learning

When using traditional generative models, such as hidden Markov models, the unsupervised setting lends itself well to maximizing joint log-likelihood, leading to a model that performs well (Snyder et al., 2008). However, as we show in the following analysis, maximizing joint log-likelihood for a joint Markov random field with arbitrary features suffers from serious issues which are related to the complexity of the optimized objective surface.

4.1 MLE with Gradient Descent

For notational simplicity, we assume a single pair of sentences $\mathbf{s}$ and $\mathbf{t}$; generalizing to multiple training instances is straightforward. The marginalized log-likelihood of the data given $\mathbf{w}$ is

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{s}, \mathbf{t} \mid \mathbf{w}) = \log \frac{\sum_{\mathbf{x},\mathbf{y}} S(\mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{t} \mid \mathbf{w})}{\sum_{\mathbf{s}',\mathbf{t}'} \sum_{\mathbf{x},\mathbf{y}} S(\mathbf{x}, \mathbf{y}, \mathbf{s}', \mathbf{t}' \mid \mathbf{w})}.$$

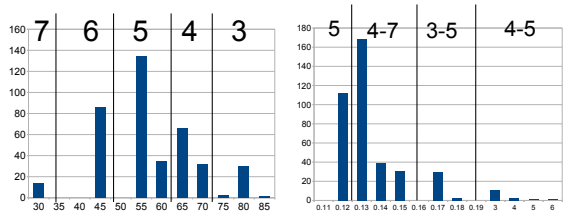
In general, maximizing marginalized log-likelihood is a non-concave optimization problem. Iterative hill-climbing methods (e.g., expectation-maximization and gradient-based optimization) will lead only to local maxima, and these may be quite shallow. Our analysis suggests that the problem is exacerbated when we move from directed to undirected models. We next describe a simple experiment that gives insight into the problem.

We created a small synthetic monolingual data set for sequence labeling. Our synthetic data consists of the following five sequences of observations: {(0 1 2 3) , (1 2 3 0) , (2 3 0 1) , (3 0 1 2) , (0 1 2 3)}. We then

maximized the marginalized log-likelihood for two models: a hidden Markov model and an MRF. Both use the same set features, only the MRF is globally normalized. The number of hidden states in both models is 4.

The global maximium in both cases would be achieved when the emission probabilities (or feature weights, in the case of MRF) map each observation symbol to a single state. When we tested whether this happens in practice, we noticed that it indeed happens for hidden Markov models. The MRF, however, tended to use fewer than four tags in the emission feature weights, i.e., for half of the tags, all emission feature weights were close to 0. This effect also appeared in our real data experiments.

The reason for this problem with the MRF, we believe, is that the parameter space of the MRF is underconstrained. HMMs locally normalize the emission probabilities, which implies that a tag cannot "disappear"—a total probability mass of 1 must always be allocated to the observation symbols. With MRFs, however, there is no such constraint. Further, effective deletion of a state y requires zeroing out transition probabilities from all other states to y, a large number of parameters that are completely decoupled within the model.



(a) likelihood (b) contrastive objective

Figure 4: Histograms of local optima found by optimizing the length neighborhood objective (a) and the contrastive objective (b) on a synthetic dataset with 8 sentences of length 7. The weights are initialized uniformly at random in the interval $[-1, 1]$. We plot frequency versus negated log-likelihood (lower horizontal values are better). An HMM always finds a solution that uses all available tags. The numbers at the top are numbers of tags used by each local optimum.

Our bilingual model is more complex than the above example, and we found in preliminary experiments that the effect persists there, as well. In the following section, we propose a remedy to this problem based on contrastive estimation (Smith and Eisner, 2005).

4.2 Contrastive Estimation

Contrastive estimation maximizes a modified version of the log-likelihood. In the modified version, it is the normalization constant of the log-likelihood that changes: it is limited to a sum over possible elements in a *neighborhood* of the observed instances. More specifically, in our bilingual tagging model, we would define a neighborhood function for sentences, $N(\mathbf{s}, \mathbf{t})$ which maps a pair of sentences to a set of pairs of sentences. Using this neighborhood function, we maximize the following objective function:

$$\begin{aligned}
&\mathcal{L}_{\text{ce}}(\mathbf{w}) \\
&= \log p(\mathbf{S} = \mathbf{s}, \mathbf{T} = \mathbf{t} \mid \mathbf{S} \in N_1(\mathbf{s}), \mathbf{T} \in N_2(\mathbf{t}), \mathbf{w}) \\
&= \log \frac{\sum_{\mathbf{x},\mathbf{y}} S(\mathbf{s}, \mathbf{t}, \mathbf{x}, \mathbf{y} \mid \mathbf{w})}{\sum_{\mathbf{s}',\mathbf{t}' \in N(\mathbf{s},\mathbf{t})} \sum_{\mathbf{x},\mathbf{y}} S(\mathbf{s}', \mathbf{t}', \mathbf{x}, \mathbf{y} \mid \mathbf{w})}.
\end{aligned} \tag{1}$$

We define the neighborhood function using a cross-product of monolingual neighborhoods: $N(\mathbf{s}, \mathbf{t}) = N_1(\mathbf{s}) \times N_1(\mathbf{t})$. N_1 is the "dynasearch" neighborhood function (Potts and van de Velde, 1995; Congram et al., 2002), used for contrastive estimation previously by Smith (2006). This neighborhood defines a subset of permutations of a sequence $\mathbf{s}$, based on local transpositions. Specifically, a permutation of $\mathbf{s}$ is in $N_1(\mathbf{s})$ if it can be derived from $\mathbf{s}$ through swaps of any adjacent pairs of words, with the constraint that each word only be moved once. This neighborhood can be compactly represented with a finite-state machine of size $O(N_s)$ but encodes a number of sequences equal to the N_sth Fibonacci number.

Monolingual Analysis To show that contrastive estimation indeed gives a remedy to the local maximum problem, we return to the monolingual synthetic data example from §4.1 and apply contrastive estimation on this problem. The neighborhood we use is the dynasearch neighborhood. In Figure 4b

we compare the maxima identified using MLE with the monolingual MRF model to the maxima identified by contrastive estimation. The results are conclusive: MLE tends to get stuck much more often in local maxima than contrastive estimation.

Following an analysis of the feature weights found by contrastive estimation, we found that contrastive estimation puts more weight on the transition features than emission features, i.e., the transition features weights have larger absolute values than emission feature weights. We believe that this could explain why contrastive estimation finds better local maximum that plain MLE, but we leave exploration of this effect for future work.

It is interesting to note that even though the contrastive objective tends to use more tags available in the dictionary than the likelihood objective does, the maximum objective that we were able to find does not correspond to the tagging that uses all available tags, unlike with HMM, where the maximum that achieved highest likelihood also uses all available tags.

4.3 Optimizing the Contrastive Objective

To optimize the objective in Eq. 1 we use a generic optimization technique based on the gradient. Using the chain rule for derivatives, we can derive the partial derivative of the log-likelihood with respect to a weight w_i:

$$\frac{\partial \mathcal{L}_{\text{ce}}(\mathbf{w})}{\partial w_i} = \mathbb{E}_{p(\mathbf{X},\mathbf{Y}|\mathbf{s},\mathbf{t},\mathbf{w})}[f_i] - \mathbb{E}_{p(\mathbf{S},\mathbf{T},\mathbf{X},\mathbf{Y}|\mathbf{S}\in N_1(\mathbf{s}),\mathbf{T}\in N_1(\mathbf{t}),\mathbf{w})}[f_i]$$

The second term corresponds to a computationally expensive inference problem, because of the loops in the graphical model. This situation is different from previous work on linear chain-structured MRFs (Smith and Eisner, 2005; Haghighi and Klein, 2006), where exact inference is possible. To overcome this problem, we use Gibbs sampling to obtain the two expectations needed by the gradient. This technique is closely related to methods like stochastic expectation-maximization (Andrieu et al., 2003) and to contrastive divergence (Hinton, 2000).

The training algorithm iterates between sampling part-of-speech tags and sampling permutations of words to compute the expected value of features. To sample permutations, the sampler iterates through the sentences and decides, for each sentence, whether to swap a pair of adjacent tags and words or not. The Markov blanket for computing the probability of swapping a pair of tags and words is shown in Figure 5. We run the algorithm for a fixed number (50) of iterations. By testing on a development set, we observed that the accuracy may increase after 50 iterations, but we chose this small number of iterations for speed.

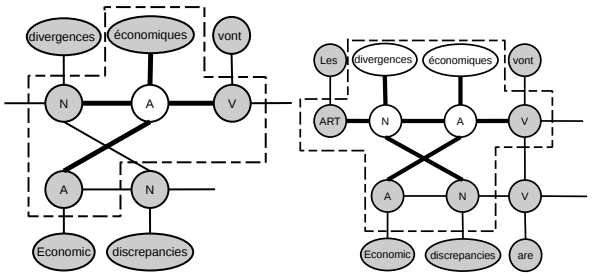


Figure 5: Markov blanket of a tag (left) and of a pair of adjacent tags and words (right).

In preliminary experiments we considered stochastic gradient descent, with online updating. We found this led to low-accuracy local optima, and opted for gradient descent with batch updates in our implementation. The step size was chosen to limit the maximum absolute value of the update in any weight to 0.1. Preliminary experiments showed only harmful effects from regularization, so we did not use it. These issues deserve further analysis and experimentation in future research.

5 Experiments

We next describe experiments using our undirected model to unsupervisedly learn POS tags.

With unsupervised part-of-speech tagging, it is common practice to use a full or partial dictionary that maps words to possible part-of-speech tags. The goal of the learner is then to discern which tag a word should take among the tags available for that word. Indeed, in all of our experiments we make use of a tag dictionary. We consider both a *complete* tag dictionary, where all of the POS tags for all words in the data are known,[2] and a smaller tag dictionary that only provides possible tags for the 100

[2]Of course, additional POS tags may be possible for a given word that were not in evidence in our finite dataset.

most frequent words in each language, leaving the other words completely ambiguous. The former dictionary makes the problem easier by reducing ambiguity; it also speeds up inference.

Our experiments focus on the Orwell novel *1984* dataset for our experiments, the same data used by Snyder et al. (2008). It consists of parallel text of the *1984* novel in English, Bulgarian, Slovene and Serbian (Erjavec, 2004), totalling 5,969 sentences in each language. The *1984* datset uses fourteen part-of-speech tags, two of which denote punctuation. The tag sets for English and other languages have minor differences in determiners and particles.

We use the last 25% of sentences in the dataset as a test set, following previous work. The dataset is manually annotated with part-of-speech tags. We use automatically induced word alignments using Giza++ (Och and Ney, 2003). The data show very regular patterns of tags that are aligned together: words with the same tag in two languages tend to be aligned with each other.

When a complete tag dictionary derived from the Slavic language data is available, the level of ambiguity is very low. The baseline of choosing random tags for each word gives an accuracy in the low 80s. For English, we use an extended tag dictionary built from the Wall Street Journal and the *1984* data. The English tag dictionary is much more ambiguous because it is obtained from a much larger dataset. The random baseline gives an accuracy of around 56%. (See Table 1.)

In our first set of experiments (§5.1), we perform a "sanity check" with a monolingual version of the MRF that we described in earlier sections. We compare it against plain HMM to assure that the MRFs behave well in the unsupervised setting.

In our second set of experiments (§5.2), we compare the bilingual HMM model from Snyder et al. (2008) to the joint MRF model. We show that using an MRF has an advantage over an HMM model in the partial tag dictionary setting.

5.1 Monolingual Experiments

We turn now to two monolingual experiments that verify our model's suitability for the tagging problem.

Language	Random	HMM	MRF
Bulgarian	82.7	88.9	93.5
English	56.2	90.7	87.0
Serbian	83.4	85.1	89.3
Slovene	84.7	87.4	94.5

Table 1: Unsupervised monolingual tagging accuracies with complete tag dictionary on *1984* data.

Supervised Learning As a very primitive comparison, we trained a monolingual supervised MRF model to compare to the results of supervised HMMs. The training procedure is based on sampling, just like the unsupervised estimation method described in §4.3. The only difference is that there is no need to sample the words because the tags are the only random variables to be marginalized over. Our model and HMM give very close performance with difference in accuracy less than 0.1%. This shows that the MRF is capable of representing an equivalent model represented by the HMM. It also shows that gradient descent with MCMC approximate inference is capable of finding a good model with the weights initialized to all 0s.

Unsupervised Learning We trained our model under the monolingual setting as a sanity check for our approximate training algorithm. Our model under monolingual mode is exactly the same as the models introduced in §2. We ran our model on the *1984* data with the complete tag dictionary. A comparison between our result and monolingual directed model is shown in Table 1. "Random" is obtained by choosing a random tag for each word according to the tag dictionary. "HMM" is a Bayesian HMM implemented by (Snyder et al., 2008). We also implemented a basic (non-Bayesian) HMM. We trained the HMM with EM and obtained results similar to the Bayesian HMM (not shown).

5.2 Bilingual Results

Table 2 gives the full results in the bilingual setting for the *1984* dataset with a partial tag dictionary. In general, MRFs do better than their directed counterparts, the HMMs. Interestingly enough, removing crossing links from the data has only a slight adverse effect. It appears like the prefix and suffix features are more important than having crossing links. Re-

Language pair	HMM	MRF	MRF w/o cross.	MRF w/o spell.
English	71.3	73.3$_{\pm 0.6}$	73.4$_{\pm 0.6}$	67.4$_{\pm 0.9}$
Bulgarian	62.6	62.3$_{\pm 0.3}$	63.8$_{\pm 0.4}$	55.2$_{\pm 0.5}$
Serbian	54.1	55.7$_{\pm 0.2}$	54.6$_{\pm 0.3}$	47.7$_{\pm 0.5}$
Slovene	59.7	61.4$_{\pm 0.3}$	60.4$_{\pm 0.3}$	56.7$_{\pm 0.4}$
English	66.5	73.3$_{\pm 0.3}$	73.4$_{\pm 0.2}$	62.3$_{\pm 0.5}$
Slovene	53.8	59.7$_{\pm 2.5}$	57.6$_{\pm 2.0}$	52.1$_{\pm 1.3}$
Bulgarian	54.2	58.1$_{\pm 0.1}$	56.3$_{\pm 1.3}$	58.0$_{\pm 0.2}$
Serbian	56.9	58.6$_{\pm 0.3}$	59.0$_{\pm 1.2}$	55.1$_{\pm 0.3}$
English	68.2	72.8$_{\pm 0.6}$	72.7$_{\pm 0.6}$	65.7$_{\pm 0.4}$
Serbian	54.7	58.5$_{\pm 0.6}$	57.7$_{\pm 0.3}$	54.2$_{\pm 0.3}$
Bulgarian	55.9	59.8$_{\pm 0.1}$	60.3$_{\pm 0.5}$	55.0$_{\pm 0.4}$
Slovene	58.5	61.4$_{\pm 0.3}$	61.6$_{\pm 0.4}$	58.1$_{\pm 0.6}$
Average	59.7	62.9	62.5	56.5

Table 2: Unsupervised bilingual tagging accuracies with tag dictionary only for the top 100 frequent words. "HMM" is the result reported by (Snyder et al., 2008). "MRF" is our contrastive model averaged over ten runs. "MRF w/o cross." is our model trained without crossing links, like Snyder et al.'s HMM. "MRF w/o spell." is our model without prefix and suffix features. Numbers appearing next to results are standard deviations over the ten runs.

Language	w/ cross.	w/o cross.
French	73.8	70.3
English	56.0	59.2

Table 3: Effect of removing crossing links when learning French and English in a bilingual setting.

moving the prefix and suffix features gives substantially lower results on average, results even below plain HMMs.

The reason that crossing links do not change the results much could be related to fact that most of the sentence pairs in the *1984* dataset do not contain many crossing links (only 5% of links cross another link). To see whether crossing links do have an effect when they come in larger number, we tested our model on French-English data. We aligned 10,000 sentences from the Europarl corpus (Koehn, 2005), resulting in 87K crossing links out of a total of 673K links. Using the Penn treebank (Marcus et al., 1993) and the French treebank (Abeillé et al., 2003) to evaluate the model, results are given in Table 3. It is evident that crossing links have a larger effect here, but it is mixed: crossing links improve performance for French while harming it for English.

6 Conclusion

In this paper, we explored the capabilities of joint MRFs for modeling bilingual part-of-speech models. Exact inference with dynamic programming is not applicable, forcing us to experiment with approximate inference techniques. We demonstrated that using contrastive estimation together with Gibbs sampling for the calculation of the gradient of the objective function leads to better results in unsupervised bilingual POS induction.

Our experiments also show that the advantage of using MRFs does not necessarily come from the fact that we can use non-monotonic alignments in our model, but instead from the ability to use overlapping features such as prefix and suffix features for the vocabulary in the data.

Acknowledgments

We thank the reviewers and members of the ARK group for helpful comments on this work. This research was supported in part by the NSF through grant IIS-0915187 and the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-10-1-0533.

References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a treebank for French. In A. Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. 2003. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43.

T. Berg-Kirkpatrick, A. Bouchard-Cote, J. DeNero, and D. Klein. 2010. Unsupervised learning with features. In *Proceedings of NAACL*.

S. B. Cohen and N. A. Smith. 2010. Covariance in unsupervised learning of probabilistic grammars. *Journal of Machine Learning Research*, 11:3017–3051.

S. B. Cohen, D. Das, and N. A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of EMNLP*.

R. K. Congram, C. N. Potts, and S. L. van de Velde. 2002. An iterated Dynasearch algorithm for the single-machine total weighted tardiness scheduling problem. *Informs Journal On Computing*, 14(1):52–67.

D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Procedings of ACL*.

T. Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of LREC*.

S. Goldwater and T. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceeding of ACL*.

A. Haghighi and D. Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of HLT-NAACL*.

G. E. Hinton. 2000. Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, University College London.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19:313–330.

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

C. N. Potts and S. L. van de Velde. 1995. Dynasearch–iterative local improvement by dynamic programming. Part I: The traveling salesman problem. *Technical report*.

N. A. Smith and J. Eisner. 2005. Contrastive estimation: training log-linear models on unlabeled data. In *Proceedings of ACL*.

N. A. Smith. 2006. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. Ph.D. thesis, Johns Hopkins University.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of EMNLP*.