Cross-Lingual Discriminative Learning of Sequence Models with Posterior Regularization

Kuzman Ganchev
Google Research
76 9th Avenue
New York, NY 10011
kuzman@google.com

Dipanjan Das
Google Research
76 9th Avenue
New York, NY 10011
dipanjand@google.com

Abstract

We present a framework for cross-lingual transfer of sequence information from a resource-rich source language to a resource-impoverished target language that incorporates soft constraints via posterior regularization. To this end, we use automatically word aligned bitext between the source and target language pair, and learn a discriminative conditional random field model on the target side. Our posterior regularization constraints are derived from simple intuitions about the task at hand and from cross-lingual alignment information. We show improvements over strong baselines for two tasks: part-of-speech tagging and named-entity segmentation.

1 Introduction

Supervised systems for NLP tasks are available for a handful of languages. These systems achieve high accuracy for many applications; a variety of robust algorithms to train them from labeled data have been developed. Here, we focus on learning sequence models for the languages that lack annotated resources. For a given resource-poor target language of interest, we assume that parallel data with a resource-rich source language exists. With the help of this bitext and a supervised system in the source language, we infer constraints over the label distribution in the target language, and train a discriminative model using posterior regularization (Ganchev et al., 2010).

Cross-lingual learning of structured prediction models via parallel data has been applied for several natural language processing problems, including part-of-speech (POS) tagging (Yarowsky and Ngai, 2001), syntactic parsing (Hwa et al., 2005) and named-entity recognition (Kim et al., 2012). These methods are useful in several ways. First, they help in fast prototyping of natural language systems for new languages that do not boast human annotations. Second, the output of such systems could be used to bootstrap more extensive human annotation projects (Vlachos, 2006). Finally, they are significantly more accurate than purely unsupervised systems (McDonald et al., 2011; Das and Petrov, 2011).

Recently, Täckström et al. (2013) presented a technique for coupling token constraints derived from projected cross-lingual information and type constraints derived from noisy tag dictionaries to learn POS taggers. Although this technique resulted in state-of-the-art weakly supervised taggers, the authors used a heuristic to combine the aforementioned two sources of constraints: the dictionary constraints pruned the tagger's search space, and the intersected token-level projections were treated as hard observations. On the other hand, Ganchev et al. (2009) presented a framework for learning weakly-supervised systems (in their case, dependency parsers) that incorporated alignment-based information too, but used the cross-lingual information only as soft constraints, via posterior regularization. The advantage of this framework lay in the fact that the projections were only trusted to a certain degree, determined by a strength hyperparameter, which unfortunately the authors did not have an elegant way to tune. In this paper, we exploit the better aspects of these two lines of work: first, we extend the framework of Täckström et al. by treating the alignment-based projections only as soft constraints (see §3.4); second, we choose the constraint strength by utilizing the tag ambiguity of tokens for a given resource-poor language (see §6.1).

Other than validating our framework on part-of-speech tagging, we experiment on named-entity segmentation in a cross-lingual framework. For this

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1996–2006,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

task, we present a novel method to perform high-precision phrase-level entity transfer (§5.2.2); we also provide ways to balance precision and recall with posterior regularization (§6.2) by incorporating intuitive soft constraints during learning. We measure performance on standard benchmark datasets for both of these tasks, and report improvements over state-of-the-art baselines.

2 Prior Work

Cross-lingual projection methods can be classified by their use of two very broad ideas. The first idea utilizes parallel data to create full or partial annotations in the low-resource language and trains from this data. This was popularized by Yarowsky and Ngai (2001) who applied this to POS tagging and shallow parsing. It was later applied to parsing (Hwa et al., 2005) and named entity recognition (Kim et al., 2012). The second idea, first proposed by Zeman and Resnik (2008) and applied more broadly by McDonald et al. (2011), is to train a model on a resource-rich language and apply it to a resource-poor language directly. The disparity between the languages is mitigated by the choice of features. In addition to cross-lingual projection, purely unsupervised methods have been explored but with limited success (Christodoulopoulos et al., 2010). Here, we resort to cross-lingual projection and incorporate the first idea; we also follow Li et al. (2012) and use Wiktionary to further constrain the POS tagging task.

Our learning setup is similar to that of Ganchev et al. (2009), who also use posterior regularization but focus on dependency parsing alone. Our work differs with respect to the tasks, the learning algorithm and also in that we use corpus-wide constraints, while Ganchev et al. use one constraint per sentence. For the part-of-speech tagging task, our approach is similar to that of Täckström et al. (2013), who use an almost identical learning setup but only make use of hard constraints. By relaxing these constraints, we allow the model to identify and ignore inconsistently labeled parts of sentences, and achieve better results using identical training and test data.

3 Approach

We give an overview of our approach, and present the details of our model used for cross-lingual learning.

Algorithm 1 Cross-Lingual Learning with Posterior Regularization

Require: Parallel source and target language data $\mathcal{D}^e$ and $\mathcal{D}^f$, source language model $(M)^e$, task-specific target language constraints $\mathcal{C}$.
Ensure: Θ^f, a set of target language parameters.
1: $\mathcal{D}^{e \leftrightarrow f} \leftarrow$ word-align-bitext$(\mathcal{D}^e, \mathcal{D}^f)$
2: $\widehat{\mathcal{D}^e} \leftarrow$ label-supervised$(\mathcal{D}^e)$
3: $\widehat{\mathcal{D}^f} \leftarrow$ project-and-filter-labels$(\mathcal{D}^{e \leftrightarrow f}, \widehat{\mathcal{D}^e})$
4: $\Theta^f \leftarrow$ learn-posterior-constrained$(\widehat{\mathcal{D}^f}, \mathcal{C})$
5: Return Θ^f

3.1 General Overview

The general overview of our framework is provided in Algorithm 1. The process of learning parameters for a target language for a given task involves four subtasks. First, we run word alignment over a large corpus of parallel data between the resource-rich source language and the resource-impoverished target language (see §4.3). In the second step, we use a supervised model to label the source side of the parallel data (see §5.1.1 and §5.2.1). The third step involves a task-specific word-alignment filtering step; this step involves heuristics for which we use cues from prior state-of-the-art (Das and Petrov, 2011; Täckström et al., 2013, see §5.1.2) and also introduce some novel ones for the NE segmentation problem (see §5.2.2). In the fourth step, we train a linear chain conditional random field (Lafferty et al., 2001, CRF henceforth) using posterior regularization. In the next subsection, we turn to a brief summary of this final step of estimating parameters of a discriminative model with posterior regularization.

3.2 Learning with Posterior Regularization

In this work, we utilize discriminative CRF models, and use posterior regularization (PR) to optimize their parameters. As a framework, posterior regularization is described in detail by Ganchev et al. (2010). However in our work, we adopt a different optimization technique; in what follows, we summarize the optimization algorithm in the context of CRF models.

Let $\mathbf{x}$ be an input sentence with a set of possible labelings $\mathcal{Y}(\mathbf{x})$ and let $\mathbf{y} \in \mathcal{Y}(\mathbf{x})$ be a particular labeling for sentence $\mathbf{x}$. We use bold capital letters $\mathbf{X} = \{\mathbf{x}_1 \dots \mathbf{x}_n\}$ and $\mathbf{Y} = \{\mathbf{y}_1 \dots \mathbf{y}_n\}$ to denote

a corpus of sentences and labelings for the corpus respectively. A CRF models the probability distribution over possible labels for a sentence $p_\theta(\mathbf{y}|\mathbf{x})$ as:

$$p_\theta(\mathbf{y} \mid \mathbf{x}) \propto \exp(\theta \cdot \mathbf{f}(\mathbf{x}, \mathbf{y})) \quad (1)$$

where θ are the model parameters and $\mathbf{f}(.)$ is a feature function. The model examines sentences in isolation, and the probability of a particular labeling for a corpus is defined as a product over the individual sentences:

$$p_\theta(\mathbf{Y} \mid \mathbf{X}) = \prod_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})} p_\theta(\mathbf{y} \mid \mathbf{x}). \quad (2)$$

Traditionally, CRF models have been trained to optimize the regularized log-likelihood of the training data

$$\max_\theta \ \mathcal{L}(\theta) = \max_\theta \ \log(p_\theta(\mathbf{Y} \mid \mathbf{X})) - \gamma \, ||\theta|| \quad (3)$$

In our setting, we do not have a fully labeled corpus, but we have constraints on the distribution of labels. For example, we may know that a particular token could be labeled only by a label inventory licensed by a dictionary, or that a labeling projected from a source language is usually (but not always) correct. We define these constraints in terms of feature expectations. Let $q(\mathbf{Y})$ be a distribution over all possible labelings of our corpus $\mathcal{Y}(\mathbf{X})$. Let $\mathcal{Q}$ be a set of distributions defined by:

$$\mathcal{Q} = \{q(\mathbf{Y}) : \mathbf{E}_q[\phi(\mathbf{X}, \mathbf{Y})] \leq \mathbf{b}\}, \quad (4)$$

where ϕ is a *constraint* feature function and $\mathbf{b}$ is a vector of non-negative values that serve as upper bounds to the expectations of every constraint feature. The vector $\mathbf{b}$ is used to encode our prior knowledge about desirable distributions $q(\mathbf{Y})$. Note that the constraint features ϕ are not related to the model features $\mathbf{f}$. The model features, together with the model parameters θ define the CRF model; the model features need to be computed at inference time for prediction. By contrast, the constraint features and their corresponding constraint values are used to define our training objective function (and are only used during learning). The PR objective with no labeled data is defined with respect to $\mathcal{Q}$ as:

$$\textbf{PR: } \max_\theta \ \mathcal{J}_\mathcal{Q}(\theta) = \max_\theta \ -\mathbf{KL}(\mathcal{Q}||p_\theta(\mathbf{Y} \mid \mathbf{X})) - \gamma \, ||\theta|| \quad (5)$$

where $\mathbf{KL}(\mathcal{Q}||p) = \min_{q\in\mathcal{Q}} \mathbf{KL}(q||p)$ is the KL-divergence (Kullback and Leibler, 1951) from a set to a point. Note that as we add more constraints, $\mathcal{Q}$ becomes a smaller set. In the limit, $\mathcal{Q} = \{q(\mathbf{Y}) : q(\hat{\mathbf{Y}}) = 1\}$ contains just one distribution concentrated on a single labeling $\hat{\mathbf{Y}}$. In this limit, posterior regularization degenerates into the convex log-likelihood objective normally used for supervised data $\mathcal{J}_\mathcal{Q}(\theta) = \mathcal{L}(\theta)$. However, in the general case, the PR objective $\mathcal{J}_\mathcal{Q}$ is not necessarily convex. Prior work, including that of Ganchev et al. propose an algorithm similar to Expectation-Maximization (Dempster et al., 1977, EM henceforth) to optimize $\mathcal{J}_\mathcal{Q}$, but we follow Liang et al. (2009) in using a schochastic update-based algorithm described below.

Note: To make it easier to reason about constraint values $\mathbf{b}$, we scale constraint features $\phi(\mathbf{X}, \mathbf{Y})$ to lie in $[0, 1]$ by computing $\max_\mathbf{Y} \phi(\mathbf{X}, \mathbf{Y})$ for the corpus to which ϕ is applied.

3.3 Optimization

The optimization procedure proposed by Ganchev et al. is similar to the EM algorithm, and computes the minimization $\min_{q\in\mathcal{Q}} \mathbf{KL}(q||p)$ at each step, using its dual form; this minimization is convex, so there is no duality gap. They show that the optimal primal variables $q^*(\mathbf{Y})$ are related to the optimal dual variables λ^* by:

$$q^*(\mathbf{Y}) = \frac{p_\theta(\mathbf{Y}|\mathbf{X})e^{-\lambda^*\cdot\phi(\mathbf{X},\mathbf{Y})}}{Z(\lambda^*)}. \quad (6)$$

where $Z(\lambda^*)$ is the normalizer. The dual problem is given by:

$$\max_{\lambda\geq 0} -\mathbf{b}\cdot\lambda - \log Z(\lambda). \quad (7)$$

Substituting Eq. 7 into the objective in Eq. 5, we get the saddle-point problem:

$$\max_\theta \ \min_{\lambda\geq 0} \ \mathbf{b}\cdot\lambda + \log \sum_\mathbf{Y} p_\theta(\mathbf{Y}|\mathbf{X})e^{-\lambda^*\cdot\phi(\mathbf{X},\mathbf{Y})} - \gamma \, ||\theta|| \, . \quad (8)$$

To optimize the above objective function, we need to compute partial derivatives with respect to both θ and λ. First, to compute the partial derivatives of Eq. 8

with respect to θ, we need to find expectations of the model features $\mathbf{f}$ given the current distribution p_θ and the constraint distribution q. To perform tractable inference, a linear-chain CRF model assumes that the feature function factorizes according to smaller parts; in particular the factorization uses the following structure:

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{f}(\mathbf{x}, y_i, y_{i-1}) \quad (9)$$

where i ranges over the tokens in the sentence. This factorization allows us to efficiently compute expectations over the labels y_i and label-pairs (y_i, y_{i+1}). To compute the partial gradient of Eq. 8 with respect to λ, we need to find the expectations of the constraint features ϕ. In order to be tractable here too, we ensure that ϕ also factorize according to the same structure as $\mathbf{f}$. Therefore, the gradient computation w.r.t. λ turns out to be straightforward.

For all the experiments in this paper, we optimize Eq. 8 using stochastic projected gradient. For each training sentence, we compute the gradient of θ and λ with respect to Eq. 8, take a gradient step in each one, and truncate the negative entries in λ to zero. We use a step size of 1 for all experiments.[1]

3.4 Relationship with Täckström et al. (2013)

In this subsection, we focus briefly on the relationship between this work and the work of Täckström et al. (2013), who focused on constrained learning of POS taggers. Täckström et al. define constrained lattices and train by optimizing marginal conditional log-likelihood. In our notation, they define their objective as:

$$\max_\theta \log \sum_{\mathbf{Y} \in \widehat{\mathcal{Y}}(\mathbf{X})} p_\theta(\mathbf{Y}|\mathbf{X}) - \gamma\|\theta\| \quad (10)$$

where $\widehat{\mathcal{Y}}(\mathbf{X})$ are the constrained lattices of label sequences that agree with both a dictionary and cross-lingually projected POS tags for each sentence of the training corpus. Let us define a constraint feature $\phi(\mathbf{X}, \mathbf{Y})$ which counts the number of tags in $\mathbf{Y}$ which are outside the constraint set $\widehat{\mathcal{Y}}(\mathbf{X})$ and require $\phi(\mathbf{X}, \mathbf{Y}) \leq 0$. Note that,

$$\arg\min_q \mathbf{KL}(q||p_\theta(\mathbf{Y}|\mathbf{X})) \text{ s.t. } \phi(\mathbf{X}, \mathbf{Y}) \leq 0$$

gives the same distribution as Eq. 10. Given this equivalence, it is easy to see that the gradient of Eq. 5 with respect to θ is the same as that of Eq. 10. By using such constrained lattices, Täckström et al. avoid maintaining a parameter for the constraint, but lose the ability to relax the constraint value and allow some probability mass outside the pruned lattice. Their paper also differs from ours in that they use L-BFGS (Liu and Nocedal, 1989), while we use an online optimization procedure. Since the objectives are non-convex, the two optimization techniques could lead to different local optima even when the constraint is not relaxed ($\mathbf{b} = 0$).

4 Tasks and Data

In this section, we focus on the nature of the two tasks that we attempt to solve, describe the source language datasets we use to train our supervised models for transfer, the target language datasets on which we evaluate our models and the parallel data we use for cross-lingual transfer.

4.1 Part-of-Speech Tagging

First, we focus on the task of part-of-speech tagging. Following previous work on cross-lingual POS tagging (Das and Petrov, 2011; Täckström et al., 2013), we adopt the POS tags of Petrov et al. (2012), version 1.03;[2] we use the October 2012 version of Wiktionary[3] as our tag dictionary.

After pruning the search space with the dictionary, we place soft constraints derived by projecting POS tags across word alignments. The alignments are filtered for confidence (see §5.1.2), but we also filter any projected tags that are not licensed by the dictionary. The example in Figure 1 illustrates why this dictionary filtering step is important. Consider the English-Spanish phrase pair from Figure 1, which we observed in our training data. Our supervised tagger correctly tags *Asian* with the ADJ tag as shown in the figure. *Asian* is aligned to the Spanish word *Asia*, which should be tagged NOUN. Because the Spanish Wiktionary only allows the NOUN tag for *Asia*, we do not project the ADJ tag from the English word *Asian*. By contrast, we do project the NOUN tag from the English word *sponges* to the Spanish

[1] Note that we did not implement regularization of θ in the stochastic optimizer, hence our PR objective (Eq. 8) was unregularized; however, the baseline models use ℓ_2 regularization.

[2] http://code.google.com/p/universal-pos-tags

[3] http://meta.wikimedia.org/wiki/Wiktionary

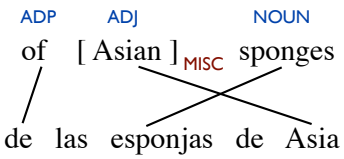

Figure 1: An English (top) – Spanish (bottom) phrase pair from our parallel data. The correct POS tags and NER annotations are shown for the English phrase. Word alignments are shown as links between English and Spanish words.

word *esponjas* because this tag is in our dictionary for the latter word.

For all our POS experiments, we evaluate on seventeen target languages. Fifteen of these languages were part of the experiments conducted by Täckström et al. (2013); we add Arabic and Hungarian to the set. The first column of Table 1 lists all seventeen languages using their two-letter abbreviation codes from the ISO 639-1 standard. The evaluation datasets correspond to the test sets from the CoNLL shared tasks on dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007). For French we use the treebank of Abeillé et al. (2003). English serves as our source language and we use the Penn Treebank (Marcus et al., 1993, with tags mapped to the universal tags) to train our supervised source-side model.

4.2 Named-Entity Segmentation

Second, we investigate the task of named-entity segmentation. The goal of this task is to identify the boundaries of named-entities for a given language without classifying them by type. This is the *unlabeled* version of named-entity recognition, and is more amenable to cross-lingual supervision. To understand why that is, consider again the example from Figure 1. The English supervised NE tagger correctly identifies *Asian* as a named entity of type MISC (miscellaneous). The word-alignments suggest we should transfer this annotation to the Spanish word *Asia* which is also an entity. However, this should be labeled LOC (location) according to the CoNLL annotation guidelines (Tjong Kim Sang and De Meulder, 2003). Because syntactic variations of this kind are common, it makes cross-lingual detection of NE boundaries as well as types hard.[4] In this paper, we focus on named-entity segmentation alone, consider the full NER task out of scope. We use English as a source language and train a supervised English named-entity tagger with the labels in place, using the CoNLL 2003 shared task data (Tjong Kim Sang and De Meulder, 2003). We project the spans using the maximal-span heuristic (Yarowsky and Ngai, 2001). We project into Dutch, German and Spanish and evaluate on the standard CoNLL 2002 and 2003 shared task data sets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

4.3 Parallel Data

For both tasks we use parallel data gathered automatically from the web using the method of Uszkoreit et al. (2010), as well as data from Europarl (Koehn, 2005) and the UN parallel corpus (UN, 2006), for languages covered by the latter two corpora. The parallel sentences are word aligned with the aligner of DeNero and Macherey (2011). The size of the parallel corpus is larger than we need for our tasks, so we follow Täckström et al. (2013) in sampling 500k tokens for POS tagging and 10k sentences for named-entity segmentation (see §5.1.2 and §5.2.2).

5 Experimental Details

In this section, we provide details about task-specific implementations of the supervised source-side model and the word-alignment filtering techniques (steps 2 and 3 in Algorithm 1 respectively); we also briefly describe the setup of the cross-lingual experiments for each task.

5.1 Part-of-Speech Tagging

We first focus on the experimental setup for the POS tagging task. When describing feature sets we refer to features conjoined with just a single tag as *emission features* and with consecutive tag pairs as *transition features*.

[4] We tried using English and German gazetteers from the CoNLL 2002 and 2003 shared tasks as a label dictionary similar to the way we use Wiktionary for POS tagging. This did not work well because the CoNLL gazetteers do not have good coverage on our parallel datasets, which we use for training.

5.1.1 Supervised Source-Side Model

We tag the English side of our parallel data with a supervised first-order linear-chain CRF POS tagger. We use standard features for tagging. Our emission features are a bias feature, the current word, its suffixes up to length 3, its capitalization shape, whether it contains a hyphen, digit or punctuation and its cluster identity. Our transition features are a bias feature and the cluster identities of each word in the transition. For the cluster-based features, we use monolingual word clusters induced with the exchange algorithm of Uszkoreit and Brants (2008), which implements the same objective as Brown et al. (1992); these clusters have shown improvements for sequence labeling tasks (Turian et al., 2010; Täckström et al., 2012). We set the number of clusters to 256 for both the source side tagger and all the other languages. On Section 23 of the WSJ section of the Penn Treebank, the source side tagger achieves an accuracy of 96.2%.

5.1.2 Word Alignment Filtering

Following Täckström et al. (2013), we tag the English side of our parallel data using the source-side POS tagger, intersect the word alignments and filter alignments with confidence below 0.95. We sample 500,000 tokens of target side sentences for each language, and use this as training data for learning weakly-supervised taggers.

5.1.3 Setup for Cross-Lingual Experiments

Following Täckström et al. (2013) we use a reduced feature set for the cross-lingual models. The emission features are the same as the supervised model but without the punctuation feature,[5] and we use only the bias transition feature. Because this limits the ability of the model to use context, we also experiment with an extended feature set that has transition features for the clusters of each word in the transition, and their suffixes up to length 3. We refer to the extended-feature models as "BASE+" and "PR+" to distinguish them from the models with fewer features, labeled "BASE" and "PR".

We train BASE and BASE+ using L-BFGS with an ℓ_2 regularization weight of 1 for 100 iterations to reproduce the setup used by Täckström et al. (2013). We have only one constraint feature in our posterior regularization models that fires for the unpruned projected tags on words x_i. This feature controls how often our model trusts a projected tag; we explain how its strength is chosen in §6.1. The PR and PR+ models are trained using the stochastic gradient method described in §3.3.

[5]The dictionary licenses punctuations, only by the '.' tag.

5.2 Named-Entity Segmentation

In this subsection, we turn to the experimental details of the named-entity segmentation system.

5.2.1 Supervised Source-Side Model

To train our supervised source-side NER model, we implemented a linear-chain first order CRF model. Our feature set was inspired by the model of Kazama and Torisawa (2007, §6.1); we used all the local features from their model except the gazetteer features, and added cluster emission features for offsets in the range [-2, 2] and transition features for offsets in the range [-1, 1] as well as a sentence-start feature. We use automatic POS tags for all the experiments.

We use a BIO encoding of the four NER labels (PER, LOC, ORG and MISC). We also experimented with omitting the NE labels from the tagger, still with a BIO encoding for segments, but the results were worse on average than what we report in Table 2. We train the source-side model on the CoNLL 2003 English training set with log-loss using L-BFGS for 100 iterations with ℓ_2 regularization weight of 0.1. The model gets 90.9% and 87.5% labeled F_1 on the CoNLL development and test sets respectively.[6]

5.2.2 Word-Alignment Filtering

Projecting named entities across languages can be error prone for several reasons. Mistakes introduced by the automatic word aligner is one of them. Word alignment errors are particularly problematic for entity mentions because of the garbage collector effect (Brown et al., 1993); due to differences in the word order between languages, a few alignment errors can result in many errors in the other language. Additionally, entities can occur on just one side of the bitext.[7] Another source of error is the automatic

[6]These performance values would place us among the top three competitors of the CoNLL 2003 shared task.

[7]For example, "*It's all Greek to me.*" in one language and "*I don't understand it.*" in another.

labeling on the source side, which is inaccurate if the parallel corpus is out of domain. To mitigate these errors, we aggressively filter the training data for this task. We discard sentence pairs where more than 30% of the source language tokens are unaligned, where any source entities are unaligned or where any source entities are more than 4 tokens long. We also compute a confidence score over entity annotations as the minimum posterior over the tags that comprise the entity and discard sentence pairs that have an entity with confidence below 0.9. Finally, we discard any sentences that contain no projected entities. These filtering steps allow us to keep 7.4%, 9.7% and 10.4% of the aligned sentence pairs for German, Spanish and Dutch, respectively, resulting in very high-precision named-entity projections (see Table 2). For comparison, we also perform experiments without this filtering step.

5.2.3 Setup for Cross-Lingual Experiments

We use a CRF with the same feature set and BIO encoding for the cross-lingual models as the source-side NER model. We compare our approach ("PR" in Table 2) to a baseline ("BASE" in Table 2) which treats the projected annotations as fully observed. The PR model treats the projected NE spans of a sentence as observed, and allows all labels on the remaining tokens. Since the "O" tag is never seen, an unconstrained model would learn to never predict it. We add two features that fire when the current word is tagged "O": a bias feature and a feature that fires when the automatic POS tag is a proper noun. We set up $\mathcal{Q}$ so the desired expectations are at least 0.98 and at most 0.1 for these constraint features respectively.

6 Results

In this section, we turn to our experimental results; first, we focus on POS tagging and then turn to the NE segmentation task.

6.1 Part-of-Speech Tagging

Constraint Strength: As discussed in §4.1, it is important to filter out projected annotations not licensed by Wiktionary. Thus, the quality of weakly-supervised POS taggers learned from projections is closely correlated with the coverage of the Wiktionary. To quantify the effect of Wiktionary coverage, we counted the expected number of possible tags per token (TpT) for our unlabeled corpora. Specifically, for each token, we counted the number of tags licensed by the dictionary, or all tags for word forms not in the dictionary. For each language, we also ran our system with constraint strengths in {0.7, 0.75, 0.8, 0.85, 0.9, 0.92, 0.95, 0.98, 1.00}, and computed the optimal constraint strength from this set. We found that the best constraint strength is closely correlated with the average number of tags available for each token. Figure 2 shows the best constraint strength as a function of the inverse of the number of unpruned tags per token. As observed in the figure, the relationship between the optimal strength and 1/TpT is roughly linear. Figure 2 also shows a linear approximation to the data plotted. When applying this technique to a new language, we would not be able to estimate the optimal constraint strength, but we could use the linear approximation and knowledge of 1/TpT to estimate it. For our experiments below, we perform this estimation for each language using the linear approximation computed from the remaining languages.

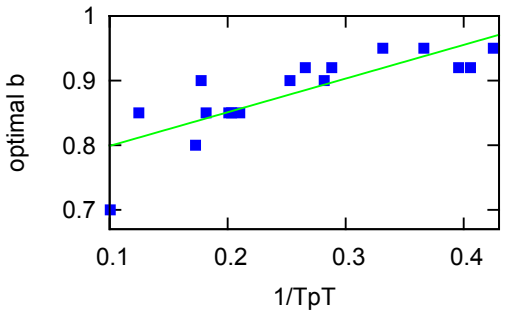

Figure 2: Correlation between optimal constraint value $\mathbf{b}$ and dictionary pruning efficiency. Each blue square is a language, the green line is a linear approximation of the data.

Results: The results for our part-of-speech tagging experiments are in Table 1. We compare our results to BASE, which corresponds to reruns of the best model of Täckström et al. (2013, Column 9 of Table 2), and closely aligns with the numbers reported by the authors. We see in Table 1 that for both feature sets (i.e., with and without the '+' extension), our estimated constraint strength is usually better than using a constraint strength of 1. The results in the column labeled PR are better than BASE for 12 out of 17 languages, and the results for PR+ are

	BASE	BASE+	PR	PR+
ar	37.84	44.96*	49.04*	**50.10***
bg	88.04	87.93	88.02	**88.42***
cs	79.67	80.01*	80.20*	**80.68***
da	88.14	87.92	**88.24***	87.90
de	90.32	89.97	**90.41***	90.29
el	90.03	89.03	**90.63***	90.24*
es	86.99	86.81	87.20*	**87.21***
fr	87.07	**87.53***	87.44*	87.48*
hu	82.05	82.05	82.14*	**83.13***
it	89.48	**89.89***	89.52	89.72*
ja	**80.63**	78.54	80.02	79.68
nl	85.89	85.77	85.59	**85.98***
pt	90.93	**91.60***	91.48*	91.56*
sl	82.46	82.08	83.16*	**83.49***
sv	89.06	88.72	**89.25***	88.77
tr	64.39	65.74*	63.88	**66.47***
zh	73.98	72.82	**74.51***	68.43
Avg	81.59	81.85*	**82.40***	82.33*
-zh-ar	85.01	84.91	85.15*	**85.40***

Table 1: POS tagging results. BASE represents the best model of Täckström et al. (2013). PR is a system with the same features but with relaxed constraints. BASE+ and PR+ add additional model features (see §5.2.3). * indicates improvements over the previous state of the art (BASE), and bold values indicate the best score for a language. "Avg" indicates averaged results for all 17 languages, while "-zh-ar" shows averaged results without Chinese and Arabic.

better than BASE+ for 13 out of 17 languages. Additionally, adding features does not tend to help the baseline model to a large extent (the wins are for 6 languages), but does tend to help the PR model (for 11 languages); however, there is a large drop in performance for Chinese.

Error Analysis: Here, we analyze the nature of improvements that the PR models get. For the languages where PR results in large improvements, it stems from the ability to allow the sentential context to sometimes override the tag projected via the parallel data. For example, the Czech word *se* can either be a reflexive pronoun (such as *ourselves* in English) or translate to the preposition *with*. The pronominal sense comprises about 95% of occurrences in the Czech annotations, but it would not appear in an English translation. For example, the phrase "*podívali jsme se*" translates to "*we looked*", and the word *jsme* would typically be aligned to *we*; *se*, which serves as a reflexive pronoun here, remains unaligned. Consequently, in our data, over 7000 occurrences of *se* appear, but only 17 instances have a tag projection that is not filtered by Wiktionary. Since the remaining are tagged with the preposition tag, the hard-constrained baseline always tags *se* as a preposition. By contrast, the soft-constrained PR model predicts the pronominal sense in cases where the context is most indicative of a pronoun – 38% of the time. It still mistags many of the pronominal cases where the contextual evidence is not strong enough. We get very similar behavior with the Hungarian word *hogy* which can translate to the conjunction *that* (as in "*I see that you are here*") or the adverb *how*.

We found that the drastic drop in performance for Chinese under the PR+ model is due to the possessive marker "的" which serves exclusively as a particle in the test data. Wiktionary also allows the noun and adverb tags. The adverbial use is actually a different token (的确 ↔ *really, truly*) containing the same character. Because the cross-lingual training data is based on machine-learned alignments, 99.4% of the training examples of 的 have no annotations, and only 0.6% have the particle annotation projected from the English *'s* possessive marker. If we remove the noun and adverb senses from the Wiktionary performance of PR+ improves to 72.87%. Alternatively, we could add another constraint to prefer closed-class words over open-class words when both are licensed by the dictionary. When we add such a constraint to Chinese with a constraint value of 0.95, we recover most of the loss (68.43 → 72.94); however, we do not report this specific change to the Chinese experimental setup in Table 1 to maintain generality.

6.2 Named-Entity Segmentation

Results: Table 2 shows the results for the named entity segmentation experiments. First, we observe that the word alignment filtering step (§5.2.2) improves results for all three languages by significant margins, for both the BASE and PR models. Both with and without filtering, we observe that the baseline models are very strongly biased towards precision. The filtering step tends to help with recall more than precison for both models. By having a soft constraint via PR and allowing some segmentations to fall outside of the transferred one, we get an increase in recall,

		No Filtering		Filtering (§5.2.2)	
Lang	Metric	BASE	PR	BASE	PR
de	Prec	74.29	73.85	75.36	**76.47**
	Recall	41.69	54.50	54.71	**64.61**
	F_1	53.41	62.71	63.39	**70.04**
es	Prec	74.53	62.10	**82.50**	70.22
	Recall	56.39	78.33	67.27	**81.10**
	F_1	64.20	69.28	74.11	**75.27**
nl	Prec	81.90	75.12	**86.39**	76.09
	Recall	50.54	76.11	65.45	**79.11**
	F_1	62.51	75.61	74.47	**77.57**
Above: dev, below: test					
de	Prec	73.23	71.67	69.90	**70.94**
	Recall	39.70	51.81	52.52	**61.42**
	F_1	51.49	60.14	59.97	**65.84**
es	Prec	75.38	65.40	**83.50**	73.68
	Recall	56.00	80.30	67.55	**83.31**
	F_1	64.26	72.09	74.68	**78.20**
nl	Prec	79.45	73.55	**86.01**	77.05
	Recall	47.45	75.37	65.16	**80.11**
	F_1	59.42	74.45	74.14	**78.55**

Table 2: Result for the named-entity segmentation experiments. The highest score in each category is shown in bold. Note that "No Filtering" still discards sentences with no projected entities.

and in turn an improved F_1 score. On average the PR model improves F-score by 3.6% on the development set and 4.6% on the test set over the baseline (when filtering is used). Note that because we focus on named entity segmentation, our results are not directly comparable to those of Täckström (2012), who train a de-lexicalized named entity recognizer on one language and apply it to other languages.

Error Analysis: In order to get a sense for the types of errors made by the baseline which are corrected by the PR model, we collected statistics about the most frequent errors in the segments extracted by the baseline and by our model. We divided the errors into missing segments, extraneous segments and overlapping segments.

From Table 2, it is clear that the most common errors for the baseline models are missing entities. From our analysis of the CoNLL development data, we found that the entities that occur with little context (such as the location and publisher of an item) at the onset of news articles are most frequently missed. For German, *dpa* (Deutsche Presse-Agentur) and *Reuter* are the two most common missing segmentations; the Spanish counterparts are *Gobierno* (Government) and *Barcelona*, while for Dutch they are *De Morgen* and *Brussel*. While filtering parallel sentences and using a soft constraint both increase recall, even our strongest model does not get enough information to predict these entities, and they continue to be major sources of error. By contrast, the names mentioned in context are the ones that are most frequently added to the analysis when PR is used. In a sense this is desirable, since a machine-learned named-entity segmentation system is most useful for the long tail of entity mentions.

If we filter the training data and use the PR model to further increase recall, precision errors tend to become relatively more frequent (this trend is observable in Table 2). For German, the most frequent precision error is *Mark* referring to the *Deutsche Mark*. For Spanish, the most frequent precision errors are due to boundary errors. The Spanish annotation guidelines include enclosing quotes as part of the entity name, and failing to include them accounts for just under 1% of the precision errors of the PR system that uses filtering. The second most frequent error is failing to segment *Inter de Milán*. The model segments out either *Inter* or *Milán* or both by themselves depending on context.

7 Conclusions

In this paper, we presented a framework for cross-lingual transfer of sequence information from a resource-rich source language to a resource-poor target language. Our framework incorporates soft constraints while training with projected information via posterior regularization. We presented the efficacy of our framework on two very useful natural language tasks: POS tagging and named-entity segmentation. The soft constraints used in our work model intuitions about a given task. For the POS tagging problem, we designed constraints that also incorporate projected token-level information, and presented a principled method for choosing the extent to which this information should be trusted within the PR framework. This approach generalizes the state of the art in cross-lingual projection work in the context of POS tagging, and improves upon it.

Across seventeen languages, our models outperform the previous state of the art by an average of 0.8% (greater than 4% error reduction), and outperforms it on twelve out of seventeen languages. For named-entity segmentation, our model results in 3.6% and 4.6% absolute improvements in F_1-score on our development and test sets respectively, when averaged across three languages.

Acknowledgments

We would like to thank Ryan McDonald, Fernando Pereira, Slav Petrov and Oscar Täckström for numerous discussions on this topic and providing detailed feedback on early drafts of this paper. We are also grateful to the four anonymous reviewers for their valuable comments.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 10. Kluwer.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty. 1993. But dictionaries are data too. In *Proceedings of the Workshop on Human Language Technology*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of EMNLP*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of ACL-HLT*.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of ACL-IJCNLP*.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Jun'ichi Kazama and Kentaro Torisawa. 2007. A new perceptron algorithm for sequence labeling with non-local features. In *Proceedings of EMNLP*.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of ACL*.

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

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86.

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

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of EMNLP-CoNLL*.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proceedings of ICML*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of NAACL-HLT*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of CoNLL*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

UN. 2006. ODS UN parallel corpus.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-HLT*.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of COLING*.

Andreas Vlachos. 2006. Active annotation. *Proceedings of EACL*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of IJCNLP Workshop: NLP for Less Privileged Languages*.