Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{yuanzh, dgaddy, regina, tommi}@csail.mit.edu

Abstract

In the absence of annotations in the target language, multilingual models typically draw on extensive parallel resources. In this paper, we demonstrate that accurate multilingual part-of-speech (POS) tagging can be done with just a few (e.g., ten) word translation pairs. We use the translation pairs to establish a coarse linear isometric (orthonormal) mapping between monolingual embeddings. This enables the supervised source model expressed in terms of embeddings to be used directly on the target language. We further refine the model in an unsupervised manner by initializing and regularizing it to be close to the direct transfer model. Averaged across six languages, our model yields a 37.5% absolute improvement over the monolingual prototype-driven method (Haghighi and Klein, 2006) when using a comparable amount of supervision. Moreover, to highlight key linguistic characteristics of the generated tags, we use them to predict typological properties of languages, obtaining a 50% error reduction relative to the prototype model.[1]

1 Introduction

After two decades of study, the best performing multilingual methods can in some cases approach their supervised monolingual analogues. To reach this level of performance, however, multilingual methods typically make use of significant parallel resources such as parallel translations or bilingual dictionaries. These resources act as substitutes for explicit annotations available in the target language for supervised methods. It is less clear what can be done without extensive parallel resources. Indeed, the motivation for our paper comes from trying to understand how little parallel data is necessary for effective multilingual transfer.

In this paper, we demonstrate that only ten word translation pairs suffice for effective multilingual transfer of part-of-speech (POS) tagging. To achieve this we make use of and integrate two sources of statistical signal. First, we enable transfer of information from the source to target languages by establishing a coarse mapping between word embeddings in two languages on the basis of the few available translation pairs. The mapping is useful because of significant structural similarity of embedding spaces across languages. Second, we leverage the potential of unsupervised monolingual models to capture language-specific syntactic properties. The two sources of signals are largely complementary. Embeddings provide a coarse alignment between languages while unsupervised methods fine tune the correspondences in service of the task at hand. While unsupervised methods are fragile and challenging to estimate in general, they can be helpful if initialized and regularized properly, which is our focus.

In order to transfer annotations, we align monolingual embeddings between languages. However, a full fine-grained alignment is not possible with only ten translation pairs due to differences between the languages and variations across raw corpora from which the embeddings are derived. Instead, we re-

[1] Our code and data are available at https://github.com/yuanzh/transfer_pos.

Proceedings of NAACL-HLT 2016, pages 1307–1317,
San Diego, California, June 12-17, 2016.

strict the initial coarse mapping to be linear and isometric (orthonormal) so as to leave lengths and angles between the word vectors invariant. One advantage is that this preserves cosine similarity between vectors, which is viewed as a proxy for syntactic/semantic similarity (Mikolov et al., 2013a; Pennington et al., 2014; Herbelot and Vecchi, 2015). The resulting coarse alignment is then used to initialize and guide an unsupervised model over the target language.

Our unsupervised model is a feature-based hidden Markov model (HMM) expressed in terms of word embeddings. By establishing a common multilingual embedding space, we can map the source HMM estimated from supervised annotations directly to the target. The resulting "direct transfer" model should be further adjusted as languages differ, and the initial alignment obtained based on embeddings is imperfect. For this reason we cast the direct transfer model as a regularizer for the target HMM, and permit the HMM to further adjust the embedding transformations and relations of embeddings to the tags both globally (overall rotation and scaling) and locally (introducing small corrections).

Our two phase approach is simple to implement, performs well, and can be adapted to other NLP tasks. We evaluate our approach on POS tagging using the multilingual universal dependency treebanks (Nivre et al., 2016). Specifically, we use English as the source language and test on three Indo-European languages (Danish, German and Spanish) and three non-Indo-European-languages (Finnish, Hungarian and Indonesian). Experimental results show that our method consistently outperforms various baselines across languages. On average, our full model achieves 8% absolute improvement over the direct transfer counterpart. We also compare against a prototype-driven tagger (Haghighi and Klein, 2006) using 14 prototypes as supervision. Our model significantly outperforms Haghighi and Klein (2006)'s model by 37.5% (67.5% vs 30%).

We also introduce a novel task-based evaluation of automatic POS taggers, where tagger predictions are used to determine typological properties of the target language. This evaluation highlights key linguistic features of the generated tags. On this task, our model achieves 80% accuracy, yielding 50% error reduction relative to the prototype model.

2 Related Work

Multilingual POS Tagging Prior work on multilingual POS tagging has mainly focused on the *tag projection* method (Yarowsky et al., 2001; Wisniewski et al., 2014; Duong et al., 2013; Duong et al., 2014; Täckström et al., 2013; Das and Petrov, 2011; Snyder et al., 2008; Naseem et al., 2009; Chen et al., 2011). All these approaches assume access to a large amount of parallel sentences to facilitate multilingual transfer. In our work, we focus on a more challenging scenario, in which we do not assume access to parallel sentences. Instead of projecting tag information via word alignment, the transfer in our model is driven by mapping multilingual embedding spaces. Kim et al. (2015) also use latent word representations for multilingual transfer. However, similarly to prior work, this representation is learned using parallel data.

The feasibility of POS tagging transfer without parallel data has been shown by Hana et al. (2004). The transfer is performed between typologically similar languages, which enables the model to directly transfer the transition probabilities from source to the target. Moreover, emission probabilities are hand-engineered to capture language-specific morphological properties. In contrast, our method does not require any language-specific knowledge on the target side.

Multilingual Word Embeddings There is an expansive body of research on learning multilingual word embeddings (Gouws et al., 2014; Faruqui and Dyer, 2014; Lu et al., 2015; Lauly et al., 2014; Luong et al., 2015). Previous work has shown its effectiveness across a wide range of multilingual transfer tasks including tagging (Kim et al., 2015), syntactic parsing (Xiao and Guo, 2014; Guo et al., 2015; Durrett et al., 2012), and machine translation (Zou et al., 2013; Mikolov et al., 2013b). However, these approaches commonly require parallel sentences or bilingual lexicon to learn multilingual embeddings. Vulic and Moens (2015) have alleviated the requirements by inducing multilingual word embeddings directly from a document-aligned corpus such as a set of Wikipedia pages on the same theme but in different languages. However, they still used about ten thousands aligned documents as parallel supervision. Our work demonstrates that useful multi-

lingual embeddings can be learned with a minimal amount of parallel supervision.

3 Multilingual POS Tagger

Our method is designed to operate in the regime where there are no parallel sentences or target annotations. We assume only a few, in our case ten, word translation pairs. This small number of translation pairs together with the tags that they carry from the source to the target do not provide sufficient information to train a reasonable supervised tagger, even for very close languages where word translations would be mostly one-to-one and tags fully preserved in translation. Other cues are necessary.

The few translation pairs provide just enough information to obtain a coarse global alignment between the source and target language embeddings. We limit the initial linear transformation between embeddings to isometric (orthonormal) mappings so as to preserve norms and angles (e.g., cosine similarities) between words. Once the embeddings are aligned, any source language model expressed in terms of embeddings can be mapped to a target language model. The approach is akin to direct transfer commonly applied in parsing (McDonald et al., 2011; Zeman and Resnik, 2008) though often with more information. We use the term "direct transfer" to mean the process where no further adjustment is performed beyond the immediate mapping via (coarsely) aligned embeddings.

Direct transfer is insufficient between languages that are syntactically (even moderately) divergent. Instead, we use the directly transferred model to initialize and regularize an unsupervised tagger. Specifically, we employ a feature-based HMM (Berg-Kirkpatrick et al., 2010) tagger for both the source and target languages with two important modifications. The emission probabilities in the source language HMM are expressed solely in terms of word embeddings (cf. skip-gram models). Such distributions can be directly transferred to the target domain. Our target language HMM is, however, equipped with additional adjustable parameters that can be learned in an unsupervised manner. These include parameters for modifying the initial global linear transformation between embeddings. Beyond this linear transformation, we also add "correction terms" to each tag-word pair that are in principle sufficient to specify any HMM. Both of these additional sets of parameters are regularized towards keeping the initial direct transfer model. As a result, our strongly governed unsupervised tagger can succeed where an unguided unsupervised tagger would typically fail.

In the remainder of this section, we describe the approach more formally, starting with the coarse alignment between embeddings, followed by the supervised feature-based HMM, and the unsupervised target language HMM.

3.1 Isometric Alignment of Word Embeddings

Here we find a linear transformation from the target language embeddings to the source language embeddings using the translation pairs. The resulting transformation permits us to directly apply any source language model on the target language, i.e., it enables direct transfer. To this end, let $\boldsymbol{V}_s \in \mathbb{R}^{n_s \times d}$ and $\boldsymbol{V}_t \in \mathbb{R}^{n_t \times d}$ be the word embeddings estimated for the source and target languages, respectively, with vocabulary sizes n_s and n_t. All the embeddings are of dimension d. The submatrices of embeddings pertaining to k anchor words (from translation pairs) are denoted as $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_t$, where $\boldsymbol{\Sigma}_s, \boldsymbol{\Sigma}_t \in \mathbb{R}^{k \times d}$.

We find a linear transformation $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ that best aligns the embeddings of the translation pairs in the sense of minimizing

$$||\boldsymbol{\Sigma}_t \boldsymbol{P} - \boldsymbol{\Sigma}_s||^2 \quad (1)$$

subject to the **isometric** (orthonormal) constraint $\boldsymbol{P}^T \boldsymbol{P} = \boldsymbol{I}$. We use the steepest descent algorithm (Abrudan et al., 2008) to solve this optimization problem.[2] Once $\boldsymbol{P}$ is available, we can map all the target language embeddings $\boldsymbol{V}_t$ to the source language space with $\boldsymbol{V}_t \boldsymbol{P}$. Note that since typically in our setting $k < d$ (e.g. $k = 10$) additional constraints such as isometry are required.

Motivation behind the Isometric Constraint We impose isometry on the linear transformation so as to preserve angles and lengths of the word vectors after the transformation. A number of recent studies have explored the use of cosine similarity of word

[2] Our implementation is based on the toolkit available at `http://legacy.spa.aalto.fi/sig-legacy/unitary_optimization/`.

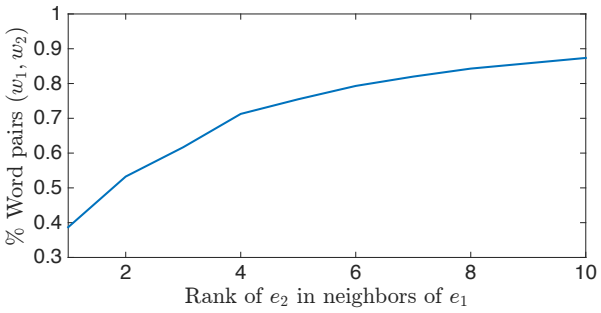

Figure 1: Cumulative fraction of word translation pairs among top 1,000 most frequent words where the nearest neighbor of a German word (vector) appears as the r^{th} nearest neighbor after translation, measured in terms of their monolingual word embeddings.

vectors as a measure of semantic relations between words. Thus, for example, if two words have high cosine similarity in German (target), the corresponding words in English (source) should also be similar. To validate our isometric constraint further, we verify whether nearest neighbors are preserved in monolingual embeddings after translation. To this end, we take the top 1,000 most frequent words in German and their translations into English and ask whether nearest neighbors are preserved if measured in terms of their monolingual embeddings. For each word vector w_1 and its nearest neighbor w_2 in German, let e_1 and e_2 be the corresponding English vectors. We compute the rank of e_2 in the ordered list of nearest neighbors of e_1. As Figure 1 shows, in more than 50% of word pairs, e_2 is among the top-2 neighbors of e_1. In over 90% of the word pairs e_2 is among e_1's top-10 closest neighbors.

For the purposes of comparison (see Section 5), we introduce also a linear transformation without isometry. In other words, we find $\boldsymbol{P}$ that minimizes $||\boldsymbol{\Sigma}_t\boldsymbol{P} - \boldsymbol{\Sigma}_s||^2$ via the **Moore–Penrose pseudoinverse** (Moore, 1920; Penrose, 1955). Specifically, let $\boldsymbol{\Sigma}_t^+$ be the pseudoinverse of $\boldsymbol{\Sigma}_t$. Then the solution takes the form $\boldsymbol{P} = \boldsymbol{\Sigma}_t^+\boldsymbol{\Sigma}_s$, and has the minimum Frobenius norm among all possible solutions.

3.2 Supervised Source Language HMM

Here we briefly describe how we train a supervised tagger on the source language. The resulting model, together with aligned embeddings, specifies the direct transfer model. It will also be used to initialize and guide the unsupervised tagger on the target language.

Our model has the same structure as the standard HMM but we replace the transition and emission probabilities with log-linear models (cf. feature-based HMM by Berg-Kirkpatrick et al. (2010)). The transition probabilities include all indicator features and therefore impose no additional constraints. The emission probabilities, in contrast, are expressed entirely in terms of word embeddings $\boldsymbol{v}_x$ as features. More formally, the emission probability of word x given tag y is given by

$$p_\theta(x|y) \propto \exp\{\, \boldsymbol{v}_x^T \boldsymbol{\mu}_y \,\} \tag{2}$$

Note that the parameters $\boldsymbol{\mu}_y$ (one vector per tag) can be viewed as tag embeddings. This supervised tagging model is trained to maximize the joint log-likelihood with l_2-regularization over parameters. We use the L-BFGS (Liu and Nocedal, 1989) algorithm to optimize the parameters.

Once the HMM has been trained, we can specify the direct transfer model. It has the same transition probabilities but the emission probabilities are modified according to $p_\theta^{dt}(x|y) \propto \exp\{\, \boldsymbol{v}_x^T \boldsymbol{P} \boldsymbol{\mu}_y \,\}$ where $\boldsymbol{v}_x$ is now the monolingual target embedding, transformed into the source space via $\boldsymbol{v}_x^T\boldsymbol{P}$. We apply the Viterbi algorithm to predict the most likely POS tag sequence.

3.3 Unsupervised Target Language HMM

Our unsupervised HMM for the target language is strictly more expressive than the direct transfer model so as to better tailor it to the target language. Let $\boldsymbol{v}_x$ again be the monolingual target embeddings estimated separately, prior to the HMMs. We map these vectors to the source language embedding space via $\boldsymbol{v}_x^T\boldsymbol{P}$ as discussed earlier, where $\boldsymbol{P}$ is already set and no longer considered a parameter. The form of the emission probabilities

$$p_\theta^t(x|y) \propto \exp\{\boldsymbol{v}_x^T \boldsymbol{P} \boldsymbol{M} \boldsymbol{\mu}_y + \theta_{x,y}\} \tag{3}$$

includes two modifications to the direct transfer model. First, we have introduced an additional global linear transformation $\boldsymbol{M}$ to correct the initial alignment represented by $\boldsymbol{P}$. Second, we include per-symbol parameters $\theta_{x,y}$ which, in principle, are capable of specifying any emission distribution on their own. The adjustable parameters in this model

(denoted collectively θ) are $\boldsymbol{M}$, $\{\boldsymbol{\mu}_y\}$, $\{\theta_{x,y}\}$, and the parameters pertaining to the transition probabilities. If we set $\boldsymbol{M} = \boldsymbol{I}$, $\theta_{x,y} = 0$ for all x and y, and borrow $\boldsymbol{\mu}_y$ and the transition parameters from the supervised HMM, then we recover the direct transfer model. Let θ_0 denote this setting of the parameters. In other words, the unsupervised HMM with initial parameters θ_0 is the direct transfer model.

Our approach include initializing $\theta = \theta_0$ and later regularizing θ to remain close to θ_0. The motivation behind this approach is two-fold. First, the initial alignment between embeddings was obtained only on the basis of the few available anchor words and may therefore need to be adjusted. Note that the linear transformation of embeddings now involves scaling and is no longer necessarily isometric. Second, the source and target languages differ and the embeddings are not strictly related to each other via any global linear transformation. We can interpret parameters $\theta_{x,y}$ as local (per word) non-linear deformations of the embedding vectors that specify the emission probabilities. We allow only small non-linear corrections by regularizing $\theta_{x,y}$ to remain close to zero, i.e., the values they have in θ_0.

Our unsupervised HMM is estimated by maximizing the regularized log-likelihood

$$L(\theta) = \sum_{i=1}^{n} \log P_\theta(\mathbf{x}_i) - \frac{\beta}{2}||\theta - \theta_0||_2^2 \quad (4)$$

where $\mathbf{x}_i$ is the i^{th} target language sentence, $P_\theta(\mathbf{x}_i)$ is the HMM with parameters θ, and n is the number of sentences in the target text to be annotated. Since all the parameters in the model are in a log-linear form, we simply use the regularization parameter β. Once estimated, we use the Viterbi algorithm to predict the most likely POS tag sequence.

Estimation Details We maximize $L(\theta)$ using the Expectation-maximization (EM) algorithm. In the E-step, we evaluate expected counts $e_{y',y}$ for tag-tag and $e_{x,y}$ for word-tag pairs, using the forward-backward algorithm. The M-step searches for θ that maximizes

$$l(\theta) = \sum_{y',y} e_{y',y} \log p_\theta^t(y'|y) + \sum_{x,y} e_{x,y} \log p_\theta^t(x|y) - \frac{\beta}{2}||\theta - \theta_0||_2^2 \quad (5)$$

The maximization can be be done via L-BFGS which involves computing the gradients of $\log p_\theta^t(y'|y)$ and $\log p_\theta^t(x|y)$ with respect to θ at every iteration. Because the conditional probabilities are expressed in a log-linear form, the gradients take on typical forms such as

$$\begin{aligned}\frac{dl(\theta)}{d\boldsymbol{\mu}_y} &= \sum_x e_{x,y}(\boldsymbol{v}_x^T \boldsymbol{P}\boldsymbol{M} - \sum_{x'} p_\theta^t(x'|y)\boldsymbol{v}_{x'}^T \boldsymbol{P}\boldsymbol{M}) \\ &\quad - \beta(\boldsymbol{\mu}_y - \boldsymbol{\mu}_{0y}) \\ \frac{dl(\theta)}{d\boldsymbol{M}} &= \sum_{x,y} e_{x,y}(\boldsymbol{P}^T \boldsymbol{v}_x \boldsymbol{\mu}_y^T - \sum_{x'} p_\theta^t(x'|y)\boldsymbol{P}^T \boldsymbol{v}_{x'} \boldsymbol{\mu}_y^T) \\ &\quad - \beta(\boldsymbol{M} - \boldsymbol{I}) \end{aligned} \quad (6)$$

where $\boldsymbol{\mu}_{0y}$ are initial values for $\boldsymbol{\mu}_y$.

4 Experimental Setup

Dataset We evaluate our method on the latest Version 1.2 of the Universal Dependencies Treebanks (Nivre et al., 2016; McDonald et al., 2013). We use English as the source language and six other languages as targets. Specifically, we choose three Indo-European languages: Danish (da), German (de), Spanish (es), and three non-Indo-European languages: Finnish (fi), Hungarian (hu), Indonesian (id). All treebanks are annotated with the same universal POS tagset. In our work, we map proper nouns to nouns and map symbol marks[3] and interjections to a catch-all tag X because it is hard and unnecessary to disambiguate them in a low-resource learning scenario. After mapping, our tagset includes the following 14 tags: noun, verb, auxiliary verb, adjective, adverb, pronoun, determiner, adposition, numeral, conjunction, sentence conjunction, particle, punctuation mark, and a catch-all tag X. Note that this universal tagset contains two more tags than the traditional universal tagset proposed by Petrov et al. (2011): auxiliary verb and sentence conjunction. We follow the standard split of the treebanks for every language. For each target language, we use the sentences in the training set as unlabeled data, and evaluate on the testing set.

Word Embeddings To induce monolingual word embeddings, we use the processed Wikipedia text dumps (Al-Rfou et al., 2013) for each language.

[3]Examples of symbol mark include "-", "/" etc.

Language	English	Danish	German	Spanish	Finnish	Hungarian	Indonesian
Tokens (10^6)	1,888	44	687	399	66	89	41

Table 1: Number of tokens of the Wikipedia dumps used for inducing word embeddings.

While Wikipedia texts may contain parallel articles, we show in Table 1 that the amount of text varies significantly across languages. Smith et al. (2010) also demonstrated that parallel information in Wikipedia is very noisy. Therefore, direct translations are difficult to get from these texts. We use the `word2vec` tool with the skip-gram learning scheme (Mikolov et al., 2013a). In our experiments we use $d = 20$ for the dimension of word embeddings and $w = 1$ for the context window size of the skip-gram, which yields the best overall performance for our model. In our analysis, we also explore the impact of embedding dimension and window size.

Word Translation Pairs For each target language, we collect English translations for the top ten most frequent words in the training corpus. Our preliminary experiments show that this selection method performs the best. The selected words are typically from closed classes, such as punctuation marks, determiners and prepositions. We find translations using Wiktionary.[4]

Model Variants Our model varies along two dimensions. On one dimension, we use two different methods for inducing multilingual word embeddings: **Pseudoinverse** and **Isometric** alignment as described in Section 3.1. On the other dimension, we experiment with two different multilingual transfer models. We use **Direct Transfer** to denote our direct transfer model, and **Transfer+EM** for our unsupervised model trained in the target language.

Baselines We also compare against the prototype-driven method of Haghighi and Klein (2006). Specifically, we use the publicly available implementation provided by the authors.[5] Note that their model requires at least one prototype for each POS category. Therefore, we select 14 prototypes (the most frequent word from each category) for the baseline, while our method only uses ten translation pairs.

[4] `https://www.wiktionary.org/`

[5] `http://code.google.com/p/prototype-sequence-toolkit/`

Evaluation Unlike other unsupervised methods, all models in our experiments can identify the label for each POS tag because of knowledge from either the source languages or prototypes. Therefore, we directly report the token-level POS accuracy for all experiments.

Other Details For all experiments, we use the following regularization weights: $\gamma = 0.001$ for supervised models learned on the source language and $\beta = 0.01$ for unsupervised models learned on the target language. During training, we also normalize the log-likelihood of labeled or unlabeled data by the total number of tokens. As a result, the magnitude of the objective value is independent of the corpus size, hence we do not need to tune the regularization weight for each target language. We run ten iterations of the EM algorithm.

5 Results

In this section, we first show the main comparison between the tagging performance of our model and the baselines. In addition, we include an experiment on typology prediction. In Section 5.2, we provide a more detailed analysis of model properties.

5.1 Main Results

Table 2 summarizes the results of the prototype baseline and different variations of our transfer model. Averaged across languages, our model significantly outperforms the prototype baseline by about 37.5% (67.5% vs 30%), demonstrating the effectiveness of multilingual transfer. Moreover, Table 2 shows that our full model (Transfer+EM with the isometric alignment mapping) consistently achieves the best performance compared to other model variations. Our model performs better on Indo-European languages than on other languages (72.9% vs. 62.1% on average), because Indo-European languages are linguistically more similar to the source language (English).

Impact of Training in the Target Language We observe that training on unlabeled data in the tar-

Method	Indo-European				Non-Indo-European			
	da	de	es	Avg.	fi	hu	id	Avg.
Prototype Model	41.3	25.5	28.7	31.8	8.2	44.5	30.1	27.6
Pseudoinverse								
Direct Transfer	56.7	49.4	68.4	58.2	54.3	60.1	57.7	57.4
Transfer+EM	64.4	65.8	74.9	68.4	57.5	**65.3**	62.7	61.8
Isometric Alignment								
Direct Transfer	59.8	55.4	67.4	60.9	54.4	61.4	57.2	57.7
Transfer+EM	**72.5**	**68.7**	**77.5**	**72.9**	**58.2**	63.4	**64.8**	**62.1**

Table 2: Token-level POS tagging accuracy (%) for different variants of our transfer model. We always use English as the source language. Target languages include Danish (da), German (de), Spanish (es), Finnish (fi), Hungarian (hu) and Indonesian (id). We average the results separately for Indo-European and non-Indo-European languages. The first row shows performance of the prototype-driven baseline (Haghighi and Klein, 2006). The rest shows results of our model when multilingual embeddings are induced with the pseudoinverse or isometric alignment method. "Direct Transfer" and "Transfer+EM" indicates our direct transfer model and our transfer model trained in the target language respectively.

get language (Transfer+EM model) consistently improves over the direct transfer counterpart. As the bottom part of Table 2 shows, running EM on unlabeled data yields an average of 12% absolute gain on Indo-European languages, while on non-Indo-European languages the gain is only 4.4%.

Impact of the Isometric Alignment Constraint As Table 2 shows, when we use Transfer+EM models, the isometric alignment method yields a 4.5% improvement over the pseudoinverse method (72.9% vs. 68.4%) on Indo-European languages. However, the improvement margin drops to 0.3% on non-Indo-European languages (62.1% vs. 61.8%). We hypothesis that this discrepancy is due to the difference in the degree of ambiguities of the anchor words across languages. For example, the anchor words of Spanish have an average of 1.5 possible translations to English, while for Indonesian the average ambiguity is 2.7. Therefore, the isometric assumption holds better and the EM algorithm finds a better local optimum for Indo-European languages than for non-Indo-European languages. We also observe a similar pattern in the direct transfer scenario.

Prediction of Linguistic Typology To assess the quality of automatically generated tags, we use them to determine typological properties of the target language. We predict values of the following five typological properties for each language: subject-verb, verb-object, adjective-noun, adposition-noun and demonstrative-noun. More specifically, the goal is to predict word ordering preferences such as whether an adjective comes before a noun (as in English) or after a noun (as in Spanish). We collect the true ordering preferences from "The World Atlas of Language Structure (WALS)" (Dryer et al., 2005). To make predictions, we train a multiclass support vector machine (SVM) classifier (Tsochantaridis et al., 2004) on a multilingual corpus using bigrams and trigrams of POS tags as features. The training data for SVM comes from a combination of the Universal Dependencies Treebanks, CoNLL-X, and CoNLL-07 datasets (Buchholz and Marsi, 2006; Nilsson et al., 2007), excluding all sentences in the target language. We train one classifier for each typological property, and make predictions for each of the six target languages. For evaluation, we directly report the overall accuracy on all 30 test cases (six languages combined with five typological prop-

Tagging Method	Typology Accuracy
Prototype	60.0
Direct Transfer	66.7
Transfer + EM	80.0
Gold	93.3

Table 3: The accuracy (%) of typological properties prediction using the outputs from different taggers. "Gold" indicates the result using gold POS annotations.

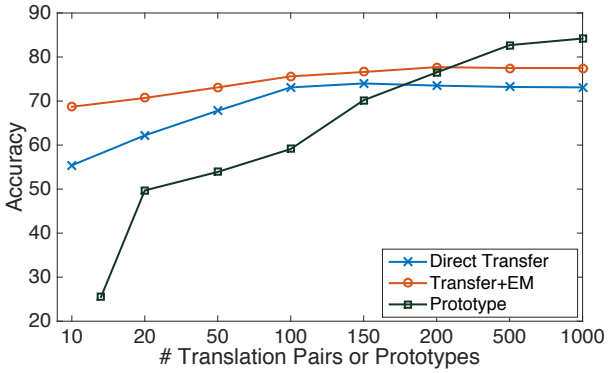

Figure 2: Accuracy of our models and the prototype baseline as a function of the amount of supervision, in German. x-axis is the number of translation pairs or prototypes used as supervision. Our models use multilingual embeddings induced with the isometric alignment method. The minimum number of prototypes used by the prototype baseline is 14.

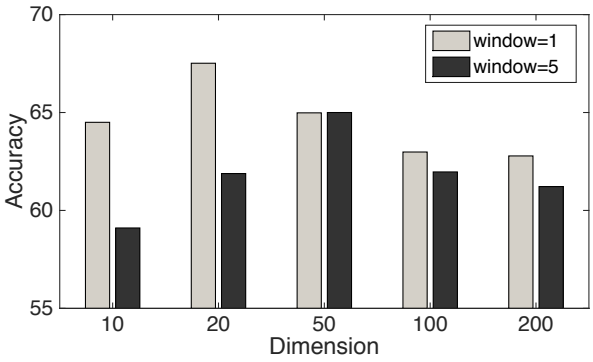

Figure 3: The average tagging accuracy (%) with different embedding dimensions and context window sizes. The model is Transfer+EM with the isometric alignment projection method.

erties).

Table 3 shows the accuracy of predicting typological properties with different tagging models. "Gold" corresponds to the result with gold POS annotations and is an upper bound of the prediction accuracy. We observe that the typology prediction accuracy correlates with the tagging quality. With the output of our best model, we predict the correct values for 80% of the typological properties. This corresponds to a 50% error reduction relative to the prototype model.

5.2 Analyses

Impact of the Amount of Supervision Figure 2 shows the accuracy of the Direct Transfer, Transfer+EM models, and prototype baseline with different amounts of supervision in German. Specifically, the x-axis is the number of translation pairs or prototypes used as supervision. The numbers with ten pairs or prototypes are the same as that in Table 2. We automatically extract more translation pairs using the Europarl parallel corpus (Koehn, 2005) and select pairs based on the word frequency in the target language. For the prototype model, we select the most frequent words as prototypes based on annotations in the training data, and guarantee that each POS category has at least one prototype. Note that the minimum number of prototypes used by the prototype model is 14.

One particularly interesting observation is that our model with ten pairs achieves an equivalent performance as that of the prototype-driven method with 150 prototypes. Multilingual transfer compensates for 15 times the amount of supervision. We also observe that the prototype-driven model outperforms our model when large amount of annotations are available. This can be explained by noise in the translation and the limitation from the linear embedding mapping process, which makes POS tags not preserve well across languages.

When comparing between our models, Figure 2 shows that Transfer+EM consistently improves over the Direct Transfer, while the gains are more profound in the low-supervision scenario. This is not surprising because with more translation pairs, we are able to induce higher quality multilingual embeddings, which is more beneficial to the direct transfer model.

Impact of Embedding Dimensions and Window Size Figure 3 shows the average accuracy across six target languages with different embedding dimensions and context window sizes. First, we observe that a small window size $w = 1$ consistently outperforms window size $w = 5$, demonstrating that smaller window sizes appear to produce word embeddings better for POS tagging. This observation is in line with the finding by Lin et al. (2015). Moreover, we obtain the best performance with dimension $d = 20$ when $w = 1$. On one hand, embeddings with smaller dimension (e.g. $d = 10$) have too little syntactic information for good POS tagging. On the other hand, if the embedding space has larger dimen-

Model	da	de	es	fi	hu	id	Average
All features	72.5	68.7	77.5	58.2	63.4	64.8	67.5
- Indicator features	70.8	64.8	73.9	53.7	62.9	56.8	63.8
- Transformation matrix M	60.2	65.6	73.2	58.6	59.6	70.8	64.7

Table 4: The accuracy (%) of our best Transfer+EM model with different feature sets, removing either indicator features or transformation matrix M at a time.

sion, the space will be more complex and mapping embedding spaces will be more difficult given only ten translation pairs. Therefore, we observe a performance drop with either smaller or larger dimensions.

Ablation Analysis on Features In our Transfer+EM model, we add indicator features and transformation matrix M to enhance the emission distribution (see Section 3.3). To analyze their contribution, we remove these features in turn and report the results in Table 4. Averaged over all languages, adding indicator features improves the accuracy by 3.7%, and adding a transformation matrix increases the accuracy by 2.8%.

6 Conclusions

In this paper, we demonstrate that ten translation pairs suffice for an effective multilingual transfer of POS tagging. Experimental results show that our model significantly outperforms the direct transfer method and the prototype baseline. The effectiveness of our approach suggests its potential application to a broader range of NLP tasks that require word-level multilingual transfer, such as multilingual parsing and machine translation.

Acknowledgments

The authors acknowledge the support of the U.S. Army Research Office under grant number W911NF-10-1-0533, and the support of the MIT EECS Super UROP program. We thank the MIT NLP group and the NAACL reviewers for their comments. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.

References

Traian E. Abrudan, Jan Eriksson, and Visa Koivunen. 2008. Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE Transaction on Signal Processing*, 56(3):1134–1147.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Desai Chen, Chris Dyer, Shay B Cohen, and Noah A Smith. 2011. Unsupervised bilingual POS tagging with markov random fields. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 64–71. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 600–609. Association for Computational Linguistics.

Matthew S Dryer, David Gil, Bernard Comrie, Hagen Jung, Claudia Schmidt, et al. 2005. The world atlas of language structures.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Increasing the quality and quantity of source language data for unsupervised cross-lingual POS tagging. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1243–1249.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? A case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 886–897.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1234–1244.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the 2006 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 320–327. Association for Computational Linguistics.

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

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32. Association for Computational Linguistics.

Young-Bum Kim, Benjamin Snyder, and Ruhi Sarikaya. 2015. Part-of-speech taggers for low-resource languages using CCA features. In *Proceedings of the Empiricial Methods in Natural Language Processing*. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 2015. Unsupervised POS induction with word embeddings. *arXiv preprint arXiv:1503.06760*.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Eliakim H. Moore. 1920. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9):394–395.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36:341–385.

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods*

in Natural Language Processing, volume 12, pages 1532–1543.

Roger Penrose. 1955. A generalized inverse for matrices. In *Proceedings of the Cambridge Philosophical Society*, pages 406–413.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Jason R Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050. Association for Computational Linguistics.

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.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1779–1785.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. *CoNLL-2014*, page 119.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Empiricial Methods in Natural Language Processing*, pages 1393–1398.