

Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages

Edoardo M. Ponti^κ, Ivan Vulić^κ, Ryan Cotterell^{κ,ζ},
Marinela Parović^κ, Roi Reichart^τ, Anna Korhonen^κ

^κUniversity of Cambridge ^ζETH Zürich ^τTechnion, IIT
^κ{ep490, iv250, rdc42, mp939, alk23}@cam.ac.uk
^τroiri@ie.technion.ac.il

Abstract

Most combinations of NLP tasks and language varieties lack in-domain examples for supervised training because of the paucity of annotated data. How can neural models make sample-efficient generalizations from task–language combinations with available data to low-resource ones? In this work, we propose a Bayesian generative model for the space of neural parameters. We assume that this space can be factorized into latent variables for each language and each task. We infer the posteriors over such latent variables based on data from seen task–language combinations through variational inference. This enables zero-shot classification on unseen combinations at prediction time. For instance, given training data for named entity recognition (NER) in Vietnamese and for part-of-speech (POS) tagging in Wolof, our model can perform accurate predictions for NER in Wolof. In particular, we experiment with a typologically diverse sample of 33 languages from 4 continents and 11 families, and show that our model yields comparable or better results than state-of-the-art, zero-shot cross-lingual transfer methods. Our code is available at github.com/cambridgeltl/parameter-factorization.

1 Introduction

The annotation efforts in NLP have achieved impressive feats, such as the Universal Dependencies (UD) project (Nivre et al., 2019) which now includes 83 languages. But, even UD covers only a meager subset of the world’s estimated 8,506 languages (Hammarström et al., 2016). What is more, the Association for Computational Linguistics Wiki¹ lists 24 separate NLP tasks. Labeled data, which is both costly and labor-intensive, is

missing for many of such task–language combinations. This shortage hinders the development of computational models for the majority of the world’s languages (Snyder and Barzilay, 2010; Ponti et al., 2019a).

A common solution is transferring knowledge across domains, such as tasks and languages (Yogatama et al., 2019; Talmor and Berant, 2019), which holds promise to mitigate the lack of training data inherent to a large spectrum of NLP applications (Täckström et al., 2012; Agić et al., 2016; Ammar et al., 2016; Ponti et al., 2018; Ziser and Reichart, 2018, *inter alia*). In the most extreme scenario, *zero-shot learning*, no annotated examples are available for the target domain. In particular, zero-shot transfer across *languages* implies a change in the data domain, and leverages information from resource-rich languages to tackle the same task in a previously unseen target language (Lin et al., 2019; Rijhwani et al., 2019; Artetxe and Schwenk, 2019; Ponti et al., 2019a, *inter alia*). Zero-shot transfer across *tasks* within the same language (Ruder et al., 2019a), on the other hand, implies a change in the space of labels.

As our main contribution, we propose a Bayesian generative model of the neural parameter space. We assume this to be structured, and for this reason factorizable into task- and language-specific latent variables.² By performing transfer of knowledge from both related tasks *and* related languages (i.e., from *seen* combinations), our model allows for zero-shot prediction on *unseen* task–language combinations. For instance, the availability of annotated data for part-of-speech (POS) tagging in Wolof and for named-entity recognition (NER) in Vietnamese supplies plenty of information to infer a task-agnostic representation for Wolof and

¹aclweb.org/aclwiki/State_of_the_art

²By latent variable we mean every variable that has to be inferred from observed (directly measurable) variables. To avoid confusion, we use the terms *seen* and *unseen* when referring to different task–language combinations.

a language-agnostic representation for NER. Conditioning on these, the appropriate neural parameters for Wolof NER can be generated at evaluation time. While this idea superficially resembles matrix completion for collaborative filtering (Mnih and Salakhutdinov, 2008; Dziugaite and Roy, 2015), the neural parameters are latent and are non-identifiable. Rather than recovering missing entries from partial observations, in our approach we reserve latent variables to each language and each task to tie together neural parameters for combinations that have either of them in common.

We adopt a Bayesian perspective towards inference. The posterior distribution over the model’s latent variables is approximated through stochastic variational inference (SVI; Hoffman et al., 2013). Given the enormous number of parameters, we also explore a memory-efficient inference scheme based on a diagonal plus low-rank approximation of the covariance matrix. This guarantees that our model remains both expressive and tractable.

We evaluate the model on two sequence labeling tasks: POS tagging and NER, relying on a typologically representative sample of 33 languages from 4 continents and 11 families. The results clearly indicate that our generative model surpasses standard baselines based on cross-lingual transfer 1) from the (typologically) nearest source language; 2) from the source language with the most abundant in-domain data (English); and 3) from multiple source languages, in the form of either a multi-task, multi-lingual model with parameter sharing (Wu and Dredze, 2019) or an ensemble of task- and language-specific models (Rahimi et al., 2019).

2 Bayesian Generative Model

In this work, we propose a Bayesian generative model for multi-task, multi-lingual NLP. We train a single Bayesian neural network for several tasks and languages jointly. Formally, we consider a set $T = \{t_1, \dots, t_n\}$ of n tasks and a set $L = \{l_1, \dots, l_m\}$ of m languages. The core modeling assumption we make is that the parameter space of the neural network is *structured*: specifically, we posit that certain parameters correspond to tasks and others correspond to languages. This structure assumption allows us to generalize to unseen task–language pairs. In this regard, the model is reminiscent of matrix factorization as applied to collaborative filtering (Mnih and Salakhutdinov, 2008; Dziugaite and Roy, 2015).

We now describe our generative model in three steps that match the nesting level of the plates in the diagram in Figure 1. Equivalently, the reader can follow the nesting level of the **for** loops in Algorithm 1 for an algorithmic illustration of the generative story.

- (1) **Sampling Task and Language Representations:** To kick off our generative process, we first sample a latent representation for each of the tasks and languages from multivariate Gaussians: $\mathbf{t}_i \sim \mathcal{N}(\boldsymbol{\mu}_{t_i}, \Sigma_{t_i}) \in \mathbb{R}^h$ and $\mathbf{l}_j \sim \mathcal{N}(\boldsymbol{\mu}_{l_j}, \Sigma_{l_j}) \in \mathbb{R}^h$, respectively. While we present the model in its most general form, we take $\boldsymbol{\mu}_{t_i} = \boldsymbol{\mu}_{l_j} = \mathbf{0}$ and $\Sigma_{t_i} = \Sigma_{l_j} = I$ for the experimental portion of this paper.
- (2) **Sampling Task–Language-specific Parameters:** Afterward, to generate task–language-specific neural parameters, we sample $\boldsymbol{\theta}_{ij}$ from $\mathcal{N}(f_\psi(\mathbf{t}_i, \mathbf{l}_j), \text{diag}(f_\phi(\mathbf{t}_i, \mathbf{l}_j))) \in \mathbb{R}^d$ where $f_\psi(\mathbf{t}_i, \mathbf{l}_j)$ and $f_\phi(\mathbf{t}_i, \mathbf{l}_j)$ are learned deep feed-forward neural networks $f_\psi : \mathbb{R}^h \rightarrow \mathbb{R}^d$ and $f_\phi : \mathbb{R}^h \rightarrow \mathbb{R}_{\geq 0}^d$ parametrized by ψ and ϕ , respectively, similar to Kingma and Welling (2014). These transform the latent representations into the mean $\boldsymbol{\mu}_{\theta_{ij}}$ and diagonal of the covariance matrix $\sigma_{\theta_{ij}}^2$ for the parameters $\boldsymbol{\theta}_{ij}$ associated with t_i and l_j . The feed-forward network f_ψ just has a final linear layer as the mean can range over \mathbb{R}^d whereas f_ϕ has a final softplus (defined in Section 3) layer to ensure it ranges only over $\mathbb{R}_{\geq 0}^d$. Following Stolee and Patterson (2019), the networks f_ψ and f_ϕ take as input a linear function of the task and language vectors: $\mathbf{t} \oplus \mathbf{l} \oplus (\mathbf{t} - \mathbf{l}) \oplus (\mathbf{t} \odot \mathbf{l})$, where \oplus stands for concatenation and \odot for element-wise multiplication. The sampled neural parameters $\boldsymbol{\theta}_{ij}$ are partitioned into a weight $W_{ij} \in \mathbb{R}^{e \times c}$ and a bias $\mathbf{b}_{ij} \in \mathbb{R}^c$, and reshaped appropriately. Hence, the dimensionality of the Gaussian is chosen to reflect the number of parameters in the affine layer, $d = e \cdot c + c$, where e is the dimensionality of the input token embeddings (detailed in the next paragraph) and c is the maximum number of classes across tasks.³ The number of hidden layers and the hidden size of f_ψ and f_ϕ are hyper-parameters discussed in Section 4.2. We tie the parameters

³Different tasks might involve different class numbers, the number of parameters hence oscillates. The extra dimensions not needed for a task can be considered as padded with zeros.

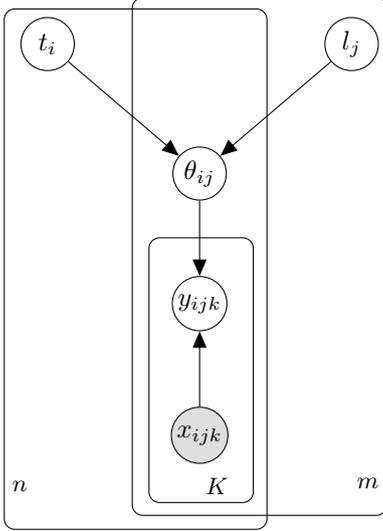


Figure 1: A graph (plate notation) of the generative model based on parameter space factorization. Shaded circles refer to observed variables.

ψ and ϕ for all layers except for the last to reduce the parameter count. We note that the space of parameters for all tasks and languages forms a tensor $\Theta \in \mathbb{R}^{n \times m \times d}$, where d is the number of parameters of the largest model.

- (3) **Sampling Task Labels:** Finally, we sample the k^{th} label y_{ijk} for the i^{th} task and the j^{th} language from a final softmax: $p(y_{ijk} | \mathbf{x}_{ijk}, \boldsymbol{\theta}_{ij}) = \text{softmax}(W_{ij} \text{BERT}(\mathbf{x}_{ijk}) + \mathbf{b}_{ij})$ where $\text{BERT}(\mathbf{x}_{ijk}) \in \mathbb{R}^e$ is the multi-lingual BERT (Pires et al., 2019) encoder. The incorporation of m-BERT as a pre-trained multilingual embedding allows for enhanced cross-lingual transfer.

Consider the Cartesian product of all tasks and languages $T \times L$. We can decompose this product into seen task–language pairs \mathcal{S} and unseen task–language pairs \mathcal{U} , i.e. $T \times L = \mathcal{S} \sqcup \mathcal{U}$. Naturally, we are only able to train our model on the seen task–language pairs \mathcal{S} . However, as we estimate all task–language parameter vectors $\boldsymbol{\theta}_{ij}$ jointly, our model allows us to draw inferences about the parameters for pairs in \mathcal{U} as well. The intuition for why this should work is as follows: By observing multiple pairs where the task (language) is the same but the language (task) varies, the model learns to distill the relevant knowledge for zero-shot learning because our generative model structurally enforces a disentangled representations—separating representations for the tasks from the representations for the languages rather than lumping them

Algorithm 1 Generative Model of Neural Parameters for Multi-task, Multi-lingual NLP.

```

1: for  $t_i \in T$ 
2:    $\mathbf{t}_i \sim \mathcal{N}(\boldsymbol{\mu}_{t_i}, \Sigma_{t_i})$ 
3: for  $l_j \in L$ 
4:    $\mathbf{l}_j \sim \mathcal{N}(\boldsymbol{\mu}_{l_j}, \Sigma_{l_j})$ 
5: for  $t_i \in T$ 
6:   for  $l_j \in L$ 
7:      $\boldsymbol{\mu}_{\theta_{ij}} = f_\psi(\mathbf{t}_i, \mathbf{l}_j)$ 
8:      $\Sigma_{\theta_{ij}} = f_\phi(\mathbf{t}_i, \mathbf{l}_j)$ 
9:      $\boldsymbol{\theta}_{ij} \sim \mathcal{N}(\boldsymbol{\mu}_{\theta_{ij}}, \Sigma_{\theta_{ij}})$ 
10:    for  $\mathbf{x}_{ijk} \in X_{ij}$ 
11:       $y_{ijk} \sim p(\cdot | \mathbf{x}_{ijk}, \boldsymbol{\theta}_{ij})$ 

```

together into a single entangled representation (Wu and Dredze, 2019, *inter alia*). Furthermore, the neural networks f_ψ and f_ϕ mapping the task- and language-specific latent variables to neural parameters are shared allowing the model to generalize across task–language pairs.

3 Variational Inference

Exact computation of the posterior over the latent variables $p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l} | \mathbf{x})$ is intractable. Thus, we need to resort to an approximation. In this work, we consider variational inference as our approximate inference scheme. Variational inference finds an approximate posterior over the latent variables by minimizing the variational gap, which may be expressed as the Kullback–Leibler (KL) divergence between the variational approximation $q(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l})$ and the true posterior $p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l} | \mathbf{x})$. In our work, we employ the following variational distributions:

$$q_\lambda = \mathcal{N}(\mathbf{m}_t, S_t) \quad \mathbf{m}_t \in \mathbb{R}^h, S_t \in \mathbb{R}^{h \times h} \quad (1)$$

$$q_\nu = \mathcal{N}(\mathbf{m}_l, S_l) \quad \mathbf{m}_l \in \mathbb{R}^h, S_l \in \mathbb{R}^{h \times h} \quad (2)$$

$$q_\xi = \mathcal{N}(f_\psi(\mathbf{t}, \mathbf{l}), \text{diag}(f_\phi(\mathbf{t}, \mathbf{l}))) \quad (3)$$

We note the unusual choice to tie parameters between the generative model and the variational family in Equation (3); however, we found that this choice performs better in our experiments.

Through a standard algebraic manipulation in Equation (4), the KL-divergence for our generative model can be shown to equal the marginal log-likelihood $\log p(\mathbf{x})$, independent from $q(\cdot)$, and the so-called evidence lower bound (ELBO) \mathcal{L} . Thus, approximate inference becomes an optimization problem where maximizing \mathcal{L} results in minimizing the KL-divergence. One derives \mathcal{L} is by expand-

$$\begin{aligned}
\text{KL}(q(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l}) \parallel p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l} \mid \mathbf{x})) &= - \mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \log \frac{p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l} \mid \mathbf{x})}{q(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l})} \\
&= - \mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} [\log p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l}, \mathbf{x}) - \log p(\mathbf{x}) - \log q(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l})] \\
&= \log p(\mathbf{x}) - \mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \log \frac{p(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l}, \mathbf{x})}{q(\boldsymbol{\theta}, \mathbf{t}, \mathbf{l})} \triangleq \log p(\mathbf{x}) - \mathcal{L} \quad (4)
\end{aligned}$$

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \left(\iiint p(\mathbf{x}, \boldsymbol{\theta}, \mathbf{t}, \mathbf{l}) \, d\boldsymbol{\theta} \, d\mathbf{t} \, d\mathbf{l} \right) \\
&= \log \left(\iiint p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l}) p(\mathbf{t}) p(\mathbf{l}) \, d\boldsymbol{\theta} \, d\mathbf{t} \, d\mathbf{l} \right) \\
&= \log \left(\iiint \frac{q_\lambda(\mathbf{t}) q_\nu(\mathbf{l}) q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})}{q_\lambda(\mathbf{t}) q_\nu(\mathbf{l}) q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})} p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l}) p(\mathbf{t}) p(\mathbf{l}) \, d\boldsymbol{\theta} \, d\mathbf{t} \, d\mathbf{l} \right) \\
&= \log \left(\mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \frac{p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l}) p(\mathbf{t}) p(\mathbf{l}) p(\mathbf{x} \mid \boldsymbol{\theta})}{q_\lambda(\mathbf{t}) q_\nu(\mathbf{l}) q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})} \right) \\
&\geq \mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \left[\log \frac{p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l}) p(\mathbf{t}) p(\mathbf{l})}{q_\lambda(\mathbf{t}) q_\nu(\mathbf{l}) q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})} \right] \triangleq \mathcal{L} \quad (5) \\
&= \mathbb{E}_{\mathbf{t} \sim q_\lambda} \mathbb{E}_{\mathbf{l} \sim q_\nu} \left[\mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \left[\log p(\mathbf{x} \mid \boldsymbol{\theta}) + \log \frac{p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})}{q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})} \right] + \log \frac{p(\mathbf{t})}{q_\lambda(\mathbf{t})} + \log \frac{p(\mathbf{l})}{q_\nu(\mathbf{l})} \right] \\
&= \underbrace{\mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \log p(\mathbf{x} \mid \boldsymbol{\theta})}_{\text{requires approximation}} - \underbrace{\left(\text{KL}(q_\lambda(\mathbf{t}) \parallel p(\mathbf{t})) + \text{KL}(q_\nu(\mathbf{l}) \parallel p(\mathbf{l})) + \text{KL}(q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l}) \parallel p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})) \right)}_{\text{closed-form solution}} \quad (6)
\end{aligned}$$

ing the marginal log-likelihood as in Equation (5) by means of Jensen’s inequality. We also show that \mathcal{L} can be further broken into a series of terms as illustrated in Equation (6). In particular, we see that it is only the first term in the expansion that requires approximation. The subsequent terms are KL-divergences between variational and true distributions that have closed-form solution due to our choice of prior. Due to the parameter-tying scheme above, the KL-divergence in Equation (6) between the variational distribution $q_\xi(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})$ and the prior distribution $p(\boldsymbol{\theta} \mid \mathbf{t}, \mathbf{l})$ is zero.

In general, the covariance matrices S_t and S_l in Equation (1) and Equation (2) will require $\mathcal{O}(h^2)$ space to store. As h is often very large, it is impractical to materialize either matrix in its entirety. Thus, in this work, we experiment with smaller matrices that have a reduced memory footprint; specifically, we consider a *diagonal* covariance matrix and a *diagonal plus low-rank* covariance structure.

A diagonal covariance matrix makes computation feasible with a complexity of $\mathcal{O}(h)$; this, however, comes at the cost of not letting parameters influence each other, and thus failing to capture their complex interactions. To allow for a more expressive variational family, we also consider a covariance matrix that is the sum of a diagonal matrix and a low-rank matrix:

$$S_t = \text{diag}(\boldsymbol{\delta}_t^2) + B_t B_t^\top \quad (7)$$

$$S_l = \text{diag}(\boldsymbol{\delta}_l^2) + B_l B_l^\top \quad (8)$$

where $B \in \mathbb{R}^{h \times k}$ ensures that $\text{rank}(BB^\top) \leq k$, and $\text{diag}(\boldsymbol{\delta})$ is diagonal. We can store this structured covariance matrix in $\mathcal{O}(kh)$ space.

By definition, covariance matrices must be symmetric and positive semi-definite. The first property holds by construction. The second property is enforced by a softplus parameterization where $\text{softplus}(\cdot) \triangleq \ln(1 + \exp(\cdot))$. Specifically, we define $\boldsymbol{\delta}^2 = \text{softplus}(\boldsymbol{\rho})$ and we optimize over $\boldsymbol{\rho}$.

3.1 Stochastic Variational Inference

To speed up the training time, we make use of *stochastic* variational inference (Hoffman et al., 2013). In this setting, we randomly sample a task $t_i \in T$ and language $l_j \in L$ among seen combinations during each training step,⁴ and randomly select a batch of examples from the dataset for the sampled task–language pair. We then optimize the parameters of the feed-forward neural networks ψ and ϕ as well as the parameters of the variational approximation to the posterior $\mathbf{m}_t, \mathbf{m}_l, \boldsymbol{\rho}_t, \boldsymbol{\rho}_l, B_t$ and B_l with a stochastic gradient-based optimizer (discussed in Section 4.2).

The KL divergence terms and their gradients in the ELBO appearing in Equation (6) can be computed in closed form as the relevant densities are Gaussian (Duchi, 2007, p. 13). Moreover, they can be calculated for Gaussians with diagonal and diagonal plus low-rank covariance structures without explicitly unfolding the full matrix. For a choice of prior $p = \mathcal{N}(\mathbf{0}, I)$ and a diagonal plus low-rank covariance structure, we have:

$$\text{KL}(q \parallel p) = \frac{1}{2} \left[\sum_{i=1}^h (m_i^2 + \delta_i^2 + \sum_{j=1}^k b_{ij}^2) - h - \ln \det(S) \right] \quad (9)$$

where b_{ij} is the element in the i -th row and j -th column of B . The last term can be estimated without computing the full matrix explicitly thanks to the generalization of the matrix–determinant lemma,⁵ which, applied to the factored covariance structure, yields:

$$\ln \det(S) = \ln \left[\det(I + B^\top \text{diag}(\boldsymbol{\delta}^{-2})B) \right] + \sum_{i=1}^h \ln(\delta_i^2) \quad (10)$$

where $I \in \mathbb{R}^k$. The KL divergence for the variant with diagonal covariance is just a special case of Equation (9) with $b_{ij} = 0$.

However, as stated before, the following expectation does not admit a closed-form solution. Thus

⁴As an alternative, we experimented with a setup where sampling probabilities are proportional to the number of examples of each task–language combination, but this achieved similar performances on the development sets.

⁵ $\det(A + UV^\top) = \det(I + V^\top A^{-1}U) \cdot \det(A)$. Note that the lemma assumes that A is invertible.

we consider a Monte Carlo approximation:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta} \sim q_\xi} \log p(\mathbf{x} \mid \boldsymbol{\theta}) &= \int q_\xi(\boldsymbol{\theta}) \log p(\mathbf{x} \mid \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx \frac{1}{V} \sum_{v=1}^V \log p(\mathbf{x} \mid \boldsymbol{\theta}^{(v)}) \quad \text{where } \boldsymbol{\theta}^{(v)} \sim q_\xi \end{aligned} \quad (11)$$

where V is the number of Monte Carlo samples taken. In order to allow the gradient to easily flow through the generated samples, we adopt the re-parametrization trick (Kingma and Welling, 2014). Specifically, we exploit the following identities $t_i = \boldsymbol{\mu}_{t_i} + \boldsymbol{\sigma}_{t_i} \odot \boldsymbol{\epsilon}$ and $l_j = \boldsymbol{\mu}_{l_j} + \boldsymbol{\sigma}_{l_j} \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$ and \odot is the Hadamard product. For the diagonal plus low-rank covariance structure, we exploit the identity:

$$\boldsymbol{\mu} + \text{diag}(\boldsymbol{\delta}^2 \odot \boldsymbol{\epsilon}) + B\boldsymbol{\zeta} \quad (12)$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^h$, $\boldsymbol{\zeta} \in \mathbb{R}^k$, and both are sampled from $\mathcal{N}(\mathbf{0}, I)$. The mean $\boldsymbol{\mu}_{\theta_{ij}}$ and the diagonal of the covariance matrix $\boldsymbol{\sigma}_{\theta_{ij}}^2$ are deterministically computed given the above samples and the parameters θ_{ij} are sampled from $\mathcal{N}(\boldsymbol{\mu}_{\theta_{ij}}, \text{diag}(\boldsymbol{\sigma}_{\theta_{ij}}^2))$, again with the re-parametrization trick.

3.2 Posterior Predictive Distribution

During test time, we perform zero-shot predictions on an unseen task–language pair by plugging in the posterior means (under the variational approximation) into the model. As an alternative, we experimented with ensemble predictions through Bayesian model averaging. I.e., for data for seen combinations \mathbf{x}_S and data for unseen combinations \mathbf{x}_U , the true predictive posterior can be approximated as $p(\mathbf{x}_U \mid \mathbf{x}_S) = \int p(\mathbf{x}_U \mid \boldsymbol{\theta}, \mathbf{x}_S) q_\xi(\boldsymbol{\theta} \mid \mathbf{x}_S) d\boldsymbol{\theta} \approx \sum_{v=1}^V p(\mathbf{x}_U \mid \boldsymbol{\theta}^{(v)}, \mathbf{x}_S)$, where V are 100 Monte Carlo samples from the posterior q_ξ . Performances on the development sets are comparable to simply plugging in the posterior mean.

4 Experimental Setup

4.1 Data

We select NER and POS tagging as our experimental tasks because their datasets encompass an ample and diverse sample of languages, and are common benchmarks for resource-poor NLP (Cotterell and Duh, 2017, *inter alia*). In particular, we opt for WikiANN (Pan et al., 2017) for the NER task and Universal Dependencies 2.4 (UD;

Nivre et al., 2019) for POS tagging. Our sample of languages is chosen from the intersection of those available in WikiANN and UD. However, we remark that this sample is heavily biased towards the Indo-European family (Gerz et al., 2018). Instead, the selection should be: i) typologically diverse, to ensure that the evaluation scores truly reflect the expected cross-lingual performance (Ponti et al., 2020); ii) a mixture of resource-rich and low-resource languages, to recreate a realistic setting and to allow for studying the effect of data size. Hence, we further filter the languages in order to make the sample more balanced. In particular, we sub-sample Indo-European languages by including only resource-poor ones, and keep all the languages from other families. Our final sample comprises 33 languages from 4 continents (17 from Asia, 11 from Europe, 4 from Africa, and 1 from South America) and from 11 families (6 Uralic, 6 Indo-European, 5 Afroasiatic, 3 Niger-Congo, 3 Turkic, 2 Austronesian, 2 Dravidian, 1 Austroasiatic, 1 Kra-Dai, 1 Tupian, 1 Sino-Tibetan), as well as 2 isolates. The full list of language ISO 639-2 codes is reported in Figure 2.

In order to simulate a zero-shot setting, we hold out in turn half of all possible task–language pairs and regard them as unseen, while treating the others as seen pairs. The partition is performed in such a way that a held-out pair has data available for the same task in a different language, and for the same language in a different task.⁶ Under this constraint, pairs are assigned to train or evaluation at random.⁷

We randomly split the WikiANN datasets into training, development, and test portions with a proportion of 80-10-10. We use the provided splits for UD; if the training set for a language is missing, we treat the test set as such when the language is held out, and as a training set when it is among the seen pairs.⁸

4.2 Hyper-parameters

The multilingual M-BERT encoder is initialized with parameters pre-trained on masked language modeling and next sentence prediction on 104 lan-

⁶We use the controlled partitioning for the following reason. If a language lacks data both for NER and for POS, the proposed factorization method cannot provide estimates for its posterior. We leave model extensions that can handle such cases for future work.

⁷See Section 5.2 for further experiments on splits controlled for language distance and sample size.

⁸Note that, in the second case, no evaluation takes place on such language.

guages (Devlin et al., 2019).⁹ We opt for the cased BERT-BASE architecture, which consists of 12 layers with 12 attention heads and a hidden size of 768. As a consequence, this is also the dimension e of each encoded WordPiece unit, a subword unit obtained through BPE (Wu et al., 2016). The dimension h of the multivariate Gaussian for task and language latent variables is set to 100. The deep feed-forward networks f_ψ and f_ϕ have 6 layers with a hidden size of 400 for the first layer, 768 for the internal layers, and ReLU non-linear activations. Their depth and width were selected based on validation performance.

The expectations over latent variables in Equation (6) are approximated through 3 Monte Carlo samples per batch during training. The KL terms are weighted with $\frac{1}{|K|}$ uniformly across training, where $|K|$ is the number of mini-batches.¹⁰ We initialize all the means m of the variational approximation with a random sample from $\mathcal{N}(0, 0.1)$, and the parameters for covariance matrices S of the variational approximation with a random sample from $\mathcal{U}(0, 0.5)$ following Stolee and Patterson (2019). We choose $k = 10$ as the number of columns of B so it fits into memory. The maximum sequence length for inputs is limited to 250. The batch size is set to 8, and the best setting for the Adam optimizer (Kingma and Ba, 2015) was found to be an initial learning rate of $5 \cdot 10^{-6}$ based on grid search. In order to avoid over-fitting, we perform early stopping with a patience of 10 and a validation frequency of 2.5K steps.

4.3 Baselines

We consider four baselines for cross-lingual transfer that also use BERT as an encoder shared across all languages.

First Baseline. A common approach is transfer from the **nearest source** (NS) language, which selects the most compatible source to a target language in terms of similarity. In particular, the selection can be based on family membership (Zeman and Resnik, 2008; Cotterell and Heigold, 2017; Kann et al., 2017), typological features (Deri and Knight, 2016), KL-divergence between part-of-speech trigram distributions (Rosa and Žabokrtský, 2015; Agić, 2017), tree edit distance of delexical-

⁹Available at github.com/google-research/bert/blob/master/multilingual.md

¹⁰We found this weighting strategy to work better than annealing as proposed by Blundell et al. (2015).

ized dependency parses (Ponti et al., 2018), or a combination of the above (Lin et al., 2019). In our work, during evaluation, we choose the classifier associated with the observed language with the highest cosine similarity between its typological features and those of the held-out language. These features are sourced from URIEL (Littell et al., 2017) and contain information about family, area, syntax, and phonology.

Second Baseline. We also consider transfer from the **largest source** (LS) language, i.e. the language with most training examples. This approach has been adopted by several recent works on cross-lingual transfer (Conneau et al., 2018; Artetxe et al., 2020, *inter alia*). In our implementation, we always select the English classifier for prediction.¹¹ In order to make this baseline comparable to our model, we adjust the number of English NER training examples to the sum of the examples available for all seen languages \mathcal{S} .¹²

Third Baseline. Next, we apply a protocol designed by Rahimi et al. (2019) for weighting the predictions of a classifier ensemble according to their reliability. For a specific task, the reliability of each language-specific classifier is estimated through a Bayesian graphical model. Intuitively, this model learns from error patterns, which behave more randomly for untrustworthy models and more consistently for the others. Among the protocols proposed in the paper, we opt for **BEA** in its zero-shot, token-based version, as it achieves the highest scores in a setting comparable to the current experiment. We refer to the original paper for the details.¹³

Fourth Baseline. Finally, we take inspiration from Wu and Dredze (2019). The **joint multi-lingual** (JM) baseline, contrary to the previous baselines, consists of two classifiers (one for POS tagging and another for NER) shared among all observed languages for a specific task. We follow the original implementation of Wu and Dredze (2019) closely adopting all recommended hyperparameters and strategies, such as freezing the parameters of all encoder layers below the 3rd for

sequence labeling tasks.

It must be noted that the number of parameters in our generative model scales better than baselines with language-specific classifiers, but worse than those with language-agnostic classifiers, as the number of languages grows. However, even in the second case, increasing the depth of baselines networks to match the parameter count is detrimental if the BERT encoder is kept trainable, which was also verified in previous work (Peters et al., 2019).

5 Results and Discussion

5.1 Zero-shot Transfer

Firstly, we present the results for zero-shot prediction based on our generative model using both of the approximate inference schemes (with diagonal covariance **PF-d** and factor covariance **PF-lr**). Table 1 summarizes the results on the two tasks of POS tagging and NER averaged across all languages. Our model (in both its variants) outperforms the four baselines on both tasks, including state-of-the-art alternative methods. In particular, PF-d and PF-lr gain 4.49 / 4.20 in accuracy (~7%) for POS tagging and 7.29 / 7.73 in F1 score (~10%) for NER on average compared to transfer from the largest source (**LS**), the strongest baseline for single-source transfer. Compared to multilingual joint transfer from multiple sources (**JM**), our two variants gain 0.95 / 0.67 in accuracy (~1%) for POS tagging and +0.61 / +1.05 in F1 score (~1%).

More details about the individual results on each task–language pair are provided in Figure 2, which includes the mean of the results over 3 separate runs. Overall, we obtain improvements in 23/33 languages for NER and on 27/45 treebanks for POS tagging, which further supports the benefits of transferring both from tasks and languages.

Considering the baselines, the relative performance of LS versus NS is an interesting finding per se. LS largely outperforms NS on both POS tagging and NER. This shows that having more data is more informative than relying primarily on similarity according to linguistic properties. This finding contradicts the received wisdom (Rosa and Žabokrtský, 2015; Cotterell and Heigold, 2017; Lin et al., 2019, *inter alia*) that related languages tend to be the most reliable source. We conjecture that this is due to the pre-trained multi-lingual BERT encoder, which helps to bridge the gap between unrelated languages (Wu and Dredze, 2019).

The two baselines that hinge upon transfer from

¹¹We include English to make the baseline more competitive, but note that this language is not available for our generative model as it is both Indo-European and resource-rich.

¹²The number of NER training examples is 1,093,184 for the first partition and 520,616 for the second partition.

¹³We implemented this model through the original code at github.com/afshinrahimi/mmer.

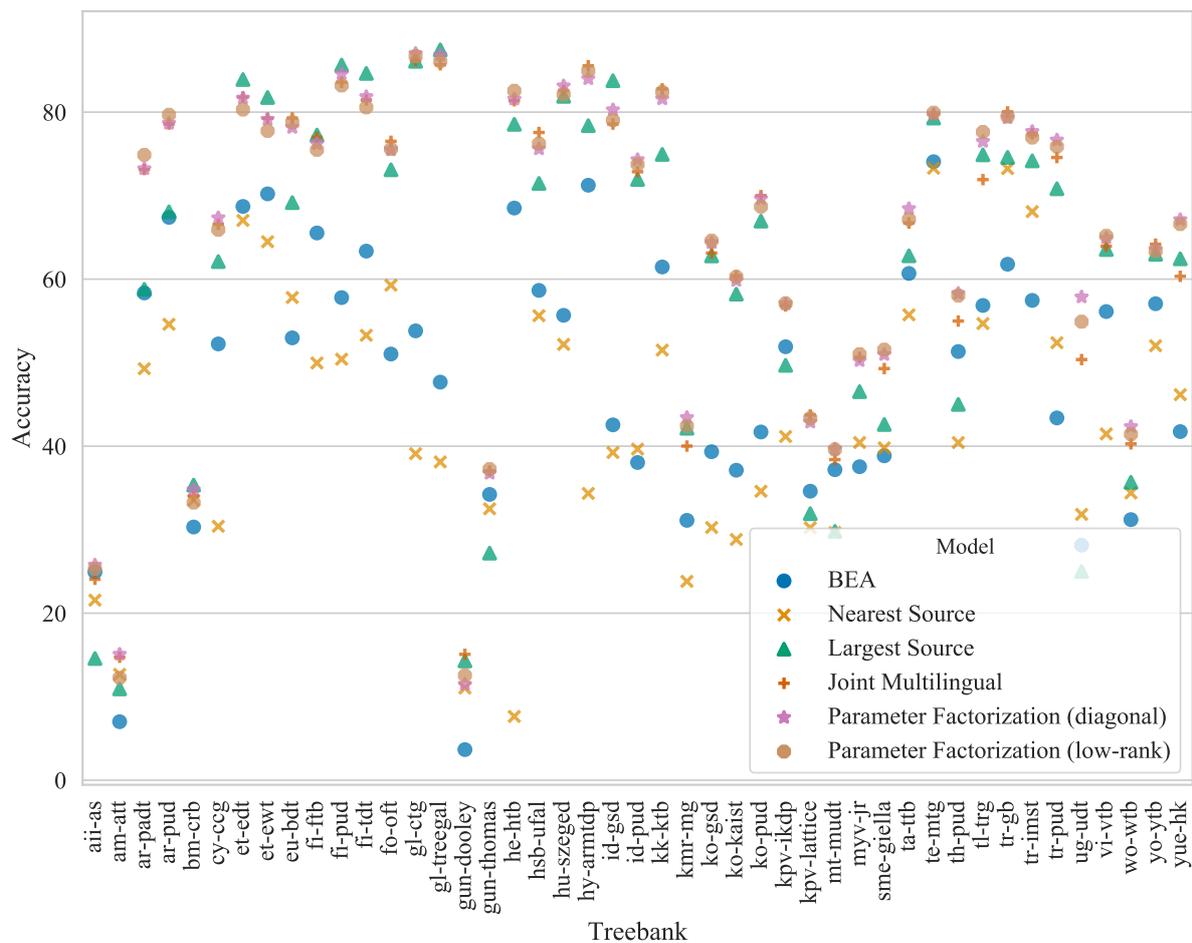
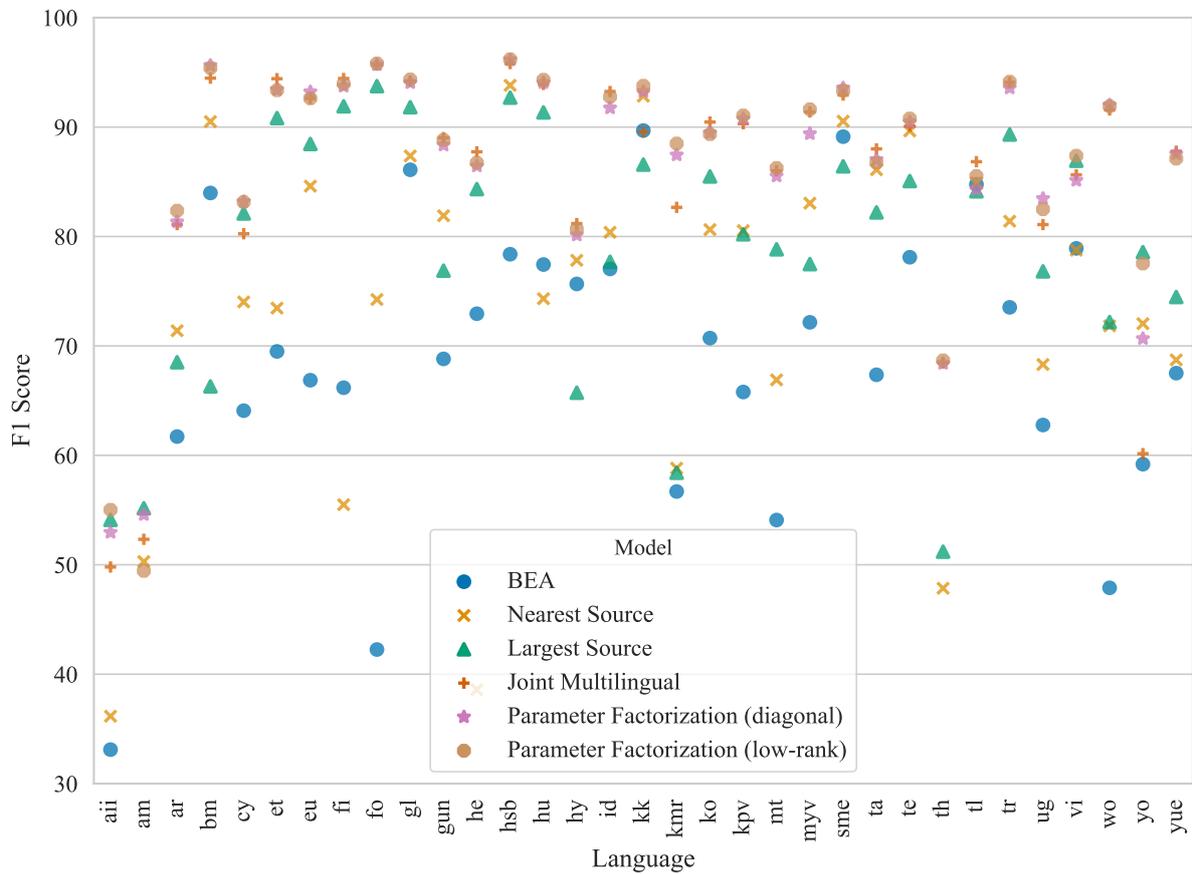


Figure 2: Results for NER (top) and POS tagging (bottom): four baselines for cross-lingual transfer compared to Matrix Factorization with diagonal covariance and diagonal plus low-rank covariance.

Task	BEA	NS	LS	JM	PF-d	PF-lr
POS	47.65 ± 1.54	42.84 ± 1.23	60.51 ± 0.43	64.04 ± 0.18	65.00 ± 0.12	64.71 ± 0.18
NER	66.45 ± 0.56	74.16 ± 0.56	78.97 ± 0.56	85.65 ± 0.13	86.26 ± 0.17	86.70 ± 0.10

Table 1: Results per task averaged across all languages.

Task	$ L = 11$		$ L = 22$	
	Sim	Dif	Sim	Dif
POS	72.44	53.25	66.59	63.22
NER	89.51	81.73	86.78	85.12

Table 2: Average performance when relying on $|L|$ similar (*Sim*) versus different (*Dif*) languages in the train and evaluation sets.

multiple sources lie on opposite sides of the spectrum in terms of performance. On the one hand, BEA achieves the lowest average score for NER, and surpasses only NS for POS tagging. We speculate that this is due to the following: i) adapting the protocol from Rahimi et al. (2019) to our model implies assigning a separate classifier head to each task–language pair, each of which is exposed to fewer examples compared to a shared one. This fragmentation fails to take advantage of the massively multilingual nature of the encoder; ii) our language sample is more typologically diverse, which means that most source languages are unreliable predictors. On the other hand, JM yields extremely competitive scores. Similarly to our model, it integrates knowledge from multiple languages and tasks. The extra boost in our model stems from its ability to disentangle each aspect of such knowledge and recombine it appropriately.

Moreover, comparing the two approximate inference schemes from Section 3.1, PF-lr obtains a small but statistically significant improvement over PF-d in NER, whereas they achieve the same performance on POS tagging. This means that the posterior is modeled well enough by a Gaussian where covariance among co-variables is negligible.

We see that even for the best model (PF-lr) there is a wide variation in the scores for the same task across languages. POS tagging accuracy ranges from 12.56 ± 4.07 in Guaraní to 86.71 ± 0.67 in Galician, and NER F1 scores range from 49.44 ± 0.69 in Amharic to 96.20 ± 0.11 in Upper Sorbian. Part of this variation is explained by the fact that the multilingual BERT encoder is not pre-trained in a subset of these languages (e.g., Amharic, Guaraní, Uyghur). Another cause is more straightforward: the scores are expected to be lower in languages

for which we have fewer training examples in the seen task–language pairs.

5.2 Language Distance and Sample Size

While we designed the language sample to be both realistic and representative of the cross-lingual variation, there are several factors inherent to a sample that can affect the zero-shot transfer performance: i) *language distance*, the similarity between seen and held-out languages; and ii) *sample size*, the number of seen languages. In order to disentangle these factors, we construct subsets of size $|L|$ so that training and evaluation languages are either maximally similar (*Sim*) or maximally different (*Dif*). As a proxy measure, we consider as ‘similar’ languages belonging to the same family. In Table 2, we report the performance of parameter factorization with diagonal plus low-rank covariance (PF-lr), the best model from Section 5.1, for each of these subsets.

Based on Table 2, there emerges a trade-off between language distance and sample size. In particular, performance is higher in *Sim* subsets compared to *Dif* subsets for both tasks (POS and NER) and for both sample sizes $|L| \in \{11, 22\}$. In larger sample sizes, the average performance increases for *Dif* but decreases for *Sim*. Intuitively, languages with labeled data for several relatives benefit from small, homogeneous subsets. Introducing further languages introduces noise. Instead, languages where this is not possible (such as isolates) benefit from an increase in sample size.

5.3 Entropy of the Predictive Distribution

A notable problem of point estimate methods is their tendency to assign most of the probability mass to a single class even in scenarios with high uncertainty. Zero-shot transfer is one of such scenarios, because it involves drastic distribution shifts in the data (Rabanser et al., 2019). A key advantage of Bayesian inference, instead, is marginalization over parameters, which yields smoother posterior predictive distributions (Kendall and Gal, 2017; Wilson, 2019).

We run an analysis of predictions based on (approximate) Bayesian model averaging. First, we

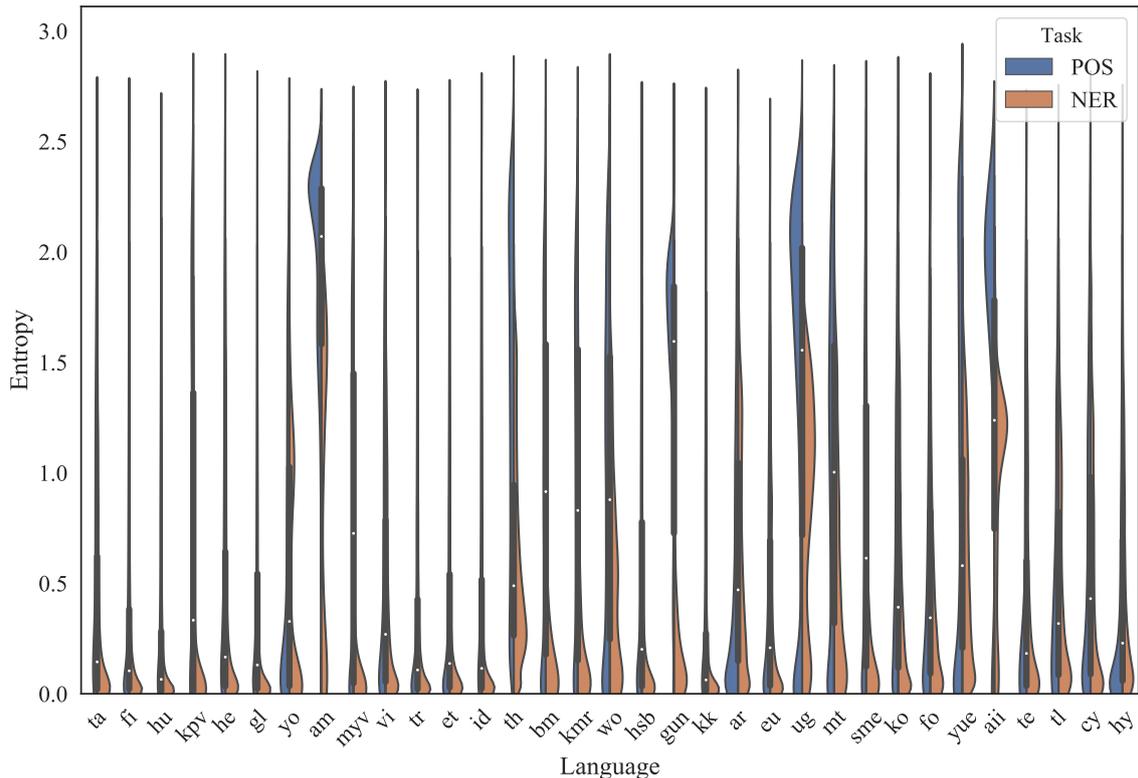


Figure 3: Entropy of the posterior predictive distributions over classes for each test example. The higher the entropy, the more uncertain the prediction.

randomly sample 800 examples from each test set of a task–language pair. For each example, we predict a distribution over classes Y through model averaging based on 10 samples from the posteriors. We then measure the prediction entropy of each example, i.e. $H(p) = -\sum_y^{|Y|} p(Y = y) \ln p(Y = y)$, whose plot is shown in Figure 3.

Entropy is a measure of uncertainty. Intuitively, the uniform categorical distribution (maximum uncertainty) has the highest entropy, whereas if the whole probability mass falls into a single class (maximum confidence), then the entropy $H = 0$.¹⁴ As it emerges from Figure 3, predictions in certain languages tend to have higher entropy on average, such as in Amharic, Guaraní, Uyghur, or Assyrian Neo-Aramaic. This aligns well with the performance metrics in Figure 2. In practice, languages with low scores tend to display high entropy in the predictive distribution, as expected. To verify this claim, we measure the Pearson’s correlation between entropies of each task–language pair in Figure 3 and performance metrics. We find

¹⁴The maximum entropy is ≈ 2.2 for 9 classes as in NER and ≈ 2.83 for 17 classes as in POS tagging.

a very strong negative correlation with a coefficient of $\rho = -0.914$ and a two-tailed p-value of 1.018×10^{-26} .

6 Related Work

Our approach builds on ideas from several different fields: cross-lingual transfer in NLP, with a particular focus on sequence labeling tasks, as well as matrix factorization, contextual parameter generation, and neural Bayesian methods.

Cross-Lingual Transfer for Sequence Labeling.

One of the two dominant approaches for cross-lingual transfer is *projecting annotations* from a source language text to a target language text. This technique was pioneered by Yarowsky et al. (2001) and Hwa et al. (2005) for parsing, and later extended to applications such as POS tagging (Das and Petrov, 2011; Garrette et al., 2013; Täckström et al., 2012; Duong et al., 2014; Huck et al., 2019) and NER (Ni et al., 2017; Enghoff et al., 2018; Aggeri et al., 2018; Jain et al., 2019). This requires tokens to be aligned through a parallel corpus, a machine translation system, or a bilingual dictio-

nary (Durrett et al., 2012; Mayhew et al., 2017). However, creating machine translation and word-alignment systems demands parallel texts in the first place, while automatically induced bilingual lexicons are noisy and offer only limited coverage (Artetxe et al., 2018; Duan et al., 2020). Furthermore, errors inherent to such systems cascade along the projection pipeline (Agić et al., 2015).

The second approach, *model transfer*, offers higher flexibility (Conneau et al., 2018). The main idea is to train a model directly on the source data, and then deploy it onto target data (Zeman and Resnik, 2008). Crucially, bridging between different lexica requires input features to be language-agnostic. While originally this implied delexicalization, replacing words with universal POS tags (McDonald et al., 2011; Dehouck and Denis, 2017), cross-lingual Brown clusters (Täckström et al., 2012; Rasooli and Collins, 2017), or cross-lingual knowledge base grounding through wikification (Camacho-Collados et al., 2016; Tsai et al., 2016), more recently these have been supplanted by cross-lingual word embeddings (Ammar et al., 2016; Zhang et al., 2016; Xie et al., 2018; Ruder et al., 2019b) and multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020).

An orthogonal research thread regards the *selection of the source language(s)*. In particular, multi-source transfer was shown to surpass single-best source transfer in NER (Fang and Cohn, 2017; Rahimi et al., 2019) and POS tagging (Enghoff et al., 2018; Plank and Agić, 2018). Our parameter space factorization model can be conceived as an extension of multi-source cross-lingual model transfer to a cross-task setting.

Data Matrix Factorization. Although we are the first to propose a factorization of the *parameter* space for unseen combinations of tasks and languages, the factorization of *data* for collaborative filtering and social recommendation is an established research area. In particular, the missing values in sparse data structures such as user-movie review matrices can be filled via probabilistic matrix factorization (PMF) through a linear combination of user and movie matrices (Mnih and Salakhutdinov, 2008; Ma et al., 2008; Shan and Banerjee, 2010, *inter alia*) or through neural networks (Dziugaite and Roy, 2015). Inference for PMF can be carried out through MAP inference (Dziugaite and Roy, 2015), Markov chain Monte Carlo (MCMC; Salakhutdinov and Mnih, 2008) or stochastic varia-

tional inference (Stolee and Patterson, 2019). Contrary to prior work, we perform factorization on latent variables (task- and language-specific parameters) rather than observed ones (data).

Contextual Parameter Generation. Our model is reminiscent of the idea that parameters can be conditioned on language representations, as proposed by Platanios et al. (2018). However, since this approach is limited to a single task and a joint learning setting, it is not suitable for generalization in a zero-shot transfer setting.

Bayesian Neural Models. So far, these models have found only limited application in NLP for resource-poor languages, despite their desirable properties. Firstly, they can incorporate priors over parameters to endow neural networks with the correct inductive biases towards language: Ponti et al. (2019b) constructed a prior imbued with universal linguistic knowledge for zero- and few-shot character-level language modeling. Secondly, they avoid the risk of over-fitting by taking into account uncertainty. For instance, Shareghi et al. (2019) and Doitch et al. (2019) use a perturbation model to sample high-quality and diverse solutions for structured prediction in cross-lingual parsing.

7 Conclusion

The main contribution of our work is a Bayesian generative model for multiple NLP tasks and languages. At its core lies the idea that the space of neural weights can be factorized into latent variables for each task and each language. While training data are available only for a meager subset of task–language combinations, our model opens up the possibility to perform prediction in novel, undocumented combinations at evaluation time. We performed inference through stochastic variational methods, and ran experiments on zero-shot named entity recognition (NER) and part-of-speech (POS) tagging in a typologically diverse set of 33 languages. Based on the reported results, we conclude that leveraging the information from tasks and languages simultaneously is superior to model transfer from English (relying on more abundant in-task data in the source language), from the most typologically similar language (relying on prior information on language relatedness), or from multiple source languages. Moreover, we found that the entropy of predictive posterior distributions obtained through Bayesian model averaging correlates al-

most perfectly with the error rate in the prediction. As a consequence, our approach holds promise to alleviating data paucity issues for a wide spectrum of languages and tasks, and to make knowledge transfer more robust to uncertainty.

Finally, we remark that our model is amenable to be extended to multilingual tasks beyond sequence labeling—such as natural language inference (Conneau et al., 2018) and question answering (Artetxe et al., 2020; Lewis et al., 2019; Clark et al., 2020)—and to zero-shot transfer across combinations of multiple modalities (e.g. speech, text, and vision) with tasks and languages. We leave these exciting research threads for future research.

Acknowledgements

We would like to thank action editor Jacob Eisenstein and the three anonymous reviewers at TAACL. This work is supported by the ERC Consolidator Grant LEXICAL (no 648909) and the Google Faculty Research Award 2018. RR was partially funded by ISF personal grants No. 1625/18.

References

- Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. 2018. [Developing new linguistic resources and tools for the Galician language](#). In *Proceedings of LREC*.
- Željko Agić. 2017. [Cross-lingual parser selection for low-resource languages](#). In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10.
- Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. [If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages](#). In *Proceedings of ACL*, pages 268–272.
- Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. [Multilingual projection for parsing truly low-resource languages](#). *Transactions of the ACL*, 4:301–312.
- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. [Many languages, one parser](#). *Transactions of the ACL*, 4:431–444.
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. [Unsupervised neural machine translation](#). In *Proceedings of ICLR*.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of ACL*.
- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the ACL*, 7:597–610.
- Johannes Bjerva and Isabelle Augenstein. 2018. [From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings](#). In *Proceedings of NAACL-HLT*, pages 907–916.
- Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. [What do language representations really represent?](#) *Computational Linguistics*, 45(2):381–389.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. [Weight uncertainty in neural networks](#). In *Proceedings of ICML*, pages 1613–1622.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. [Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities](#). *Artificial Intelligence*, 240:36–64.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of ACL*.
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI](#):

- Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.
- Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of IJNLP*, pages 91–96, Taipei, Taiwan.
- Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of EMNLP*, pages 748–759.
- Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pages 600–609.
- Mathieu Dehouck and Pascal Denis. 2017. Delexicalized word embeddings for cross-lingual dependency parsing. In *Proceedings of EACL*, pages 241–250.
- Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of ACL*, pages 399–408.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Amichay Doitch, Ram Yazdi, Tamir Hazan, and Roi Reichart. 2019. Perturbation based learning for structured nlp tasks with application to dependency parsing. *Transactions of the ACL*, 7:643–659.
- Xiangju Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of ACL*.
- John Duchi. 2007. *Derivations for linear algebra and optimization*. Technical report, University of California, Berkeley.
- 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 EMNLP*, pages 886–897.
- Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of EMNLP-CoNLL*, pages 1–11.
- Gintare Karolina Dziugaite and Daniel M. Roy. 2015. Neural network matrix factorization. *arXiv preprint arXiv:1511.06443*.
- Jan Vium Enghoff, Søren Harrison, and Željko Agić. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201.
- Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of ACL*, pages 587–593.
- Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of POS-tagger for low-resource languages. In *Proceedings of ACL*, pages 583–592.
- Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP*, pages 316–327.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. 2016. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena.
- Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.
- Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual annotation projection is effective for neural part-of-speech tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

- Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. [Entity projection via machine translation for cross-lingual NER](#). In *Proceedings of EMNLP-IJCNLP*, pages 1083–1092.
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. [One-shot neural cross-lingual transfer for paradigm completion](#). In *Proceedings of ACL*, pages 1993–2003.
- Alex Kendall and Yarin Gal. 2017. [What uncertainties do we need in Bayesian deep learning for computer vision?](#) In *Proceedings of NeurIPS*, pages 5574–5584.
- Diederik P. Kingma and Jimmy L. Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proceedings of ICLR*.
- Diederik P. Kingma and Max Welling. 2014. [Auto-encoding variational Bayes](#). In *Proceedings of ICLR*.
- Patrick S. H. Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. [MLQA: Evaluating cross-lingual extractive question answering](#). *CoRR*, abs/1910.07475.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of ACL*, pages 3125–3135.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of EACL*, pages 8–14.
- Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. 2008. [SoRec: Social recommendation using probabilistic matrix factorization](#). In *Proceedings of CIKM*, pages 931–940.
- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. [Learning language representations for typology prediction](#). In *Proceedings of EMNLP*, pages 2529–2535.
- Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. [Cheap translation for cross-lingual named entity recognition](#). In *Proceedings of EMNLP*, pages 2536–2545.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. [Multi-source transfer of delexicalized dependency parsers](#). In *Proceedings of EMNLP*, pages 62–72.
- Andriy Mnih and Ruslan Salakhutdinov. 2008. [Probabilistic matrix factorization](#). In *Proceedings of NeurIPS*, pages 1257–1264.
- Jian Ni, Georgiana Dinu, and Radu Florian. 2017. [Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection](#). In *Proceedings of ACL*, pages 1470–1480.
- Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, et al. 2019. [Universal Dependencies 2.4](#). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Robert Östling and Jörg Tiedemann. 2017. [Continuous multilinguality with language vectors](#). In *Proceedings of the EACL*, pages 644–649.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. [Cross-lingual name tagging and linking for 282 languages](#). In *Proceedings of ACL*, volume 1, pages 1946–1958.
- Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. [To tune or not to tune? Adapting pretrained representations to diverse tasks](#). In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of ACL*, pages 4996–5001.

- Barbara Plank and Željko Agić. 2018. [Distant supervision from disparate sources for low-resource part-of-speech tagging](#). In *Proceedings of EMNLP*, pages 614–620.
- Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. [Contextual parameter generation for universal neural machine translation](#). In *Proceedings of EMNLP*, pages 425–435.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal commonsense reasoning](#). In *Proceedings of EMNLP*.
- Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. [Modeling language variation and universals: A survey on typological linguistics for natural language processing](#). *Computational Linguistics*, 45(3):559–601.
- Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. [Isomorphic transfer of syntactic structures in cross-lingual NLP](#). In *Proceedings of ACL*, pages 1531–1542.
- Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. [Towards zero-shot language modeling](#). In *Proceedings of EMNLP*, pages 2900–2910.
- Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. [Failing loudly: An empirical study of methods for detecting dataset shift](#). In *Proceedings of NeurIPS*, pages 1394–1406.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. [Massively multilingual transfer for NER](#). In *Proceedings of ACL*, pages 151–164.
- Mohammad Sadegh Rasooli and Michael Collins. 2017. [Cross-lingual syntactic transfer with limited resources](#). *Transactions of the Association for Computational Linguistics*, 5:279–293.
- Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime G. Carbonell. 2019. [Zero-shot neural transfer for cross-lingual entity linking](#). In *Proceedings of AACL*, pages 6924–6931.
- Rudolf Rosa and Zdeněk Žabokrtský. 2015. [KL-cpos3 - a language similarity measure for delexicalized parser transfer](#). In *Proceedings of ACL*, pages 243–249.
- Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019a. [Transfer learning in natural language processing](#). In *Proceedings of NAACL-HLT: Tutorials*, pages 15–18.
- Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019b. [A survey of cross-lingual embedding models](#). *Journal of Artificial Intelligence Research*, 65:569–631.
- Ruslan Salakhutdinov and Andriy Mnih. 2008. [Bayesian probabilistic matrix factorization using Markov chain Monte Carlo](#). In *Proceedings of ICML*, pages 880–887.
- Hanhuai Shan and Arindam Banerjee. 2010. [Generalized probabilistic matrix factorizations for collaborative filtering](#). In *Proceedings of ICDM*, pages 1025–1030.
- Ehsan Shareghi, Yingzhen Li, Yi Zhu, Roi Reichart, and Anna Korhonen. 2019. [Bayesian learning for neural dependency parsing](#). In *Proceedings of NAACL-HLT*, pages 3509–3519.
- Benjamin Snyder and Regina Barzilay. 2010. [Climbing the tower of Babel: Unsupervised multilingual learning](#). In *Proceedings of ICML*, pages 29–36.
- Jake Stolee and Neill Patterson. 2019. [Matrix factorization with neural networks and stochastic variational inference](#). Technical report, University of Toronto.
- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. [Cross-lingual word clusters for direct transfer of linguistic structure](#). In *Proceedings of NAACL-HLT*, pages 477–487.
- Alon Talmor and Jonathan Berant. 2019. [MultiQA: An empirical investigation of generalization and transfer in reading comprehension](#). In *Proceedings of ACL*, pages 4911–4921.
- Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. [Cross-lingual named entity recognition via wikification](#). In *Proceedings of CoNLL*, pages 219–228.

- Andrew Gordon Wilson. 2019. [The case for Bayesian deep learning](#). *NYU Courant Technical Report*.
- Shijie Wu and Mark Dredze. 2019. [Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT](#). In *Proceedings of EMNLP-IJCNLP*, pages 833–844.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#). Technical report, Google.
- Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. [Neural cross-lingual named entity recognition with minimal resources](#). In *Proceedings of EMNLP*, pages 369–379.
- 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*, pages 1–8.
- Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. [Learning and evaluating general linguistic intelligence](#). *arXiv preprint arXiv:1901.11373v1*.
- Daniel Zeman and Philip Resnik. 2008. [Cross-language parser adaptation between related languages](#). In *Proceedings of IJCNLP*, pages 35–42.
- Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. [Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317, San Diego, California. Association for Computational Linguistics.
- Yftah Ziser and Roi Reichart. 2018. [Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance](#). In *Proceedings of EMNLP*, pages 238–249.

A KL-divergence of Gaussians

If both $p \triangleq \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $q \triangleq \mathcal{N}(\boldsymbol{m}, S)$ are multivariate Gaussians, their KL-divergence can be computed analytically as follows:

$$\text{KL}(q \parallel p) = \frac{1}{2} \left[\ln \frac{|S|}{|\Sigma|} - d + \text{tr}(S^{-1}\Sigma) + (\boldsymbol{m} - \boldsymbol{\mu})^\top S^{-1}(\boldsymbol{m} - \boldsymbol{\mu}) \right] \quad (13)$$

By substituting $\boldsymbol{m} = \mathbf{0}$ and $S = I$, it is trivial to obtain Equation (9).

B Visualization of the Learned Posteriors

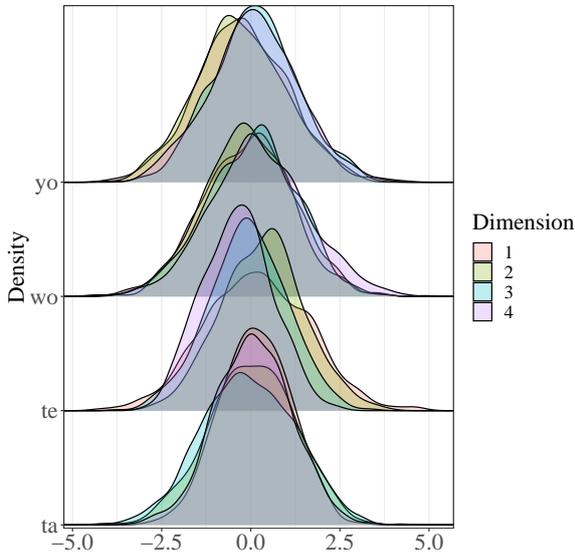


Figure 4: Samples from the posteriors of 4 languages, PCA-reduced to 4 dimensions.

The approximate posteriors of the latent variables can be visualized in order to study the learned representations for languages. Previous work (Johnson et al., 2017; Östling and Tiedemann, 2017; Malaviya et al., 2017; Bjerva and Augenstein, 2018) induced point estimates of language representations from artificial tokens concatenated to every input sentence, or from the aggregated values of the hidden state of a neural encoder. The information contained in such representations depends on the task (Bjerva and Augenstein, 2018), but mainly reflects the structural properties of each language (Bjerva et al., 2019).

In our work, due to the estimation procedure, languages are represented by full distributions rather than point estimates. By inspecting the learned representations, language similarities do not appear to

follow the structural properties of languages. This is most likely due to the fact that parameter factorization takes place *after* the multi-lingual BERT encoding, which blends the structural differences across languages. A fair comparison with previous works without such an encoder is left for future investigation.

As an example, consider two pairs of languages from two distinct families: Yoruba and Wolof are Niger-Congo from the Atlantic-Congo branch, Tamil and Telugu are Dravidian. We take 1,000 samples from the approximate posterior over the latent variables for each of these languages. In particular, we focus on the variational scheme with a low-rank covariance structure. We then reduce the dimensionality of each sample to 4 through PCA,¹⁵ and we plot the density along each resulting dimension in Figure 4. We observe that density areas of each dimension do not necessarily overlap between members of the same family. Hence, the learned representations depend on more than genealogy.

¹⁵Note that the dimensionality reduced samples are also Gaussian since PCA is a linear method.