A Probabilistic Generative Model of Linguistic Typology

Johannes Bjerva[ə] **Yova Kementchedjhieva**[ə] **Ryan Cotterell**[?,fi] **Isabelle Augenstein**[ə]

[ə]Department of Computer Science, University of Copenhagen
[?]Department of Computer Science, Johns Hopkins University
[fi]Department of Computer Science and Technology, University of Cambridge
`bjerva,yova,augenstein@di.ku.dk, rdc42@cam.ac.uk`

Abstract

In the principles-and-parameters framework, the structural features of languages depend on parameters that may be toggled on or off, with a single parameter often dictating the status of multiple features. The implied covariance between features inspires our probabilisation of this line of linguistic inquiry— we develop a generative model of language based on exponential-family matrix factorisation. By modelling all languages and features within the same architecture, we show how structural similarities between languages can be exploited to predict typological features with near-perfect accuracy, outperforming several baselines on the task of predicting held-out features. Furthermore, we show that language embeddings pre-trained on monolingual text allow for generalisation to unobserved languages. This finding has clear practical and also theoretical implications: the results confirm what linguists have hypothesised, i.e. that there are significant correlations between typological features and languages.

1 Introduction

Linguistic typologists dissect and analyse languages in terms of their structural properties (Croft, 2002). For instance, consider the phonological property of word-final obstruent decoding: German devoices word-final obstruents (*Zug* is pronounced /zuk/), whereas English does not (*dog* is pronounced /dɒg/). In the tradition of generative linguistics, one line of typological analysis is the principles-and-parameters framework (Chomsky, 1981), which posits the existence of a set of universal parameters, switches as it were, that languages toggle. One arrives at a kind of factorial typology, to borrow terminology from optimality theory (Prince and Smolensky, 2008), through different settings of the parameters. Within the principle-and-parameters research program, then, the goal is
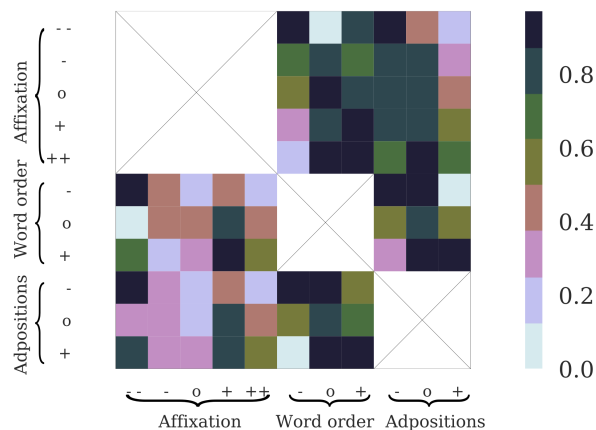

Figure 1: Correlations between selected typological parameters. Feature values are classified according to head-directionality (head-initial **+**, head-final **-**, no dominant order **o**). For instance, **++** under Affixation means strongly suffixing.

to identify the parameters that serve as axes, along which languages may vary.

It is not enough, however, to simply write down the set of parameters available to language. Indeed, one of the most interesting facets of typology is that different parameters are correlated. To illustrate this point, we show a heatmap in Figure 1 that shows the correlation between the values of selected parameters taken from a typological knowledge base (KB). Notice how head-final word order, for example, highly correlates with strong suffixation. The primary contribution of this work is a probabilisation of typology inspired by the principles-and-parameters framework. We assume a given set of typological parameters and develop a generative model of a language's parameters, casting the problem as a form of exponential-family matrix factorisation. We observe a binary matrix that encodes the settings of each parameter for each language. For example, the Manchu head-final entry of this matrix would be set to 1, because Manchu

is a head-final language. The goal of our model is to *explain* each entry of matrix as arising through the dot product of a language embedding and a parameter embedding passed through a sigmoid.

We test our model on The World Atlas of Language Structures (WALS), the largest available knowledge base of typological parameters at the lexical, phonological, syntactic and semantic level. Our contributions are: (i) We develop a probabilisation of typology inspired by the principles-and-parameters framework. (ii) We introduce the novel task of typological collaborative filtering, where we observe some of a language's parameters, but hold some out. At evaluation time, we predict the held-out parameters using the generative model. (iii) We develop a semi-supervised extension, in which we incorporate language embeddings output by a neural language model, thus improving performance with unlabelled data. Indeed, when we partially observe some of the typological parameters of a language, we achieve near-perfect (97%) accuracy on the prediction of held-out parameters. (iv) We perform an extensive qualitative and quantitative analysis of our method.

2   Typology in Generative Grammar

What we will present in this paper is a generative model that corresponds to a generative tradition of research in linguistic typology. We first outline the technical linguistic background necessary for the model's exposition. Chomsky famously argued that the human brain contains a prior, as it were, over possible linguistic structures, which he termed **universal grammar** (Chomsky, 1965). The connection between Chomsky's Universal Grammar and the Bayesian prior is an intuitive one, but the earliest citation we know for the connection is Eisner (2002, §2). As a theory, universal grammar holds great promise in explaining the typological variation of human language. Cross-linguistic similarities and differences may be explained by the influence universal grammar exerts over language acquisition and change. While universal grammar arose early on in the writtings of Chomsky, early work in generative grammar focused primarily on English (Harris, 1995). Indeed, Chomsky's *Syntactic Structures* contains exclusively examples in English (Chomsky, 1957). As the generative grammarians turned their focus to a wider selection of languages, the **principles and parameters** framework for syntactic analysis

rose to prominence. Given the tight relationship between the theory of universal grammar and typology, principles and parameters offers a fruitful manner in which to research typological variation.

The principles and parameters takes a parametric view of linguistic typology. The structure of human language is governed by a series of **principles**, which are hard constraints on human language. A common example of a principle is the requirement that every sentence has a subject, even if one that is not pronounced; see the discussion on the pro-drop parameter in Carnie (2013). Principles are universally true for all languages. On the other hand, languages are also governed by **parameters**. Unlike principles, parameters are the parts of linguistic structure that are allowed to vary. It is useful to think of parameters as attributes that can take on a variety of values. As Chomsky (2000) himself writes "we can think of the initial state of the faculty of language as a fixed network connected to a switch box; the network is constituted of the principles of language, while the switches are the options to be determined by experience. When the switches are set one way, we have Swahili; when they are set another way, we have Japanese. Each possible human language is identified as a particular setting of the switches-a setting of parameters, in technical terminology."

What are possible parameters? Here, in our formalisation of the parameter aspect of the principles-and-parameters framework, we take a catholic view of parameters, encompassing all areas of linguistics, rather than just syntax. For example, as we saw before, consider the switch of devoicing word-final obstruents a parameter. We note that while principle-and-parameters typology has primarily been applied to syntax, there are also interesting applications to non-syntactic domains. For instance, van Oostendorp (2015) applies a parametric approach to metrical stress in phonology; this is in line with our view. In the field of linguistic typology, there is a vibrant line of research, which fits into the tradition of viewing typological parameters through the lens of principles and parameters. Indeed, while earlier work due to Chomsky focused on what have come to be called macro-parameters, many linguists now focus on micro-parameters, which are very close to the features found in the WALS dataset that we will be modelling (Baker, 2008; Nicolis and Biberauer, 2008; Biberauer et al., 2009). This justifies our viewing

ℓ	SOV	$\cdots$	SVO	Str. Pref.	$\cdots$	Str. Suff.
eng	1	0	0	0	0	1
nld	0	1	0	0	0	1
deu	0	1	0	0	0	1
$\cdots$	$\cdots$			$\cdots$		
vie	0	0	1	0	1	0
tur	0	1	0	0	0	1
mrd	1	0	0	–	–	–

Figure 2: Example of training and evaluation portions of a feature matrix, with word order features (81A, SOV, $\cdots$, SVO) and affixation features (26A, Strongly Prefixing, $\cdots$, Strongly Suffixing). We train on the examples highlighted in blue, evaluate on the examples highlighted in red, and ignore those which are not covered in WALS highlighted in green.

WALS through the lens of principles and parameters, even though the authors of WALS adhere to the functional-typological school.[1]

Notationally, we will represent the parameters as a vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]$. Each typological parameter π_i is a binary variable; for instance, does the language admit word-final voiced obstruents?

3 A Generative Model of Typology

We now seek a probabilistic formalisation of the linguistic theory presented in §2; specifically, for every language ℓ, we seek to explain the observed binary vector of parameters $\boldsymbol{\pi}^\ell$: $\pi_i^\ell = 1$ indicates that the ith parameter is "on" in language ℓ. The heart of our model will be quite simple: every language ℓ will have a language embedding $\boldsymbol{\lambda}^\ell \in \mathbb{R}^d$ and every parameter will have a parameter embedding $\mathbf{e}_{\pi_i} \in \mathbb{R}^d$. Now, $\pi_i^\ell \sim sigmoid(\mathbf{e}_{\pi_i}^\top \boldsymbol{\lambda}^\ell)$.

This model also takes inspiration from work in relation extraction (Riedel et al., 2013). Writing the joint distribution over the entire binary vector of parameters, we arrive at

$$p(\boldsymbol{\pi}^\ell \mid \boldsymbol{\lambda}^\ell) = \prod_{i=1}^{|\boldsymbol{\pi}|} p(\pi_i^\ell \mid \boldsymbol{\lambda}^\ell) \qquad (1)$$

$$= \prod_{i=1}^{|\boldsymbol{\pi}|} sigmoid\left(\mathbf{e}_{\pi_i}^\top \boldsymbol{\lambda}^\ell\right) \qquad (2)$$

$$= \prod_{i=1}^{|\boldsymbol{\pi}|} \frac{1}{1 + \exp(-\mathbf{e}_{\pi_i}^\top \boldsymbol{\lambda}^\ell)} \qquad (3)$$

---

[1] For an overview of differences between these schools, we refer the reader to Haspelmath (2008).

We define the the prior over language embeddings:

$$p(\boldsymbol{\lambda}^\ell) = \mathcal{N}\left(\mathbf{0}; \sigma^2 I\right) \qquad (4)$$

where $\boldsymbol{\mu}$ is the mean of the Gaussian whose covariance is fixed at I. Now, give a collection of languages $\mathcal{L}$, we arrive at the joint distribution

$$\prod_{\ell \in \mathcal{L}} p(\boldsymbol{\pi}^\ell, \boldsymbol{\lambda}^\ell) = \prod_{\ell \in \mathcal{L}} p(\boldsymbol{\pi}^\ell \mid \boldsymbol{\lambda}^\ell) \cdot p(\boldsymbol{\lambda}^\ell) \qquad (5)$$

Note that $p(\boldsymbol{\lambda}^\ell)$ is, spiritually at least, a universal grammar: it is the prior over what sort of languages can exist, albeit encoded as a real vector. In the parlance of principles and parameters, the prior represents the principles.

Then our model parameters are $\boldsymbol{\Theta} = \{\mathbf{e}_{\pi_1}, \ldots, \mathbf{e}_{\pi_{|\boldsymbol{\pi}|}}, \boldsymbol{\lambda}^1, \ldots, \boldsymbol{\lambda}^{|\mathcal{L}|}\}$. Note that for the remainder of the paper, we will never shorten 'model parameters' to simply 'parameters' to avoid ambiguity. We will, however, refer to 'typological parameters' as simply 'parameters.'

We can view this model as a form of exponential-family matrix factorisation (Collins et al., 2001). Specifically, our model seeks to explain a binary matrix of parameters. We consider such matrices as the one in Figure 2, which depicts some of the binarised feature values for word order and affixation for English, Dutch, German, Vietnamese, Turkish, and Marind. We will have some parameters seen during training (highlighted in blue), some we use for evaluation (highlighted in red), and some which are unknown due to the nature of WALS (highlighted in green). Crucially, the model in eq. (5) allows us to learn the correlations between typological parameters, as illustrated in Figure 1. We train the model over 10 epochs with a batch size of 64, using the Adam optimiser (Kingma and Ba, 2014) and L_2 regularisation (0.1), which corresponds to the Gaussian prior with variance $\sigma^2 = 10$.

4 A Semi-Supervised Extension

A natural question we might ask is if our model can exploit unlabelled monolingual data to improve its performance. We explain how we can induce language embeddings from unlabelled data below and then incorporate these into our model through the prior eq. (4). This results in a semi-supervised model, as we incorporate an unsupervised pre-training step. This is motivated by the fact that related languages tend to exhibit correlations between each other. Figure 3 shows the distribution
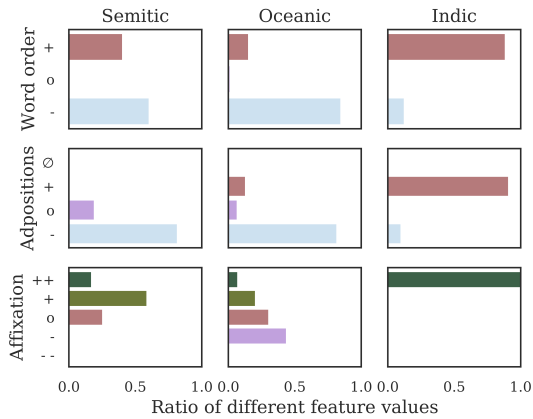
Figure 3: Distribution of feature values across three of the biggest language families in WALS. The representation of feature values is described in Figure 1.

Figure 4: Char LSTM LM with language embeddings concatenated to char embeddings at every time step.

of a few features within the Semitic, Oceanic, and Indic language branches. Notice, for instance, the skewed distribution of feature values within the Indic branch: languages in that branch are almost exclusively head-initial with respect to word order, order of adposition and noun, and affixation.

4.1 Distributional Language Embeddings

Words can be represented by distributed word representations, currently often in the form of word embeddings. Similarly to how words can be embedded, so can languages, by associating each language with a real-valued vector known as a *language embedding*. Training such representations as a part of a multilingual model allows us to infer similarities between languages. This is due to the fact that in order for multilingual parameter sharing to be successful in this setting, the neural network needs to use the language embeddings to encode features of the languages. Previous work has explored this type of representation learning in various tasks, such as NMT (Malaviya et al., 2017), language modelling (Tsvetkov et al., 2016; Östling and Tiedemann, 2017), and tasks representing morphological, phonological, and syntactic linguistic levels (Bjerva and Augenstein, 2018a).

In the context of computational typology, representations obtained through language modelling have been the most successful (Östling and Tiedemann, 2017). This approach is particularly interesting since unlabelled data is available for a large portion of the world's languages, meaning that high quality language embeddings can be obtained for more than 1,000 of the world's languages.
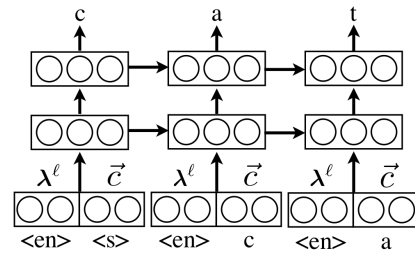
4.2 Language Embeddings through LMs

In this work, we use a language modelling objective to pre-train language embeddings; we train a character-level neural language model with a distributed language embedding of language ℓ. Specifically, we use the model of Östling and Tiedemann (2017), visualised in Figure 4. The model is a stacked character-based LSTM (Hochreiter and Schmidhuber, 1997) with two layers, followed by a softmax output layer. In order to accommodate the language embeddings, this relatively standard model is modified such that language embeddings are concatenated to the character embeddings at each time step. This method returns a language embedding, which we denote $\tilde{\boldsymbol{\lambda}}^{\ell}$ to distinguish it from the language embeddings $\boldsymbol{\lambda}^{\ell}$ discussed in the previous section. We use the same hyperparameter settings as Östling and Tiedemann (2017), with 1024-dimensional LSTMs, 128-dimensional character embeddings, and 64-dimensional language embeddings. Training is done with Adam (Kingma and Ba, 2014), and using early stopping.

In the semi-supervised regime, we use the estimated language embedding $\tilde{\boldsymbol{\lambda}}^{\ell}$ from the language model and define the model as follows

$$p(\boldsymbol{\pi}^{\ell} \mid \tilde{\boldsymbol{\lambda}}^{\ell}) = \prod_{i=1}^{|\boldsymbol{\pi}|} sigmoid(\mathbf{e}_{\pi_i}^{\top} \tilde{\boldsymbol{\lambda}}^{\ell}) \qquad (6)$$

omitting the learned language embedding $\boldsymbol{\lambda}$ in the matrix factorisation. The likelihood of this model is now convex in the parameter embeddings. In contrast to the full matrix factorisation setting, here, all language-specific knowledge must come from an external source, namely, the unlabelled text.

5 A Novel Task: Typological Collaborative Filtering

In this section, we introduce a novel task for linguistic typology, which we term **typological collaborative filtering**. Typological KBs such as WALS

are known to be incomplete. In other words, not all parameters π_i^ℓ are observed for all languages. Thus, a natural question we may ask how well our models can predict unobserved (or held-out) parameters. We view this problem as analogous to collaborative filtering (CF). Our task is similar to knowledge base population or completion in that we start off with a partly populated KB which we aim to complete, but differs in the aspect that we attempt to infer values from correlations between features and similarities between languages, rather than inferring these from a collection of texts.

CF is a common technique for recommender systems (see §9). Consider the task of recommending a new movie to a customer. Given which movies different users have liked (equivalent to a typological parameter being 'on') and which movies the user has disliked (equivalent to the typological parameter being 'off'), a CF model tries to figure out the (latent) preferences of each user and the latent genre of each movie. In our setting, the languages are analogous to users and movies are analogous to parameters. Our model in eq. (5) seeks to learn what latent properties of languages and what latent properties of parameters explain their correlations.

6 Data

WALS. The World Atlas of Language Structures (WALS) is a large knowledge base of typological properties at the lexical, phonological, syntactic and semantic level on which we will run our experiments. The documentation of linguistic structure is spread throughout a wide variety of academic works, ranging from field linguistics to grammars describing the nuances of individual grammatical uses. KB creation is a laborious tasks as it involves distilling knowledge into a single, standardised resource, which, naturally, will be incomplete, prompting the need for methods to complete them (Min et al., 2013). In the case of WALS, few languages have all values annotated for all of the properties. In this section, we offer a formalisation of typological KBs to allow for our development of a probabilistic model over vectors of properties. WALS, for instance, contains $n = 202$ different parameters (Dryer and Haspelmath, 2013).

Binarisation of WALS. Many common typological KBs, including WALS, the one studied here, contain binary as well as non-binary parameters. To deal with this, we binarise the KB as follows: Whenever there is a typological parameter that

takes ≥ 3 values, e.g., 'Feature 81A: Order of Subject, Object and Verb' which takes the 7 values 'SOV', 'SVO', 'VSO', 'VOS', 'OVS', 'OSV', 'No dominant order', we introduce that many binary parameters. At test time, we get non-binary predictions by using a simple decoder that returns the arg max over the predicted probabilities for the binary features.

As each typological parameter with $n \geq 3$ feature values is coded as a one-hot binary vector of length n, we need to make sure that we do not mix a single typological parameter for a language into the training and test sets. This is visualised in Figure 2, where we train on the blue squares, i.e., the binarised 81A feature for for English, and the 26A feature for Dutch, as well as all features for all non-Germanic languages. The model is then evaluated on the held-out features for Germanic highlighted in red, i.e., 26A for English, 81A for Dutch, and both of these for German. This is important, as knowing that a language is SVO would make it trivial to infer that it is not, e.g., SOV.

Unlabelled multilingual data. To induce language embeddings, we need a considerable amount of multilingual unlabelled data. We use an in-house version of the massively multilingual Bible corpus, so as to have comparable data for all languages, although parallel data is not a strict necessity for our method.[2] We train the multilingual language model on a collection of 975 languages, each with approximately 200,000 tokens available. We only train on languages for which the symbol size is relatively modest, a criterion which we fulfil by only using translations with Latin, Cyrillic, and Greek alphabets.

Language used in the bible differs substantially from most modern language use, which would be a challenge if one were interested in transferring the language model itself. Here, we are only interested in the distributed language embeddings for each language ℓ. It is safe to assume that the typological features underlying the texts we use will be representative of those coded in WALS, hence the domain should not matter much and the method should work equally well given any domain of input texts for the unsupervised training of language embeddings.

---

[2]Some languages available via `http://homepages.inf.ed.ac.uk/s0787820/bible/`
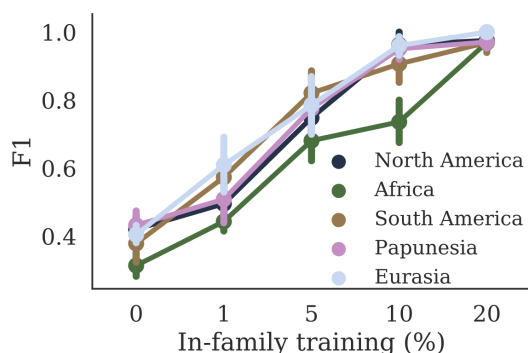
Figure 5: F1 score for feature prediction separated by macroarea with varying in-branch training data.

7 Experiments

7.1 General Experimental Setup

As described in §6, we binarise WALS. In order to compare directly with our semi-supervised extension, we limit our evaluation to the subset of languages which is the intersection of the languages for which we have Bible data, and the languages which are present in WALS. Finally, we observe that some languages have very few features encoded, and some features are encoded for very few languages. For instance, feature 10B (Nasal Vowels in West Africa), is only encoded for a total of 40 languages, and only one feature value appears for more than 10 languages. Because of this, we restrict our evaluation to languages and feature values which occur at least 10 times. Note that we evaluate on the original parameters, and not the binarised ones.

Our general experimental set-up is as follows. We first split the languages in WALS into each language branch (*genus* using WALS terminology). This gives us, e.g., a set of Germanic languages, a set of Romance languages, a set of Berber languages, and so on. (We note that this does not correspond to the notion of a language family, e.g., the Indo-European language family.) We wish to evaluate on this type of held-out set, as it is both relatively challenging: If we know the parameters of Portuguese, predicting the parameters for Spanish is a much easier task. This setup will both give us a critical estimate of how well we can predict features overall, in addition to mimicking a scenario in which we either have a poorly covered language or branch, which we wish to add to WALS.

7.2 Typological Collaborative Filtering

We evaluate our method for typological collaborative filtering on each language branch in WALS in a series of experiments. Given a branch B (a set of languages), we randomly select 80% of the feature-language combinations from the languages $\ell \in B$, which we use for evaluation (e.g. those highlighted in red in Figure 2). The remaining 20% is either not considered, or (partially) used for training, as we run experiments in which we train on $(0, 1, 5, 10, 20)\%$ relative of the held-out data. The idea behind this is that it should be very difficult to predict features if nothing is known for a language at all, whereas knowing a few features of a language, or of related languages, should allow the model to take advantage of the strong correlations between features (Figure 1) and between languages (Figure 3). We train on all data in WALS for languages which are not in the current evaluation branch under consideration.

Each experiment is then an evaluation of how well we can predict features for a completely or relatively unseen language family. Evaluation is done across the branches in WALS with more than four languages represented, after filtering away languages for which we have fewer than 10 features available. This amounts to a total of 36 branches, and 448 languages. We repeat each experiment 5 times per language branch, for each proportion of in-branch training data in $(0, 1, 5, 10, 20)\%$, yielding a total of 900 experiment runs. The results reported are the mean across these runs.

Figure 5 shows the micro F1 score we obtain averaged across macroareas. The bars indicate 95% confidence intervals. We can see that, with access to 20% of the in-branch training data, we can predict features at above 90% F1 score regardless of macroarea. Prediction generally is more challenging for languages in the macroarea Africa. This can be explained by, e.g., contrasting with the Eurasian macroarea. Whereas the latter includes branches which are relatively uncontroversial, such as Germanic and Slavic languages, this is not the case with the former. One such example is Bongo-Bagirmi (one of the evaluation branches, spoken in Central Africa), for which there is poor agreement in terms of classification (Bender, 2000).

7.3 Semi-supervised extension

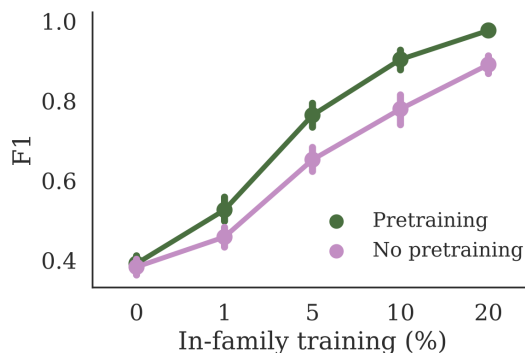We next evaluate the semi-supervised extension, which requires unlabelled texts for a large amount

Figure 6: F1 scores across all held-out language families, comparing pretraining with no pretraining.

In-branch train %	Freq. F1	Indiv. pred. F1	T-CF F1	SemiSup F1
0.00	0.2950	0.2990	0.3998	0.3916
0.01	0.2949	0.2976	0.4578	0.5263
0.05	0.2947	0.2970	0.6552	0.7641
0.10	0.2945	0.2971	0.7807	0.9040
0.20	0.2938	0.2973	0.8835	**0.9767**

Table 1: Aggregate results w. 0-20% relative use of in-branch training data. Columns show: most frequent class (Freq.), individual prediction per feature with language-embeddings (Individual pred.), typological collaborative filtering (T-CF), semi-supervised extension (SemiSup). For standard deviations across runs and per-branch results, see appendix.

of languages, although the domain of these texts should not matter much. This allows us to take advantage of correlations between similar languages to point the model in the right direction. Even with 1% training data, it may be very useful for a model to know that, e.g., German and Dutch are grammatically very similar. Hence, if the 1% training data contains features for Dutch, it should be quite easy for the model to learn to transfer these to German.

Figure 6 shows results with pre-training. Without any in-branch training data, using pretrained embeddings does not offer any improvement in prediction power. This can be explained by the fact language embeddings are updated during training, which leads to a drift in the representations present in the training material. This was chosen since not updating the representations yielded poor performance in all settings explored. We hypothesise that this is because, although a language modelling objective offers a good starting point in terms of encoding typological features, it is not sufficient to explain the typological diversity of languages. For instance, a language model should hardly care about the phonological nature of a language. This is in line with previous work which shows that the linguistic nature of the target task is important when predicting typological features with language embeddings (Bjerva and Augenstein, 2018a). However, once we have access to $> 5\%$ of in-branch training data, language embeddings offers a substantial improvement, e.g. an F1 error reduction of more than 50% with access to 10% of the in-branch data (see Table 1 and Table 3 for per-branch results in the appendix). This shows that we can partially aid a typologist's work by utilising unannotated data.

7.4 Quantitative comparison

In addition to evaluating our method of typological CF, we compare to some baselines drawn from earlier work. First, we report a most frequent value baseline. As many typological features are heavily skewed, this is quite high already. For instance, defaulting to the most frequent value for word order (i.e. SVO) would yield an accuracy of 41% (**Freq.** in Table 1). A more involved baseline is Bjerva and Augenstein (2018a), who use pre-trained language embeddings in a k-NN classifier trained on individual WALS features (**Individual pred.** in Table 1). For the baseline reported here, we only use one nearest neighbour for this prediction. The scores we obtain here are quite low compared to Bjerva and Augenstein (2018a), which is explained by the fact that we have access to very little training data in the current setting, and highlights the importance of taking advantage of correlations between languages and features, and not simply looking at these factors in isolation. Finally, we compare our typological collaborative filtering approach, as well as our semi-supervised extension (**T-CF** and **SemiSup** in Table 1).

8 Analysis

Accuracy for several experimental settings is visualised in Figure 7, broken down by the linguistic category of the predicted features. Since results change little between the 5% in-branch setting and higher percentages, we only look at 0%, 1% and 5% here. We also visualise accuracy without (Figure 7, left) and with our semi-supervised extension (Figure 7, right) in each setting.

8.1 Typological collaborative filtering with variable amounts of data

Focussing on Figure 7 (left) alone first we observe an expected pattern: using increasingly more in-branch data boosts performance across all feature groups. This increase in accuracy can be attributed to the model having more knowledge about each query language in itself and about how languages relate to one another, based on similarities in their parameter configurations.

Making a prediction about the order of adposition and noun phrase in a given language, lacking any other information about that language, is basically a shot in the dark. In-branch training data, in our experiments, includes in-language training data, too. Having one piece of information about the word order of that language, its ordering of relative clause and noun, or even its affixational properties, immediately makes the prediction informed rather than random: in many other languages within and outside this particular language family the model would have likely observed a strong correlation between these features and the order of adposition and noun phrase, which are all subject to the more general headedness parameter.

Certain features and feature configurations may not be as abundant cross-linguistically as the set of headedness features. In those cases, access to in-branch data is crucial. Consider e.g. the feature 10B Nasal Vowels in West Africa: a handful of language branches exhibit this feature and at least one of its values, *no nasal vs. oral vowel contrast*, is characteristic predominantly of Niger-Congo languages. Without *any* in-branch training data, the model's knowledge of this feature value is extremely limited, making its correct prediction for a Niger-Congo language virtually impossible. A small amount of in-branch training data thus increases the chance of a correct prediction greatly.

8.2 Semi-supervised extension

Comparing Figure 7 to Figure 6 reveals a crucial finding. While we see very little improvement from pretraining for 0% in-branch training overall, for individual linguistic categories, it mostly benefits prediction: seven out of nine feature groups are predicted more accurately with pretraining. Phonological and morphological predictions experience moderate deterioration, however, counterbalancing much of the improvement in other categories, which leads to the overall result of seemingly lit-

Language	Genus	Fixed Stress Location	Weight-Sensitive Stress
English	Germanic	?	Right-oriented
Icelandic	Germanic	Initial	Fixed stress

Table 2: In-branch training data in example scenario

tle improvement from pretraining. The limited effect of pretraining on prediction of phonological and morphological features can be explained with reference to the richness and complexity of these linguistic domains, which makes for data sparsity and generally makes them harder to learn based on distributional information alone. Moreover, a number of phonological features refer to aspects of language that may not be reflected in writing, such as stress and devoicing. All other categories concern syntactic and semantic information, which is known to be learnable from word distribution, and therefore benefit from the knowledge carried by language embeddings.

Figure 7 (right) shows an unsteady interaction between pretraining and the addition of increasing amounts of in-branch data. While pretraining alone helps for predicting most features, as pointed out above, an extra 1% of in-branch data in the pretrained setting has a rather variable impact across feature groups. For a few groups it helps, as is expected, for a few it has no effect and for two groups, 'Word Order' and 'Simple Clauses', it makes for quite a drop in accuracy. We speculate that this effect, while negative, is indicative of the general power of language embeddings in associating related languages. Consider the test query 'Fixed Stress Location' in English, where the 1% of in-branch training data contains the information in Table 2. Based on feature correlation alone, the model should predict 'No fixed stress' for English, since this value always co-occurs with 'Right-oriented stress'. Yet, due to the proximity in the English and Icelandic embeddings, the model may copy the value of Icelandic and falsely predict 'Initial stress' for English, too. The risk of this happening decreases with more in-branch training data, since the model can generalise over more in-branch features.

Lastly, notice that accuracy for phonological features remains low even with 5% of in-branch data, and it is lower in the pretrained setting compared to the no-pretraining one. This brings us to the conclusion that using pretrained embeddings which are fine-tuned for specific tasks which encode different linguistic levels, as in Bjerva and Augen-
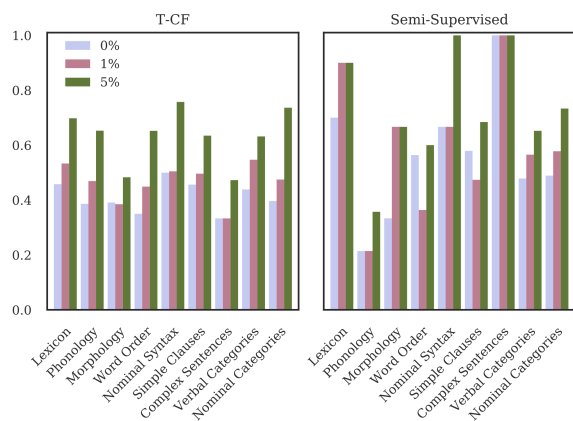
Figure 7: Accuracy per feature group (Germanic).

stein (2018a), might also be useful in our semi-supervised extension of typological collaborative filtering.

9   Related Work

Computational Typology   The availability of unlabelled datasets for hundreds of languages permits inferring linguistic properties and categories (Östling, 2015; Asgari and Schütze, 2017). Individual prediction of typological features has been attempted in conjunction with several NLP tasks (Malaviya et al., 2017; Bjerva and Augenstein, 2018a,b). Our work is most similar to Murawaki (2017), who presents a Bayesian approach to utilising relations between features and languages for feature prediction. However, our work differs on several important counts, as we (i) include language information obtained through unsupervised learning, which allows us to take advantage of raw data and predict features for completely unannotated languages, (ii) analyse the effects of varying amounts of known features, especially in situations with and without in-branch training data, and (iii) view the problem of typological features through the lens of parameters from principles and parameters (Chomsky, 2000). Deep generative models have also been explored previously for modelling phonology (Cotterell and Eisner, 2017). Our work builds on these research directions, by (i) developing a deep generative model which (ii) takes advantage of correlations, rather than predicting features individually, and (iii) exploits unlabelled data. This work is also related to linguistic representations encoded in neural models (Kádár et al., 2017) and language embeddings (Bjerva et al., 2019), multilingual relations between languages in various representational levels (Beinborn and Choenni, 2019), as well as the related problem of phylogenetic inference (Farach et al., 1995; Nichols and Warnow, 2008). For a survey of typology in NLP, see Ponti et al. (2018).

Matrix Factorisation   Collaborative Filtering was popularised in the early 1990s as a technique for recommender systems with applications such as mail filtering (Goldberg et al., 1992), and article (Resnick et al., 1994) and movie recommendation (Dahlen et al., 1998). Model-based algorithms soon became popular (Breese et al., 1998) to overcome the cold start problem arising for unseen users or items at test time. The most successful one of these, in turn, is matrix factorisation, as applied in this paper, which represents users and items as (dense) vectors in the same latent feature space and measures their compatibility by taking the dot product between the two representations (Koren et al., 2009; Bokde et al., 2015). Beyond recommender systems, matrix factorisation has shown successes in a wide variety of subareas of NLP (Riedel et al., 2013; Rocktäschel et al., 2015; Levy and Goldberg, 2014; Lei et al., 2014; Augenstein et al., 2018).

10   Conclusion

We introduce a generative model inspired by the principles-and-parameters framework, drawing on the correlations between typological features of languages to solve the novel task of typological collaborative filtering. We further show that raw text can be utilised to infer similarities between languages, thus allowing for extending the method with semi-supervised language embeddings.

Acknowledgements

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. 2018. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906, New Orleans, Louisiana. Association for Computational Linguistics.

Mark C Baker. 2008. The macroparameter in a microparametric world. *The Limits of Syntactic Variation*, 132.

Lisa Beinborn and Rochelle Choenni. 2019. Semantic drift in multilingual representations. *arXiv preprint arXiv:1904.10820*.

M Lionel Bender. 2000. Nilo-Saharan. In B. Heine and Nurse. D., editors, *African languages: An introduction*, pages 43–73. Cambridge University Press Cambridge.

Theresa Biberauer, Anders Holmberg, Ian Roberts, and Michelle Sheehan. 2009. *Parametric variation: Null subjects in minimalist theory*. Cambridge University Press.

Johannes Bjerva and Isabelle Augenstein. 2018a. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018b. Tracking typological traits of uralic languages in distributed language representations. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 76–86, Helsinki, Finland. Association for Computational Linguistics.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do Language Representations Really Represent? *Computational Linguistics*. In press, arXiv preprint arXiv:1901.02646.

Dheeraj Bokde, Sheetal Girase, and Debajyoti Mukhopadhyay. 2015. Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey. *Procedia Computer Science*, (49):136–146.

John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52.

Andrew Carnie. 2013. *Syntax: A generative introduction*. John Wiley & Sons.

Noam Chomsky. 1957. *Syntactic structures*. Walter de Gruyter.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press.

Noam Chomsky. 1981. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Noam Chomsky. 2000. *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. 2001. A generalization of principal components analysis to the exponential family. In *NIPS*, pages 617–624.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

William Croft. 2002. *Typology and Universals*. Cambridge University Press.

B.J. Dahlen, J.A. Konstan, J.L. Herlocker, N. Good, A. Borchers, and J. Riedl. 1998. Jump-starting movieLens: User benefits of starting a collaborative filtering system with "dead-data". In . University of Minnesota TR 98-017.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Jason Eisner. 2002. Discovering syntactic deep structure via Bayesian statistics. *Cognitive Science*, 26(3):255–268.

Martin Farach, Sampath Kannan, and Tandy Warnow. 1995. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155–179.

David Goldberg, David Nichols, Brian Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70. Special Issue on Information Filtering.

Randy Allen Harris. 1995. *The Linguistics Wars*. Oxford University Press.

Martin Haspelmath. 2008. Parametric versus functional explanations of syntactic universals. *The limits of syntactic variation*, 132:75–107.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural network. *Computational Linguistics*, 43:761–780.

Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391, Baltimore, Maryland. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia. Association for Computational Linguistics.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Johanna Nichols and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2:760–820.

Marco Nicolis and Theresa Biberauer. 2008. The null subject parameter and correlating properties. *The limits of syntactic variation*, 132:271.

Marc van Oostendorp. 2015. Parameters in phonological analysis: stress. In A. Fabregas, J. Mateu, and M. Putnam, editors, *Contemporary Linguistic Parameters*, Contemporary Studies in Linguistics, pages 234–258. London: Bloomsbury.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics, arXiv preprint arXiv:1807.00914*.

Alan Prince and Paul Smolensky. 2008. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 Conference on Computer Supported Collaborative Work*, pages 175–186.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia. Association for Computational Linguistics.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1119–1129, Denver, Colorado. Association for Computational Linguistics.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

A   Appendices

Macroarea	Branch	In-branch training (%)	Freq. F1	Individual pred. F1	T-CF F1	SemiSup F1
Africa	Bantoid	0.00	0.3510 (0.0042)	0.3540 (0.0057)	0.3527 (0.0260)	0.3287 (0.0292)
		0.01	0.3510 (0.0042)	0.3643 (0.0077)	0.4100 (0.0196)	0.4013 (0.0260)
		0.05	0.3510 (0.0036)	0.3523 (0.0053)	0.4997 (0.0446)	0.5467 (0.0266)
		0.10	0.3507 (0.0031)	0.3570 (0.0029)	0.5687 (0.1065)	0.6107 (0.0963)
		0.20	0.3480 (0.0057)	0.3450 (0.0071)	0.8397 (0.0475)	**0.8620** (0.0645)
Papunesia	Chimbu	0.00	0.2860 (0.0182)	0.2940 (0.0142)	0.3460 (0.0771)	0.3357 (0.0708)
		0.01	0.2860 (0.0182)	0.2887 (0.0120)	0.3600 (0.0805)	0.2773 (0.0890)
		0.05	0.2850 (0.0171)	0.2990 (0.0236)	0.3833 (0.1012)	0.4560 (0.2013)
		0.10	0.2850 (0.0171)	0.2937 (0.0161)	0.6140 (0.0884)	0.6950 (0.0318)
		0.20	0.2873 (0.0181)	0.2893 (0.0135)	0.7800 (0.1181)	**0.8267** (0.1605)
Africa	Kwa	0.00	0.2417 (0.0111)	0.2540 (0.0099)	0.3163 (0.0381)	0.2720 (0.0490)
		0.01	0.2407 (0.0104)	0.2550 (0.0051)	0.3807 (0.0775)	0.3843 (0.0796)
		0.05	0.2417 (0.0111)	0.2597 (0.0095)	0.4927 (0.1181)	0.5290 (0.0978)
		0.10	0.2417 (0.0111)	0.2453 (0.0076)	0.6827 (0.1092)	0.6827 (0.1092)
		0.20	0.2417 (0.0111)	0.2443 (0.0033)	**0.9960** (0.0057)	**0.9960** (0.0057)
North America	Mixtecan	0.00	0.3507 (0.0076)	0.3427 (0.0047)	0.4387 (0.1244)	0.4747 (0.0766)
		0.01	0.3507 (0.0076)	0.3360 (0.0065)	0.4593 (0.0514)	0.5010 (0.0304)
		0.05	0.3500 (0.0083)	0.3460 (0.0065)	0.6867 (0.2274)	0.7803 (0.0988)
		0.10	0.3490 (0.0094)	0.3400 (0.0114)	0.8610 (0.1288)	0.8830 (0.1441)
		0.20	0.3490 (0.0094)	0.3383 (0.0130)	0.9883 (0.0165)	**1.0000** (0.0000)
Papunesia	Oceanic	0.00	0.3837 (0.0048)	0.4030 (0.0096)	0.3677 (0.0416)	0.3663 (0.0433)
		0.01	0.3837 (0.0048)	0.3957 (0.0066)	0.3783 (0.0324)	0.4047 (0.0217)
		0.05	0.3833 (0.0041)	0.3993 (0.0068)	0.6193 (0.0382)	0.6753 (0.0838)
		0.10	0.3830 (0.0051)	0.3983 (0.0071)	0.7243 (0.1116)	0.7337 (0.1245)
		0.20	0.3817 (0.0034)	0.4057 (0.0113)	0.8393 (0.0662)	**0.9200** (0.0641)
All	All	0.00	0.2950 (0.0453)	0.2990 (0.0448)	0.3998 (0.1095)	0.3916 (0.1077)
		0.01	0.2949 (0.0453)	0.2976 (0.0457)	0.4578 (0.1298)	0.5263 (0.1605)
		0.05	0.2947 (0.0452)	0.2970 (0.0436)	0.6552 (0.1726)	0.7641 (0.1572)
		0.10	0.2945 (0.0451)	0.2971 (0.0434)	0.7807 (0.1782)	0.9040 (0.1324)
		0.20	0.2938 (0.0446)	0.2973 (0.0451)	0.8835 (0.1278)	**0.9767** (0.0464)

Table 3: Aggregate results with 0-20% relative use of in-branch training data. The columns indicate the most frequent class (Freq.), individual prediction per feature with language-embeddings (Individual pred.), typological collaborative filtering (T-CF), semi-supervised extension (SemiSup).