6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)
29-31 August 2018, Gurugram, India

# Predicting the Features of World Atlas of Language Structures from Speech

*Alexander Gutkin, Tatiana Merkulova, Martin Jansche*

Google AI, London, United Kingdom

{agutkin,merkulova,mjansche}@google.com

## Abstract

We present a novel task that involves prediction of linguistic typological features from the World Atlas of Language Structures (WALS) from multilingual speech. We frame this task as a multi-label classification involving predicting the set of non-mutually exclusive and extremely sparse multi-valued WALS features. We investigate whether the speech modality has enough signals for an RNN to reliably discriminate between the typological features for languages which are included in the training data as well as languages withheld from the training. We show that the proposed approach can identify typological features with the overall accuracy of 91.6% for the 16 in-domain and 71.1% for 19 held-out languages. In addition, our approach outperforms language identification-based baselines on all the languages. Also, we show that correctly identifying all the typological features for an unseen language is still a distant goal: for 14 languages out of 19 the prediction error is well above 30%.

**Index Terms**: linguistic typology, speech, low-resource languages, multi-label classification, neural networks

## 1. Introduction

The field of linguistic typology organizes the world's languages according to their structural and functional features and helps to describe and explain their linguistic diversity [1]. In recent years there has been a growing interest in employing linguistic typology resources in natural language processing [2], where linguistic typology is used to scale up existing language technologies to the long tail of the world's languages [3] for which the traditional resources are very scarce or missing altogether. Typological resources such as PHOIBLE [4], Glottolog [5], and PanPhon [6] have been successfully used in diverse speech and language tasks such as grapheme-to-phoneme conversion [7], multilingual language modeling [8] and dependency parsing [9].

In this study we present a novel task of learning linguistic typological information from multilingual speech corpora. We frame this problem as a classification task where, given the speech utterance, one needs to determine the structural features of a corresponding language. The source for the features is the World Atlas of Language Structures (WALS) [10] that contains phonological, lexical, grammatical and other attributes gathered from descriptive materials for 2,679 languages. The main motivation for this study is to discover structural typological properties of the low-resource languages and dialects for which very little or no training data is available. This task is different from spoken language identification [11]: As a hypothetical example, when applied to the spoken Scots, the language identification is likely to detect it as English. This is not very helpful because, for this task, one hopes to discover the features that make Scots unique, such as its hypothesized phonological link to Scandinavian languages [12, 13]. Moreover, this task is compelling because having an accurate linguistic feature detector can aid development work. For example, correctly identifying broad phonetic features of an unknown language can help one build crude grapheme-to-phoneme rules and phoneme inventories for automatic speech recognition and text-to-speech.

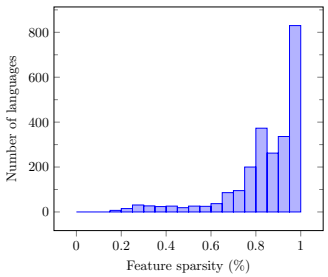


Figure 1: *WALS features sparsity estimated as percentage of the features not attested for each language.*

This paper is organized as follows: we introduce our approach to predicting WALS features as a sparse multi-label classification problem in Section 2. The goal of the experiments, described in Section 3, is to determine which features and groups thereof can be reliably predicted for the in-domain and out-of-domain languages. For the out-of-domain languages for which the data was withheld when training our model, we compare the results against a language identification baseline that we train. This baseline uses the predicted language as a lookup key into WALS. We show that the network does not simply memorize the features for language identification, but is able to generalize over the held-out set. We also show that the speech modality offers enough signal to discriminate well between some groups of features. Section 4 concludes the paper. It is important to note that we treat this task as discriminating between WALS feature tags, not making any claims about the discovery of actual structure of underlying phenomena. Since this is a novel task, our goal is to provide a baseline, a very likely crude one, but one that can be improved upon over time.

## 2. Problem Formulation and Models

### 2.1. World Atlas of Language Structures

We use WALS [10] as a source of typological information for 2,679 languages. The 192 WALS typological multi-valued language features are organized into 152 chapters, each chapter corresponding to a particular phonological, morphological or syntactic linguistic property. For example, there are 18 chapters corresponding to "Word Order" syntactic property where most chapters contains one typological feature, such as "Order of Genitive and Noun", while other chapters, such as "Order of Negative Morpheme and Verb" contain 7 features, such as "Obligatory Double Negation" [14]. The dataset is very sparse, as demonstrated by Figure 1. For example, for 1,801 languages out of 2,679 only 20% (or less) of the WALS features are attested. Only 149 languages have 50% (or more) coverage. We



10.21437/SLTU.2018-52

prune WALS by removing the set of 57 languages for which no ISO 639-3 code is defined. In addition, we remove three languages for which no WALS features are attested.

### 2.2. Multi-label classification

Given an example variable-length speech sequence $\mathbf{x}$ in an input feature space $X$, the classification task consists of selecting a set of multiple applicable WALS feature labels $\{\lambda_i\}$ from a finite set of labels $L = \{\lambda_1, \lambda_2, \ldots, \lambda_{N_L}\}$, where $N_L = 192$ is the number of WALS features. Each candidate label takes its value from a set of disjoint classes $Y_i = \{y_j^i\}, 1 \leq i \leq N_L$, corresponding to the values of a particular WALS feature $\lambda_i$. For example, a language may or may not have a "number of genders" feature label present, but if it is present this feature cannot take the values of "None" and "Four" simultaneously. The cardinality $C$ of the set of all unique feature values is 1316. This scenario fits the *multi-label multi-class* classification problem [15, 16].

### 2.3. Model architecture

The model architecture is an extension of automatic language identification architecture described in [17]. Given the input speech parametrization, the input layer is a convolutional neural network (CNN) [18]. The CNN employs one-dimensional temporal convolutions similar to the architecture described in [19]: There are three one-dimensional convolution layers containing 20, 40 and 60 filters, respectively. Receptive field sizes $r$ for each layer are 5, 5 and 3. The stride parameter is set to 1. Rectified linear units (ReLU) are used in each layer [20]. Batch normalization is applied before each layer [21]. Each convolution layer is followed by a max-pooling layer with filter size $r'$ set to 2. Dropout [22] with probability 0.5 is applied to the last max-pooling layer. In our experiments we found that using this configuration of the input layers aided feature extraction and improved the overall accuracy compared to a simpler feed-forward layer or no input layer at all.

The CNN outputs are fed into recurrent neural network (RNN) which is a bidirectional variant [23] of a long short term memory (LSTM) model [24]. The RNN consists of two layers of bidirectional LSTMs with 128 forward and 128 backward cells in each layer. Dropout is applied to each layer with a probability of 0.5. The uniform weight initialization scheme from [25] was used. The forward and backward outputs corresponding to the last time step of the last layer of bidirectional LSTM are concatenated together and provided to the single fully-connected linear activation layer. Each output of this layer corresponds to a particular value of a WALS feature. There are 1316 outputs in total.

### 2.4. Dealing with data imbalance

As noted in [26, 27], many typological databases are designed to suit the needs of theoretical linguistic typology, resulting in a sparse representation of features across languages (mostly due to intentional statistical balancing of features across language families and geographic areas). Also, for certain languages the maintainers are sometimes unable to obtain a reliable description of linguistic attributes from the available linguistic sources [28]. This situation is problematic for statistical modeling because it results in heavy *data imbalance* between different types of features and complicates construction of machine learning models. A classifier constructed without regard to data imbalance leans towards correctly predicting the majority class, which in case of WALS corresponds to missing or intentionally undefined features, while the "interesting" features with low coverage are heavily underestimated.

Table 1: *Languages used for training, development and testing.*

| Languages | Type | Train | Dev | Test |
|---|---|---|---|---|
| Egyptian Arabic, Bulgarian, Czech, Danish, Dutch, French, German, Greek, Spanish, Hebrew, Hindi, Italian, Japanese, Korean, Latvian, Nepali | *I* | ✓ | ✓ | ✓ |
| Amharic, Basque, Bengali, Burmese, Gujarati, Nigerian English, Lao, Lithuanian, Kannada, Maithili, Malayalam, Marathi, Romanian, Russian, Sinhala, Swedish, Tamil, Telugu, Urdu | *H* | | | ✓ |

Approaches to data imbalance have been extensively studied in the literature [29, 30]. In this study we employ the reciprocal frequency approach inspired by [31]. We apply sigmoid non-linearities to the outputs of fully-connected layer and optimize all predictions $\hat{\mathbf{y}}$ against the true labels $\mathbf{y}$ all at once using the weighted variant of a cross-entropy function [32] defined as

$$L(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} w(i)\mathbf{y}_n^i \log\left(\hat{\mathbf{y}}^i(\mathbf{x}_n, \boldsymbol{\theta})\right) + r(\boldsymbol{\theta}) \,, \quad (1)$$

where $\boldsymbol{\theta}$ represents network parameters, $\mathbf{x}$ is the training set, $C$ is the dimension of the prediction vector, $r$ is an $l_2$-norm regularization term and $w(c)$ is the weight function associated with class $c$ which, for the observed classes, is defined as $N/N_c$, where $N$ is the number of training sequences and $N_c$ is the feature value count ($c \in \mathcal{L}$) for feature $\mathcal{L}$. The counts are computed solely from the training data. For the unobserved classes $w(c) = 0$. The purpose of the function is to penalize the frequent classes and boost the rare ones. The unattested classes do not contribute to the overall loss.

## 3. Experiments and Discussion

Our experiments focus on four aspects: First, we aim to establish the baselines that correspond to the lower performance bounds for the neural network-based WALS feature classifier. In particular, we investigate the majority class prediction and what happens when the network learns by chance. Second, we investigate a regular scenario where the generalization over the unseen examples from a language observed in the training data is tested. Third, we present the results for the languages withheld from training. Finally, we compare our approach with a feature lookup guided by language identification.

When computing the various metrics we ignore the undefined WALS feature values focusing on attested features only. For the attested features we rely on the weighted loss function from equation (1) to alleviate the inherent imbalance between the WALS classes.

### 3.1. Dataset details

Our dataset consists of a small subset of an in-house corpus of mobile speech collected over the years using Datahound - a spoken utterance data collection application running on Android devices [33]. The speech comes from a variety of speakers of both genders. The recording conditions range from the quiet indoor recordings to very noisy recordings in public places. The languages used for training, development and testing are shown in Table 1. The training and development sets consist of 122,610 and 18,006 utterances from 16 languages from diverse language groups (Semitic, Balto-Slavic, Germanic, Indo-Aryan, Romance, Koreanic and Japonic). Each language in

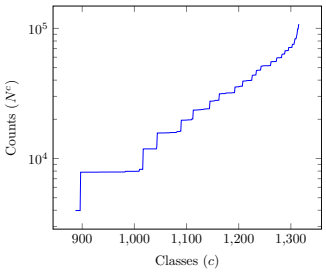


Figure 2: *WALS feature value (class) counts displayed on a logarithmic scale. Classes are sorted by their counts. 886 classes out of 1316 are unobserved.*

training and development set has roughly the same number of utterances. The test set for the above 16 languages consists of 14,437 utterances. Because this set contains the data for the languages observed in training, we denote this test set as *in-domain* ($I$). We keep an additional test set of 16,125 utterances corresponding to the 19 languages excluded from the training. We denote this test set as *held-out* ($H$). Most of the languages in this set have some (possibly remote, as is the case between Amharic and Hebrew) relation to the languages in the training data apart from the languages belonging to Dravidian and Tai-Kadai language families, as well as Basque.

In Section 2.1 we provided initial analysis of WALS feature sparsity based on feature value counts computed solely from the WALS corpus (Figure 1). The dimension of difficulty involved in training a neural network on our speech dataset is demonstrated in Figure 2, which shows the counts for all possible WALS feature values (corresponding to classes that a classifier has to predict) encountered in the training data. A significant proportion of the classes (886 out of 1316) are not encountered in the training data for the 16 languages. The distribution of counts for the majority of the remaining 430 classes is approximately log-linear.

### 3.2. Model training details

The 16 kHz speech was parameterized into HTK-style Mel Frequency Cepstral Coefficients (MFCC) [34] using a 10 msec frame shift. The dimension of the MFCC parameters is 39 (13 static + $\Delta$ + $\Delta\Delta$ coefficients). We compute the mean and standard deviation over all the parameters in the training set and use these values to scale the training parameter sequences to have zero mean and unit variance. The same values of mean and standard deviation are used to scale the parameter sequences in the development and test sets. The parameters serve as inputs to our hybrid CNN-LSTM classifier network, which we introduced in Section 2.3. No language identifying features are used. The models are trained using AdaDelta [35] with $\rho = 0.95$ and $\epsilon = 10^{-8}$. We use an exponential learning rate decay, with an initial learning rate set to $10^{-3}$, a reasonably slow decay factor of $0.95$ and the number of decay steps is set to $4 * 10^5$. In our experiments we found that the above settings for learning rate lead to better convergence. $L_2$ regularization is applied to the recurrent layer weights, with the weight scaling factor set to $0.05$. In addition, the global gradient clipping limit is set to 10 and the training mini-batch size is 10. We tuned the above hyper-parameters manually.

### 3.3. Baselines: majority class and learning by chance

Our first baseline corresponds to the simplest scenario where the classifier makes the decisions based on majority class labels only. In this case, instead of predicting the class corresponding to the value of a particular WALS feature using a neural network, we select the most frequent value of that feature. The frequency is computed solely from the WALS corpus using the information from 2,679 languages. The overall accuracies for all the WALS features computed over the in-domain ($I$) and held-out ($H$) test sets are shown in Table 2. The results for the second baseline correspond to the lower bound on the capacity of our network to learn by chance. In this scenario, before training the network the input acoustic parameters in our training set are sampled from a normal distribution with zero mean and unit standard deviation. The network is then tested against original unpermuted test sets. This baseline significantly outperforms the majority class approach, which implies that the network is learning *something* about the sparse WALS labels from the totally corrupt data. The result is intriguing because robustness property has so far been only confirmed to the extremely noisy labels rather than features [36]. Since in our models we batch-normalize the inputs, we hypothesize that this result corresponds to the lower bound of what can be learned from the inputs scaled to zero mean and unit variance.

Table 2: *Baseline system accuracies.*

| Baseline Type | Accuracy (%) | |
|---|---|---|
| | In-domain ($I$) | Held-out ($H$) |
| Majority class (WALS) | 54.4 | 60.1 |
| Randomized inputs | 71.8 | 68.0 |

### 3.4. WALS chapter types

For the following experiments we introduce an additional baseline system in order to investigate whether our feature predictor is simply memorizing the properties of language identification. This baseline, denoted LANGID, consists of the language identification network trained using an unweighted softmax cross-entropy loss on the same training data and similar topology to WALS feature predictor. The system is evaluated on the in-domain ($I$) and held-out ($H$) languages by using the predicted language code as a lookup key into the WALS features. The results are shown in Table 3. This baseline is evaluated on the same data against WALS feature predictor, denoted $M$. All the results are aggregated by the WALS feature chapter types, where, for each of the four evaluations, the total number of comparisons between the predicted and ground-truth values ($N_e$), accuracy ($A$), precision ($P$), recall ($R$) and $F_1$ scores are displayed. The overall feature prediction accuracy (rather than the average over chapter types) is displayed at the bottom row of the table.

As can be seen from the table, the feature predictor $M$ outperforms language identification-based approach LANGID on both the in-domain and held-out languages, leading us to conclude that in our case the language identification approach has simply not enough signal to robustly discriminate between languages and, as a result, their features, which is especially true for the held-out scenario. The network $M$ also strongly outperforms the majority class baseline from Table 2. In the held-out scenario, the improvement over the network trained on randomized data amounts to 3.1%, which is not as high as was originally hoped and indicates an overall model confusion when faced with languages not seen during the training. The highest accuracy (76.5%) on the held-out languages is obtained for the WALS features belonging to the phonological chapter. This implies that, compared to other types of features, rather unsurprisingly the model has greater leeway to utilize the signals present in speech. The second and third most accurate chapters ("Word

Table 3: *Metrics for WALS features grouped by WALS chapter type.*

| Chapter Type | $M(I)$ | | | | | LangID($I$) | | | | $M(H)$ | | | | | LangID($H$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N_e$ | $A$ (%) | $P$ | $R$ | $F_1$ | $A$ (%) | $P$ | $R$ | $F_1$ | $N_e$ | $A$ (%) | $P$ | $R$ | $F_1$ | $A$ (%) | $P$ | $R$ | $F_1$ |
| Complex Sentences | 40,739 | 87.2 | 0.87 | 0.88 | 0.87 | 84.4 | 0.71 | 0.89 | 0.79 | 49,006 | 68.3 | 0.77 | 0.55 | 0.65 | 36.7 | 0.35 | 0.58 | 0.43 |
| Lexicon | 51,975 | 88.1 | 0.78 | 0.80 | 0.79 | 81.8 | 0.79 | 0.85 | 0.82 | 61,401 | 66.9 | 0.42 | 0.44 | 0.43 | 23.6 | 0.34 | 0.59 | 0.43 |
| Morphology | 77,679 | 91.2 | 0.87 | 0.89 | 0.88 | 79.4 | 0.79 | 0.84 | 0.81 | 119,664 | 71.6 | 0.49 | 0.50 | 0.49 | 29.0 | 0.25 | 0.47 | 0.33 |
| Nominal Categories | 23,189 | 89.1 | 0.89 | 0.86 | 0.88 | 80.1 | 0.84 | 0.83 | 0.83 | 326,369 | 67.4 | 0.55 | 0.52 | 0.53 | 32.6 | 0.31 | 0.43 | 0.36 |
| Nominal Syntax | 63,564 | 91.6 | 0.86 | 0.89 | 0.87 | 80.4 | 0.78 | 0.79 | 0.78 | 104,968 | 71.0 | 0.63 | 0.47 | 0.54 | 41.1 | 0.37 | 0.40 | 0.33 |
| Other | 1,917 | 100.0 | 1.00 | 1.00 | 1.00 | 85.6 | 0.77 | 0.79 | 0.78 | - | - | - | - | - | - | - | - | - |
| Phonology | 140,124 | 91.4 | 0.82 | 0.87 | 0.84 | 79.6 | 0.81 | 0.82 | 0.81 | 230,796 | 76.5 | 0.50 | 0.60 | 0.54 | 36.9 | 0.29 | 0.41 | 0.34 |
| Sign Languages | 4,198 | 97.2 | 0.98 | 0.97 | 0.97 | 88.3 | 0.66 | 0.93 | 0.77 | 6,640 | 14.1 | 0.61 | 0.42 | 0.50 | 7.2 | 0.59 | 0.38 | 0.46 |
| Simple Clauses | 192,493 | 91.1 | 0.89 | 0.79 | 0.84 | 79.9 | 0.82 | 0.79 | 0.80 | 274,105 | 74.3 | 0.64 | 0.63 | 0.63 | 39.0 | 0.37 | 0.36 | 0.36 |
| Verbal Categories | 172,194 | 87.3 | 0.85 | 0.74 | 0.79 | 83.9 | 0.79 | 0.83 | 0.81 | 220,197 | 65.3 | 0.44 | 0.58 | 0.50 | 38.8 | 0.21 | 0.54 | 0.30 |
| Word Order | 361,012 | 91.2 | 0.90 | 0.88 | 0.89 | 84.6 | 0.87 | 0.86 | 0.86 | 461,693 | 75.0 | 0.65 | 0.55 | 0.60 | 43.5 | 0.36 | 0.56 | 0.44 |
| *Total* | 1,129,084 | **91.6** | | | | 82.6 | | | | 1,854,839 | **71.1** | | | | 36.3 | | | |

Table 4: *Selected features from the "Phonology" area.*

| Id | $R$ | Name | $N_e$ | Accuracy (%) |
|---|---|---|---|---|
| 7A | 9 | Glottalized Consonants | 9,211 | 89.5 |
| 10A | 10 | Vowel Nasalization | 4,631 | 88.9 |
| 5A | 11 | Voicing and Gaps in Plosive Systems | 9,211 | 88.4 |
| 11A | 14 | Front Rounded Vowels | 9,211 | 86.8 |
| 8A | 36 | Lateral Consonants | 8,220 | 66.4 |
| 12A | 48 | Syllable Structure | 8,214 | 59.7 |

Table 5: *WALS feature accuracies aggregated by language.*

| In-domain $M(I)$ | | | Held-out $M(H)$ | | |
|---|---|---|---|---|---|
| Language | $N_e$ | Accuracy (%) | Language | $N_e$ | Accuracy (%) |
| Latvian | 127,836 | 98.3 | Romanian | 640 | 82.0 |
| Japanese | 135,447 | 97.0 | Russian | 132,957 | 76.5 |
| Nepali | 61,620 | 95.1 | Swedish | 72,450 | 74.9 |
| Hebrew | 139,731 | 95.1 | Lithuanian | 75,316 | 73.1 |
| Czech | 53,070 | 92.5 | Sinhala | 41,538 | 70.7 |
| Arabic (Egypt) | 27,390 | 92.1 | Maithili | 18,642 | 67.4 |
| French | 150,416 | 91.5 | Kannada | 109,909 | 65.9 |
| Greek | 142,135 | 91.1 | English (Nigeria) | 147,339 | 60.4 |
| Hindi | 132,912 | 89.9 | Basque | 127,380 | 57.2 |
| Bulgarian | 77,364 | 88.3 | Telugu | 52,140 | 54.3 |
| Danish | 54,694 | 87.9 | Amharic | 70,125 | 52.4 |
| Italian | 78,084 | 83.3 | Bengali | 48,363 | 51.5 |
| German | 95,732 | 75.9 | Marathi | 52,140 | 50.9 |
| Dutch | 85,885 | 75.7 | Lao | 39,640 | 50.7 |
| Korean | 69,285 | 72.8 | Gujarati | 33,520 | 50.3 |
| Spanish | 962 | 60.0 | Urdu | 34,727 | 49.7 |
| | | | Burmese | 109,450 | 47.5 |
| | | | Tamil | 72,048 | 47.3 |
| | | | Malayalam | 60,099 | 44.2 |

Order" and "Simple Clauses") are related to syntax. The result for these two chapters is likely explained by the relatively low class sparsity of the features in these two chapters in our training data, which results in a model which is relatively robust in the held-out scenario.

### 3.5. Individual WALS features

Table 4 shows the six most accurate phonological features predicted for the held-out languages. For each feature, the corresponding WALS feature identifier (Id), its accuracy rank ($R$) among 192 features, name and the number of predictions ($N_e$) are shown. The "Glottalized Consonants" feature tops the table. It's interesting to note, that this feature is only defined for Korean (as "Ejectives only") and explicitly set to "None" for the rest of the languages. Hence, to correctly predict this feature, the network essentially learns to distinguish Korean speech from the rest. Similarly, the "Vowel Nasalization" is only defined for French and Hindi (as "Contrast present"), and the network learns to contrast this class with the majority class ("None"). Given this analysis, we hypothesize that these two features are the easiest to learn from our data.

### 3.6. Aggregation by languages

Table 5 shows feature prediction accuracies aggregated by language for both 16 in-domain and 19 held-out languages, where $N_e$ is the overall number of feature predictions performed for each language. The list of held-out languages is topped by the Indo-European languages for which a related language is available for training (e.g., Romanian and Italian, Lithuanian and Latvian). Rather unsurprisingly, the tonal languages Burmese and Lao do not get accurate predictions because no related languages are present during the training and F0 information is not used in acoustic features. Predictions for Dravidian languages (with a surprising exception of Kannada) are also totally inaccurate. The model is very inaccurate for Indo-Aryan languages apart from Sinhala and Maithili, which seem to benefit from the presence of Nepali and Hindi in the training data.

### 3.7. Alternatives considered

Since WALS features are non-exclusive but their values corresponding to our predictions are not, we can only hope that the network described above learns that within each feature its values are independent. To address this potential shortcoming we tested an alternative multi-head strategy which constrains the universe of predicted values for each individual WALS feature to be mutually independent. We broke the problem down into 192 tasks, one for each WALS feature, effectively replacing a single sigmoid cross-entropy loss in equation (1) with the linear combination of weighted softmax cross-entropy loss functions $L(\boldsymbol{\theta}_i)$ corresponding to each feature. The weight function for the observed classes $c^i$ for label $i$ is defined as $N^i_{\mathcal{L}}/N^i_c$, where $N^i_{\mathcal{L}}$ denotes the count of a WALS feature in the training data and $N^i_c$ is the class count. This model performed slightly worse than our default model described in this paper.

## 4. Conclusions and Future Work

In this study we approached the problem of predicting the very sparsely populated WALS features from speech as a multi-label classification problem. We have shown that a reasonably standard recurrent neural network utilizing a reciprocal class frequency weighting optimization loss significantly outperforms the language identification-based feature lookup approach in both in-domain and held-out scenarios. The model generalizes well over the languages observed during the training. Although it is evident that the network can generalize over unseen languages as well (the accuracies are above the majority class, randomized and language identification-based lookup approaches), the predictions for individual features, groups of features and language-based aggregation are not very accurate. On the one hand, this is disappointing because the focus of this work is on the unseen language typology prediction. On the other hand, there is significant room for several improvements: First, our training corpus is proprietary, very small and does not cover some of the major language families (e.g., Tai-Kadai). Second, the mobile speech is used as is, without any speaker, volume or noise normalization. Third, MFCC are not the best acoustic representation for hybrid CNN-LSTMs. Finally, the WALS labels may be too sparse for our task. Filling the gaps using an approach suggested in [26] will be beneficial.

# 5. References


[1] J. J. Song, *The Oxford Handbook of Linguistic Typology*. Oxford University Press, 2013.

[2] E. Asgari and H. Schütze, "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*. Association for Computational Linguistics, 2017, pp. 113–124.

[3] H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, and A. Korhonen, "Survey on the Use of Typological Information in Natural Language Processing," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, Japan, 2016, pp. 1297–1308.

[4] S. Moran, D. McCloy, and R. Wright, *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2014. [Online]. Available: http://phoible.org/

[5] H. Hammarström, S. Bank, R. Forkel, and M. Haspelmath, "Glottolog 3.2," Max Planck Institute for the Science of Human History, Jena, 2018. [Online]. Available: http://glottolog.org

[6] D. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors." in *Proc. COLING 2016: 26th International Conference on Computational Linguistics*, Japan, December 2016, pp. 3475–3484. [Online]. Available: https://github.com/dmort27/panphon/

[7] B. Peters, J. Dehdari, and J. van Genabith, "Massively Multilingual Neural Grapheme-to-Phoneme Conversion," in *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Denmark: Association for Computational Linguistics, 2017, pp. 19–26.

[8] Y. Tsvetkov, S. Sitaram, M. Faruqui, G. Lample, P. Littell, D. Mortensen, A. W. Black, L. Levin, and C. Dyer, "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*. San Diego, California: Association for Computational Linguistics, 2016, pp. 1357–1366.

[9] W. Ammar, G. Mulcaire, M. Ballesteros, C. Dyer, and N. Smith, "Many languages, one parser," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 431–444, 2016.

[10] M. S. Dryer and M. Haspelmath, "WALS online," *Leipzig: Max Planck Institute for Evolutionary Anthropology*, 2013. [Online]. Available: http://wals.info

[11] H. Li, B. Ma, and K. A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[12] M. Lorvik, "Mutual intelligibility of timber trade terminology in the North Sea countries during the time of the 'Scottish Trade'," *Nordic Journal of English Studies*, vol. 2, no. 2, pp. 223–243, 2003.

[13] D. Heddle, "The Norse element in the Orkney dialect," *Northern Lights, Northern Words: Selected papers from the FRLSU Conference*, pp. 48–57, 2010.

[14] M. S. Dryer, "Order of negative morpheme and verb," in *The World Atlas of Language Structures Online*, M. S. Dryer and M. Haspelmath, Eds. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. [Online]. Available: http://wals.info/chapter/143

[15] E. Gibaja and S. Ventura, "Multi-label learning: a review of the state of the art and ongoing research," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444, 2014.

[16] M. Zhang and Z. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[17] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic Language Identification using Long Short-Term Memory Recurrent Neural Networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2155–2159.

[18] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[19] Y. Xiao and K. Cho, "Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers," *arXiv preprint arXiv:1602.00367*, February 2016.

[20] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks." in *Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, ser. JMLR Proceedings, vol. 15, 2011, pp. 315–323.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005.

[24] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[26] C. Malaviya, G. Neubig, and P. Littell, "Learning Language Representations for Typology Prediction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Denmark: Association for Computational Linguistics, 2017, pp. 2529–2535.

[27] P. Littell, D. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "URIEL and `lang2vec`: Representing languages as typological, geographical, and phylogenetic vectors," in *Proc. EACL 2017*, Spain, April 2017, pp. 8–14.

[28] B. Comrie, "Maltese and the World Atlas of Language Structures," *Selected papers from the 1st International Conference on Maltese Linguistics*, pp. 3–12, October 2009.

[29] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[30] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[31] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[33] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Japan, September 2010, pp. 1914–1917.

[34] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task," in *Proceedings of the SPECOM*, vol. 1, 2005, pp. 191–194.

[35] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[36] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep Learning is Robust to Massive Label Noise," *arXiv preprint arXiv:1705.10694*, 2017.