NUIG: Multitasking Self-attention based approach to SigTyp 2020 Shared Task

Chinmay Choudhary
National University of Ireland, Galway
c.choudhary1@nuigalway.ie

Colm O'Riordan
National University of Ireland, Galway
colm.oriordan@nuigalway.ie

Abstract

The paper describes the *Multitasking Self-attention based approach* to constrained sub-task within Sigtyp 2020 Shared task. Our model is simple neural network based architecture inspired by Transformers (Vaswani et al., 2017) model. The model uses Multitasking to compute values of all WALS features for a given input language simultaneously.
Results show that our approach performs at par with the baseline approaches, even though our proposed approach requires only phylogenetic and geographical attributes namely *Longitude*, *Latitude*, *Genus-index*, *Family-index* and *Country-index* and do not use any of the known WALS features of the respective input language, to compute its missing WALS features.

1 Introduction

In this paper we describe our *Multitasking Self-attention based approach* to Sigtyp 2020 Shared task (Constrained Sub-task) (Bjerva et al., 2020) which involves prediction of values of features from WALS Typology database for various low-resourced languages.
Linguistic typology is the classification of human languages according to their syntactic, phonological and semantic features. WALS (Haspelmath, 2009) is the most popular and comprehensive database which provides list of typological features and their possible values, as well as the respective feature-values for most of the world's languages.
However all the popular typological databases (Haspelmath, 2009; Collins and Kayne, 2009; Maddieson et al., 2013; Hartmann and Bradley Taylor, 2013; Bickel et al., 2017; Michaelis and Magnus Huber, 2013) (including WALS) suffer from a major shortcoming of limited coverage. In fact, values of many important typological features for most languages (specially less documented ones) are missing in these databases. This sparked a line of research on automatic acquisition of such missing typology knowledge (Malaviya et al., 2017; Coke et al., 2016; Daumé III, 2009; Daumé III and Campbell, 2009; Littell et al., 2017; Bjerva et al., 2019).
Our proposed model is a neural network architecture which takes in as input, the phylogenetic and geographical attributes of a language. The model subsequently predicts values of all its typology features simultaneously using Multitask learning setup (Ruder, 2017).

2 Model

Figure 1 depicts the architecture of our proposed model that computes values of all WALS typology features for a given language simultaneously. As evident in Figure 1, our proposed model architecture comprises of three components namely *Input Network Component*, *Self-attention Network Component* and *Multitasking Output Networks Component* described as section 2.1, 2.2 and 2.3 respectively.

2.1 Input Network Component

The input component is a simple two layered feed-forward neural network. The input of the network is a 5-dimensional vector x comprising of values of five key attributes of any language, namely *Longitude*, *Latitude*, *Genus-index*, *Family-index* and *Country-index* as these are the attributes provided by train and test datasets (for all languages within the datasets) for Sigtyp 2020 Shared Task. We computed Genus-index, Family-index and Country-index from *genus*, *family* and *countryCode* attributes provided within dataset using respective name-index dictionaries.
This two layered feed forward network computes output vector $o \in R^{T*d}$ where T is the total number of WALS typology features to be predicted by

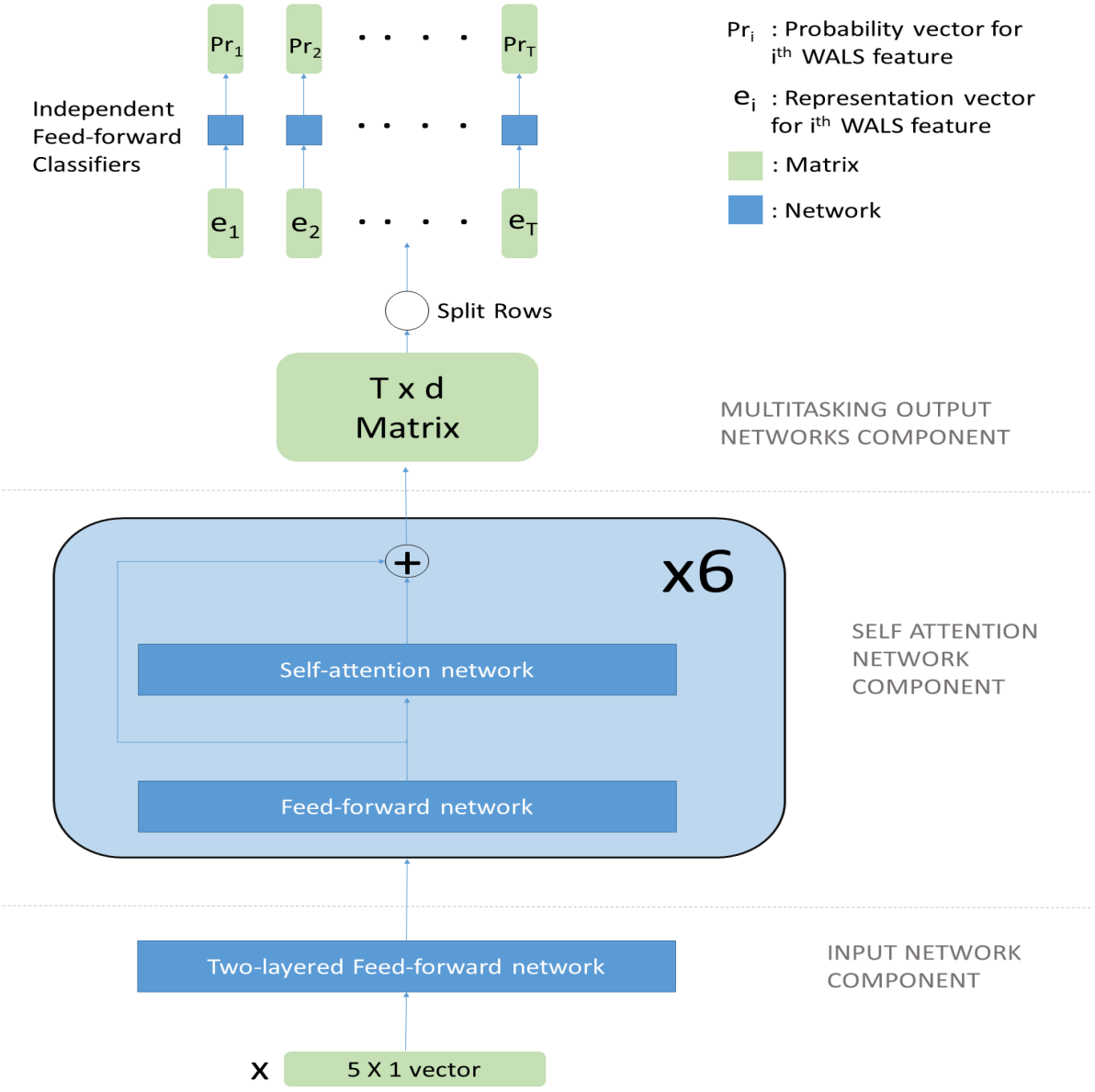

Figure 1: Architecture of proposed model

applying equations 1 and 2.

$$\hat{o} = tanh(A_1 * x + a_1) \tag{1}$$

$$o = tanh({A_2}^T * \hat{o} + a_2) \tag{2}$$

Here $A_1 \in R^{d*5}$, $A_2 \in R^{T*1}$ are weights and $a_1 \in R^d$ and $a_2 \in R^{T*d}$ are biases.

2.2 Self-attention Network Component

The architecture of this component is inspired by Transformers (Vaswani et al., 2017) model. The model architecture comprises a stack of $N = 6$ identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple fully connected feed-forward network. Hence input to layer i is always the output from layer $i-1$. Input to the first layer is the output of the previous Input Network Component.
For i^{th} layer within architecture, its *Feed-forward* and *self-attention* sub-layers are given by equations 3 and 4.

$$h_i = tanh(W_i * y_{i-1} + b_i) \tag{3}$$

$$k_i = attention(h_i, h_i) \tag{4}$$

Here $h_i \in R^d$ and $k_i \in R^d$ are outputs of *feed-forward* and *self-attention* layers respectively. We used same attention mechanism as used by (Vaswani et al., 2017). Final output of i^{th} layer

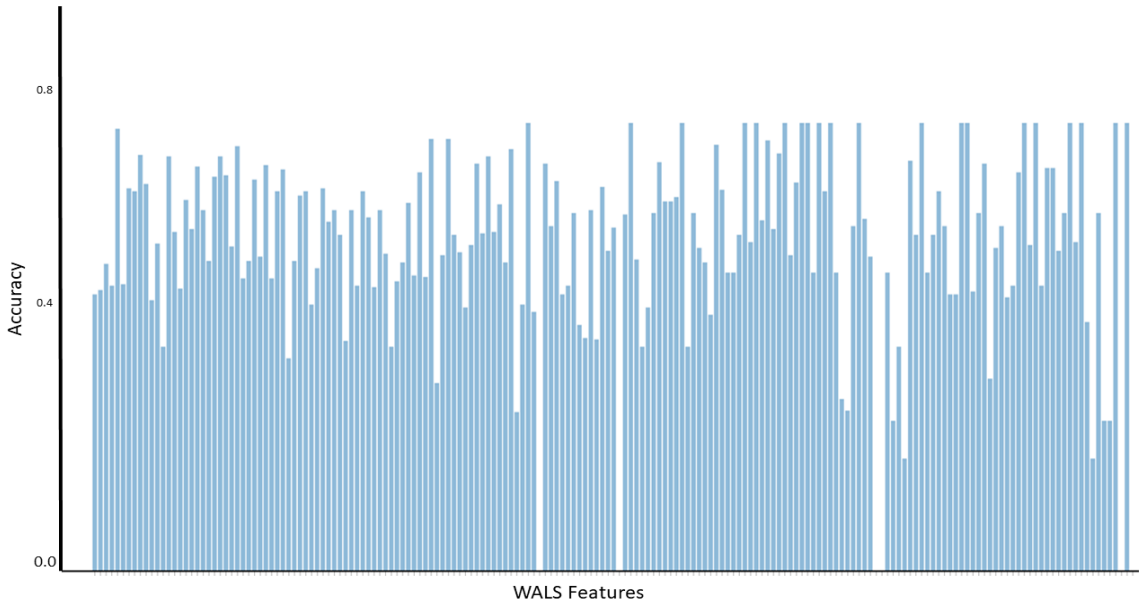

Figure 2: Plot depicting trend in accuracy values achieved on all WALS features

y_i is computed by adding h_i and k_i (equation 3).

$$y_i = h_i + k_i \tag{5}$$

Input to the first layer y_0 is the output from previous *Input Network Component*. Output of *Self-attention Network Component* is the output of final layer y_N.

It is been observed that there is a correlation between various WALS typology features. Thus, to predict the missing value of a particular typology feature for a specific languages, knowledge about other typology features for that languages would be useful. Such knowledge is ensured by the self-attention layers.

2.3 Multitasking Output Networks Component

Multitasking Output Networks Component comprises of T independent feed-forward neural-network classifiers. The component splits the output of previous Self-attention Network Component i.e $y_6 \in R^{T*d}$ into T d-dimensional vectors $e_1, e_2, \ldots . e_T$. each corresponds to one of the T typology features to be predicted.

Value of the j^{th} typology feature is computed by applying equation 6.

$$Pr_j = Softmax(W_j * e_j + c_j) \tag{6}$$

Here $1 <= j <= T$, Pr_j provides probability of each of the possible values for j^{th} typology feature being the true-value. Dimensions of weights and biases are unique for each classifier as number of possible values for each of the typology features is unique.

Hyper-parameter	**Value**
d	548
drop_out probability	0.1
learning_rate	0.1
reduce lr on plateau	Yes
reduce factor	0.001
batch-size	20
steps-per-epoch	50
epochs	200
Number of features (T)	185

Table 1: Hyper-parameters

3 Training

The parameters of model described in section 2 are trained by optimizing the loss function given by equation 7.

$$Loss = \Sigma_{t=1}^{T} CE(Pr_t, OH_t) \tag{7}$$

Here OH_t is the one-hot encoding of true-value for t^{th} typology feature. CE is the Cross-entropy loss. Table 1 lists the hyper-parameters used during training. These are computed by minimizing the loss over Validation set.

Model	Accuracy
frequency-baseline_constrained	0.514
knn-imputation-baseline_constrained	0.508
NUIG_constrained	0.487

Table 2: Overall Accuracy of baseline and proposed models

4 Results

Table 2 compared the accuracy achieved by our proposed model with two baselines provided namely *frequency-baseline-constrained* and *knn-imputation-baseline-constrained*.

It is evident from table that our model performs at par with baselines, even though it utilizes only five attributes of the input language, namely *Longitude*, *Latitude*, *Genus-index*, *Family-index* and *Country-index* (model doesn't utilize any known WALS feature values, provided within test dataset for various languages).

Figure 2 is bar-plot that depicts the trend in accuracy achieved by our model on various WALS features. Precise accuracy score achieved by our model on all 185 WALS typology features is provided in Appendix.

5 Conclusion

In this paper we described our *Multitasking Self-attention based approach* to Sigtyp 2020 Shared task, Constrained Sub-task. Our model is simple neural network based approach which computes values of all WALS features for a given input language simultaneously in Multitasking settings. The architecture of our network is inspired by Transformers (Vaswani et al., 2017).

Results show that our approach performs at par with the baseline approaches, even though our approach uses only five linguistic and geographical attributes namely *Longitude*, *Latitude*, *Genus-index*, *Family-index* and *Country-index* and do not use any of the known WALS features of the respective input language, to compute its missing WALS features.

References

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Fernando Hildebrandt, Kristine, and John B Lowe. 2017. The autotyp typological databases. *Version 0.1. 0. Online: https://github.com/autotyp/autotyp-data/tree/0.1.0*.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. *arXiv preprint arXiv:1903.10950*.

Johannes Bjerva, Elizabeth Salesky, Sabrina Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo M. Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 Shared Task: Prediction of Typological Features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*. Association for Computational Linguistics.

Reed Coke, Ben King, and Dragomir Radev. 2016. Classifying syntactic regularities for hundreds of languages. *arXiv preprint arXiv:1603.08016*.

Chris Collins and Richard Kayne. 2009. Syntactic structures of the world's languages. http://sswl.railsplayground.net/..

Hal Daumé III. 2009. Non-parametric bayesian areal linguistics. *arXiv preprint arXiv:0906.5114*.

Hal Daumé III and Lyle Campbell. 2009. A bayesian model for discovering typological implications. *arXiv preprint arXiv:0907.0785*.

Martin Haspelmath Hartmann, Iren and editors Bradley Taylor. 2013. Valency Patterns Leipzig.

Martin Haspelmath. 2009. *The typological database of the World Atlas of Language Structures*. Berlin: Walter de Gruyter.

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 the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Ian Maddieson, Sébastien Flavier, Egidio Marsico, Christophe Coupé, and François Pellegrino. 2013. Lapsyd: lyon-albuquerque phonological systems database. In *INTERSPEECH*, pages 3022–3026.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. *arXiv preprint arXiv:1707.09569*.

Philippe Maurer Martin Haspelmath Michaelis, Susanne Maria and editors Magnus Huber. 2013. Atlas of Pidgin and Creole Language Structures Online.

Sebastian Ruder. 2017. An overview of multitask learning in deep neural networks. corr abs/1706.05098. *arXiv preprint arXiv:1706.05098*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

A Appendix 1

Table 3: Feature-wise accuracy.

Begin of Table	
WALS Feature	Accuracy
Order_of_Person_Markers_on_the_Verb	0.4307692307692308
Order_of_Subject,_Object_and_Verb	0.43855421686746987
Order_of_Adposition_and_Noun_Phrase	0.4805970149253731
Position_of_Case_Affixes	0.4445945945945946
Minor_morphological_means_of_signaling_negation	0.6905405405405406
Position_of_Tense-Aspect_Affixes	0.5986842105263158
Order_of_Degree_Word_and_Adjective	0.5923076923076923
Postnominal_relative_clauses	0.65
Postverbal_Negative_Morphemes	0.6054054054054054
Person_Marking_on_Adpositions	0.5115384615384615
Weight_Factors_in_Weight-Sensitive_Stress_Systems	0.35
Presence_of_Uncommon_Consonants	0.6487804878048781
Preverbal_Negative_Morphemes	0.5297297297297298
Negative_Morphemes	0.44
Absence_of_Common_Consonants	0.5804878048780487
Polar_Questions	0.5345454545454545
Glottalized_Consonants	0.6317073170731707
Voicing_in_Plosives_and_Fricatives	0.5634146341463415
Position_of_Negative_Word_With_Respect_to_Subject,_Object,_and_Verb	0.4846153846153846
Passive_Constructions	0.616
Front_Rounded_Vowels	0.6487804878048781
Gender_Distinctions_in_Independent_Personal_Pronouns	0.6192307692307693
Rhythm_Types	0.506896551724138
Tone	0.6631578947368421
Position_of_Polar_Question_Particles	0.45769230769230773
Order_of_Numeral_and_Noun	0.4830985915492958
Pronominal_and_Adnominal_Demonstratives	0.6125
Fixed_Stress_Locations	0.49
Finger_and_Hand	0.6348837209302325
Verbal_Person_Marking	0.45769230769230773
Third_Person_Zero_of_Verbal_Person_Marking	0.5923076923076923
Order_of_Subject_and_Verb	0.6263157894736842
Weight-Sensitive_Stress	0.3325
Order_of_Object_and_Verb	0.48404255319148937
Order_of_Relative_Clause_and_Noun	0.586046511627907
Alignment_of_Verbal_Person_Marking	0.5923076923076923
Position_of_Pronominal_Possessive_Affixes	0.4148148148148148
Voicing_and_Gaps_in_Plosive_Systems	0.5975609756097561
Consonant-Vowel_Ratio	0.5463414634146342
Expression_of_Pronominal_Subjects	0.5630434782608695
Intensifiers_and_Reflexive_Pronouns	0.525
Consonant_Inventories	0.35853658536585364
Vowel_Quality_Inventories	0.5634146341463415
Syllable_Structure	0.44545454545454544

Continuation of Table 3	
WALS Feature	Accuracy
Order_of_Genitive_and_Noun	0.5936708860759494
Order_of_Adjective_and_Noun	0.5530864197530864
Third_Person_Pronouns_and_Demonstratives	0.4421052631578947
Lateral_Consonants	0.5634146341463415
Prefixing_vs._Suffixing_in_Inflectional_Morphology	0.49482758620689654
The_Position_of_Negative_Morphemes_in_Object-Initial_Languages	0.35
Distance_Contrasts_in_Demonstratives	0.4529411764705883
Order_of_Negative_Morpheme_and_Verb	0.4824324324324324
Hand_and_Arm	0.574
Uvular_Consonants	0.4609756097560976
Position_of_Interrogative_Phrases_in_Content_Questions	0.6234375000000001
SONegV_Order	0.45862068965517244
NegSOV_Order	0.6758620689655173
The_Position_of_Negative_Morphemes_in_SOV_Languages	0.2930232558139535
The_Associative_Plural	0.49411764705882355
SNegOV_Order	0.6758620689655173
Order_of_Adverbial_Subordinator_and_Clause	0.525
The_Prohibitive	0.49677419354838714
SOVNeg_Order	0.4117647058823529
Coding_of_Nominal_Plurality	0.5101694915254237
The_Morphological_Imperative	0.6363636363636364
Order_of_Demonstrative_and_Noun	0.5272727272727273
Comitatives_and_Instrumentals	0.6481481481481481
Ditransitive_Constructions:_The_Verb_'Give'	0.5303030303030303
'Want'_Complement_Subjects	0.5727272727272728
Order_of_Object,_Oblique,_and_Verb	0.48125
Noun_Phrase_Conjunction	0.6588235294117647
Predicative_Possession	0.24705882352941178
Definite_Articles	0.41621621621621624
Languages_with_two_Dominant_Orders_of_Subject,_Object,_and_Verb	0.7
Zero_Copula_for_Predicate_Nominals	0.40384615384615385
Languages_with_different_word_order_in_negative_clauses	0.0
Tea	0.6363636363636364
Nominal_and_Locational_Predication	0.5384615384615385
Predicative_Adjectives	0.4307692307692308
Nominal_and_Verbal_Conjunction	0.44545454545454544
Cultural_Categories_of_Languages_with_Identity_of_'Finger'_and_'Hand'	0.56
Inclusive/Exclusive_Forms_in_Pama-Nyungan	0.385
Occurrence_of_Nominal_Plurality	0.364
Indefinite_Pronouns	0.564516129032258
Indefinite_Articles	0.3612903225806452
The_Optative	0.6
Ordinal_Numerals	0.5
Semantic_Distinctions_of_Evidentiality	0.5352941176470588
Multiple_Negative_Constructions_in_SOV_Languages	0.0
Coding_of_Evidentiality	0.5558823529411765
Politeness_Distinctions_in_Pronouns	0.7
Systems_of_Gender_Assignment	0.4869565217391304

Continuation of Table 3	
WALS Feature	Accuracy
Locus_of_Marking:_Whole-language_Typology	0.35
Asymmetrical_Case-Marking	0.4117647058823529
M_in_First_Person_Singular	0.56
Sex-based_and_Non-sex-based_Gender_Systems	0.6391304347826087
Number_of_Cases	0.5764705882352941
Reduplication	0.5764705882352941
Numeral_Classifiers	0.5833333333333334
Number_of_Possessive_Nouns	0.7
Applicative_Constructions	0.35
M_in_Second_Person_Singular	0.56
Locus_of_Marking_in_Possessive_Noun_Phrases	0.5055555555555555
Possessive_Classification	0.48125
Other_Roles_of_Applied_Objects	0.4
M-T_Pronouns	0.665
N-M_Pronouns	0.595
Locus_of_Marking_in_the_Clause	0.4666666666666667
Adjectives_without_Nouns	0.4666666666666667
Antipassive_Constructions	0.525
Zero_Marking_of_A_and_P_Arguments	0.7
Productivity_of_the_Antipassive_Construction	0.5133333333333333
Obligatory_Possessive_Inflection	0.7
Number_of_Genders	0.5478260869565218
Nonperiphrastic_Causative_Constructions	0.6719999999999999
Plurality_in_Independent_Personal_Pronouns	0.5333333333333333
Purpose_Clauses	0.6533333333333333
Imperative-Hortative_Systems	0.7
Vowel_Nasalization	0.49411764705882355
Prenominal_relative_clauses	0.6066666666666667
SVNegO_Order	0.7
SVONeg_Order	0.7
Number_of_Non-Derived_Basic_Colour_Categories	0.4666666666666667
Red_and_Yellow	0.7
NegSVO_Order	0.5923076923076923
Green_and_Blue	0.7
Number_of_Basic_Colour_Categories	0.4666666666666667
SNegVO_Order	0.2692307692307693
The_Position_of_Negative_Morphemes_in_SVO_Languages	0.25
Negative_Indefinite_Pronouns_and_Predicate_Negation	0.5384615384615385
Utterance_Complement_Clauses	0.7
'When'_Clauses	0.55
The_Velar_Nasal	0.49
Optional_Double_Negation_in_SOV_languages	0.0
Optional_Double_Negation	0.0
Para-Linguistic_Usages_of_Clicks	0.4666666666666667
Distributive_Numerals	0.23333333333333334
Verb-Initial_with_Preverbal_Negative	0.35
The_Position_of_Negative_Morphemes_in_Verb-Initial_Languages	0.175
Reason_Clauses	0.6416666666666666

Continuation of Table 3	
WALS Feature	Accuracy
Verb-Initial_with_Clause-Final_Negative	0.7
Alignment_of_Case_Marking_of_Pronouns	0.4666666666666667
Alignment_of_Case_Marking_of_Full_Noun_Phrases	0.525
Inclusive/Exclusive_Distinction_in_Verbal_Inflection	0.5923076923076923
Syncretism_in_Verbal_Person/Number_Marking	0.5384615384615385
Numeral_Bases	0.4307692307692308
Case_Syncretism	0.4307692307692308
Inclusive/Exclusive_Distinction_in_Independent_Pronouns	0.7
Relativization_on_Subjects	0.7
Action_Nominal_Constructions	0.4375
Periphrastic_Causative_Constructions	0.56
Situational_Possibility	0.6363636363636364
Relativization_on_Obliques	0.3
Symmetric_and_Asymmetric_Standard_Negation	0.5055555555555555
Epistemic_Possibility	0.5384615384615385
Subtypes_of_Asymmetric_Standard_Negation	0.4277777777777778
Overlap_between_Situational_and_Epistemic_Modal_Marking	0.44545454545454544
Comparative_Constructions	0.6222222222222222
The_Future_Tense	0.7
Perfective/Imperfective_Aspect	0.5090909090909091
The_Perfect	0.7
The_Past_Tense	0.44545454545454544
Suppletion_in_Imperatives_and_Hortatives	0.63
Exponence_of_Selected_Inflectional_Formatives	0.63
Genitives,_Adjectives_and_Relative_Clauses	0.5
Verbal_Number_and_Suppletion	0.56
Suppletion_According_to_Tense_and_Aspect	0.7
Reciprocal_Constructions	0.5133333333333333
Inflectional_Synthesis_of_the_Verb	0.7
Fusion_of_Selected_Inflectional_Formatives	0.38888888888888895
Conjunctions_and_Universal_Quantifiers	0.175
Exponence_of_Tense-Aspect-Mood_Inflection	0.56
Obligatory_Double_Negation_in_SOV_languages	0.23333333333333334
Obligatory_Double_Negation	0.23333333333333334
Multiple_Negative_Constructions_in_SVO_Languages	0.7
Double_negation_in_verb-initial_languages	0.0
Internally-headed_relative_clauses	0.7
Optional_Double_Negation_in_SVO_languages	0.0
End of Table	