KMI-Panlingua-IITKGP @SIGTYP2020: Exploring Rules and Hybrid Systems for Automatic Prediction of Typological Features

Ritesh Kumar[1], Deepak Alok[2], Akanksha Bansal[3],
Bornini Lahiri[4], Atul Kr. Ojha[5,3],
[1]Dr. Bhimrao Ambedkar University, Agra, [2]Rutgers University, USA,
[5]NUIG, Galway, [3]Panlingua Language Processing LLP, New Delhi,
[4]Indian Institute of Technology, Kharagpur,
ritesh78_llh@jnu.ac.in, bornini@hss.iitkgp.ac.in
(deepak06alok,akanksha.bansal15,shashwatup9k)@gmail.com

Abstract

This paper enumerates SigTyP 2020 Shared Task on the prediction of typological features as performed by the KMI-Panlingua-IITKGP team. The task entailed the prediction of missing values in a particular language, provided, the name of the language family, its genus, location (in terms of latitude and longitude coordinates and name of the country where it is spoken) and a set of feature-value pair are available. As part of fulfillment of the aforementioned task, the team submitted 3 kinds of system - 2 rule-based and one hybrid system. Of these 3, one rule-based system generated the best performance on the test set. All the systems were 'constrained' in the sense that no additional dataset or information, other than those provided by the organisers, was used for developing the systems.

1 Introduction

This paper is a detailed documentation of the KMI-Panlingua-IITKGP team's system submission at the SigTyP 2020 Shared Task on the prediction of typological features. The objective behind this task is to develop a computational model that predicts (missing) linguistic features of a language, given its location, language family, genus, and a set of feature-value pair. The shared task organisers provided the dataset used for this purpose, which has been extracted from Worlds Atlas of Language structures (WALS) (see section 2 for details). Since the provided dataset was not large and comprised of unevenly distributed features, we prepared and experimented with three different systems and compare them with each other to provide the best model. Of these three systems, 2 are rule-based, in which, one is frequency-based system (see subsection 3.1) and the other is statistical system (see subsection 3.2). The third one is a hybrid (see subsection 3.3). The statistical system provides the best results among the three (see section 4). This paper promises two major contributions. First, it provides an automatic system that enables extraction of feature-value pair of a given language - a tedious job if done manually. Second, it compares three different systems and provides evidence that a statistical model gives better results for the given data set.

2 Dataset

The dataset[1] used for this experiment was extracted from World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). It covered the typological features of close to 2,000 languages (Bjerva et al., 2020). These typological features were organised in 8 columns (including Language ID, Language name, Latitude, Longitude, Genus, Family, Country

[1] https://github.com/sigtyp/ST2020/tree/master/data

Codes, and feature-value. The feature values were separated by '|' for each language). The task was divided into two sub-tasks: (a) Constrained and (b) Unconstrained. The dataset was sub-divided into training, dev, and test sets. Out of 1,125 training instances, Genus had 280 different features, Country had 258 different features, Family had 127 different features and Features had 185 different feature-values. Out of 84 dev instances, Genus had 50 different features, Country had 32, Family had 36, while Features had 182 different feature-values. Out of 149 test instances, Genus had 42 different features, Country had 30, Family had 35, while Features had 183 different feature-values (see Table 1).

	Genus	**Family**	**Country**	**Features**
Training	280	127	258	185
Dev	50	36	32	182
Test	42	35	30	183

Table 1: Statistics of the dataset at Genus, Family, Country and Features

3 Experiments

Our task consisted of 3 experiments and this section enumerates and discusses each one of the systems in detail. All the 3 systems are based on the notion of shared typological properties of languages belonging to the same language family (or sub-family, represented as genus in the dataset) and shared areal properties of languages belonging to different language families but being in regular contact by virtue of being in close contact, mainly because of speakers residing in close geographical proximity.

3.1 Baseline System

Our baseline system is a frequency-based system that makes use of the language family and genus-based typological properties to predict the grammatical features of a given language. In the training phase, for each feature, frequency of each of its value in each of the language family and genus is calculated and stored. During the prediction, for a specific feature in a given language (under the given language family and genus), the feature value with the maximum frequency within that genus (or language family, if the genus does not occur in the training data) is predicted as the value for the concerned feature. If neither the language family nor the genus occurs during the training phase then a default value of the feature is predicted by the system.

3.2 Statistical System

The statistical system is an extension of our baseline system where the absence of both the language family and the genus in the training data is handled in a more principled way. In such cases, we made use of the 'distance' between two languages to decide on the feature values. The training phase for this system is exactly the same as that of our baseline system. However, during prediction, the following steps are taken -

1. Step 1: As in the case of the baseline system, if the genus of the language whose feature is to be predicted is seen during the training phase and if the feature that is to be predicted was seen in that genus during the training phase then the value that was most frequent for that genus-feature combination is predicted as the value for the current case. In case the genus of the language or the concerned feature was not seen in the genus then the same step is carried out with the language family. If neither genus-feature nor family-feature combination is seen in the training phase then we move to Step 2.

2. Step 2: In the second step, based on the latitude and longitude position of the

given language, the Haversine distance between the language for which the feature value is to be predicted, and all the other languages in the training set, is calculated. Then we take the language families and genus of the four closest languages. We look at the frequency of each feature value across these four closest language families and the value with the maximum frequency is predicted as the correct feature value. The choice of four closest language families is established experimentally, by looking at numbers from 1 - 10 and deciding on the basis of best performance with the train set. If the feature is not found in these four closest language families then we move to step 3.

3. Step 3: In the third step, we look at the closest language family and genus which has the feature that we are trying to predict. The system takes the feature value with the maximum frequency and predict that, as the value for the feature. In this and the previous step, it is to be noted that each feature may have multiple values in a specific language family and genus - as such the value which occurs in the maximum number of languages of that family and genus is the one that is considered most frequent and, hence, predicted.

3.3 Hybrid System

For the hybrid system, we trained 180 different classifiers for the 180 features, which were present in the training set. Since it was not a huge dataset, with quite uneven distribution of each features and the features for training were also limited, we used SVM (Pedregosa et al., 2011), (Buitinck et al., 2013) for training each of the classifier. We experimented with different c-values from 0.001 - 10. For each feature, there were 1,100 training data points (each data point for each language in the dataset). We used the normalised Haversine distance (calculated using the coordinates), language family, genus, country and the other 179 linguistic features as features for training the classifiers. All the features not listed for a specific language was considered absent in that language; otherwise its assigned value was used for training. As mentioned earlier, since the dataset was imbalanced and some features were adequately represented while others occurred only a few times in the dataset, the performance of the classifiers accordingly varied from approx 0.30 - 0.98 (F-score). Clearly, it would not have been possible to use the classifiers that performed too low. As such we decided to use only those classifiers which had an F-score of 0.6 or above; for the other features, the statistical method (outlined in the previous section) was employed. This F-score was again experimentally deduced by looking at the best performance for multiple systems ranging from an all classifier-based system to using only those classifiers with 0.9 or above F1 score. The performance was measured by predicting features in the train set for different languages i.e. the train accuracy was taken as the benchmark for deciding this value.

4 Results

Among the three systems, our statistical system performed the best on the test set with a micro-average F-score (see Table 2) of slightly under 0.61 (a 10-point gain over the knn-imputation baseline, 9 point gain over the frequency-based baseline). The hybrid system performed the worst among the three systems with a score of slightly above 0.56.

While we were expecting the hybrid system to work better than the statistical system (since we assumed that it combined the best of both worlds), in the final results, it is the statistical systems (even the most naive one) that perform better than the hybrid systems. This

Systems	Score
kmi-panlingua-iitkgp_constrained_rule	0.607
kmi-panlingua-iitkgp_constrained_hybrid	0.562
kmi-panlingua-iitkgp_constrained	0.574
frequency-baseline_constrained	0.513
knn-imputation-baseline_constrained	0.507

Table 2: Overall accuracy of the KMI-Panlingua-IITKGP Systems

shows that even in those cases where SVMs have performed reasonably well, the statistical systems have performed equally well or better than the SVMs. There are two takeaways from this -

- **Size of Datatset:** For some of the features (especially those 2 or 3 values), such as absence of common consonants' or 'number of possessive nouns', SVMs seems to perform quite well. However, other features such as 'Action Nominal Constructions', which have a high number of classes (9 in this case), there were just not sufficient instances of each class to get sufficient discriminating features for adequate classification. One way of handling this could be by looking at the correlation among features and giving higher weightage to the features that are more likely to co-occur with each other.

- **Typological and Areal Features:** The typological and areal features are derived via systematic study of multiple languages and prior linguistic studies have shown that language families as well as geographically closer languages share certain linguistic features. The statistical system makes use of these generalisations about human language and manages to capture, at least partially, these properties of human languages. This could be one of the reasons why the statistical system performs better than the hybrid system, where sufficient information was not available to the classifier to make this kind of generalisation. This also provides some kind of explicit validity to the arguments related to the use of typological and areal features for augmenting the NLP systems, especially in low-resource situations. In this case even with minimal data and a rather naive approach our statistical system has outperformed a SVM-based hybrid system - this itself attests the fact that typological and areal features capture a significant generalisation about human languages and they could prove to be very valuable, if used judiciously, for low-resource NLP.

5 Conclusion

In this paper, we presented two rule-based systems and one hybrid system to predict typological features of a given language. We demonstrated that the statistical, a rule-based system, gave the best performance on the test set. Only the data set provided by the organisers was used for developing the systems.

References

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.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.