Typological Feature Prediction with Matrix Completion

Annebeth Buis
Department of Linguistics
University of Colorado Boulder
anne.buis@colorado.edu

Mans Hulden
Department of Linguistics
University of Colorado Boulder
mans.hulden@colorado.edu

1 Introduction

Currently, there are approximately 7,000 non-extinct languages in the world (Lewis, 2009). Linguistic typology aims at studying and classifying these languages in a systematic way, based on their structural and functional features. The World Atlas of Language Structures (WALS, Dryer and Haspelmath, 2013) is an online database that describes typological features—phonological, syntactic, lexical, word order features, etc.—and records the value of 144 such features for 2,679 languages. However, even for this small set of languages and features, the value of 80% of the language-feature combinations is undefined. Previous research has shown that WALS feature values can be predicted based on the existing data in WALS (e.g., Takamura et al., 2016). Predicted values for missing language-feature combinations in WALS can be useful both for downstream NLP tasks (e.g., Naseem et al., 2012; Daiber et al., 2016) and for typological research.

We discuss predicting WALS data using matrix completion. In our first experiment, we use a simple set-up and a leave-one-out-cross-validation to predict feature values in the database. We compare the results to a majority class baseline and a logistic regression classifier. In further experiments, we test the robustness of our method by leaving out (1) features in the same domain and (2) languages within the same language family To our knowledge, matrix completion approaches have not been used for typological prediction tasks previously; we show that they outperform our two baselines on the WALS data set.

2 Related Work

For space reasons, we point the reader to Ponti et al. (2018) for a comprehensive overview of research on typological information in NLP/computational linguistics. We have included a comparison to accuracies obtained in other WALS prediction experiments in Table 1. Note that it is difficult to objectively compare performance between different projects because of the wide disparity in methods and subsets of the WALS data used. Because of this, our focus point as regards comparison will be the two baselines evaluated on the same data set used for matrix completion.

3 WALS and Preprocessing

Features in WALS are split up in 11 different domains: *phonology, sign languages,*[1] *morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon and other*. The data set also includes 10 meta-features (isocodes, language family, genus, etc.), which are not included in the data for prediction.

Our method requires very little preprocessing.[2] The original WALS matrix contains categorical feature values, which were binarized before running matrix completion. We excluded 214 languages for which only 1 feature value has been recorded in WALS.

4 Matrix completion

Matrix completion algorithms are not yet part of the standard computational linguistics toolkit. However, there are several reasons why matrix completion is potentially a good method for our task. First, these algorithms have been used extensively with sparse matrices. Second, since they learn from the entire matrix at once, we expect

[1] Both sign languages and features related to sign languages have been excluded from the data in this project.

[2] All data and code used to obtain the results in this paper is available at https://github.com/annebeth/wals-matrix-completion.

	Accuracy	Method
Georgi et al. (2010)	65.5%	Language clustering
Takamura et al. (2016)	75.5%	Logistic regression
without language family	73.0%	Logistic regression
Murawaki (2017)	74.5%	Bayesian model
Baseline 1	53.1%	Majority class
Baseline 2	65.7%	Logistic regression
Matrix completion	74.3%	IterativeSVD
without domain	61.6%	IterativeSVD
without language family	71.2%	IterativeSVD

Table 1: Matrix completion experiment results compared with results obtained in previous work.

them to be able to learn more holistic patterns in the data than individual local predictors (such as our logistic regression baseline).

The matrix completion algorithm used in this paper is IterativeSVD,[3] based on Troyanskaya et al. (2001). This method attempts to learn a low-rank approximation of the original matrix by using Singular Value Decomposition (SVD).

5 Experimental set-up

In our experiment, we are predicting each language × feature-combination that currently has a value in WALS (i.e., it is not *undefined*) separately by using leave-one-out cross validation (LOOCV).

First, we calculate results for two baselines. The first baseline predicts a feature value by simply assigning the majority class for each feature. The second baseline consists of logistic regression classifiers that are trained to predict a specific feature based on all other features. Besides our basic matrix completion setting, we have run the experiment in two additional settings: (1) *without domain*-setting: all features from the same domain as the feature that is being predicted are excluded from the matrix, and (2) *without family*-setting: when predicting a feature value for a certain language, all other languages that are in the same language family are excluded from the matrix.

6 Results

Table 1 shows the prediction results obtained with matrix completion on the WALS data and compares them to results obtained in related work.[4] Matrix completion significantly outperforms our two baselines and also improves on the baselines in the *without language family* setting.

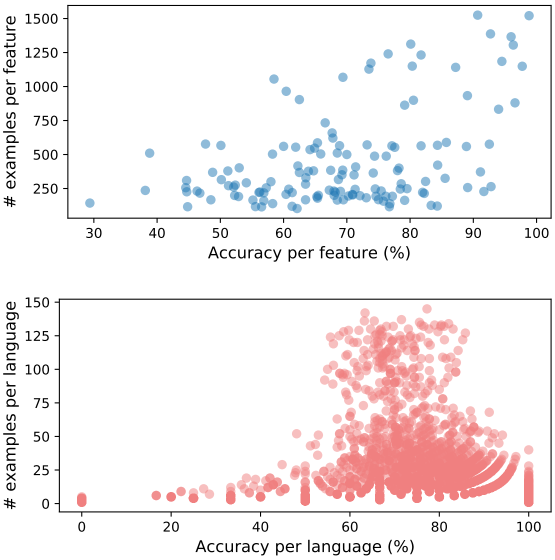

Figure 1: Comparison of (top) number of examples of each feature to the prediction accuracy for that feature and (bottom) the number of examples of each language to the prediction accuracy for that language.

Figure 1 shows different distributional patterns in the comparison of the number of examples with the obtained accuracy. The prediction accuracies calculated per feature vary much more than those calculated per language. The number of examples shows no correlation with the prediction accuracy per language. For feature accuracy, however, having more examples of a feature can result in better predictions.

7 Conclusion

Matrix completion outperforms the baselines on the WALS data and performs on par with previous work. This shows that matrix completion captures holistic patterns in the data that cannot be learned in a traditional classifier approach. Furthermore, our method requires minimal preprocessing and can easily be used with any typological database.

We leave the discussion of other matrix completion algorithms to future work. Our work has shown that treating WALS as a matrix is an effective approach. That idea could be exploited in future work on WALS. Non-negative matrix factorization (Lee and Seung, 1999) could be used to improve the clustering of languages or the analysis of typological implications (such as *VSO order* → *Noun-Adjective order*, Greenberg, 1963).

[3] IterativeSVD as implemented in the FancyImpute Python package: `https://github.com/iskandr/`.

[4] We included papers that use only WALS as training data and evaluate on all domains in WALS.

References

Joachim Daiber, Miloš Stanojević, and Khalil Sima'an. 2016. Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176.

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

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. Cambridge: MIT Press.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

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.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. 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. *arXiv preprint arXiv:1807.00914*.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 69–76.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.