Colex2Lang: Language Embeddings from Semantic Typology

Yiyi Chen[1], Russa Biswas[2], Johannes Bjerva[1]
[1]Aalborg University, Denmark [2]FIZ Karlsruhe, Germany

Overview

Objective

- Investigate the potential of incorporating semantic typology, such as colexifications, to multilingual NLP.

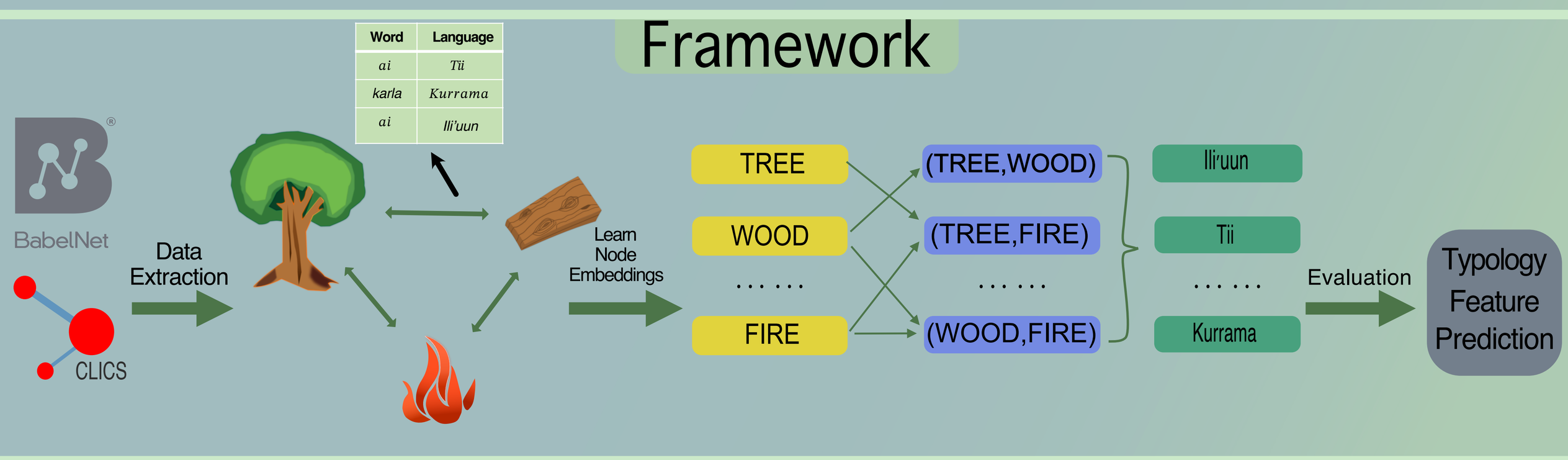
Hypothesis

- Language representations learned from semantic typology encapsulate a distinct language signal

Approach and Evaluation

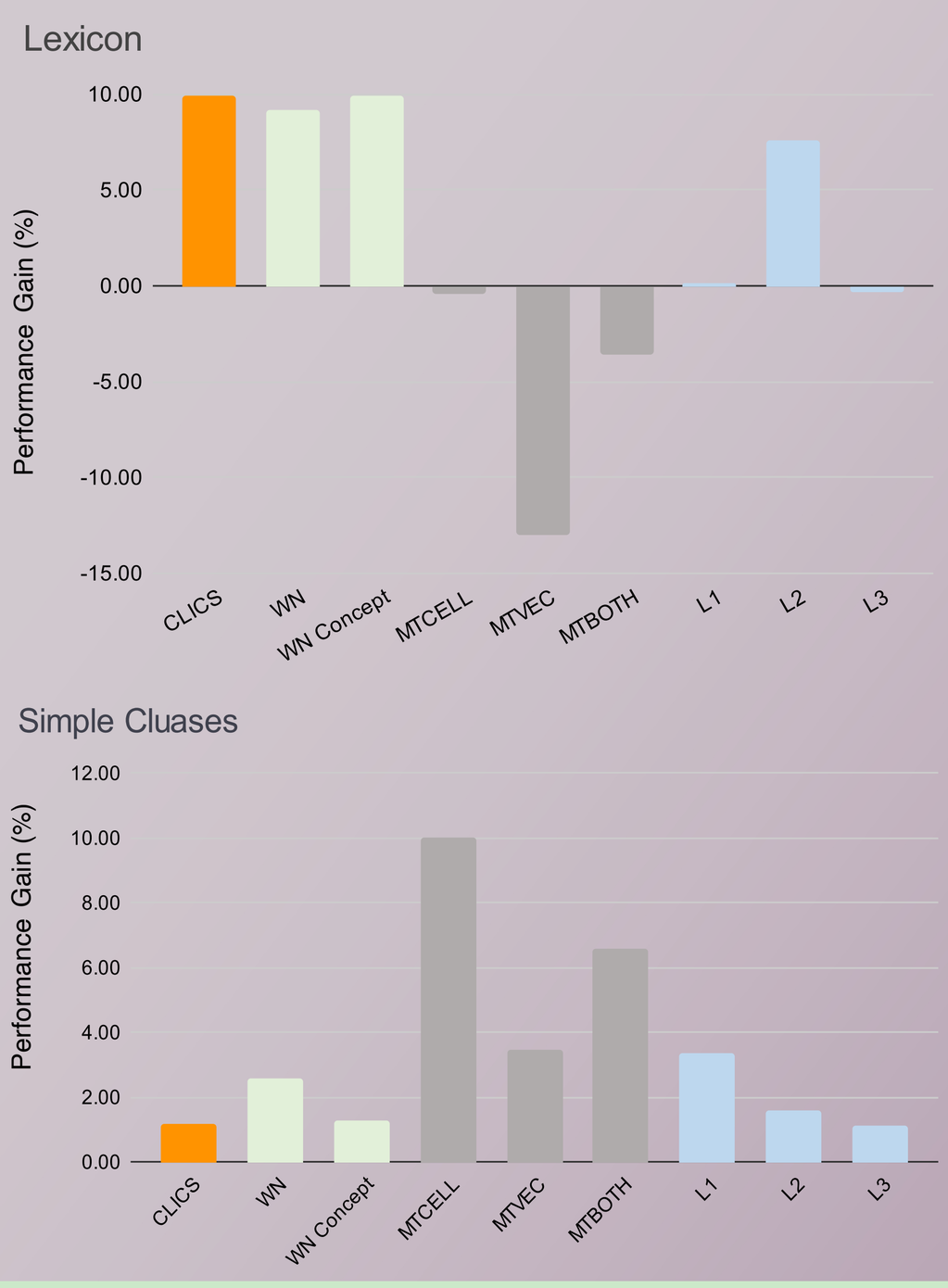
- Build large-scale synset graphs
- Learn synset embeddings using node embedding algorithms
- Construct colexification and language embeddings
- Evaluating the colexification-informed language embeddings using typological feature prediction

Framework

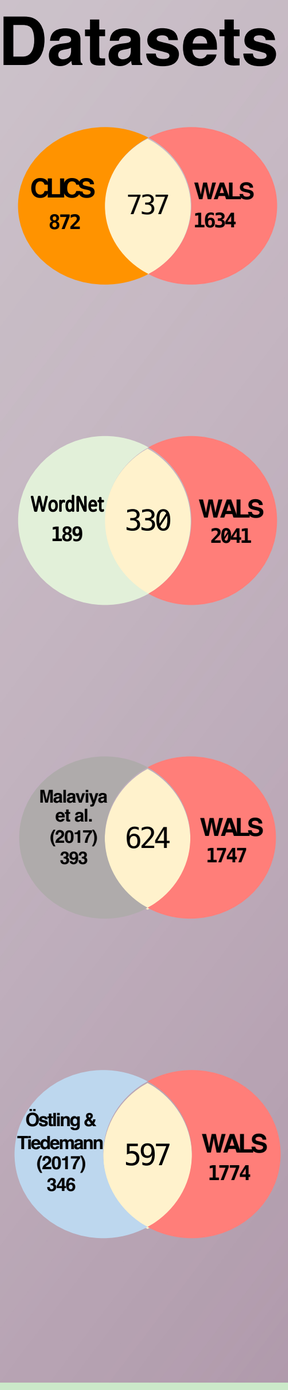


Results

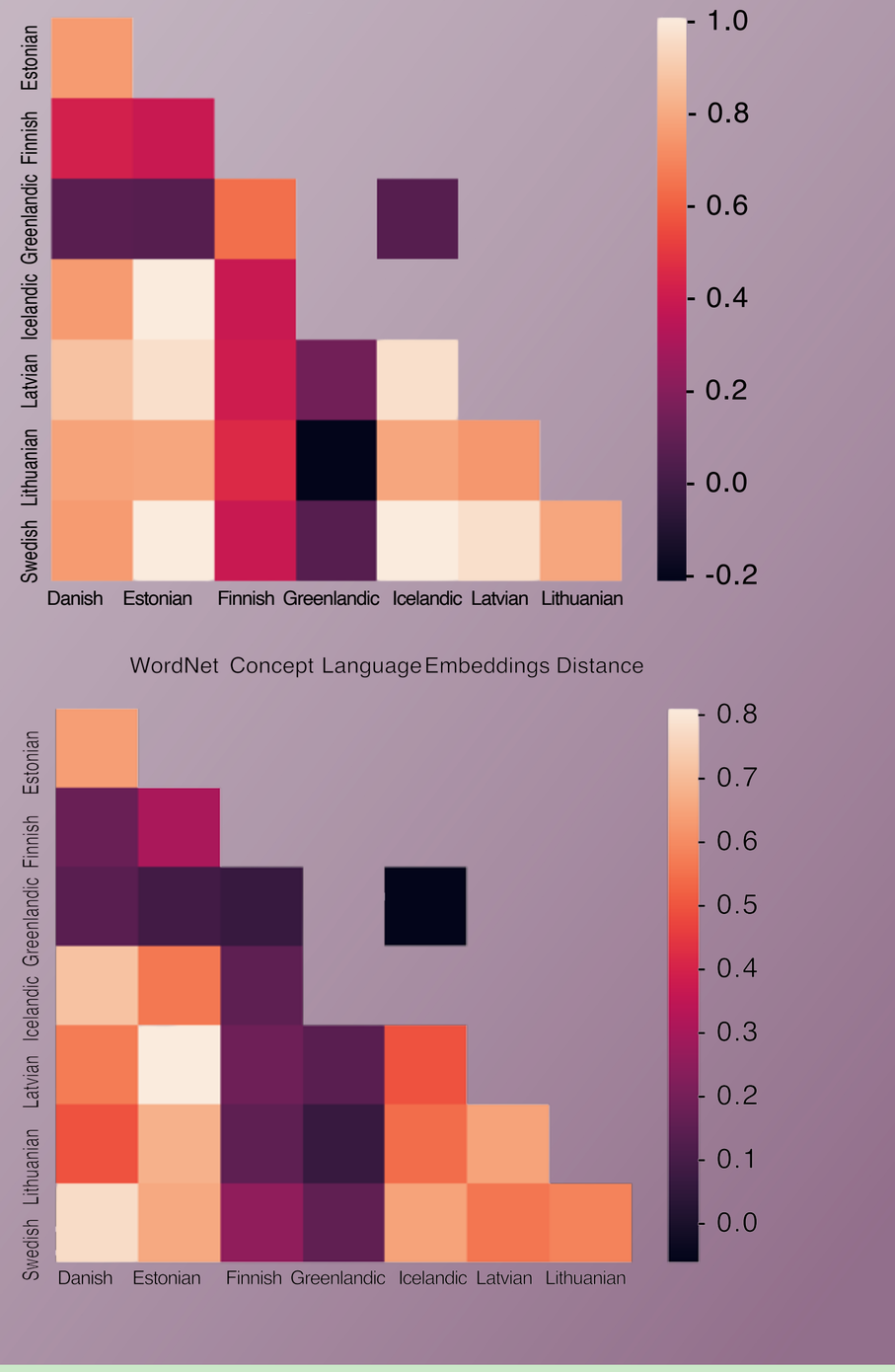
Typological Feature Prediction



Datasets



Language Similarity



Conclusions

Conclusion

- The first attempt to learn language representations and model language similarities using semantic typology at a large-scale

Future Work

- Apply colexification-informed language embeddings to further multilingual tasks and transfer learning from high-resource to low-resource languages

References

Y. Chen, R. Biswas, J. Bjerva, 2023, Colex2Lang: Language Embeddings from Semantic Typology. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa).*

Funding

This project is supported by the Carlsberg foundation, under the grant 'Multilingual Modelling for Resource-Poor Languages' (CF21-0454).