Using Typological Information in WALS to Improve Grammar Inference

Youyun Zhang, Tifa de Almeida, Kristen Howell, Emily M. Bender

University of Washington

Introduction

Using implemented grammars to model low-resource languages can assist the process of language documentation (Bender et al., 2012), but such grammars are expensive to build and require different expertise to that required for linguistic field work.

The AGGREGATION Project aims to automatically generate grammars for low resource languages, taking advantage of the linguistic information incoded in Interlinnear Glossed Text, generalizations in the typological literature and stored syntactic syntactic analyses in the Grammar Matrix customization system.

The Grammar Matrix is a cross-linguistic grammar customization toolkit that creates precision grammars for a language based on a users' specification of its linguistic properties (Bender et al., 2002, 2010). Linguistic phenomena such as sentential negation (Crowgey, 2012) are modeled for customization .

The World Atlas of Language Structures (WALS) is a typological database that includes about 200 structural features of over 2,500 languages, which also schematizes the typological features of languages (Dryer and Haspelmath, 2013).

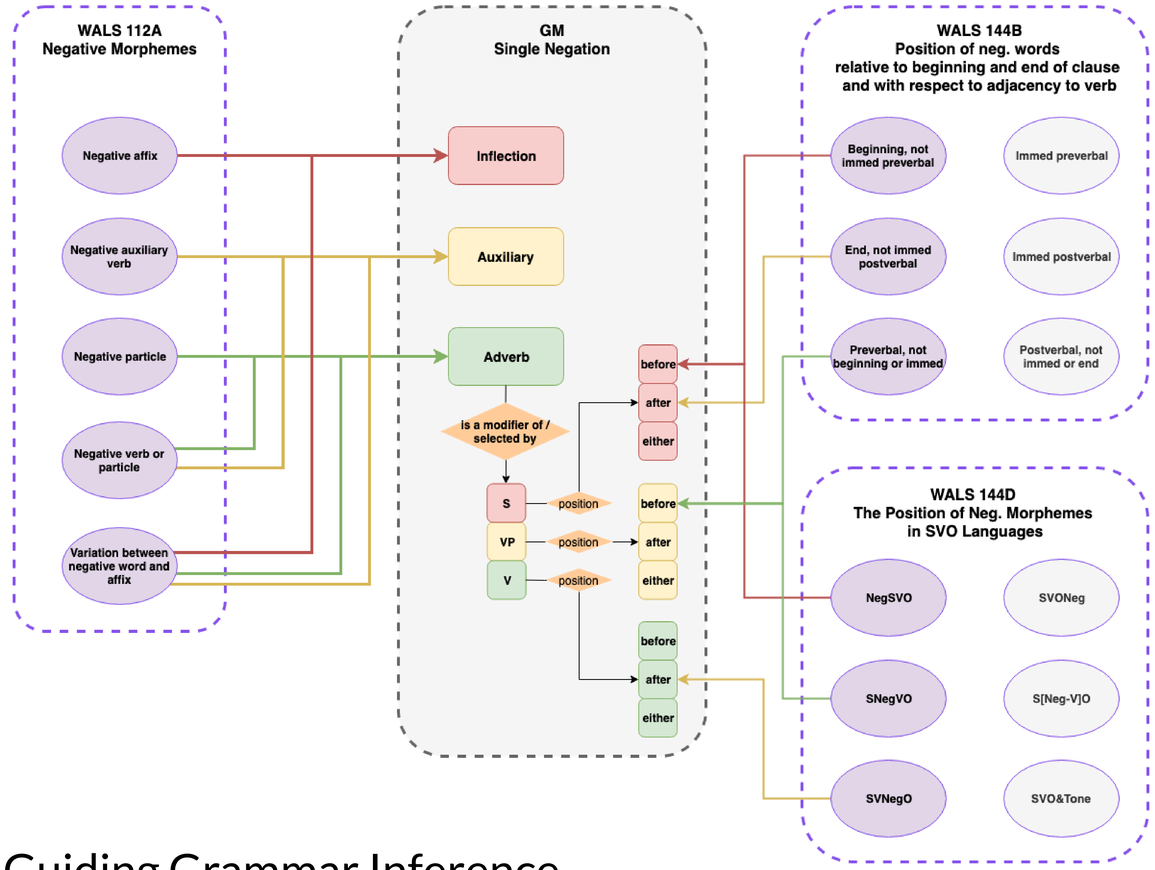
OVERLAP on linguistic typology:

de Almeida et al. (2019) concludes that about 10.4% of WALS features can be imported into the Matrix.

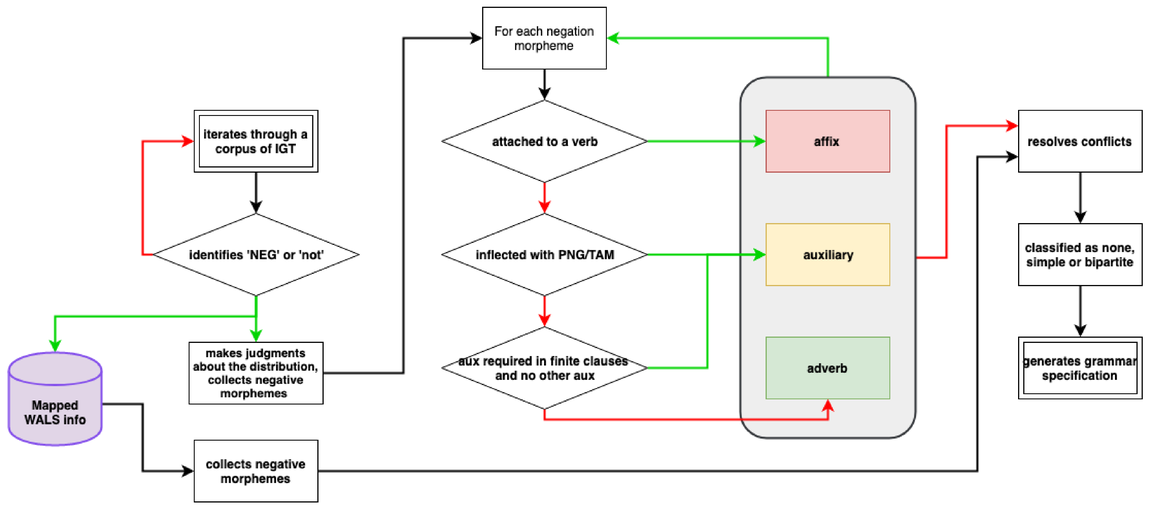
Goal

We consider how the **mapping** of features between WALS and the Matrix can be used to improve the quality of grammar inference, as set forth by Bender et al. (2014) and Zamaraeva et al. (2019). We illustrate with a case study of sentential negation.

Mapping WALS features to the Grammar Matrix



Guiding Grammar Inference



Evaluation

We plan to evaluate this method for improving grammar inference by using the same coverage and ambiguity based evaluation strategy of Zamaraeva et al. (2019):

- create grammars with the Matrix customization system using inferred grammar specifications for 5-10 different languages
- use those grammars to parse held out data not used in grammar inference
- compare grammar specifications inferred with and without guidance from mapped WALS features

We predict that the guidance will result in grammars that have higher coverage, lower ambiguity, or both.

Evaluating Grammar Specification:

- baseline
- grammar inference
- grammar inference + WALS info
- grammar inference + WALS (use core guidance only)
- grammar inference + WALS (use the most common values of features to fill in missing info)
- grammar inference WALS (use info from languages in the same family to fill in missing info)

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Tifa de Almeida, Youyun Zhang, Kristen Howell, and Emily M. Bender. 2019. Feature comparison across typological resources. Unpublished paper, submitted to TypNLP.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. http://www.aclweb.org/anthology/W14-2206 Learning grammar specifications from IGT: A case study of Chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA. Association for Computational Linguistics.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8:1–50.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Emily M Bender, Robert Schikowski, and Balthasar Bickel. 2012. Deriving a lexicon for a precision grammar from language documentation resources: A case study of Chintang. *Proceedings of COLING 2012*, pages 247–262.

Joshua David Crowgey. 2012. The syntactic exponence of sentential negation: A model for the LinGO Grammar Matrix. Master's thesis.

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

Olga Zamaraeva, Michael Wayne Goodman, Emily M. Bender, and Kristen Howell. 2019. Improving toolbox IGT using the Xigt data model. To be presented at The 6th International Conference on Language Documentation Conservation (ICLDC) Technology Showcase.

DEPARTMENT OF LINGUISTICS
UNIVERSITY of WASHINGTON