Using Typological Information in WALS to Improve Grammar Inference

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

Department of Linguistics, University of Washington, Seattle, WA, USA

{youyunzh, trda, kphowell, ebender}@uw.edu

1 Introduction

Using implemented grammars to model low-resource languages largely aids 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 rich annotation provided in the form of Interlinear Glossed Text (IGT), POS tags and dependency parses projected from English translations (Georgi, 2016), as well as stored syntactic analyses from the LinGO Grammar Matrix (GM) customization system (Bender et al., 2002, 2010).

The GM is a cross-linguistic grammar customization toolkit that creates precision grammars for a language based on a users' specification of its linguistic properties. A variety of linguistic phenomenon are modeled for customization, such as word order (Bender et al., 2010), adnominal possession (Nielsen and Bender, 2018) and sentential negation (Crowgey, 2012), among others.

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

While the goals of the two projects are different, because they both focus on linguistic typology, there is some overlap. de Almeida et al. (2019) concludes that about 10.4% of WALS features can be imported into the GM. Here, we consider how the mapping of features between WALS and the Grammar 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.

2 Background

The GM Setential Negation library (Crowgey 2012), based on Miestamo 2003, requires specification of how many markers of negation are required (zero, one, two), and the form each marker takes (affix, auxiliary, adverb). Each marker can have further features specified, such as for adverbs whether they attach to V, VP or S and to the left or the right.

2.1 Negation Inference

The negation inference module in the AGGREGATION project iterates through a corpus of IGT and collects negative morphemes from the glossed line of the IGT by identifying those glossed as 'NEG' or 'not'.

For each morpheme, the inference system first identifies whether the the morpheme is a verbal affix by checking the POS tag (provided by INTENT; Georgi 2016) of the verb it's a part of. If the negative morpheme is not attached to a verb, the inference system predicts whether it is an auxiliary by checking to see if it is inflected with person, number or gender agreement or with tense, aspect or mood features. If it is, it is classified as an auxiliary. If not, the system checks to see if the language requires an auxiliary in finite clauses and if there is no other auxiliary in the negated clause: in that case, it is also classified as an auxiliary. If the morpheme is not classified as a verbal affix or auxiliary, it is classified as an adverb.

The inference system also makes judgments about the distribution (before/after the verb, attaching to VP or S), and collects the orthographies. The negation strategy is then classified as none, simple or bipartite based on the average number of neg morphemes in each negated sentence.

Value of WALS Feature 112A	Negation Strategy	Negation Strategy sub-type
a. Negative affix	Single	"inflectional"
b. Negative particle	Single	"adverb"
c. Negative auxiliary verb	Single	"auxiliary"
d. Negative word, unclear if verb or particle	Single	"adverb / auxiliary"
e. Variation between negative word and affix	Single	"adverb / auxiliary"&"inflectional"
f. Double negation	Bipartite	

Table 1: Mapping from WALS Feature 112A to choices in Sentential Negation Library in GM

2.2 Negation Features in WALS

Features related to sentential negation in WALS are 112A (*Negative Morphemes*; Dryer 2013a), 143 (*Order of Negative Morpheme and Verb*; Dryer 2013b), and 144 (*Position of Negative Morpheme with Respect to Subject, Object, and Verb*; Dryer 2013c). Features 143 and 144 collectively have 32 subfeatures.

3 Methodology

3.1 Mapping

Feature 112A (*Negative Morphemes*) provides information of a language's negative morphemes in clausal negation (Dryer, 2013a). Table 1 shows how we map this feature (when available) to the GM's specifications. Feature 143A (*Order of Negative Morpheme and Verb*) (Dryer, 2013b) increases the number of languages that can be mapped to Sentential Negation Library in the same style by 74.

Feature 143B (*Obligatory Double Negation*) further details the relative position of double negative morphemes and the verb, such as "NegVNeg" (two free morphemes around the verb) and "[V-Neg]Neg" (one bound morpheme attached to the verb and a free morpheme on the right). Feature 143C (*Optional Double Negation*) provides information of double negative morphemes in a similar style with a pair of parentheses surrounding one of the negative morpheme indicating it is optional (e.g. "Neg[V(-Neg)]").

3.2 Guiding Grammar Inference

The WALS information alone doesn't provide enough information to create a GM grammar, because it does not provide specific forms (words, affixes). However, it does provide knowledge that can be used to guide grammar inference. Specifically, it can tell the inference system where to look for negative morphemes by already knowing a language's negation strategy from WALS.

However, it is likely that a feature value in WALS is defined overlapping two types (or subtypes) that exist mutually exclusively in the GM, for example all languages in Feature 143C use both simple and bipartite negation strategy, or not specified enough to be categorized, such as value d in Table 1. This ambiguity is expressed when the inference looks up the mapped negation strategy. The inference then makes a decision among these possible strategies based on the nature of the majority of negative morpheme(s) that is found in the corpus.

It is also possible for a language to use a combination of different sub-types under a category (e.g. value e in Table 1). The GM can handle this situation in the customized grammar. Therefore all sub-types should be provided when the inference looks up the mapped negation strategy and collects the orthographies.

4 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): we will use the inferred choices files for 5-10 different languages to create grammars with the Grammar Matrix customization system, and then use those grammars to parse held out data not used in grammar inference. We will compare choices files 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.

5 Future Work

Other than the features related to sentential negation, there are more features such as case, adnominal possession, etc. in WALS that can be mapped to the GM (de Almeida et al., 2019). We plan to apply this same methodology for using mapped WALS features to guide grammar inference for several such features and then evaluate them to explore which ones improve grammar inference the most.

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. 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. 2013a. Negative morphemes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013b. Order of negative morpheme and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer. 2013c. Position of negative morpheme with respect to subject, object, and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Ryan Georgi. 2016. *From Aari to Zulu: Massively Multilingual Creation of Language Tools Using Interlinear Glossed Text*. Ph.D. thesis, University of Washington.

Matti Miestamo. 2003. Clausal negation: A typological study.

Elizabeth Nielsen and Emily M. Bender. 2018. Modeling adnominal possession in multilingual grammar engineering. In *Proceedings of the 25th International Conference on Head-Driven Phrase Structure Grammar, University of Tokyo*, pages 140–153, Stanford, CA. CSLI Publications.

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.