Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud, Linda Wiechetek

Takeaway

Machine-readable grammars can be more easily applied to new langauges if they are working with higher levels of analysis. Working with morphophonology, the grammatical differences between languages preclude the reuse of analyses.

> We argue that portability here takes the form of reusing smaller modules of the grammar

Hopefully the paper expands on that because that statement doesn't make any sense

Languages

- North Lule and South Sami
  - Uralic language
  - Not very agglutinative
- Faroese
  - Germanic language

    - Four case system
- Greenlandic
    - Eskimo-Aleut language
    - Polysynthetic

Technical background

- Using existing resources developed by the University of Tromso.
    - Morphological analyzers
    - Constraint Grammar parsers

Reusing grammar

- Blick (2006) argues for using bootstrapping techniques to reuse grammar instead of appealing to statistical systems. *This fell by the wayside, everyone uses statistical methods now*

The bottom of the analysis

- The level of analysis that is close to the language substance cannot be directly used
-
    > “ Even though different languages do not have the eact same morphological processes, they may have the same process *types*

- Rules are written in a modular fashion so they can easily be adapted to new languages
    - For example, consonant gradiation processes are very common, the particulars of the rule may need to change but the module design helps guide the changes that need to be made.

Disambiguation

Mapping of syntactic tags

- Large number of tags needed due to the free word order of Sami languages
  - For example, four different subject tags needed specifying whether the verb is finite, whether elipsis of verb has occured, whether the finite verb is to the left or to the right etc.
-

The top of the analysis

This is the part that's relevant to me

- Using a constraint grammar module
-

> Syntactic tags for verbs are substituted by other tags (according to clause-type) in order to make it easier to annotate dependency across clauses

- Descibes difficulties finding the "head" of the sentence (think they mean root), when dealing with ellipses. This is definitely an issue as well in UD

> Still the analyzer retains very good accuracy for the dependency analysis: 0.99

- This is for Sami
- Table 5 say this is actually f-score?
- How is this scored? Are they scoring the flat descriptors in the visl format (e.g. #5->0)
- Use pairs of substitution and setparent rules

Bootstrapping

- Go through small modifications to the rules to consider Faroese specific phenomenon.
- Show the specific increases in performance with each new difference that is considered (e.g. when substituting the Relative pronouns that begin subordinate clauses in Sami with the CS that begins relative clauses in Faroese, the accuracy goes up to 96)

Revision #3

Created Thu, Apr 23, 2020 6:30 PM by kenneth

Updated Thu, Apr 23, 2020 9:16 PM by kenneth