Morphological Processing of Low-Resource Languages: Where We Are and What's Next

Adam Wiemerslage*, Miikka Silfverberg*, Changbing Yang,
Arya D McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann

For a survey of recent work in computational morphology for low-resource languages, please refer to our paper!

Task: *tUMPC*

Predict the full paradigm for a given word in context

"My best friend broke our lamp"

break	broke	breaking
breaks	broken	

"Some geese are flying over my head"

goose	geese

Questions

1. **How to handle *all* words (and POS) in the corpus when learning?**
2. **Which existing systems work best in a pipeline?**
3. **What kind of corpus is easiest to learn from?**

Learning

1. Cluster (partial) paradigms

Prior work

laugh
laughed
laughs

cat
cats

fast
fastest

2. Align slots (and cluster POS)

X0 X0+ed X0+s

1_1. laugh
1_2. laughed
1_3. laughs

X0+y X0+ies X0+ying

1_1. cry
1_3. cries
1_4. crying

X0+ew X0+ying

1_2. flew
1_4. flying

X0 X0+s

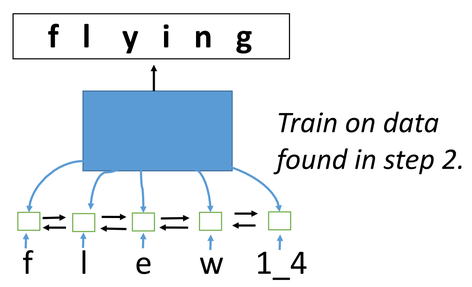
2_1 cat
2_2 cats

X0+oo+X1 X0+ee+X1

goose
geese

3. Train inflector

Prior work



4. Train tagger

...I laughed so hard... → Transformer → POS=1

...Don't pet the geese... → Transformer → POS=2

Inference

1. Predict slots

"My best friend broke our lamp"

→ Transformer →

1_1, 1_2, 1_3, 1_4

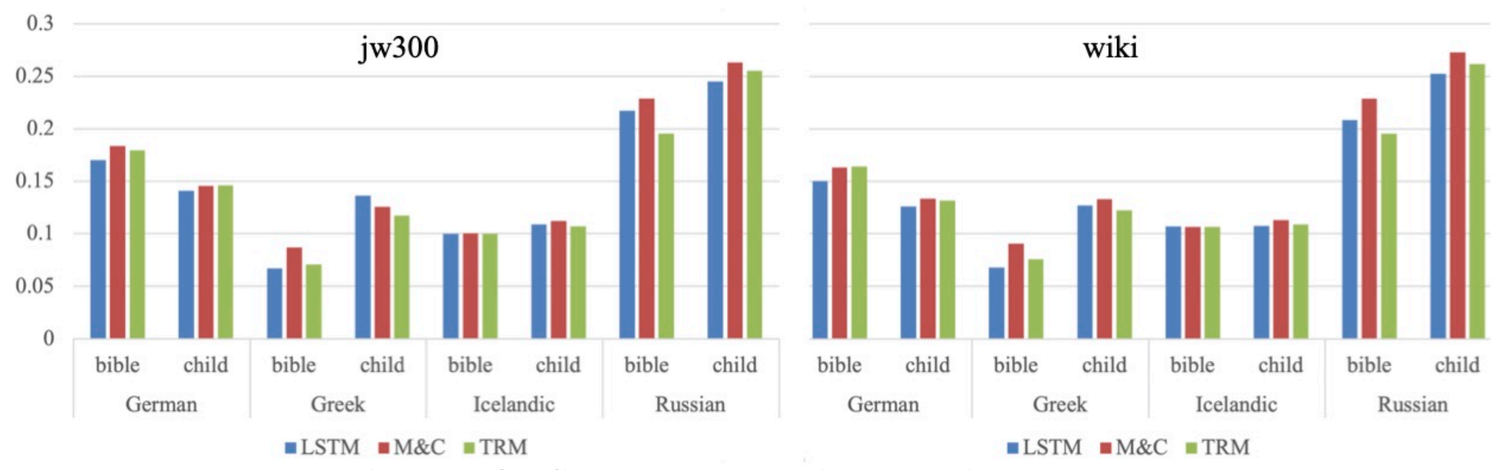
2. Inflect

broke + 1_1 → Inflector → ?

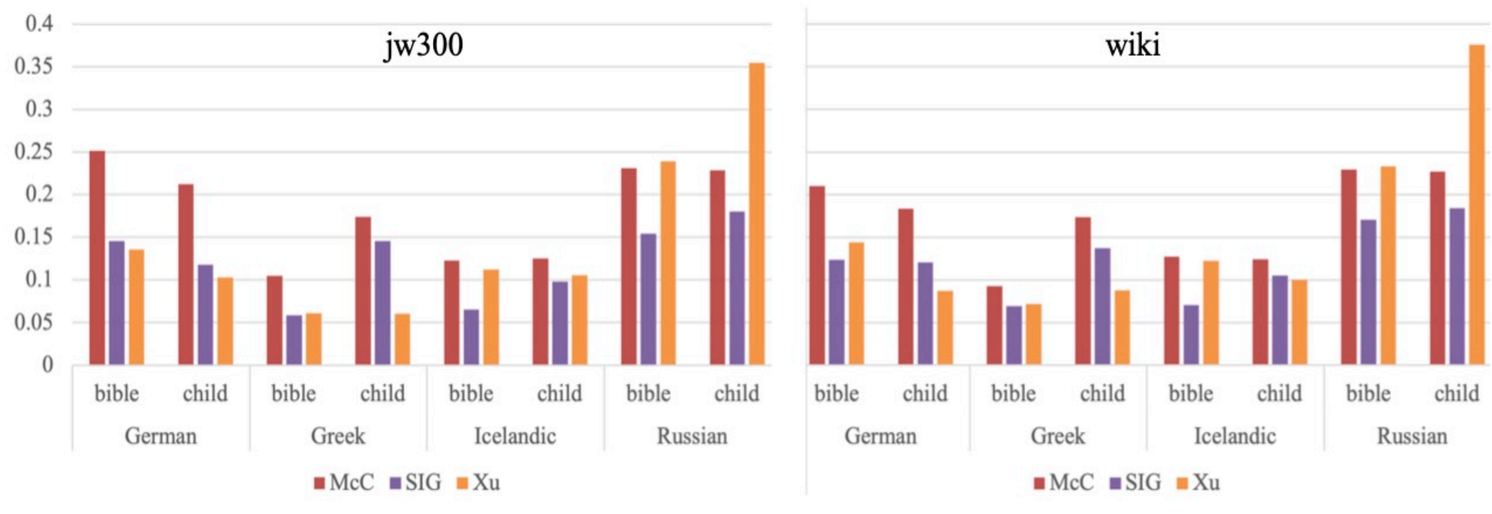
broke + 1_3 → Inflector → ?

broke + 1_4 → Inflector → ?

Results



Choice of inflection system leads to low variation



Choice of paradigm clustering algorithm leads to high variation