PALI: A Language Identification Benchmark for Perso-Arabic Scripts

Sina Ahmadi, Milind Agarwal, Antonios Anastasopoulos

{sahmad, magarwa, antonis} @ gmu.edu

Research Question

Can hierarchical approaches:

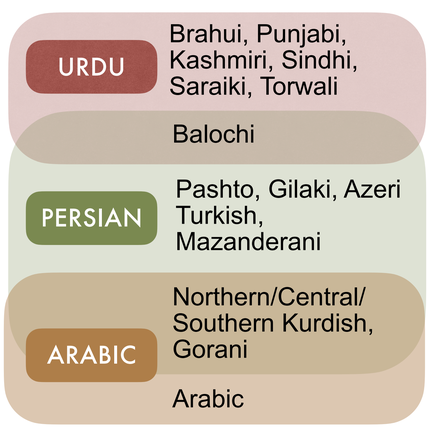
1. Identify languages that use unconventional writing systems in low-resource settings?
2. Resolve confusion among languages in custom-trained or off-the-shelf models?

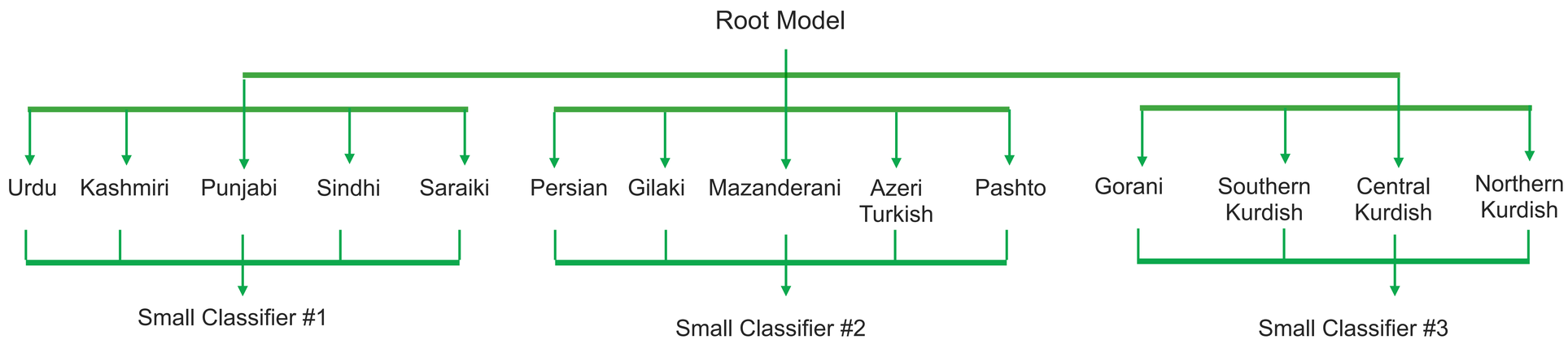
Data

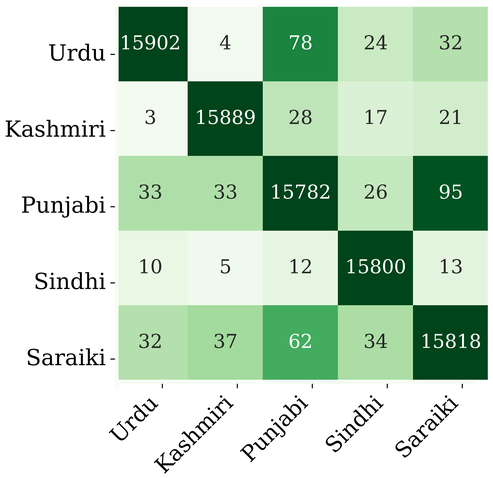
We focus on bilingual language communities that use the Urdu, Persian, or Arabic scripts

Data Sources:

- Wikipedia dumps
- Local news websites
- Small pre-published corpora







Noise %	Sentence
Clean	دووەمین پێشانگەها فۆتۆگرافەرێن کورد ل بەلجیکا Second Kurdish photographers' exhibition in Belgium
20	دووهمین پێشانکهها فۆتۆکرافهرێن کورد ل بهلجیکا
40	دووه مین بشانکه ها فطکرافه رن کورد ل به لجیکا
60	دووة مین بشانکة ها فوتوکرافة رن کورد ل بة لجیکا
80	دووەمین بیشانکةها فۆتۆکرافةرین کورد ل بةلجیکا
100	دووهمین بیشانکهها فوتوکرافهرین کورد ل بهلجیکا

Methodology

- Manually create script maps and generate data at different noise levels
- Inspect classification systems to identify highly-confused language clusters
- Off of the best root system, train specialized hierarchical classifiers to resolve confusion

		F1
OURS (HIER)	**Hier**	**0.95**
CUSTOM-TRAINED	Root	0.94
	MNB	0.10
	MLP	0.10
	fastText	0.27
OFF-THE-SHELF	CLD3	0.09
	langid.py	0.13
	Franc	0.13

Result Highlights

- Off-the-shelf and custom trained systems generally perform worse in noisy settings
- A confusion-based hierarchical classification can help build on top of a reasonable root system
- Such modeling can help identify languages that use unconventional writing systems

Check out our GitHub!

- Clean, noisy, merged data
- Compressed trained models for 19 languages
- All preprocessing and training scripts