NLPnorth

- Methods
  - Rule-based
  - Baseline
  - CRF layer
  - Multi-dataset training
  - Language modeling
- How to pick the right LM
- Remaining errors

Starting point

- MaChAmp toolkit
- Map labels to uniform set
- 37 language models

Extentions

- CRF
- Multi-dataset
- Language modeling objective

Extentions

- CRF + -
- Multi-dataset
- Language modeling objective

Extentions

- CRF + -
- Multi-dataset + -
- Language modeling objective

Extentions

- CRF + -
- Multi-dataset + -
- Language modeling objective + -

So why the high scores?

So why the high scores?

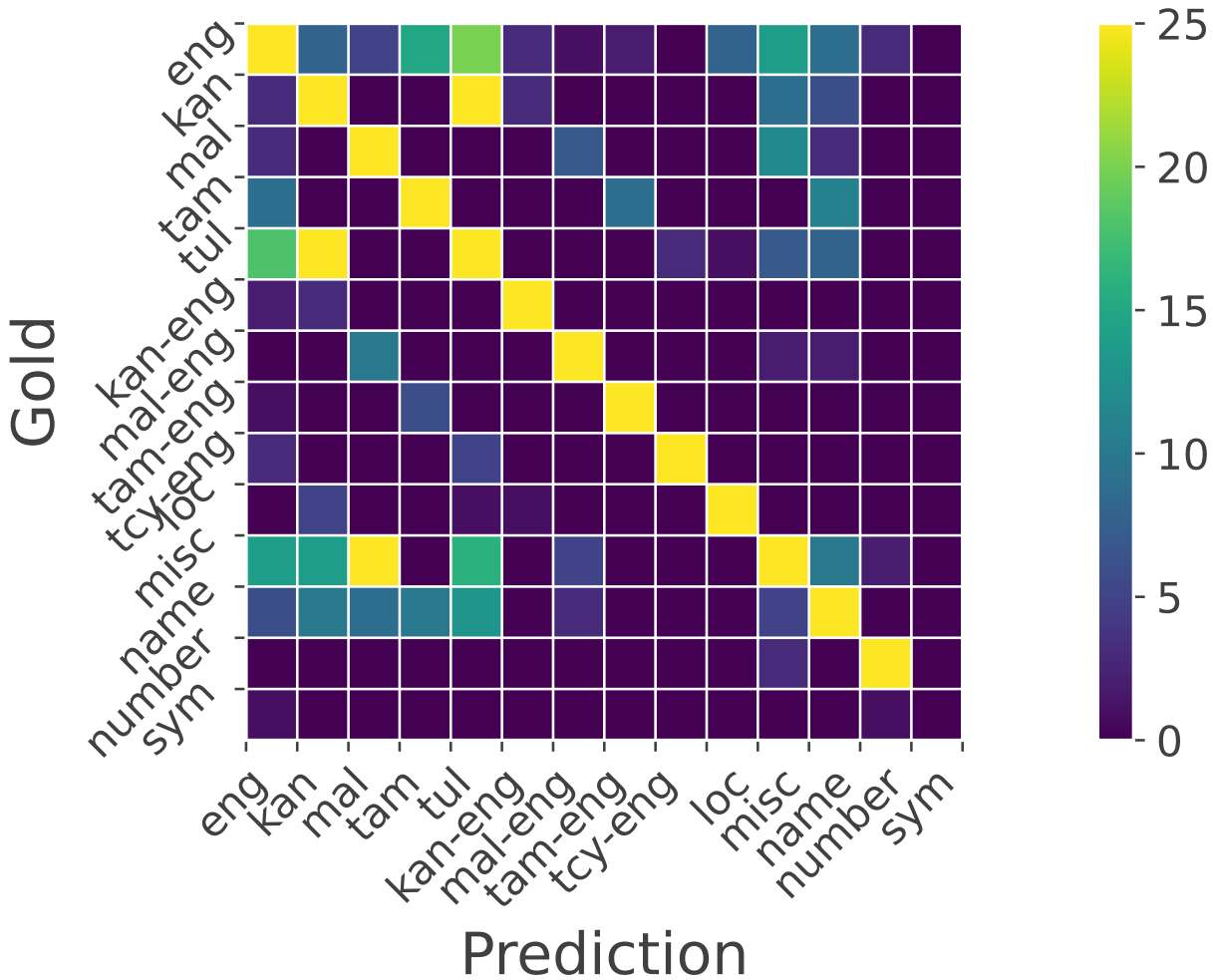
37 language models!

How to pick the right language model?

Variable	Pearson
# weights	0.2559
Vocab size	0.0693
# languages	0.1668
% used	-0.0153
Avg. word len	-0.1703

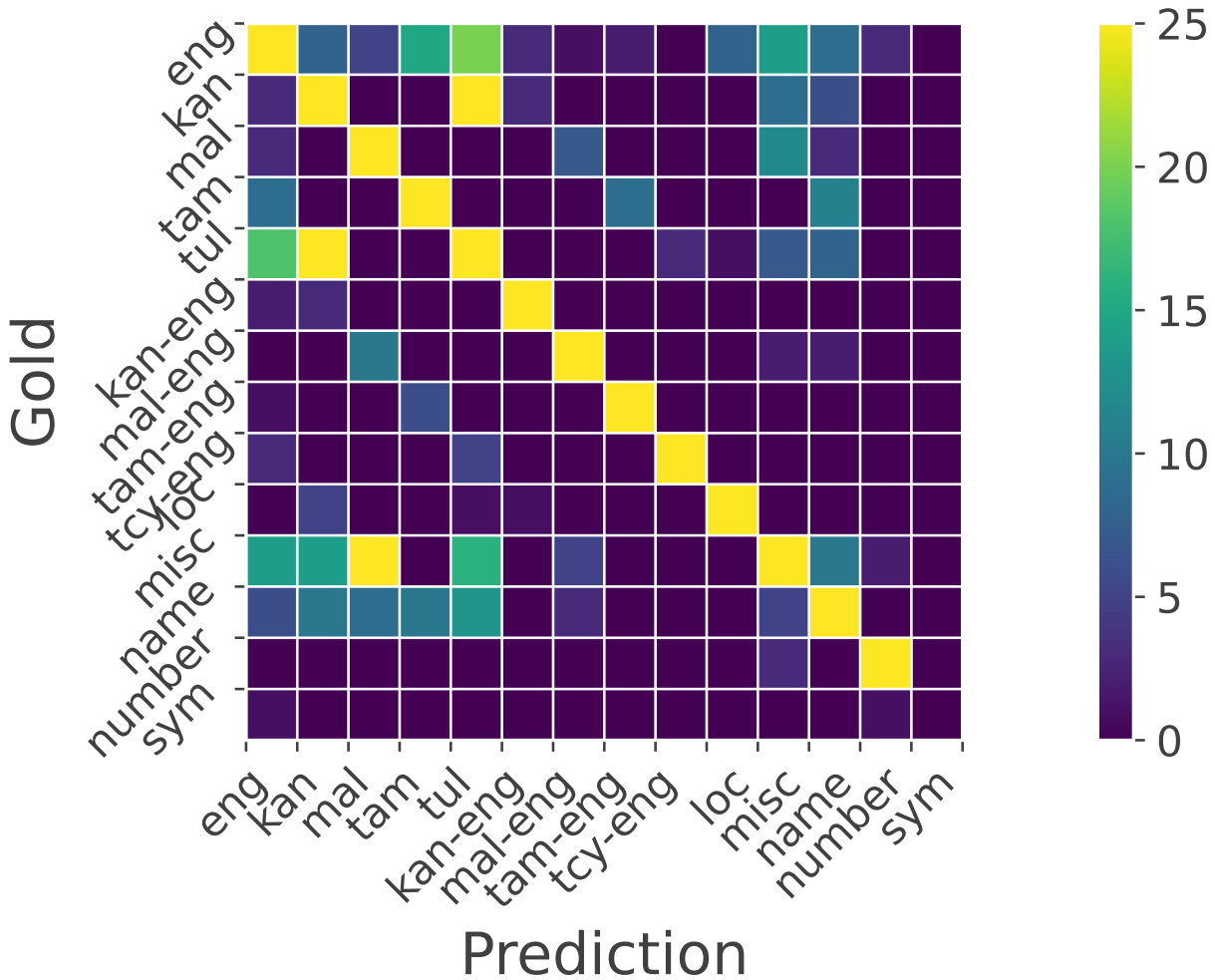
Table: Pearson correlation to macro-f1.

Remaining errors



- Kannada and Tulu occur in same dataset (ambiguous words)
- Underprediction misc
- Mixed: only inflections in English

Remaining errors



- English:
  - Interjections (ah, hahha)
  - Typos and slang (Tha = Tulu?)
  - Annotations (e.g. padike, Bakrid)

Thanks

Thanks to the shared task organizers and the ITU HPC (Lottie)!