



ParaNames 1.0: Creating an Entity Name Corpus for 400+ Languages using Wikidata

Jonne Sälevä, Constantine Lignos

Summary

Problem: Where/how to obtain entity name lists for e.g. transliteration, NER research?

Approach: Ingest Wikidata, assign NER types (PER/LOC/ORG), standardize scripts per language

Result: Freely available name list of over 140 million names across 400+ languages

Data Extraction and Filtering

Assign entities to LOC/PER/ORG based on instance-of information and type hierarchy

- Q5 (human) → PER
- Q82794 (geographic region) → LOC
- Q43229 (organization) → ORG

Entity type	Count	Percentage
PER	10,002,138	59.41%
LOC	3,880,088	23.05%
ORG	2,631,350	15.63%
Mixed	320,961	<2%
Total	16,834,537	100.00%

Challenge: Script mixing within languages

- Natural variation: languages can use multiple scripts for various reasons
- Unnatural variation: bots dumping copied English names into other languages
- Use Unicode script properties to develop distribution of scripts in a language
- Use Wikipedia to identify standard scripts in each language and filter out names written in others

Named Entity Recognition

Use case: ParaNames as a gazetteer for NER

Model: LSTM-CRF + soft gazetteer features[1]

Data: MasakhaNER[2], HiNER[3], Turku NER[4]

Evaluation metric: F1 score (span-level)

Results

- Gazetteers are useful: $\Delta > 0$ for each language
- Wide variation in SD (σ) across languages
- Performance mixed in terms of Δ/σ
 - High: Swahili, Finnish, Hausa, and Yoruba
 - Low: Amharic, Kinyarwanda, Hindi, and Wolof

Language	Best F1	Diff. (Δ)	SD (σ)	Δ/σ
Swahili	80.06	2.25	1.15	1.93
Finnish	65.18	2.14	1.57	1.37
Hausa	84.40	0.80	0.67	1.19
Yoruba	67.74	1.44	1.30	1.11
Igbo	79.75	0.54	1.03	0.52
Luganda	74.76	0.64	1.39	0.46
Wolof	59.58	0.45	1.99	0.23
Amharic	52.73	0.36	1.67	0.21
Hindi	92.09	0.02	0.14	0.11
Kinyarwanda	63.14	0.12	1.32	0.09
Median	71.25	1.04	1.31	0.80
Mean	71.05	0.87	1.22	0.71

Canonical Name Translation

Task: Translate names between English and 17 languages representing a variety of scripts and language families

Model: Character-level transformer

Evaluation metrics

- Accuracy: how often exactly correct?
- LCS avg. F1 score: how much overlap is there?[5]
- Character error rate: how many edits?

Results

- Accuracy varies wildly by language

Language	To English	From English
Swedish	90.34	88.31
Vietnamese	87.02	78.17
Lithuanian	80.56	79.30
Latvian	75.26	73.81
Tajik	51.56	56.82
Kazakh	49.14	58.30
Russian	45.65	43.26
Thai	39.59	15.02
Armenian	38.76	47.95
Georgian	32.67	51.00
Korean	32.28	42.57
Arabic	31.88	46.77
Greek	31.67	31.22
Japanese	29.97	27.30
Urdu	27.02	17.96
Persian	26.92	42.10
Hebrew	18.46	37.83
Micro-avg.	46.40	49.27

- Performance by script: Latin > Cyrillic > other
- Intuition: worse source-target alignment
- Challenge: information asymmetry
 - Source side may not convey all information
 - Vowels (e.g. Hebrew, Persian, Arabic)
 - Tones (e.g. Thai, Vietnamese)

Releases

- ParaNames is freely available under the Creative Commons Attribution 4.0 International License
- Goal: regular releases with new Wikidata exports

Future Applications

- ParaNames lends itself to many more applications, especially in the modern LLM era
- We are excited to see what you build on it!
- GitHub: <https://github.com/bltllab/paranames>

References

- [1] [Soft Gazetteers for Low-Resource Named Entity Recognition](#) (Rijhwani et al., ACL 2020)
- [2] [MasakhaNER: Named Entity Recognition for African Languages](#) (Adelani et al., TACL 2021)
- [3] [HiNER: A large Hindi Named Entity Recognition Dataset](#) (Murthy et al., LREC 2022)
- [4] [A Broad-coverage Corpus for Finnish Named Entity Recognition](#) (Luoma et al., LREC 2020)
- [5] [NEWS 2018 Whitepaper](#) (Chen et al., NEWS 2018)