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

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: Δ > 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

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

<table>
<thead>
<tr>
<th>Language</th>
<th>To English</th>
<th>From English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>90.34</td>
<td>88.31</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>87.02</td>
<td>78.17</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>80.56</td>
<td>79.30</td>
</tr>
<tr>
<td>Latvian</td>
<td>75.26</td>
<td>73.81</td>
</tr>
<tr>
<td>Tajik</td>
<td>51.56</td>
<td>56.82</td>
</tr>
<tr>
<td>Kazakh</td>
<td>49.14</td>
<td>58.30</td>
</tr>
<tr>
<td>Russian</td>
<td>45.65</td>
<td>43.26</td>
</tr>
<tr>
<td>Thai</td>
<td>39.59</td>
<td>15.02</td>
</tr>
<tr>
<td>Armenian</td>
<td>38.76</td>
<td>47.95</td>
</tr>
<tr>
<td>Georgian</td>
<td>32.67</td>
<td>51.00</td>
</tr>
<tr>
<td>Korean</td>
<td>32.28</td>
<td>42.57</td>
</tr>
<tr>
<td>Arabic</td>
<td>31.88</td>
<td>46.77</td>
</tr>
<tr>
<td>Greek</td>
<td>31.67</td>
<td>31.22</td>
</tr>
<tr>
<td>Japanese</td>
<td>29.97</td>
<td>27.30</td>
</tr>
<tr>
<td>Urdu</td>
<td>27.02</td>
<td>17.96</td>
</tr>
<tr>
<td>Persian</td>
<td>26.92</td>
<td>42.10</td>
</tr>
<tr>
<td>Hebrew</td>
<td>18.46</td>
<td>37.83</td>
</tr>
</tbody>
</table>

Micro-avg. 46.40 49.27

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/bltlab/paranames

References
[3] HiNER: A large Hindi Named Entity Recognition Dataset (Murthy et al., LREC 2022)