ParaNames: A Massively Multilingual Entity Name Corpus

Jonne Sälevä and Constantine Lignos
Brandeis University

SIGTYP Workshop, NAACL 2022

Brandeis

Introduction & motivation

- ➔ Lists of named entities can be beneficial for many NLP tasks (e.g. NER, MT), especially for lower-resourced languages
- ➔ Existing resources like Wikidata have serious data quality issues
- ➔ We address these and release *ParaNames*, a massively multilingual collection of entity names
  - ◆ Over 118 million names / 14 million entities / 400 languages
- ➔ Extended abstract @ SIGTYP 2022 ➡ 
- ➔ Full preprint available on Arxiv ➡ 

Data extraction & challenges

- ➔ Ingestion: Wikidata JSON dump ➡ MongoDB
  - ◆ Store a subset of the fields to save disk space
- ➔ Challenge 1: Script mixing within language codes
  - ◆ Many language codes indicate scripts used, but real data does not conform to it
- ➔ Solution: Approximate script identification & filtering
  - ◆ PyICU → Unicode script properties for each character in a name
  - ◆ Create script histogram →use argmax as estimate for script
  - ◆ Filter out names whose scripts are incorrect given language
  - ◆ Information on correct scripts manually collected from Wikipedia

Data extraction & challenges

- Challenge 2: How to assign types to each name?
  - Often very context-dependent
- Solution: Use Wikidata knowledge graph & instance-of relation
  - Instance of Q5 (human)? → PER
  - Instance of Q82794 (geographic region)? → LOC
  - Instance of Q43229 (organization)? → ORG
- Transitive, so being instance of a subclass is enough
  - Caveat: for PER, no inheritance allowed to reduce errors
- Not one-to-one: ~2.5% of entities get assigned multiple types.
- We leave these as-is since disambiguation requires context

Experiments

- Sample use case for ParaNames: *canonical name translation*
- Translating entity names from English to 17 languages and vice versa
  - Arabic, Armenian, Georgian, Greek, Hebrew, Japanese, Kazakh, Korean, Latvian, Lithuanian, Persian (Farsi), Russian, Swedish, Tajik, Thai, Vietnamese and Urdu
- Sample of languages represents variation in geographic location, orthographic systems, language families and typological features
- Parallel data: 80% train / 10% dev / 10% test, split by Wikidata IDs
- Prepend a “special token” to each name to indicate language
- Model: Character-level Transformer, trained for 90,000 updates
- Evaluation: accuracy, character error rate, mean “fuzziness” in F1 score

Language	Accuracy
Swedish	88.25 ± .02
Vietnamese	80.75 ± .02
Latvian	67.86 ± .02
Kazakh	55.38 ± .04
Tajik	49.62 ± .05
Lithuanian	47.39 ± .03
Thai	43.94 ± .05
Armenian	39.92 ± .05
Georgian	34.44 ± .02
Korean	33.27 ± .05
Russian	32.81 ± .06
Urdu	31.92 ± .03
Japanese	29.00 ± .04
Persian	28.68 ± .05
Arabic	25.74 ± .03
Greek	24.70 ± .03
Hebrew	15.24 ± .07
Overall	42.88 ± .02

X → English

- ➔ Best accuracy on Swedish, Vietnamese and Latvian. Sensible as all use Latin script.
- ➔ Latvian accuracy notably lower than Vietnamese → challenges with inflection?
- ➔ Next: Kazakh and Tajik. Both use Cyrillic script → nearly one-to-one with Latin
- ➔ Performance consistently worst on Hebrew
- ➔ Most likely caused by lack of vowels which the model must infer

Language	Accuracy
Swedish	85.60 ± .04
Vietnamese	48.86 ± .01
Latvian	69.28 ± .07
Kazakh	58.69 ± .09
Tajik	54.38 ± .02
Lithuanian	50.76 ± .09
Thai	14.80 ± .04
Armenian	50.45 ± .05
Georgian	51.82 ± .04
Korean	38.63 ± .05
Russian	44.59 ± .04
Urdu	14.14 ± .08
Japanese	28.70 ± .01
Persian	22.90 ± .05
Arabic	41.70 ± .02
Greek	29.67 ± .06
Hebrew	35.71 ± .03
Overall	43.57 ± .02

English → X

- ➔ When translating from English, performance rankings are quite similar to X → En
- ➔ Highest accuracy: Swedish and Latvian
  - ◆ Both Latin script, relatively few diacritics
- ➔ Followed by Kazakh and Tajik
  - ◆ Cyrillic script, nearly 1-to-1 with Latin
- ➔ Notable changes in accuracy
  - ◆ Hebrew: ⬆ 134% (no vowels)
  - ◆ Arabic: ⬆ 62% (no vowels)
  - ◆ Georgian: ⬆ 50% (phon. orthography)
  - ◆ Thai: ⬇ 66% (tone and vowel diacritics)
  - ◆ Vietnamese: ⬇ 39% (diacritics)

Conclusion

- ➔ We introduce *ParaNames*, the largest collection of entity names to date, covering approx. 14m entities in over 400 languages
- ➔ Many potential applications. We experiment with *canonical name translation* as an example use case
- ➔ We release our resource on GitHub under a CC BY 4.0 license, along with the code used to construct it (MIT licensed).
  - ◆ Target: quarterly updated releases
- ➔ For more, see www.github.com/bltlab/paranames ➡ 
- ➔ Also see our preprint on Arxiv ➡ 

Thank you!

Jonne Sälevä
jonnesaleva@brandeis.edu