Clustering of Analogies for Inter-Language Similarities

Software project - Final presentation

Justine Diliberto, Cindy Pereira, Anna Nikiforovskaja

Université de Lorraine, IDMC

03.02.2022

Summary

Subject

Aim of this project: Finding language similarities using...

- Morphological rules
- Analogies
- Clustering

```
cat : cats :: dog : dogs
```

Figure: Positive analogy

```
cat : dogs != cats : dog
```

Figure: Negative analogy

Dataset

SIGMORPHON 2020 shared task

- Task 0: Typologically Diverse Morphological Inflection
- Task 1: Multilingual Grapheme-to-Phoneme
- Task 2: Unsupervised Morphological Paradigm Completion

- 90 languages:
  - ⟨`lemma inflected_form features`⟩
  - ⟨`respect respected V.PTCP;PST`⟩

Language classification research

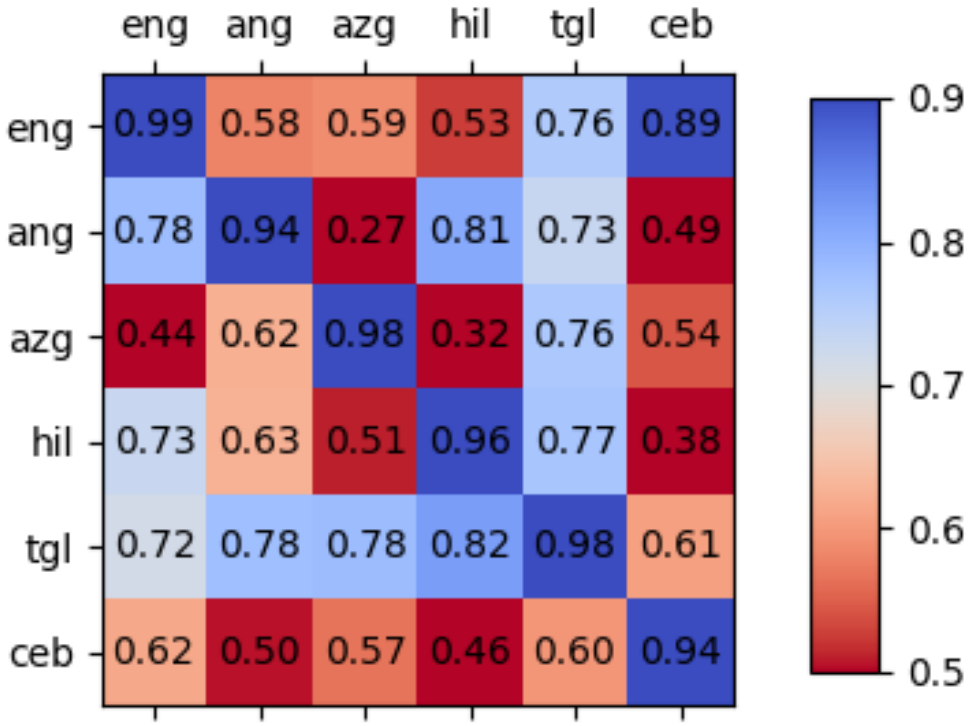
Different approaches:

- Lexico-statistical
  - Hierarchy of Indo-European languages
  - Analysis of lexical data (basic vocabulary)
- Computational
  - Dependencies
  - n-grams
- Phonetics, genetics, archaeology

First transfers

Issues and solutions:

- Dataset size influences performance
- Balanced training fixed
- Introduced f-score for better results representation



First transfers: F-score explanation

$$F = \frac{2 \cdot a_{\text{pos}} \cdot a_{\text{neg}}}{a_{\text{pos}} + a_{\text{neg}}}$$

F – the final score of the model on the language,
a_{pos} – accuracy of predicting positive analogies (a ratio of correct answers on positive analogies),
a_{neg} – accuracy on negative analogies.

Language analyses

- Germanic family
  - English
  - German
  - Swedish
- Uralic family
  - Finnish
  - Karelian
- Oto-Manguean family
  - Mezquital Otomi

Language analyses

- Germanic family
  - English
  - German
  - Swedish
- Uralic family
  - Finnish
  - Karelian
- Oto-Manguean family
  - Mezquital Otomi

Language analyses

- Germanic family
  - English
  - German
  - Swedish
- Uralic family
  - Finnish
  - Karelian
- Oto-Manguean family
  - Mezquital Otomi

Language analyses

Language	#Inflections
English	5
Mezquital Otomi	18
Swedish	34
German	35
Finnish	91
Karelian	161

Table: Number of different inflections per language

Language comparisons

Close language pairs:

- Finnish & Karelian
  - 51 similar inflections
  - indicative present 1st person sg: -*”n”* vs -*”an”*
  - 29 exact
- English & Swedish
  - some similar inflections
  - past participle: -*”ed”* / -*”en”* / *irreg* vs -*”ed”* / -*”en”*
  - 1 exact
- German & Finnish + Karelian
  - some similar inflections
  - nominative and accusative sg: *same inflected form*

Language comparisons

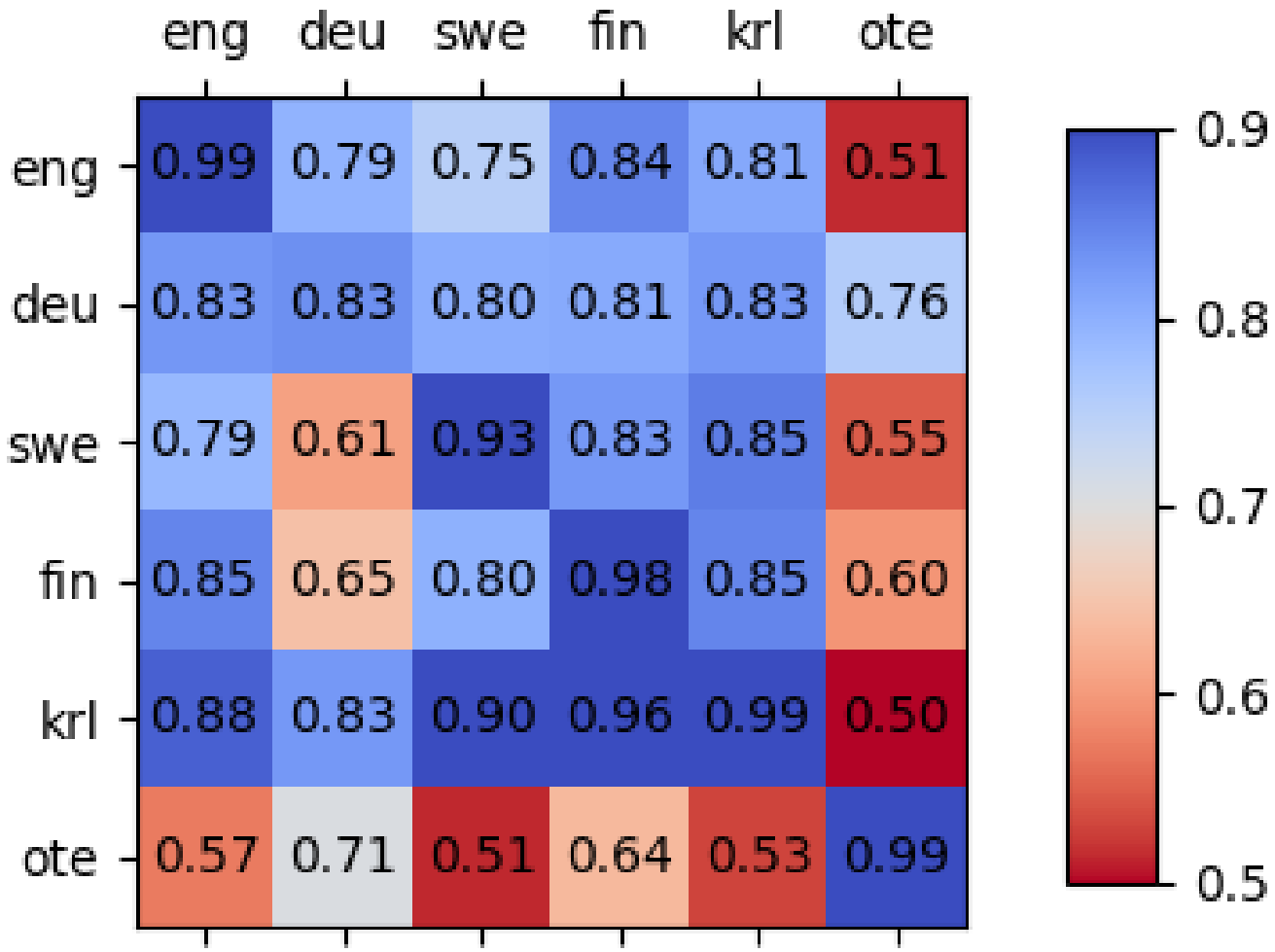
Far language pairs:

- English & German
- English & Finnish
- English & Karelian
- Swedish & Finnish
- Swedish & Karelian

Very far language pairs:

- Mezquital Otomi & the rest

Partial transfer performance



English (eng), German (deu), Swedish (swe), Finnish (fin), Karelian (krl), and Mezquital Otomi (ote)

Transfer analysis: expected

- Finnish and Karelian transfer really well – corresponds to rule analysis
- Karelian transfers well on all languages except for Otomi – highest number of different inflection rules
- Mezquital Otomi transfers poorly with every language – furthest language group

Transfer analysis: unexpected

- German doesn't transfer well with itself, but approximately same results in transfer from German to other languages
- English transfers better with Finnish and Karelian than with German or Swedish – unexpected, different groups
- Swedish also transfers better to Finnish and Karelian, than to its group

Website

Functionality:

- Show extracted rules
- Show exact rules
- Show transfer results
- Show well performing analogies

Why:

- For better results representation
- Easier rule comparison

Used microframework Flask.

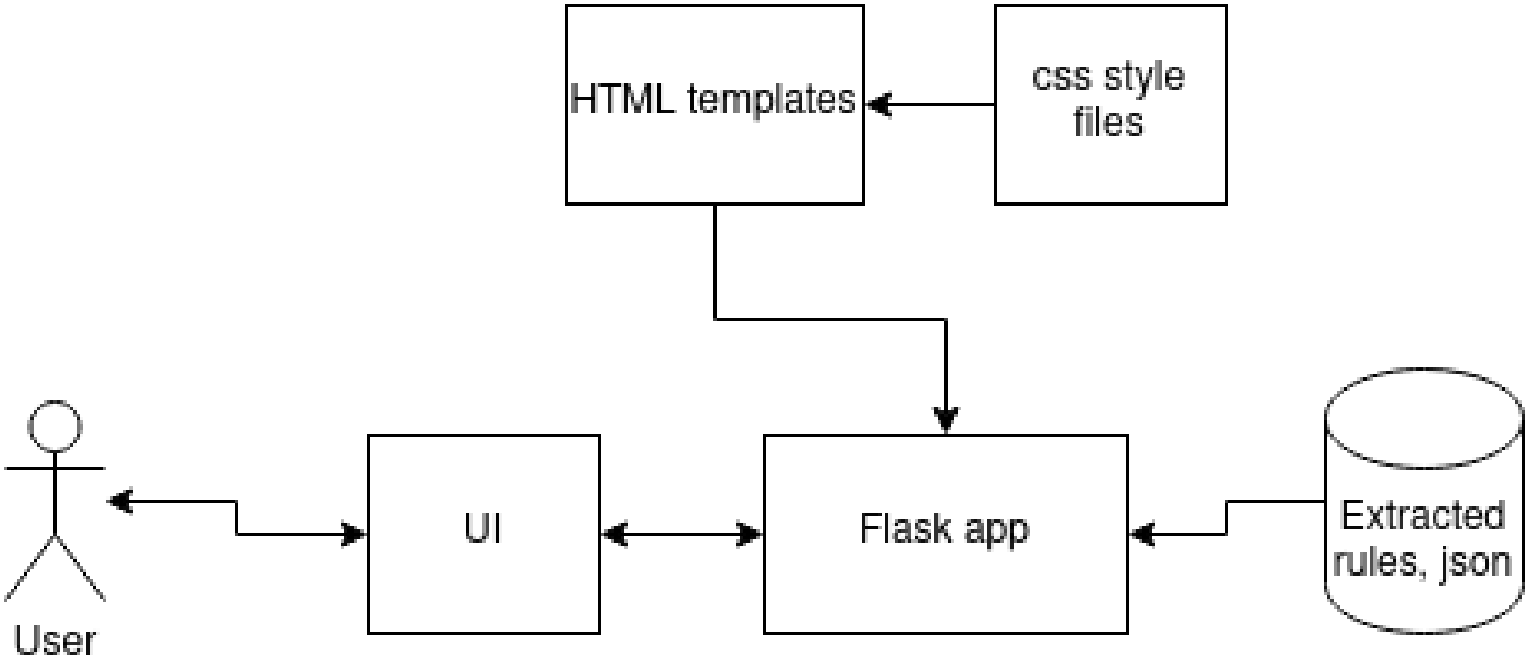
Rules format

Rules are in JSON format:

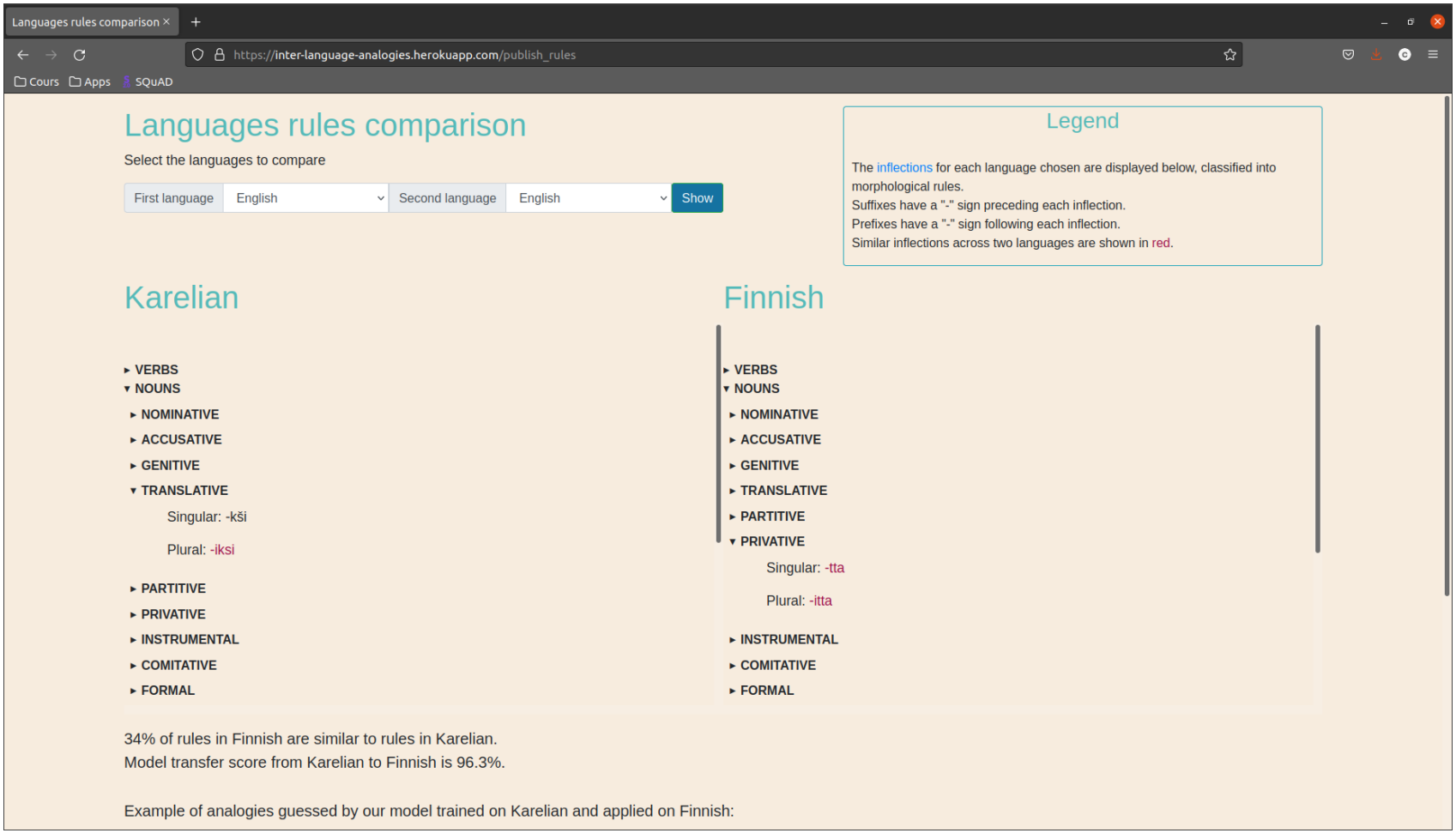
```
{"VERBS":
        {"INFINITIVE": "no change",
        "INDICATIVE":
                {"PRESENT":
                        {"Singular":
                                {"3rd person": "-s"}}},
        "PARTICIPLE":
                {"PRESENT": "-ing",
                "PAST": "-ed"},
        "PAST": "-ed"}}
```

Comparison based on tags sets

Website architecture



Demonstration time!



Conclusion

- Morphological rules have been studied
- Language similarity has been computed
- A website is available to display results

To continue:

- Further research about some results
- Add other languages

Thanks

Thank you for your attention!

Thanks



Thank you
for your attention!