Clustering of Analogies for Inter-Language Similarities

Software project - 5^{th} presentation

Justine Diliberto, Cindy Pereira, Anna Nikiforovskaja

Université de Lorraine, IDMC

10.12.2021

Summary of the project

Subject: Analogies between morphological rules

Main goal: find out about the closeness of languages and if they have common rules

Final product: Predict if two languages will transfer well, based on the rules they share

What was done before

- Listed possibly close and far pairs of languages
- Started to study some rules in these pairs
- Trained and ran the model on 6 languages

- Looked for new ways to simplify rule extraction
- Experimented with 2 tools
- Trained the multilingual model on old data and transferred to the new data

Rule extraction - issues

- Very time consuming
- Complex to do manually

n⁵ndʔɛʔ¹²	bá⁵tʔɛʔ¹²hâ³¹	V;IPFV;PL;1+EXCL;PRS
n⁵ndí⁵tzhwenʔ³	n⁵ndí⁵tzhwenʔ³	V;IRR;SG;3;FUT
n⁵ndí⁵khé⁵	kwí⁵khé⁵ha³	V;SBJV;SG;1
n⁵ndí⁵chiʔ³⁵	chi⁵³	V;IPFV;SG;1;PRS
n⁵ndí⁵ ʔ⁵han³	kwí⁵ʔ⁵han³ʔu³	V;SBJV;SG;2
n⁵ndi³nkiʔ³ xken³	n⁵ndi³nkiʔ³nkèn¹oʔ³	V;IRR;PL;3;FUT
n⁵ndí⁵hnduáʔ⁵	tí⁵hnduáʔ⁵	V;PFV;SG;3;PST
n⁵ndí⁵tzhaʔ⁵³xken³	to³ndí⁵tzhaʔ⁵³nkên³¹	V;PROG;PL;1+EXCL;PST
n⁵ndyio³ xú⁵	tyio³ʔo³xú⁵	V;PFV;PL;3;PST
n⁵ndì⁵be³⁴n⁵no³	n⁵ndí⁵be³⁴nↃ⁵³	V;IRR;SG;1;FUT

Figure: Some morphological inflections of Amuzgo

Rule extraction - ideas

- Heard about ALEA and Lepage during meeting
- Read papers about morphological rules extraction
- Found 2 tools to try:
  - AutoLEX webpage
  - NLG module (Lepage)

Rule extraction - AutoLEX

AutoLEX Project:

- Create descriptive grammars automatically
- Focus on ”agreement” rules between head and dependent token
- All languages from Universal Dependencies
- Still in progress

Rule extraction - AutoLEX

AutoLEX Language Descriptions Explorer [1]

Output: agreement + case marking + word order

Early results:

- Only 8 languages are completely extracted
- 14 languages from our dataset are partially done
- Rest is to be filled

[1] https://neulab.github.io/autolex/index.html

Rule extraction - AutoLEX

Examples of rules on the Finnish language:

```
Tense  relation,head-pos,child-pos
conj,VERB,VERB  conj,AUX,VERB  conj,VERB,AUX  conj,AUX,AUX  mod,AUX,VERB  mod,VERB,VERB
mod,VERB,AUX  mod,AUX,AUX
```

Figure: Some agreement rules about the Tense

```
Person  relation,head-pos,child-pos
conj,VERB,VERB  subj,VERB,PRON  conj,AUX,AUX  subj,AUX,PRON  conj,AUX,VERB  parataxis,AUX,VERB
parataxis,AUX,AUX  mod@relcl,VERB,VERB  comp:aux,AUX,VERB  mod@relcl,VERB,AUX
```

Figure: Some agreement rules about the Person

- Few data available
- About word relations
- Maybe combine with NLG?

Rule extraction - NLG

Creates analogical grids:

walk	walks	walking	walked
show	shows	showing	showed
open	opens	opening	
study		studying	
play		playing	played

Figure: Example of an analogical grid obtained by NLG - example taken from their article

Rule extraction - NLG

- Creates vectors with morphosyntactic description as features

```
n⁵ndí⁵chi?³⁵      chi⁵³      V;IPFV;SG;1;PRS
```

Figure: Example of morphosyntactic description in Amuzgo

- Extracts analogical rules using these vectors

```
kwí⁵ntyén⁵?u³ : kwí⁵sì¹?u³ :: to³ndí⁵ntyén⁵hâ³¹ : to³ndí⁵sì¹hâ³¹
kwí⁵ndà?¹hâ³¹ : n⁵ngo³ndí⁵ndà?¹?o³ :: kwí⁵kí⁵chì¹hâ³¹ : n⁵ngo³ndí⁵kí⁵chì¹?o³
to³nchhe³o?³ : to³hndɔ?¹²b?i¹² :: nchhe³o?³ : hndɔ?¹²b?i¹²
```

Figure: Example of analogical rules in Amuzgo

Rule extraction - NLG

Problem: not always readable and easy to analyse

- Print analogical rules with other format

```
n⁵ngo³nchhe³     : ki³nchhe³
n⁵ngo³ndui³⁴ʔu³  : ki³ndui³⁴ʔu³
n⁵ngo³ndui ʔ³o ʔ³ : ki³ndui ʔ³o ʔ³
n⁵ngo³tyiɔ³⁴     : ki³tyiɔ³⁴
```

- Better but still some problems
  - Many gaps in these rules
  - Need manual analysis and comparison

Running multilingual model

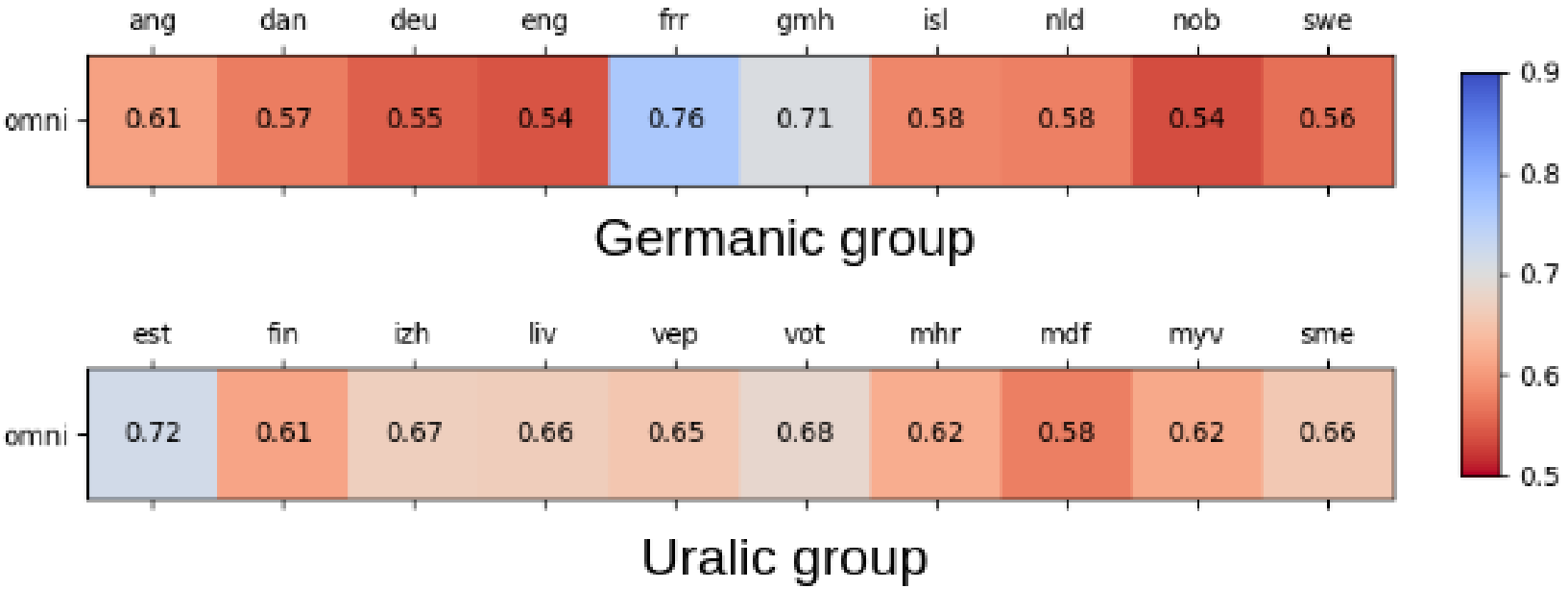
- Multilingual model trained on old data:
  - Arabic, Finnish, Georgian, German, Hungarian, Japanese, Maltese, Navajo, Russian, Spanish, Turkish
- Transfer to new languages.
  - Germanic group, Uralic group

Multilingual model: transfer results

Full transfer, 50000 analogies.

$$F_1 = 2 \cdot \frac{p \cdot n}{p + n}, \text{ where:}$$

- p — accuracy on positive analogies
- n — accuracy on negative analogies



Observations of the transfer

Three languages with the best results:

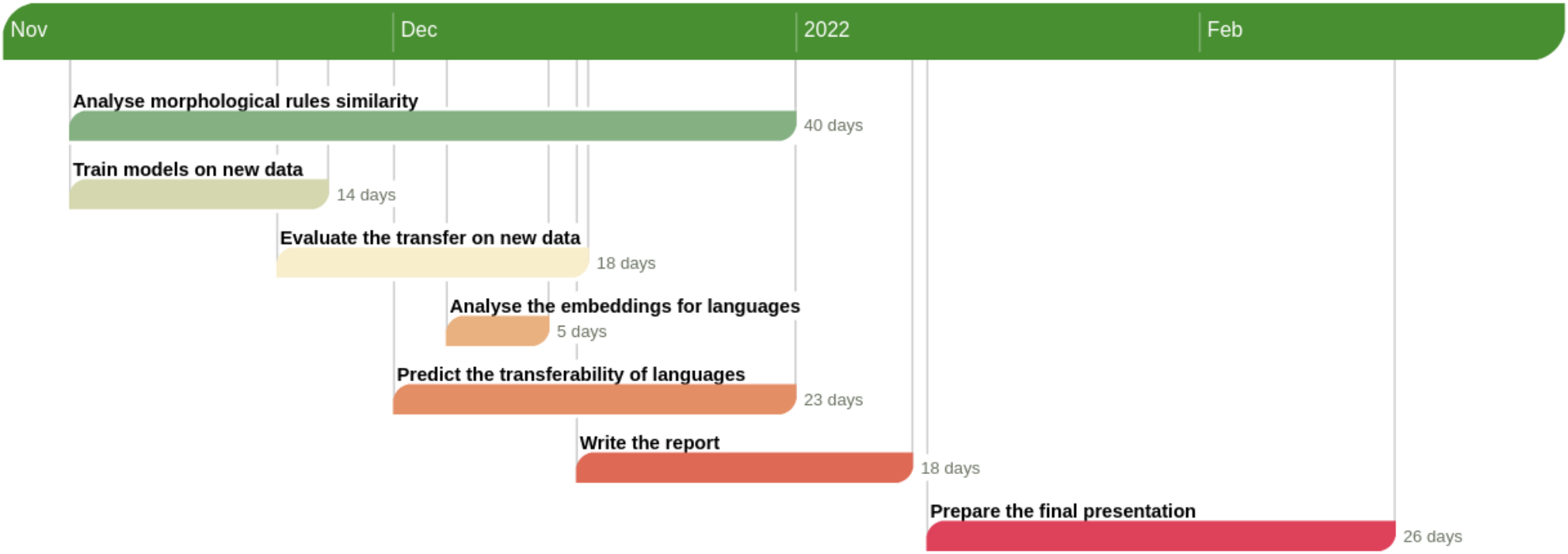
- North Frisian – small dataset
- Middle High German – small dataset
- Estonian
  - Close to Finnish from the old dataset
  - Why did not transfer well to the new Finnish?
    - New Finnish dataset: many adjectives
    - Estonian: only nouns and verbs
    - Old Finnish dataset: mainly noun and verbs

Observations of the transfer

In general – Uralic group performed better, than Germanic group of languages.

Probable reason: Finnish and Hungarian in the old dataset, while only German from Germanic group. Data imbalance.

Timeline


Nov
Dec
2022
Feb
Analyse morphological rules similarity
40 days
Train models on new data
14 days
Evaluate the transfer on new data
18 days
Analyse the embeddings for languages
5 days
Predict the transferability of languages
23 days
Write the report
18 days
Prepare the final presentation
26 days

Thanks

Thank you for your attention!