

Clustering of Analogies for Inter-Language Similarities

Software project - 4th presentation

Justine Diliberto, Cindy Pereira, Anna Nikiforovskaja

Université de Lorraine, IDMC

22.11.2021



UNIVERSITÉ
DE LORRAINE



Institut des
sciences du Digital
Management & Cognition

Summary of the project

Subject: Analogies between morphological rules
to stay -> stayed PST
to play -> ? PST
= PLAYED

Main goal: continuation of work done by Safa et al. to find out more 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

- adapted new dataset to baseline
 - ran a full transfer on new languages
 - read articles about language similarities
 - listed many possibly close pairs of languages
 - started studying some rules in these pairs
-
- listed possibly far pairs of languages
 - trained and ran the model on 6 languages (from development)
 - studied some morphological rules in these pairs

Far language pairs

Methodology:

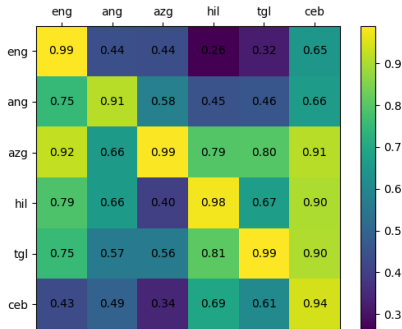
- find graphs symbolizing language families
 - many opposite families
- look at the world map from SIGMORPHON showing where the languages come from
 - many opposite geographical areas
- search for papers about specific pairs of languages
 - few studies include the languages of our dataset

Balanced training

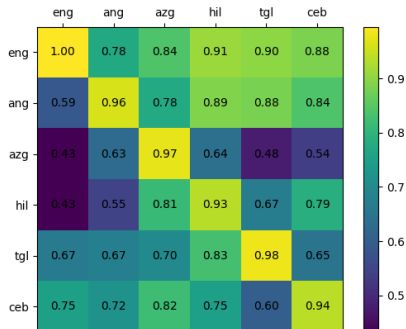
- Found a bug, which decreased a model quality, in the script for balanced training of the models. Fixed the generation of negative analogies in the case of $a : a$ as $c : c$.
- Ran training on English, Old English, Amuzgo, Cebuano, Hiligaynon and Tagalog from the new dataset on 50000 analogies.
- Performed full model transfer and partial model transfer for the above languages.

Full transfer results

Positive 50000 analogies



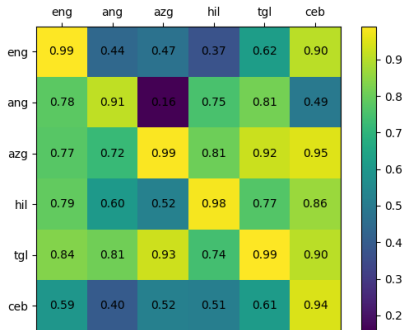
Negative 50000 analogies



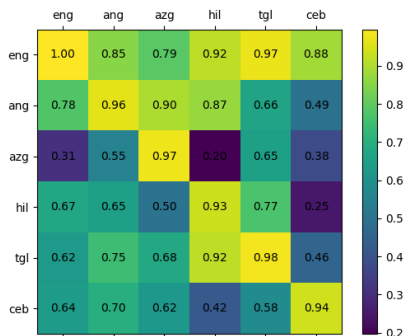
English (eng), Old English (ang), Amuzgo (azg), Hiligaynon (hil), Tagalog (tag), Cebuano (ceb)

Partial transfer results

Positive 50000 analogies



Negative 50000 analogies



English (eng), Old English (ang), Amuzgo (azg), Hiligaynon (hil), Tagalog (tag), Cebuano (ceb)

Observations

- English is not transferable to Old English, while Old English is very well transferable to English.
- Transfer from English to Cebuano works well
- Hiligaynon and Tagalog are well transferable to each other.

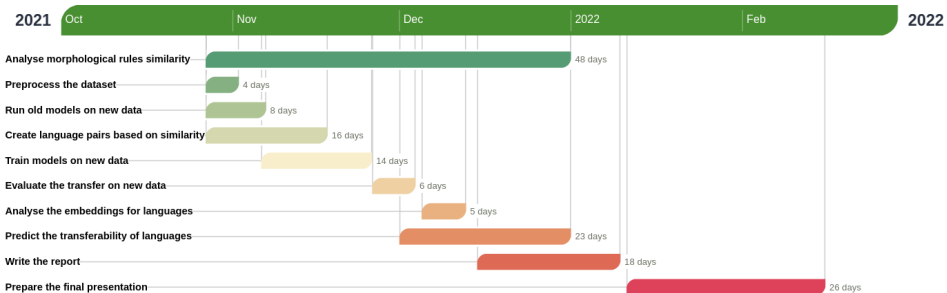
Explanations

- Old English: high variation in our data
 - Presence of verbs, adjectives, nouns
- English: poor variation in our data
 - Presence of only verbs
- 20%-22% of English and Cebuano data is infinitive
 - No change between lemma and inflected form
- In e.g. 21% of the cases, analogy is easy because it doesn't change

Comparison between participle past between English and Old English

- English: In most cases, add -ed in the end
 - defend defended V.PTCP;PST
- Old English: Replacement of the end of words
 - -ian becomes -od
 - talian talod V.PTCP;PST
 - -an becomes -en or -ed
 - bacan bacen V.PTCP;PST
 - cepan cēped V.PTCP;PST

Timeline



Thank you for your attention.