Demonym gazetteers

David Sánchez & Alex Pardo

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Content

- Introduction
- Description of the problem
- Our system
  - Evaluation
  - Results
  - Future work
- Conclusions

Introduction

- Demonym or gentilic, is a term for the residents of a locality.

  Spain -> Spanish
  Africa -> African

- Build an automatic system capable to make transformations and generate the demonym from a country or city.

Description of the problem

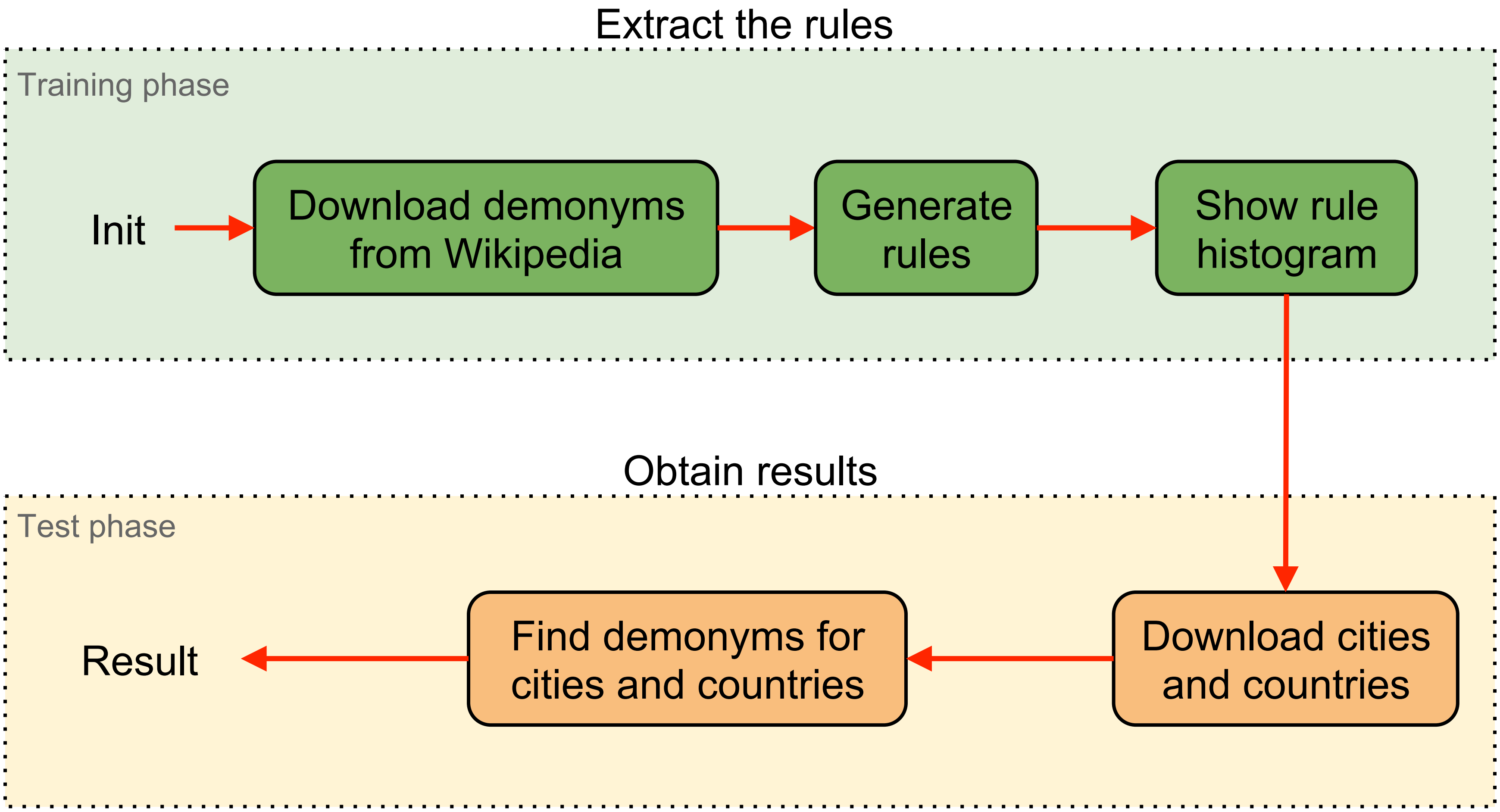
- Different types of transformations:
    - Adding and substitution

- A lot of irregular cases.
- One country or city can have more than one demonym.
    - Iran -- Iranian (also "Irani" or "Persian")

Our system

- Two phases
    - Download demonym from Wikipedia, extract the rules and build the system.
    - Download unknown countries and cities and generate demonyms.

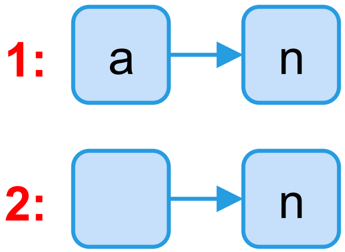
Our system


Extract the rules
Training phase
Init
Download demonyms from Wikipedia
Generate rules
Show rule histogram
Obtain results
Test phase
Download cities and countries
Find demonyms for cities and countries
Result

Extract rules

- Extract rules
  - Two types of rules:
    - **1:** Remove the current suffix and add another one.
    - **2:** Add a suffix to the word.

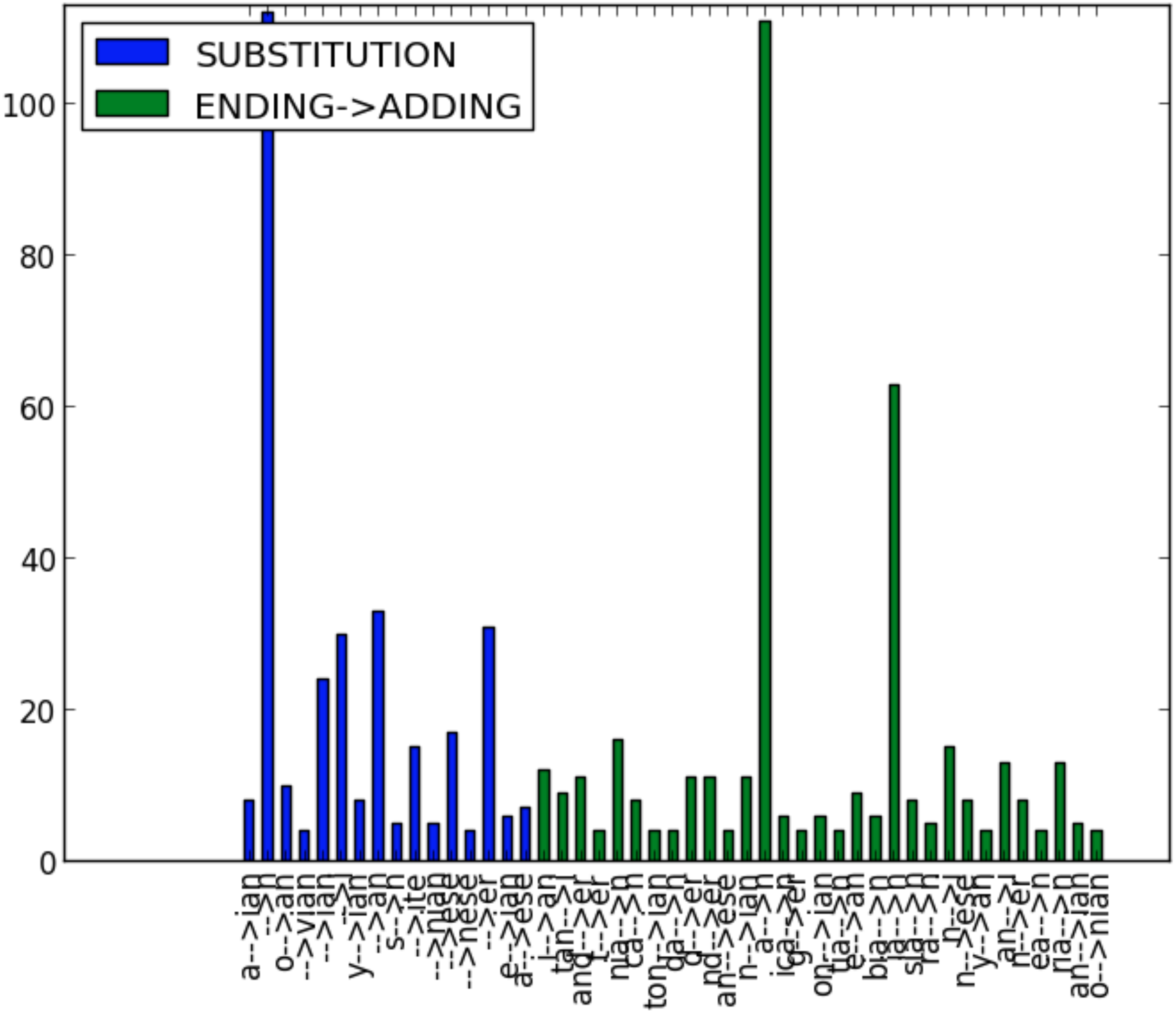
Example: Africa -> African



1: a → n

2: → n

Extract rules - Histogram



Obtain results

- Download new countries and cities.
- Apply all the possible rule.
- Find all the occurrences on the WP page of the location.

Evaluation

In order to evaluate the system we compute the accuracy:

$$AC = \frac{TP}{SAMPLES}$$

Results

- 410 training samples
- 31 add and 16 replace rules
- total of 47 rules (TH = 3), originally 791
- 55 TP for countries, 89 samples: 61.8%
- 89 TP for cities, 1751 samples: 5.08%

Discussion

- Good values for countries
- Bad results for cities:
  - System trained with countries
  - Some WP pages for cities are incomplete (the demonym does not appear)
  - High irregularity

Future work

- Use a larger number of training examples
- Use more complex models (e.g. analyse the words in terms of lexemas and its derivations)
- Try to generalize the rules

Conclusions

- We needed a lot of regular examples and discard the irregular forms.
- The combination of the two types rules is a good option since explains the main cases.

- Difficult to obtain good results because exist a lot of irregular cases.

Questions?