Laboratory Case ATNLP 2013-2014

We are interested on automatically building a gazetteer of demonyms for allowing associating to a country or city the name given to their inhabitants (France: French, Peru: Peruvian, and so). We will use the Wikipedia, WP, as knowledge source.

Searching the English WP by "demonym" the information put in the appendix (summarized) is obtained. In the retrieved page about 500 demonyms occur. A sample of relevant lines (a lot of noise occurs) follows:

```
* Africa -> African
* Croatia -> Croatian (also "Croat"),
* [North / South] Korea -> [North / South] Korean,
* Hanoi (Vietnam) -> Hanoian
* Iran -> Iranian (also "Irani" or "Persian")
* Florence -> Florentine (also Latin "Florentia")
* Ann Arbor -> Ann Arborite
* Israel -> Israelite (also "Israeli", depending on the usage; see
below)
* Netherlands -> Netherlander (though see below; Irregular forms)
```

We propose using FS technology for facing the problem of getting from a location the possible set of valid demonyms (note that LOCATION -> DEMONYM can be seen as a derivation) in order to collect as much as possible pairs for building the gazetteer.

A FS machine for performing LOCATION → DEMONYM transformations could be organized into a concatenation of three FS machines:

- FSA_{LOC} able to recognize/generate a LOCATION
- FST_{PC} able to perform phonological changes (PC) at the end of LOCATION string.
- FSA_{SUF} able to recognize/generate DEMONYM suffixes

Note that the first and last components are automata while the second is a transducer. Additionally a morphotactic filter has to be applied (not all the phonological changes or suffixes are appropriate for a given LOCATION).

You should propose and implement a way of learning automatically the basic components of the system (roots, affixes, phonological changes, morphotactics) and later apply the system to WP in order to build the gazetteer of demonyms.

FSA_{LOC} can be straightforwardly got from a geographic gazetteer (as Geonames or GNIS), filtering out not Geopolitical entities (countries, states, provinces, cities, …). Alternatively you could use the categories of WP and their subcategory links for getting for a given LOCATION type the set of sub-categories (for instance the category COUNTRY in English WP has 21 sub-categories, one of them COUNTRIES_BY_CONTINENT contains in turn 12 categories, and so on. This structure is far to be taxonomic and has a lot of noise but most of the pages linked to its members are really countries. In this way the set of countries, states, provinces, cities, … in WP can be extracted and easily converted into a FSA, FSA$_{LOC}$.

For obtaining FSA$_{SUF}$ we will assume that the set of suffixes is exhaustively provided in the DEMONYM WP page and has to be extracted from it. The procedure can be the following:

- From the text of the DEMONYM WP page we can easily extract most of the occurring pairs <LOCATION, DEMONYM> using hand made RE patterns as
    - ‘^* (.+) \-\> (.+) .*$’ for extracting the pair <”Africa”, “African”> from the line “* Africa -> African”.
    - '^* (.+) \-\> (.+) \(also \"?([^\"]+)\"?\).*$' for extracting the pairs <”Croatia”, “Croatian”> and <”Croatia”, “Croat”> from the line “* Croatia -> Croatian (also "Croat")”.

  I guess that about twenty of such RE are enough for collecting most of the pairs occurring in the text. Discard exceptions or rare pairs.

- From the collected set of pairs we extract all the ending letter ngrams of the DEMONYM not occurring as endings of the corresponding LOCATION, for instance from the pair <”Africa”, “African”> the suffix “n” is extracted, from the pair <”Iran”, “Iranian”>, the suffixes “ian”, “an”, and “n” are extracted. We can discard the less frequent suffixes and manually revise the remainder. Only a couple of dozens of suffixes are enough for our purposes. The set of suffixes can be then straightforwardly transformed into a the FSA FSA$_{SUF}$.

Building FST$_{PC}$ is not so simple. First we have to decide which changes are allowed and how to represent them. We can accept, for instance, insertions, changes and deletions of characters and represent the phonological changes as simply a pair <lower string, upper string>. Then for each triple <LOCATION, DEMONYM, SUFFIX> the phonological changes can be extracted. For instance:

- For the triple <”Hanoi”, ”Hanoian”, ”ian”> the PC is <”i”,””>, i.e. the derivation will be ”Hanoi” -> {replacing the final “i” by “”} -> ”Hano” -> {adding the suffix “ian”} -> ”Hanoian”.
- For the triple <”Hanoi”, ”Hanoian”, ”an”> there are no phonological changes, so, PC is <””,””> and the derivation is simply ”Hanoi” -> {adding the suffix “an”} -> ”Hanoian”.

I guess than less than 50 PC can be obtained in this way from the set of triples. As in the case of suffixes we can discard the less frequent PC and manually revise the remainder. From these data $\mathrm{FST}_{\mathrm{PC}}$ can be built.

We can, thus, obtain $\mathrm{FSA}_{\mathrm{LOC}}$, $\mathrm{FSA}_{\mathrm{SUF}}$, and $\mathrm{FST}_{\mathrm{PC}}$ with very limited human intervention.

The most difficult task is, however, getting the morphotactics, i.e. deciding which PCs can be applied and which suffixes added to a specific LOCATION. The most natural way of implementing it is classifying each LOCATION by its ngram endings excluding the initial letter, e.g. for "Hanoi" the possible classes could be "anoi", "noi", "oi" and "i". We can limit the size of these ngrams to, for instance, 3 letters.

Assigning to each class the set of pairs <PC, suffix> cannot be done manually because there is no clear linguistic theory explaining the derivations, that are basically idiosyncratic. So we need to learn the assignments using ML techniques. There is, however, a problem. We have not enough training material. The possible assignments are the Cartesian product of PCs and Suffixes, i.e. the order is of hundreds. The examples are the triples <DEMONYM, PC, SUFFIX>, i.e. about one thousand. The assignee are combinations of up to 3 letters occurring in LOCATION' endings (the upper bound could be $28 + 28^2 + 28^3$, i.e about 22,000. It is clear that we need more examples, although not so clean as initial ones.

A simple way of getting additional examples could be the following:

- We collect all WP pages corresponding to LOCATIONs. This can be done easily using WP categories (COUNTRY, CITY, STATE, PROVINCE, and so.) using the procedure described above for building $\mathrm{FSA}_{\mathrm{LOC}}$.
- For each of these pages we apply to the name of the page the FS machine built by concatenating $\mathrm{FSA}_{\mathrm{LOC}}$, $\mathrm{FST}_{\mathrm{PC}}$, and $\mathrm{FSA}_{\mathrm{SUF}}$ without morphotactics filtering. In this way we obtain a huge set of possible DEMONYMs, most of them wrong.
- We look in the text of the page for the occurrences of any of the possible DEMONYMs we collected in the previous step. We select as most likely DEMONYM the most frequent (if any). We hope that the accuracy of the procedure could be high (wrong elements of the set of possible DEMONYMs are not likely to occur in the WP page). For instance for the country "France" the possible demonyms could be "French", "Francean", "Francian", etc. It is likely that in the page "France" "French" could occur but not "Francean" or "Francian".

- I guess that we can collect with this simple procedure several thousands of examples with a limited amount of noise and without human intervention.

I guess that with the initial small set of clean examples and the huge set of new examples, slightly noisy, a classifier could be learn.

Appendix Extract of the page “DEMONYM” of English WP

Suffixation
The English language uses several models to create demonyms. The most common is to add a suffix to the end of the location's name, slightly modified in some instances. These may be modeled after Late Latin, Semitic, Celtic or Germanic suffixes, such as:
-(a)n
(countries / continents:
* Africa -> African
...
* Armenia -> Armenian*
...
* Croatia -> Croatian (also "Croat"),
...
* [North / South] Korea -> [North / South] Korean,
...
* Serbia-> Serbian (also "Serb"),
...
cities / states:
* Alaska -> Alaskan
...
* Hanoi (Vietnam) -> Hanoian
...
* Wallachia -> Wallachian)
...
-ian
countries:
* Bahamas -> Bahamian
...
* Iran -> Iranian (also "Irani" or "Persian")
...
cities / states:
* Adelaide -> Adelaidian
...
* Brisbane -> Brisbanian (also "Brisbanite")
...
-nian
* Bendigo -> Bendigonian
...
-in(e)
* Argentina -> Argentine

* Florence -> Florentine (also Latin "Florentia")
* Montenegro -> Montenegrin
...
-ite
* Ann Arbor -> Ann Arborite
...
* Israel -> Israelite (also "Israeli", depending on the usage; see below)
...
-(e)r
* Amsterdam -> Amsterdammer
...
* Netherlands -> Netherlander (though see below; Irregular forms)
...
* New Zealand -> New Zealander (Kiwi)
...
-(en)o
* Los Angeles -> Angeleno or Los Angeleno,
* Philippines -> Filipino
adapted from a standard Spanish suffix -(eñ/n)o, as in salvadoreño, madrileño, malagüeño, Zamboanga City -> Zamboangueño, andorrano, or chino
-ish
* Åland -> Ålandish
...
"-ish" is usually only proper as an adjective. Thus many common "-ish" forms have irregular demonyms, e.g. Britain/British/Briton; Denmark/Danish/Dane; England/English/Englishman; Finland/Finnish/Finn; Flanders/Flemish/Fleming; Ireland/Irish/Irishman; Kurdistan/Kurdish/Kurd; Poland/Polish/Pole; Scotland/Scottish/Scot; Spain/Spanish/Spaniard; Sweden/Swedish/Swede; Turkey/Turkish/Turk.
-ene
* Cairo -> Cairene
...
-ensian
* Kingston-upon-Hull (UK) -> Hullensian
-ard
* Spain -> Spaniard
* Savoy -> Savoyard
-(l)ese
* Aragon -> Aragonese
...
* Guangdong ("Canton") -> Cantonese
...
* Macao -> Macanese/Chinese
...
* South Sudan -> (South) Sudanese
...
"-ese" is usually considered proper only as an adjective, or to refer to the entirety. Thus, "a Chinese person" is used rather

than "a Chinese". Often used for East Asian and Francophone locations, from the similar-sounding French suffix -ais(e), which is originally from the Latin adjectival ending -ensis, designating origin from a place: thus Hispaniensis (Spanish), Danensis (Danish), etc.

-i

* Afghanistan -> Afghanistani

...

* Israel -> Israeli (in the Modern State of Israel)

...

Mostly for Middle Eastern and South Asian locales and in Latinate names for the various people that ancient Romans encountered (e.g. Allemanni, Helvetii)

-ic

* Hispania -> Hispanic
* Turk -> Turkic

Derives from a Latinate suffix widely used outside ethnonyms (e.g., chemical compounds), which with regard to people is mostly used adjectivally (Semite vs. Semitic Arab/Arabian vs. Arabic) to refer to a wider ethnic or linguistic group (Turkic vs. Turkish, Finnic vs. Finnish).

-iot(e)

* Corfu -> Corfiot

...

Used especially for Greek locations.

-asque

* Menton -> Mentonasque
* Monaco -> Monégasque

-gian

* Galloway -> Galwegian

...

-onian

* Aberdeen -> Aberdonian

...

-vian

* Kraków -> Krakovian

...

Irregular forms

There are many irregular demonyms for recently formed entities, such as those in the New World. There are other demonyms that are borrowed from the native or another language.

In some cases, both the location's name and the demonym are produced by suffixation, for example England and English and English(wo)man (derived from the Angle tribe). In some cases the derivation is concealed enough that it is no longer morphemic: France -> French (or Frenchman/Frenchwoman) or Flanders -> Flemish or Wales -> Welsh.

In some of the latter cases the noun is formed by adding -man or -woman, for example English/Englishman/Englishwoman; Irish/Irishman/Irishwoman; Chinese/Chinese man/Chinese woman (versus the archaic or derogatory terms Chinaman/Chinawoman).

From Latin or Latinization

* Ashbourne -> Ashburnian (Essiburn)

...

* Newcastle -> Novocastrian (Novum Castrum)

...

From native or other languages

* Aberteifi -> Cardi
* Andhra -> Andhraite
* Aguascalientes (lit. "hot waters") -> Hidrocálido, from Mexico's state and city.
* Barbados -> Bajan A colloquial term a shortened form of Barbadian -> Bar-bajan -> Bajan
* Birmingham -> Brummie

...

* Fontainebleau -> Bellifontain (from French)

...

Germanic: *þeudiskaz (all three meaning "national/popular"))

* Nice -> Niçois (from French)

...

* The Hague -> Hagenees (people born in the inner city), Hagenaar (people born elsewhere)
* Twente -> Tukker
* Vanuatu -> ni-Vanuatu

Irregular singular forms

* Bali -> Balinese

...

* Connecticut -> Connecticuter (uncommon), Nutmegger (common)

...

* Maryland -> Marylander (pron.: /->mær?l?nd?r/ MARR-i-l?nd-?r, sometimes /?m??rl?nd?r/ MAIR-l?nd-?r)
* Massachusetts -> Bay Stater

...

* Oklahoma -> Okie (derogatory), Oklahoman (formally)

...