Introduction to Natural Language Processing (INLP) - MAI

TERM 2012-2013 Final Exam Time: 2hours

We are interested on automatically building a gazetteer of *demonyms* for allowing associating of a country or city to the name given to their inhabitants (France, French, Peru, Peruvian, and so). By now we want to work in English but the idea is that the system should be applied (after tuning) to other languages. We will use the Wikipedia as knowledge source.

In the annex of this document you can find (a summary) of the information obtained when searching the English Wikipedia by "demonym". In the retrieved web page 500 demonyms occur. All but the most significant instances haven removed and replaced by '...'.

You can see that general rules are easy and highly productive but a lot of exceptions occur and the way of expressing the pairs includes many variants (what it occurs frequently in Wikipedia).

You have to answer to the following questions:

1) Discuss the problems of applying the rules described in the text of the annex for automatically processing English texts.
2) How the pairs <LOCATION, DEMONYM> could be represented?
3) Discuss the feasibility of building a fully automatic extraction system or a system with limited human intervention. In the second case try to quantify the cost (hours/person) of the human intervention.
4) In the case of using a grammar for extracting the patterns (ways of expressing <LOCATION, DEMONYM> pairs) answer the following:
    a. Type of the grammar (RG, CFG, DCG, ...).
    b. Justify the choice.
    c. The grammar should be learned or written manually?.
    d. Propose your approach for learning or writing the grammar.
    e. Include some significant examples of the grammar rules.
5) If you are not using a grammar explain in depth your approach.
6) Discuss the language dependence and the way of transporting a system written for a specific language to another language.

Annex Summary of the page "DEMONYM" of English Wikipedia

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)

...