Co-funded by the
Erasmus+ Programme
of the European Union

[NAMED] ENTITY RECOGNITION

Prof. Jesús Vilares
jesus.vilares@udc.es

UNIVERSIDADE DA CORUÑA

Purpose

- Identifying spans of text corresponding to:
  - **Named entities**: proper names denoting people, geographical locations, organizations, etc.
  - **Temporal expressions**: dates, times, etc.
  - **Numerical expressions**: measurements, counts, prices, etc.

Citing high fuel prices, [$_{\text{ORG}}$ United Airlines] said [$_{\text{TIME}}$ Friday] it has increased fares by [$_{\text{MONEY}}$ $6] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ American Airlines], a unit of [$_{\text{ORG}}$ AMR Corp.], immediately matched the move, spokesman [$_{\text{PER}}$ Tim Wagner] said. [$_{\text{ORG}}$ United], a unit of [$_{\text{ORG}}$ UAL Corp.], said the increase took effect [$_{\text{TIME}}$ Thursday] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ Chicago] to [$_{\text{LOC}}$ Dallas] and [$_{\text{LOC}}$ Denver] to [$_{\text{LOC}}$ San Francisco].

[ORGanization] [amount of MONEY]
[geo. LOCation] [TIME expr.] [PERson]

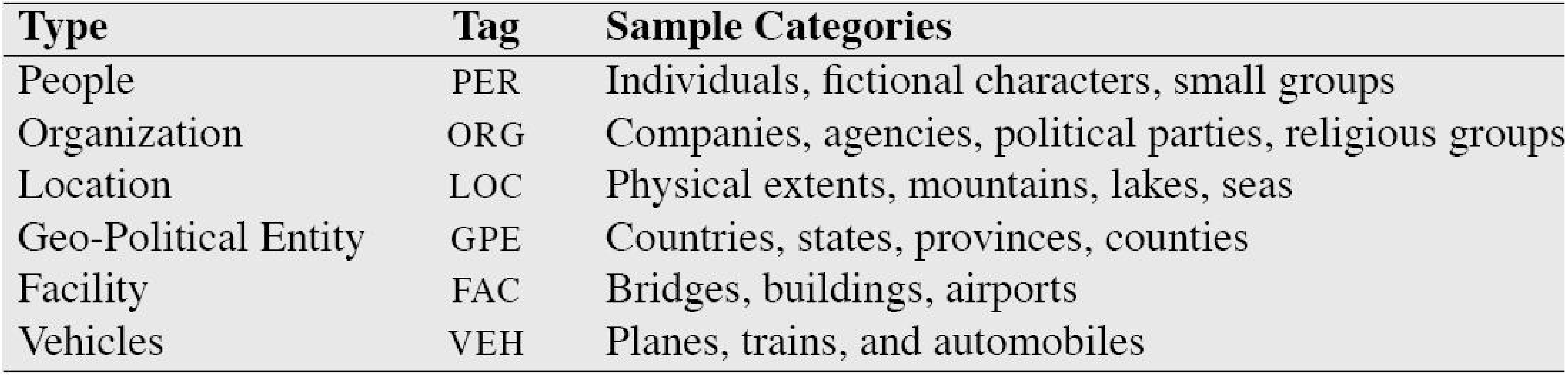
Process: Detection

- Consists of two phases (not necessarily separated):

1. **Detection**; i.e. to identify the piece of text that forms an entity based on (among others):
    - Shape differences: all caps (e.g. “EU”), presence of digits (e.g. “U2”), mixed case (e.g. “eBay”), etc.
    - Use of *gazetteers*: specialized dictionaries of [sur]names of people (e.g. “María”, “Alexei”, “Obama”), locations (e.g. “Spain”, “Paris”, “Beverly Hills”), organizations (e.g. “United Nations”, “UNICEF”, “Manchester United”, “Amazon”), etc.)
    - Predictive words: words denoting an entity type (e.g. “company”), a position (e.g. “president”), a title (e.g. “Mr.”), commercial abbreviations (e.g. “Ltd.”), etc.
    - Presence of symbols: $, €, %, etc.

Process: Classification

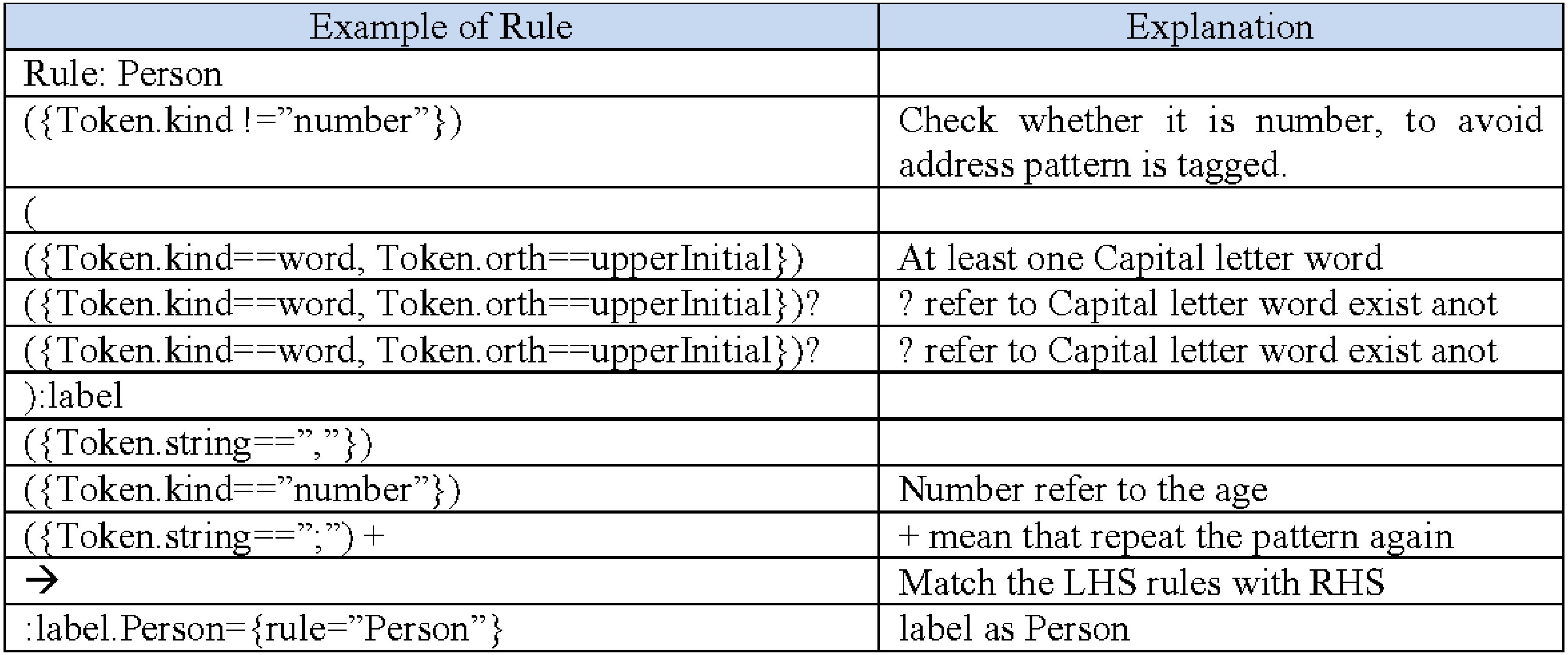
2. **Classification**; i.e. if that expression is a person name, a geographical location, organization, etc.
   - With respect to the taxonomy to be used according to the requirements of the system and the application context



Type	**Tag**	**Sample Categories**
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Process: Classification

- Example of rules for PERson identification



Example of Rule	Explanation
Rule: Person	
({Token.kind !=”number”})	Check whether it is number, to avoid address pattern is tagged.
(	
({Token.kind==word, Token.orth==upperInitial})	At least one Capital letter word
({Token.kind==word, Token.orth==upperInitial})?	? refer to Capital letter word exist anot
({Token.kind==word, Token.orth==upperInitial})?	? refer to Capital letter word exist anot
):label	
({Token.string==”,”})	
({Token.kind==”number”})	Number refer to the age
({Token.string==”;”) +	+ mean that repeat the pattern again
→	Match the LHS rules with RHS
:label.Person={rule=”Person”}	label as Person

Process: Normalization

- Temporal/numerical expr. may require an extra step:

3. **Normalization**; i.e. mapping them to a given format:

   - “seven o'clock in the morning” → 07:00:00
   - “yesterday” (text published on 29th June, 2018) → 28/06/2018
   - “half million dollars” → 500,000 USD

Bibliography

- **[Jurafsky & Martin, 2009]** Jurafsky, D. & Martin, J.H. (2009). Chapter 22: Information Extraction. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.)*. Pearson-Prentice Hall.
- **[Nadenau & Sekine, 2009]** Nadenau, D. & Sekine, S. (2009). A survey of named entity recognition and classification. In Sekine, S. & Ranchhod, E. (Eds.), *Named Entities. Recognition, classification and use,* vol. 19 of Benjamins Current Topics series, pp. 3-27. John Benjamins Publishing Co.
- **[Nugues, 2006]** Nugues, P.M. (2006). Chapter 9: Partial Parsing. *An Introduction to Language Processing with Perl and Prolog*. Springer.