Named Entity Recognition of the XIP Italian Grammar

Author : Giovanni Depau; July 2008

Introduction

The purpose of this document is to list the Named Entities that are extracted by the XIP Italian grammar.
Named Entity Recognition aims to "semantically" detect and categorize proper nouns, which can be single nouns like "*Italia*", or names like "*Confederazione Generale Italiana dei Lavoratori*".

Dependencies defined to identify Nes:

The different types of named entities that can currently be recognized are the following:

- 1) PERSON (Person names)
- 2) ORG (Organizations)
- 3) LOC (Locations)
- 4) DATE (Dates)
- 5) ARTEFACT

List of entities :

1. PERSON:
2. This unary dependency characterizes person names.

- people, first name, last name, kings (queens), princes (princesses), popes.

  NB: the category "saints" is not defined (yet): ambiguity with "LOC".

    - The Features:

- PERSON Il Principe Carlo d'Inghilterra; Il Re Luigi XV; Papa Giovanni Paolo II; Aldo Rossi; Miguel Angel Perez (y) Cruz; Aldo Di Napoli; il nostro inviato Marek Klondinsky.
- PERSON_ firstname: Italian and foreign first names.
- PERSON_ lastname: very spread last names: Rossi, Bianchi, Brambilla, etc.

Example

Re Luigi LII si accorse troppo tardi del tradimento del Principe Giorgio di Sardegna.

PERSON(Re Luigi LII)
PERSON(Giorgio di Sardegna)

3. ORG:

   This unary dependency characterizes organization names.

- companies, business societies, political parties and movements, (customers, workers) associations and organizations, political institutions ("ministero", "regione", "comune di...", etc.).

The Features:

- ORG_inst: Comune di Portoscoglio; Ministero della Salute Privata; Partito Nazionale Tedesco; Confederazione Nazionale dei Camionisti; Associazione Italiana Librai.
- ORG_bank: Banca Nazionale di Elmas.
- ORG_soc: Acqua Buona Spa
- ORG_award: il Music Award 2008
- ORG_sport: La Juventus, il Cagliari Calcio, l'AS Pabillonis
- NB: Societies: when a name is identified the first time as "org_company", then if it is found again in the same text, it is recognized as a "society" > ORG; ex: La Putzu Acque Spa = Soc > ORG (Putzu Acque Spa); bla bla Putzu Acque = Soc > ORG(Putzu Acque)

Example

La Gigi Puddu Spa ha evitato il fallimento grazie alla fusione con Metalmangimi Srl.

ORG_SOC(Gigi Puddu Spa)
ORG_SOC(Metalmangimi Srl)

4. LOC:

   This unary dependency characterizes places.

- streets (avenue, etc.), cities, regions, countries, continents, lakes, seas, rivers, mountains, seas.

The Features:

- LOC_city: Italian cities, World capitals and important cities
- LOC_country: world countries;
- LOC_continent: America, Europa, America Latina, etc.
- LOC_mountain: Monte Bianco, [Monte Bellighé], monte [Bellighé].
- LOC_lake: il Garda, il Lago di Garda.
- LOC_river: Flumendosa, Fiume Giallo, [Fiume/Rio Bellighé], fiume/rio [Bellighé].
- LOC_sea: Mediterraneo, Mar dei Caraibi.
- LOC_province: nella provincia di Xionghiajnang
- LOC_region: nella regione di Xionghiajnang; nell'Africa Meridionale
- LOC_state: U.S. State (Wisconsin, Nuovo Messico, etc.)

Example

Giovanni viene da Cagliari, il capoluogo della Sardegna, che è una regione d'Italia.

LOC_CITY(Cagliari)
LOC_REGION(Sardegna)
LOC_COUNTRY(Italia)

5. ARTEFACT:

   This unary dependency characterizes 'products' names.

- Newspapers and magazines (“information”), awards, laws, political agreements.

The Features:

- ARTEFACT_award: il Music Award 2008
- ARTEFACT_newspaper: Il Corriere di Cagliari
- ARTEFACT_agreement: Il Trattato di Londra

Example

Il Trattato di Cagliari sarà firmato domani da Italia e Francia. La notizia si trova sul Gazzettino di Sardegna.

ARTEFACT_AGREEMENT(Trattato di Cagliari)
ARTEFACT_NEWSPAPER(Gazzettino di Sardegna)

6. DATE:

   This unary dependency characterizes temporal expressions which are dates.

- days, months, seasons, period, years, centuries, dates.

Examples

Roma, giovedi 13 marzo 2008;

Vanessa è nata il 7 novembre 1989;

DATE(7 novembre 2008)

Maria è nata nel 1998;

il 2008;

DATE(2008)

dal XV secolo a.C.;

DATE(XV secolo a.C.)

Sono nato nel [XX secolo], nel [20° secolo];

nel [500 a.C.] ;

il [24 ottobre];

gli [anni 70];

negli [anni Settanta].