General description of XIP Spanish NER system

C. Hagège
December 2008

This document describes the XIP Spanish NER (Named Entities recognition) system in its first version (December 2008) developed at XRCE in collaboration with UCM Madrid.
For a good understanding of this document, a general knowledge of XIP grammars output organization is needed.

Description of dependencies for Spanish Named Entities

Named entities are delimited with the unary dependency designated as NE.
To each NE dependency, features are associated. These features correspond to the type of the NE extracted by the system. In this document we describe and exemplify the Named Entities types that we extract.

Type PERSONA

The dependency NE_PERSONA delimits all NE corresponding to Person names.

For instance in:
`el **rey Bhumibol** hable por primera vez de una crisis.`

"rey Bhumibol" is recognized as a person name expressed by the dependency
`NE_PERSONA(rey Bhumibol)`

Type GEOGRAFICO

The dependency NE_GEOGRAFICO delimits all NE corresponding to location names.
The notion of location in this version of NER system is very general and covers a wide range of names (countries, cities, administrative territories, but also name of streets, mountains, rivers etc.). However, features associated to the linguistic expressions (nodes) considered as location Named Entities can further specify the sub-type of this NE. (For instance, features city:+, country:+, street:+, water:+ are associated to the node

corresponding to the NE for city, country, adresses, seas, lakes and river respectively). They do not appear in the output of the NE dependency.

Examples:
Parecen siempre preparados para escalar al **Everest**.
gives the following dependency:
NE_GEOGRAFICO(Everest)

Una tienda del entorno de la **plaza Hackescher Markt** en **Berlín**
gives the following dependencies
NE_GEOGRAFICO(plaza Hackescher Markt)
NE_GEOGRAFICO(Berlín)

Type FECHA and HORA

These types are used in NE corresponding to dates (FECHA) and to times (HORA).
Note that in this version only absolute dates and hours are considered.
Expressions like “el lunes” (on monday) which correspond to relative dates are not taken into account.

Examples:

lunes 01/12/2008 21:11 (CET)
gives the following dependency
NE_FECHA(lunes 01/12/2008 21:11 (CET))

while an expression like
21:10 CET
gives the dependency
NE_HORA(21:10 CET)

Type ORGANIZACION

This type is used to type organization names. We consider as organization both public and private institution. Political parties, companies, governement organizations are considered as organization in our system.

Examples:
El Gobierno de Tailandia ha caído
gives:
NE(Gobierno de Tailandia)

De ser aprobada, el **Consejo Nacional Electoral** (**CNE**) tiene 30 días para someterla a consulta popular.
gives the two following dependencies:
NE_ORGANIZACION(Consejo Nacional Electoral)
NE_ORGANIZACION(CNE)

Type EVENTO

This type is used for NE corresponding to an event name.
These events can be of different domains like sport, politics, religious etc.

Examples:
Rodeados de jóvenes y en el marco de la **Feria del Libro de Guadalajara** ...
gives: NE_EVENTO(Feria del Libro de Guadalajara)

Condujo a su club a la **Copa de la UEFA** en la temporada 2005-06.
gives: NE_EVENTO(Copa de la UEFA)

Type MONUMENTO

NE of this type corresponds to monuments and human constructions. We only consider as NE the string that are capitalized (expressed as a proper noun).

Example:
Martín Araguz, especialista del **Hospital General de la Defensa**
NE_MONUMENTO(Hospital General de la Defensa)

Type DOCUMENTO

This type is used to delimit documents. We do not consider here title of books but only documents that have an official role.

Examples:
Cualquier tipo de propaganda nazi está terminantemente prohibida en la **Carta Magna** del país.
NE_DOCUMENTO(Carta Magna)

No figure en ningún lugar de la **Constitución americana**
NE_DOCUMENTO(Constitución americana)

General remarks on NE delimitation

Embedment of NE

`Feria del Libro de Guadalajara`, which is considered as an event (NE_EVENT) we do not annotate the embedded NE Guadalajara as a location (NE_GEOGRAFICO).

Inclusion of person name introducers

Titles as “Ministro, Rey, etc. are included within the Named Entity NE_PERSONA (even if these titles do not appear in upper case).