Adaptation of the General Spanish Grammar for the Pocket Engineer Application

Introduction

The aim of this document is to show all the adaptations made to the general Spanish Grammar for the online information retrieval application called Pocket Engineer.

1. Used files

Grammar files:

```
spanish_PE.grm (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/GRMFILES)
spanish_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/GRMFILES)
```

Rules' files:

```
disam_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/DISAMB)
localgrams_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/LOCALGRAMS)
normalize_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/DEPENDENCIES)
dependency_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/DEPENDENCIES)
```

Lexicons:

```
lexicon_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/LEXICONS)
```

Declaration:

```
features_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/FEATURES)
categories_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/FEATURES)
functions_PE.xip (~elopez/parSem/xip/dev/GRAMMARS/SPANISH/DEPENDENCIES)
```

2. PEU creation

The PEU are the basic units of the Pocket engineer application. We can find simple or complex PEU. In order to extract them, we have created the file **dependency_PE.xip**.

The simple PEU are:

- An NP
- A verbal chunk
- A PP
- An AP

The complex PEU are:

- Following NPs (even when the first one is included in a PP)
  PEU[de papel formato 216mm * 279mm]

- One NP and the following PPs if the preposition is 'de' (even when the first NP is included in a PP)
  PEU[el alimentador de documentos de la impresora]

- One NP and the following APs
  PEU[una intervención posterior]

- Coordinated items
  PEU[el alimentador y la bandeja de entrada]

- One verb and its direct object (even if there is an adverb in between)
  PEU[sitúe el original]
  PEU[abra cuidadosamente la puerta]

- A passive verb with its subject:
  PEU[la luz se enciende]

- The acronyms in brackets and the previous long form:
  PEU[Limpieza del Cristal de transporte de velocidad constante (CVT)]

- Multiple nouns separated by "/" or "|"
  PEU[el Fiery X12|XP12|EX12]

- Some PPs with a noun or a verb (if they are arguments of the verb or the noun)
  PEU[acceso al Auditrón]

- Negation
  PEU[no encienda la impresora]

In order to build complex PEU, we follow different steps. First of all, we create the COMBINE dependency. This is a binary dependency between nodes that might be unified into one. When the COMBINE dependency is created, we can build the PEU nodes.
Once we have built every PEU (simple or complex) we create the PEUNIT dependency. The PEUNIT dependency is a unary dependency that contains the PE unit.

3. Normalization

For the normalization, there have been deleted the determiners, the conjunctions, the pronouns, the punctuation, the prepositions, the auxiliary and modal verbs.

The words will be replaced by their lemmas

[desactivación de la portada] => (desactivación portada)

4. Features

For the construction of the PEU :

pp_de : the node is a PP where the preposition is 'de'
non_pp_de : verbs that apply for PPs starting by 'de' (e.g. ‘retirar’). When a verb with this feature is present, PPs starting by 'de' will not be attached to the previous NPs.

Added dependencies:
PEUNIT

5. Corpus

The corpus for the Pocket engineer Application is located in ~elopez/PE/ed_x_main_es.txt. To adapt the Spanish grammar to this application, we have used 100 sentences extracted from this corpus which are aligned with the English, German, French and Italian grammar. This new and training corpus is located in ~elopez/PE/Corpus_100_esp.
The original corpus has HTML tags that have not been deleted. For the training corpus, only the italic tag appears and it has no incidence on the analysis of the sentence.
We have also created from the original corpus another one without the HTML tags and it is located in ~elopez/PE/corpus_sans balises_2 parties.txt. This other corpus was meant to contain different kind of sentences (the 100 sentences correspond all of them to short titles) to help the development of the grammar and to test our grammar.

6. Problems

- The choice of deleting the prepositions in the normalization process is provisional (it would require a deeper lexical study).
- All the past participles are lemmatized as infinitives, even when the past participle has been analyzed as an adjective.
- For passive forms, the normalization does not take into account for now the inversion of the past participle and the noun.
- We can consider the possibility of deleting some adverbs from the normalized form (e.g. ‘únicamente’…)