DepPattern: a Multilingual Dependency Parser [*]

Pablo Gamallo Otero[1] and Isaac González[2]

[1] Centro de Investigação em Tecnologias da Informação (CITIUS)
Universidade de Santiago de Compostela, Galiza, Spain
`pablo.gamallo@usc.es`
[2] Cilenis S.L. (Language Technology)
Santiago de Compostela, Galiza, Spain
`isaac.gonzalez@cilenis.com`

1 Introduction

DepPattern is an open-source suite of multilingual syntactic analysis[3]. The suite includes the following main modules:

- Basic grammars for five languages: Portuguese, Spanish, English, French, and Galician
- A grammar compiler to build parsers in Perl from those grammars.
- Five syntactic parsers generated from the grammars.

Parsers takes as input the result of a PoS-tagger, either FreeLing [1] or Tree-Tagger[4]. The whole process is robust and fast. It takes 3000 words per second on a Linux platform with 2.4GHz CPU and 2G memory.

The basic grammars of DepPattern contain rules for many types of linguistic phenomena, from noun modification to more complex structures such as apposition or coordination [4]. No direct avaliation was made, however different DepPattern parsers were used in many applications on Information Extraction, namely, extraction of bilingual terminology from comparable corpora [5, 3], and word similarity acquisition [2].

2 Grammar Description

One of the main advantages of DepPattern is its easy-to-use grammar formalism.

A specific DepPattern grammar is constituted by a set of context dependent rules[5]. Every rule aims at identifying a specific dependent-head relation by means of a pattern of part-of-speech (PoS) tags. A rule is always constituted by two elements:

[3] Freely available at `http://gramatica.usc.es/tools/deppattern.html`
[4] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`
[5] A tutorial to learn the grammar formalism is available at `http://gramatica.usc.es/pln/tools/html_tutorial/tutorialGrammar.html`
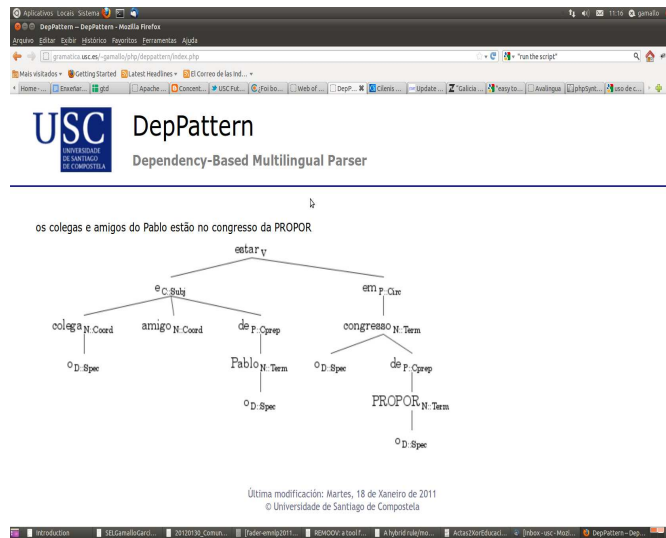
Fig. 1. Analysis of a sentence

- a patttern of PoS tags
- the name of a dependent-head relation found within the pattern

Let's see an example:

```
DobjR: VERB [DT]? [ADJ]* NOUN
%
```

The two elements are separated by a semicolon and ending with the symbol '%' . The first element is 'DobjR', which stands for the name of a specific dependency relation, namely a direct object (Dobj) appearing to the right (R) of the verbal head. The second element is a sequence of PoS tags. Two of them, VERB and NOUN, respectively represent the head and the dependent of the relation, while those between square brackets stand for the context of the relation. Rules take as input any text previously PoS tagged and lemmatized with Tree-Tagger or Freeling. The names of both dependencies and PoS tags are declared in two setting files: *dependencies.conf* and *tagset.conf*, respectively.

Further elements can be used to elaborate different aspects of a rule, namely morphological features and operations such as agreement, recursivity, inheritance, modifying values of features, adding new feature-value pairs, etc. It is also possible to define lexical classes by defining word lists in the source file *lexical_classes.conf*.

3 Demonstration

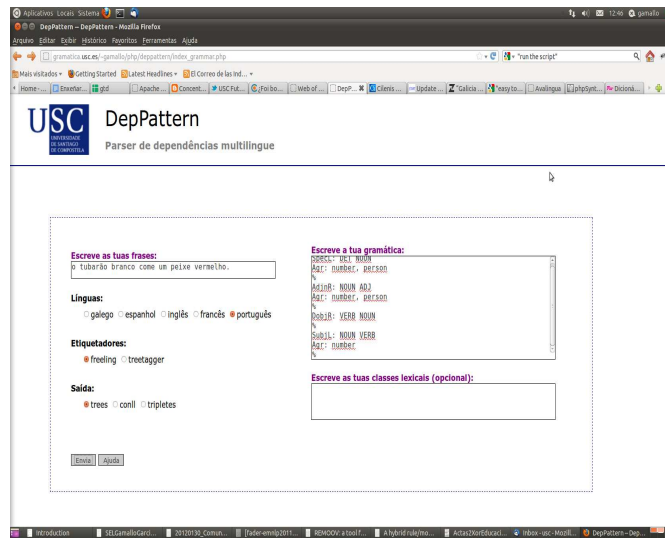The demonstration of DepPattern will consist of three tasks:

Fig. 2. Web form to write and analyse grammars

Task 1: In the command line, we run DepPattern on different large documents. The objective is to show efficiency, robustness and speed of the system.

Task 2: Using a Web interface (see Figure 1), several phrases in different languages are analysed. The objective is to show the quality of the system as well as its multilingual funcionality.

Task 3: In another Web interface (Figure 2), a small grammar is written on the fly and compiled, in order to analyse very simple phrases. The objective is to show the easy-to-write grammar formalism underlying DepPattern.

References

1. X. Carreras, I. Chao, L. Padró, and M. Padró. An Open-Source Suite of Language Analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004.
2. Pablo Gamallo. Comparing different properties involved in word similarity extraction. In *14th Portuguese Conference on Artificial Intelligence (EPIA'09), LNCS, Vol. 5816*, pages 634–645, Aveiro, Portugal, 2009. Springer-Verlag.
3. Pablo Gamallo and Stefan Bordag. Is singular value decomposition useful for word simalirity extraction. *Language Resources and Evaluation*, 45(2):95–119, 2011.
4. Pablo Gamallo and Isaac González. A grammatical formalism based on patterns of part-of-speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71, 2011.
5. Pablo Gamallo and Jos Ramom Pichel. Automatic generation of bilingual dictionaries using intermediary languages and comparable corpora. In *CICLING, LNCS, Vol. 6008*, pages 473–483, Iasi, Romania, 2010. Springer-Verlag.