PANACEA Project
Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
Grant Agreement no. 248064

Documentation of MaltParser Web Service[1]

Authors: Idan Sadan, Muntsa Padró. Barcelona, 2012
Contact: nuria.bel@upf.edu, muntsa.padro@upf.edu

1 Overview

This web service performs dependency parsing in Spanish using a Malt Parser instance. It parses plain texts introduced by the user and generates linguistically annotated Treebank instances based on a data-driven parsing model. The parsing model is induced from de dependency-annotated IULA Treebank (Marimon et al, 2012) using the language-independent MaltParser[2] system as a dependency model trainer (Nivre et al, 2007). This Treebank contains 589,542 tokens in 42,099. In order to achieve optimal performance, the training corpus was previously analyzed with MaltOptimizer[3] (Ballesteros and Nivre, 2012), a tool developed to set the best parameters for MaltParser.

2 Inputs, outputs and formats

2.1 Inputs

The input to be parsed is a plain text encoded in UTF-8. It can be introduced directly as a text instance in the dialogue box, as a text file or as a URL. The input language available at this moment in the web service is Spanish (es).

2.2 Outputs

The output generated by the webservice is a parsed tree in the CoNLL-X format, a ten-column format that holds linguistic and structural information (http://ilk.uvt.nl/conll/#dataformat). See example below:

Webservice output for the sentence:
En el tramo de Telefónica un toro descolgado ha creado peligro tras embestir contra un grupo de mozos.

```
1       En      en      s       SPS00   _       10      MOD     _       _
2       el      el      d       DA0MS0  _       3       SPEC    _       _
3       tramo   tramo   n       NCMS000 _       1       COMP    _       _
4       de      de      s       SPS00   _       3       MOD     _       _
```


[2] http://www.maltparser.org/
[3] http://nil.fdi.ucm.es/maltoptimizer/index.html

```
5       Telefónica          NP00000  n       NP00000  _       4       COMP    _       _       _
6       un          un      z        Z       _        7       SPEC    _       _
7       toro        toro    n        NCMS000 _        10      SUBJ    _       _
8       descolgado          descolgar v      VMP00SM  _       7       MOD     _       _       _
9       ha          haber   v        VAIP3S0 _        10      AUX     _       _
10      creado      crear   v        VMP00SM _        0       ROOT    _       _
11      peligro     peligro n        NCMS000 _        10      SUBJ    _       _
12      tras        tras    s        SPS00   _        10      MOD     _       _
13      embestir    embestir v       VMN0000 _        12      COMP    _       _
14      contra      contra  s        SPS00   _        13      MOD     _       _
15      un          un      z        Z       _        16      SPEC    _       _
16      grupo       grupo   n        NCMS000 _        14      COMP    _       _
17      de          de      s        SPS00   _        16      COMP    _       _
18      mozos       mozo    n        NCMP000 _        17      COMP    _       _
19      .           .       f        Fp      _        18      punct   _       _
```

2.3 Data format

Each one of the fields or columns of the output format matches a field in the following table taken from the CoNLL-X Shared Task web site. Empty fields are marked with underscore signd (`_`). The Malt Parser ignores columns 9 and 10 by definition and leaves them unmarked.

Field number:	Field name:	Description:
1	ID	Token counter, starting at 1 for each new sentence.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma or stem (depending on particular data set) of word form, or an underscore if not available.
4	CPOSTAG	Coarse-grained part-of-speech tag, where tagset depends on the language.
5	POSTAG	Fine-grained part-of-speech tag, where the tagset depends on the language, or identical to the coarse-grained part-of-speech tag if not available.
6	FEATS	Unordered set of syntactic and/or morphological features (depending on the particular language), separated by a vertical bar (|), or an underscore if not available.
7	HEAD	Head of the current token, which is either a value of ID or zero ('0'). Note that depending on the original treebank annotation, there may be multiple tokens with an ID of zero.
8	DEPREL	Dependency relation to the HEAD. The set of dependency relations depends on the particular language. Note that depending on the original treebank annotation, the dependency relation may be meaningfull or simply 'ROOT'.
9	PHEAD	Projective head of current token, which is either a value of ID or zero ('0'), or an underscore if not available. Note that depending on the original treebank annotation, there may be multiple tokens an with ID of zero. The dependency structure resulting from the PHEAD column is guaranteed to be projective (but is not available for all languages), whereas the

		structures resulting from the HEAD column will be non-projective for some sentences of some languages (but is always available).
10	PDEPREL	Dependency relation to the PHEAD, or an underscore if not available. The set of dependency relations depends on the particular language. Note that depending on the original treebank annotation, the dependency relation may be meaningfull or simply 'ROOT'.

Table 1: Information encoded in each column in a CoNLL-X formatted corpus

The POS tags in column 5 are taken from the Eagles Tagset used by FreeLing for Spanish (http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html).

The possible values for the dependency relation tags in column 8 are the following:

Tag	**Grammatical function**
ROOT	Root
SUBJ	Subject
DO	Direct Object
IO	Indirect Object
OBLC	Oblique Object
ADV	Advervial Object
AUX	Auxiliar verb
BYAG	By agent complement
ATR	Attribute
PRD	Predicative complement
OPRD	Object predicative complement
PP-LOC	Locative prepositional complement
PP-DIR	Directional prepositional complement
SUBJ-GAP	Subject in a gapping construction
COMP-GAP	Complement in a gapping construction
MOD-GAP	Modifier in a gapping construction
VOC	Vocative
MIMPERS	Impersonal marker
MPAS	Passive marker (“se”)
MPRON	Pronominal marker
COMP	Complement (of N, ADJ, ADV, PREP)
MOD	Modifier
NEG	Negation

SPEC	Specifier
COORD	Coordination
CONJ	Conjunction
punct	Punctuation

Table 2: Possible values for the dependency relation label.

3 Evaluation

We performed some evaluation experiments using IULA Treebank corpus and 10-fold cross-validation, obtaining more than 92% of accuracy for labeled attachment score close to 96% for unlabeled attachment score.

References:

Miguel Ballesteros and Joakim Nivre. 2012. MaltOptimizer: An Optimization Tool for MaltParser. In Proceedings of the System Demonstration Session of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics (EACL). Avignon, France. April 2012.

Montserrat Marimon, Beatriz Fisas, Núria Bel, Marta Villegas, Jorge Vivaldi, Sergi Torner and Mercè Lorente. 2012. The IULA Treebank. *In Proceedings of LREC 2012*. Istanbul, Turkey. May 2012.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov and E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2), 95-135. 2007.

Acknowledgements

The creation of the corpus used to train the Malt Parser was Funded by METANET4U project and IULA. METANET4U: Enhancing the European Linguistic Infrastructure, (2011-2013), funded by UNER - Competitiveness and Innovation Framework Program, (CIP-PSP-270893)

Related Web Service:

MaltParser Web Service: http://lod.iula.upf.edu/resources/225