UN PARSER PER IL RICONOSCIMENTO E LA NORMALIZZAZIONE DI ESPRESSIONI TEMPORALI IN ITALIANO

AN ITALIAN PARSER FOR TERN

LORIS FAINA · STEFANIA SPINA

SOMMARIO/*ABSTRACT*

In questo report viene descritto lo svolgimento del *Temporal Expression Recognition and Normalization* task, previsto nell'ambito di Evalita 2007, effettuato con un parser sviluppato dall' Università di Perugia.

In this report, we describe the results obtained in the Temporal Expression Recognition and Normalization task (EVALITA 2007); the task was performed using a parser developed at University of Perugia.

Keywords: temporal expression recognition, italian parser.

1. Introduction

The task was performed using UniPg_Faina_TIME an Italian parser written in Ruby programming language on a GNU-Linux system. The parser is being developed in the last 15 months at the University of Perugia, Department of Mathematics and Computer Science, within a Natural Language Processing and Question answering project, and it is still far to be completed.

2. System description

UniPg_Faina_TIME is an Italian parser that combines two separate levels of parsing:
- a constituent parsing, that entails a category annotation of morphosyntactic constituents [3];
- a dependency parsing, that implies a functional annotation of relations such as subject, complement, etc... [1], [2] .

In the first step, the system tokenizes the input text and POS tags it on the basis of a tagset that includes nine major grammatical categories (noun, verbs, adjectives, articles, adverbs, pronouns, prepositions, interjections and conjunctions). The system also recognizes a number of multiword expressions (mainly prepositional and adverbial).

In the next step, the system splits the input text in sentences, basing on punctuations, and each sentence in clauses, which are described as main, coordinate or subordinate.

Each clause is then divided in seven possible types of phrases (noun, verb, prepositional, adjectival, adverbial, interrogative, relative). In each phrase, head and modifiers are identified.

Finally, each phrase is described in its functional value (subject, direct object or other complements).

3. Temporal expressions

Temporal expressions can be contained in three different elements: noun phrases, prepositional phrases and adverbial phrases. Some grammatical items have been classified in a slightly arbitrary way: temporal adverbs like "oggi" or "domani", for example, have been tagged as nouns (temporal adverbs used as nouns).

The system builds temporal expressions incrementally: first it analyzes the phrases where a temporal expression is found and, in the meantime, the adjacent phrases. For example,

PP [prima di + NP [il5 luglio + PP [di + NP [il2007]] +
PP [a + NP[le 8 + PP [di + NP [sera]]]]]]

gives as a result:

"2007-07-05T20:00:00"

Each noun and prepositional phrase in the example has an individual temporal reference, that is modified by other post-head phrases, so as the top level PP is the result of all the temporal references that it contains.

For this test, we have tried to limit this kind of process, because the TIMEX2 standard seems to contrast with this incremental behavior. The expression "dal 2003 al 2005", for example, would be interpreted in TIMEX2 as composed by two separate expressions, whereas

UniPg_Faina_TIME would describe it as a single temporal expression.

Each temporal expression, in this system, has only three possible descriptions:

1. a date, as in 2004-12-23T20:00:00;
2. a time interval, as in [2004-12-23T20:00:00,2004-12-24T23:00:00];
3. different time intervals, with an and/or interpretation ("ieri e oggi" or "ieri od oggi").

For expressions of type 2 the left bound has been taken as *anchor_val*, AFTER has been taken as *anchor_dir* and the time interval duration or the right bound with *anchor_dir,* BEFORE as *val*.

This is why we didn't expect a good result in the evaluation of temporal expression recognition. Temporal expressions like "ex", that are evaluated in TIMEX2, are ignored in our test because they cannot be contextualized.

Once a temporal expression has been located, the system prints it as an output, without any post-head element that is not relevant to the temporal expression, and also without any initial preposition, as requested by TIMEX2. With this string, its position in the input text can be traced, so as to insert the *charseq* tag.

The time needed to perform the test is fairly high (three hours), because the system cannot separately analyze the temporal expressions, but has to perform a complete syntactical analysis of the input texts.

REFERENCES

[1] A. Bohmova, J. Hajic, E. Hajicova, B. Hladka. The Prague Dependency Treebank: Three-Level Annotation Scenario In *Treebanks: Building and Using Syntactically Annotated Corpora*, ed. Anne Abeille. Kluwer Academic Publishers, in press (available at: http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/Czech_PDT.pdf).

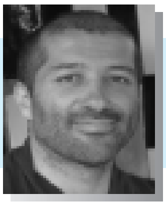
[2] T. Brants, W. Skut, and H. Uszkoreit, 1999. Syntactic Annotation of a German Newspaper Corpus. In Proceedings of the *ATALA Treebank Workshop*. Paris, France.

[3] M. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, Vol.19, 1993. Reprinted in *Using Large Corpora*, S. Armstrong (ed.), MIT Press, 1994.

CONTACTS

LORIS FAINA
Dipartimento di Matematica ed Informatica, Università degli Studi di Perugia, via Vanvitelli, 1, 06123 Perugia
E-mail: faina@unipg.it

STEFANIA SPINA
Dipartimento di Scienze del Linguaggio, Università per Stranieri di Perugia, Piazza Fortebraccio 4, 06122 Perugia
E-mail: sspina@unistrapg.it



LORIS FAINA is Researcher in Mathematics at the Faculty of Engineering of the University of Perugia. His main areas of interest are Optimization Problems, Computational Linguistics, and Artificial Intelligence.

STEFANIA SPINA is Researcher in Linguistics at the Department of Language Sciences of the University for Foreigners Perugia. Her main areas of interest are Corpus Linguistics and Computational Linguistics.