Procesamiento del lenguaje Natural, nº 41 (2008), pp. 291-292 recibido 7-05-2008; aceptado 16-06-2008

AnCoraPipe: A tool for multilevel annotation

Manuel Bertran, Oriol Borrega, Marta Recasens, Bàrbara Soriano
CLiC – Centre de Llenguatge i Computació
Universitat de Barcelona
Gran Via Corts Catalanes,585
08007 Barcelona
mbertran@lsi.upc.edu, oborrece7@fhis.ub.edu, {mrecasens,bsoriano}@ub.edu

Resumen: AnCoraPipe es una herramienta de anotación de corpus que permite etiquetar diferentes niveles lingüísticos de manera simultánea y eficiente, ya que utiliza un formato único para todas las etapas. De esta forma, se reduce el tiempo de anotación y se facilita la integración del trabajo de todos los anotadores en el proceso.
Palabras clave: Lingüística de corpus, herramienta de anotación, niveles de anotación.

Abstract: AnCoraPipe is a corpus annotation tool which allows different linguistic levels to be annotated simultaneously and efficiently, since it uses a single format for all stages. In this way, the required annotation time is reduced and the integration of the work of all annotators is made easier.
Keywords: Corpus linguistics, annotation tool, annotation levels.

1 Introduction

Corpora annotation is a very time-consuming task, and developing AnCora to its current state has meant a lot of effort by our research group. Throughout this process, different tools and formats have been used, yet always running the risk of losing data when translating from one format to another or when merging data that had been labeled with different tools. With a view to solving these problems, we present AnCoraPipe, which is based on a single XML data format. This data format allows annotation for different levels and languages. An effort was made to make the tool scalable and extensible.

Several linguists experienced in corpus annotation have participated in the process of building and testing AnCoraPipe. The interaction has made it possible to build a friendlier interface easy to use for the most usual operations. The new tool decreases the annotation time by 40% in semantic role labeling, by 60% in named entities, and by 25% in coreference.

2 Data format

In order to help concurrent annotation of different levels, the interface can associate the corpora in the local machine with a server, so users can be aware of changes made in the server and synchronize them before making their own changes in local files. Changes made in local files can then be uploaded to the server for other users to add further annotations.

Items are stored in UTF-8 encoded XML format. XML allows portability and takes advantage of the several tools and libraries available in a variety of platforms and programming languages. Besides, UTF-8 allows the format to be cross-lingual. XML has a tree structure itself, so it maps easily to the syntactic constituent structure.

Our XML is based on the following principles:

- Easy to read: the structure is intuitive.
- Easy to maintain: internal coherence can be maintained with little effort.
- Robustness: little changes do not affect overall coherence. The whole structure is maintained even when an error occurs.

These objectives are reflected in a series of design principles:

- Small set of node names: only 15 node names are possible. Thus, nodes are only generic and specificity is reached through attributes.
- Attributes are atomic: each attribute labels one and only one feature of the node. This reduces the number of possible values and makes the annotation levels independent.

ISSN: 1135-5948

- Attributes describe only their node. This makes moving, deleting and creating nodes very simple tasks, and so coherence is guaranteed.
- No redundant data.
- Easy to add new annotation levels: only the design of a new attribute and its possible values is needed.

3 Interface

This section briefly describes the **AnCoraPipe** editor.[1]

3.1 Description

The interface is organized in different panels where data are shown. Buttons and menus are available to perform operations on the corpora.

The GUI (Graphical User Interface) highlights in yellow the items to which the coder must pay attention, thus suggesting the tree nodes that should be annotated or the sentences containing such nodes depending on the annotation level.

The available panels are:

- Corpora directory tree: it shows the directory structure and allows the user to select a file.
- Sentence list: it shows the sentences within each file.
- Sentence tree: it contains the selected sentence structure. The user can also see the words and lemmas together with the data of the corresponding annotation level.
- Annotation panel: This panel is used to perform operations on the tree and annotate its nodes. The display of the tree changes according to the annotation level, which eases annotation.

Current annotation levels include morphology, syntax (changes in the tree structure, nodes grouping and splitting, etc), functions, arguments and thematic roles, named entities, WordNet synsets, and coreference.

The interface provides as well some external tools for specific levels, such as:

- Coreference annotator: coreference can be annotated in a user-friendly way, seeing the files as plain text.
- WordNet synsets: on a lemma-by-lemma basis, the external tool looks for all occurrences of the same lemma in the corpus, so they are all annotated in a row. This favors the consistency of the annotation.

The interface is extensible by creating additional tools for further annotation levels. This can be done by writing two new Java classes after having specified the new attribute and possible values.

3.2 Functionality

Many linguists have participated in the development of AnCoraPipe. This has led to a tool that is very user-oriented, focusing on usability and operational simplicity. To this end, the required mouse clicks to perform operations have been minimized, and only the relevant nodes of each annotation level are highlighted so that the risk of oversight is avoided. In this way, we have reduced annotation time up to 60%.

3.3 Installation

The requirements for AnCoraPipe are: Java 1.5 and the Java graphical library SWT. Our package includes SWT library for WindowsXP. In other platforms, this library comes with the Eclipse package or it can be obtained from http://www.eclipse.org/swt/.

4 Future work

Plans to extend the current application include: making the application available via the Web, providing methods for querying the corpus from the interface, providing methods for making statistical descriptions of the corpus, providing tools for dealing with nominal and verbal lexicons, and adding semi-automatic methods and machine learning functionalities for semi-automatic labeling.

Acknowledgments

This paper has been supported by Lang2World (TIN2006-15265-C06-06) – sub-project of TEXTMESS – and FPU-2006-08 grant from the Spanish Ministry of Education and Science.

[1] For a more detailed explanation visit http://clic.ub.edu/mbertran/tbfeditor/help.