SACR: A Drag-and-Drop Based Tool for Coreference Annotation

Bruno Oberle

LiLPa, University of Strasbourg

oberleb@unistra.fr

Context

- Statistical methods in NLP lead to an increased need for annotated corpora, but manual annotation is costly, both in terms of time and human resources. Annotation is often done by students, interns or other non-technical users.
- To speed up the annotation process, tools should be designed and optimized for a specific task. They should also be ready-to-use and provide an intuitive interface.

Presentation of SACR

- SACR ("Script d'Annotation des Chaînes de Référence") is a cross-platform tool, implemented as a single web page, running in Firefox and Chrome.
- Easy to use, with no learning time and no installation, it is well-suited for users with no specific computing skills.
- The format is text-based, with inline annotations that can be converted from or to stand-off annotation formats like the Glozz format (Widlöcher & Mathet 2012), readable by several other tools like TXM (Heiden 2010), Analec (Landragin, Poibeau & Victorri 2012).
- It is open source (MPL-2.0) and available at boberle.com/projects/sacr.

Annotating with SACR

1. Referring Expressions

- A click on the first and last words of a referring expression (*i.e.* a linguistic expression that refers to some extra-linguistic entity) is enough to mark it.
- Marked expressions are surrounded by a color frame.

2. Coreference

- Drag-and-dropping an expression over another marks a coreference link (two expressions are coreferential when they both refer to the same entity in the extra-linguistic world).
- Two coreferential expressions are assigned the same color and name.
- A pop-up window lists all the referents and their corresponding expressions. It is possible to drag and drop expressions from the main window to the pop-up and vice-versa.

3. Features

- Drop-down menus are used to annotate features, like part of speech or grammatical function.
- It is also possible to use shortcut keys (*e.g.* "d" for a definite noun phrase or "r" for a relative pronoun). The program then goes automatically to the next expression to be annotated.

4. Search

- SACR also offers to search any features and to show the corresponding expressions in the main window.

References

- Heiden S. (2010). The TXM Platform: Building Open Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, *et al.* (eds.), *24th Pacific Asia Conference on Language, Information and Computation*. Sendai, Japan. Wasea University.
- Landragin F., Poibeau T. and Victorri B. (2012). Analec: a new tool for the dynamic annotation of textual data. In Nicoletta Calzolari *et al.* (eds) *Proceedings of LREC'12*. Istanbul, Turkey. ELRA.
- Widlöcher A. and Mathet Y. (2012). The Glozz Platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, Paris, France.

Acknowledgement