Ace2txt: from Multilevel Annotation to raw Text.

Fabio Celli
CLIC, university of Trento

June 8, 2009

Abstract

This paper presents a software, called ace2txt2004, which converts the annotation format used in the ACE program into text, and places it in context. The software, written in python and released as an open source program, was tested on ACE 2004. This program is very useful for the extraction of the annotation from ACE corpora and the conversion in the .arff format used for running machine learning experiments under several environments.

1 Introduction and Related work

The ACE (Automated Content Extraction) program, promoted by the LDC (Linguistic Data Consortium), aims at developing tools to support NLP in various tasks, such as classification for machine transation (for example in [2]), and providing annotated corpora and scorers for training and evalution. The corpus I referred to is ACE 2004 [1], which is available in English, Arabic and Chinese, and includes annotation named entities, semantic relations and coreference.

2 Technical details

ACE 2004 annotation identifies three tasks: Entity Detection and Tracking (henceforth EDT), Relation Detection and Characterization (RDC), Entity Linking and Tracking (ELT).
EDT labels the entities using the following annotation schema: 1) Entity types (Person, Organization, Location, Facility, vehicle and Geo-Political Entity). 2) Several entity subtypes, 3) Entity Lexical types (named, nominal or pronominal). 4) Entity role (for example "French loves eating" is a GPE with the role of a Person). 5) Entity Semantic Class (specific, generic, attributive, negatively quantified or underspecified). The annotation is multi-leveled in the sense that both complex and nested entities are captured. Annotators also label cases of metonymy, where the name of one entity is used to refer to another entity related to it. RDC specifies the relations of the targeted types between entities. Relations are semantic links tha hold between pairs of entities. Annotators label the following: 6) Relation Type (Agent-Artifact, Person-Social, Physical, Employment, Person-Affiliation, GPE-affiliation, Discourse). 7) Relation subtype, 8) Lexical condition under which the semantic relation holds (Possessive, Premodifier, Formulaic, Preposition, Verbal). ELT keeps track of the coreference

chains of entities and pronouns defining 9) Coeferences: which entity a pronoun or a definite description is referred to.
Ace2txt2004 is an opensource python program that takes as input the .apf, .ag and .sgm files and outputs a .txt file. Basically it processes the XML tags from .apf and .ag files, converting them into python code, and then it maps them in the text (provided in the .sgm file) placing each tag beside the simple or complex entity annotated by using offsets and ids from the XML tags. In the output annotation are included the following features: "ent" (entity types), "est" (entity subtypes), "elt" (entity lexical class), "esc" (entity semantic class), "enr" (entity role), "txt" (text annotated), "rel" (relation), "rst" (relation subtype), "lxc" (relation lexical condition), "tid" (tag id), "rid" (relation id), "ref" (single antecedent in the coreference chain). The feature are combined in different ways in the output tags, table 1 reports all the feature combinations, manually classified from a sample of 163 tags.

Table 1: tag types produced by ace2txt2004.

tag type	**features included**
relation role	tid+rel+rst+enr+txt
entity role	tid+enr+txt
relation	tid+rel+rst+lxc+rid+txt
entity	tid+ent+est+elt+esc+enr+txt
empty embedded entity	tid+txt
coreference	tid+ref+enr+txt
relation mention	tid+lxc+rid+txt

3 Conclusions

The software presented here is a tool for the conversion of the rich ACE annotation from XML to raw text, and it is very useful for producing .arff files, used in machine learning under different environments such as Weka [3]. A possible expansion of this work is the adaptation of ace2txt2004 to ACE 2005, which includes events in addition to relations, entities and coreference chains.

References

[1] Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *Proceedings of LREC 2004* , pp.837-840.

[2] Ma, X., Cieri, C., 2006. Corpus Support for Machine Translation at LDC. In *proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation.*

[3] Witten, I, H., Frank, E. 2000 Data *Mining. Practical Machine Learning Tools and Techniques with Java implementations.* Morgan and Kaufman, San Francisco, CA.