CINTIL-NamedEntities - A Portuguese Named Entity Annotated Corpus

1. BASIC INFORMATION

Corpus composition:

The CINTIL-NamedEntities corpus, built upon the CINTIL International Corpus of Portuguese (Barreto et al., 2006), is composed of 30,493 sentences of written Portuguese with named entities manually disambiguated and annotated with links to appropriate pages in the Portuguese Dbpedia (Lehmann et al., 2012).

The development of the CINTIL-NamedEntities corpus has been funded by the EU project QTLeap (EC/FP7/610516) and the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012).

The corpus contains:

Corpus	CINTIL-NamedEntities
tokens	684,467
entities	26,371
linked to DBpedia	16,120 (61.13%)

Representation of the corpora:

The corpus is divided into 1,704 text files – one raw text file and one standoff annotation file for each of the 852 parts (from 420 excerpts) that the CINTIL corpus was divided into for annotation.

Character encoding:

All files in the corpora contain UTF-8 character encoding.

2. ADMINISTRATIVE INFORMATION

Contact persons:

In relation to the CINTIL-NamedEntities corpus, please contact:
--- António Branco (antonio.branco@di.fc.ul.pt)
*** NLX (Natural Language and Speech) Group,
Department of Informatics, Faculty of Sciences,
University of Lisbon, Portugal

Delivery medium:

The corpus is available from the Meta-Share repository, as well as being available for external use on demand.

Copyright statement and information on IPR:

The CINTIL-NamedEntities courpus is available for both research and commercial purposes, with attribution, and no redistribution nor derivatives allowed

3. TECHNICAL INFORMATION

Directories and files:

The corpus is divided into 420 excerpts, each of which can have numerous parts – there are 852 parts in total. In the main directory (CINTIL-NamedEntities) there exist two files for each part, totalling 1,704 files. The two files for each part are:
--- A file containing the raw text for that excerpt/part (excerpt000-part000-doc-1.txt)
--- A seperate file containing the standoff annotation (excerpt000-part000-doc-1.ann)

Data structure of an entry:

Each standoff annotation file (e.g. excerpt000-part000-doc-1.ann) contains:
- --- A line for each named entity present in the format:
  - --- <TID> <type> <start> <end> <entity>
  - --- e.g. T1 LOC 21 26 Ourém)
- --- A line for the annotation of each entity:
  - --- If the entity was annotated, in the format:
    - --- <NID> Reference <TID> <Dbpedia ID> <Dbpedia link>
    - --- e.g. N1 Reference T1 Dbpedia-PT:20076 Ourém_(Portugal)
  - --- If the entity was not annotated, in the format:
    - --- <#ID> AnnotatorNotes <TID> Not found
    - --- e.g. #2 AnnotatorNotes T2 Not found

Corpora size:

The 30,493 sentences comprising CINTIL-NamedEntities contain 684,467 tokens and require 5.8MB of disk storage.

4. CONTENT INFORMATION

Type of corpus:

The corpus is monolingual and with standoff annotation.

The natural language of the corpus:

The language of the corpus is Portuguese.

Domain / register of the corpus:

The CINTIL-NamedEntities corpus is built upon the the CINTIL International Corpus of Portuguese (Barreto et al., 2006), a linguistic resource of both written sources and transcriptions of spoken Portuguese. The written part is sourced mainly from newspaper articles and short novels.

Annotations in the corpus:

The CINTIL-NamedEntities corpus contains:

--- Manually annotated Dbpedia links for pre-recognized named entities.

Intended Application of the corpus:

The CINTIL-NamedEntities corpus is intended for use in the evaluation of and as training data for named entity recognition and disambiguation tools and applications.

5. RELEVANT REFERENCES AND OTHER INFORMATION

--- Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M. F. B., Nunes, F., and Silva, J. (2006). Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project. In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC ’06, pages 1438–1443.
--- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2012). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2):167–195.