EVALITA 2009

Local Entity Detection and Recognition (LEDR)

Guidelines for Participants[1]

Valentina Bartalesi Lenzi, Rachele Sprugnoli

CELCT, Trento 38100 Italy

{*bartalesi | sprugnoli*}*@celct.it*

Task Definition

The Local Entity Detection and Recognition (LEDR) task requires that entities (i.e. persons, organizations, geo-political entities and geographical locations) mentioned in source texts be detected, and that selected information about these entities be recognized, following the ACE-LDC standards with all the modifications needed to adapt them to the specific morphosyntactic features of Italian.

The information comprises three attributes, i.e. *type*, *subtype*, and *class* (note that each entity may have only one *class*, one *type*, and one *subtype*).

Participants should also recognise all the mentions referring to each entity. An entity, in fact, provides a representation of an object in the world, while an entity mention provides information about any textual references to that object. For instance, if "Elvis Presley" is mentioned in two different sentences of a text as "il cantante/*the singer*" and as "egli/*he*", these two expressions are considered as two mentions referring to the same entity (i.e. co-referring mentions).

Mentions are to be detected and output along with their attributes. More precisely, the output for each entity mention includes the mention type, its extent, the location of its syntactical head within the extent, and optionally the mention role (only for geo-political entities) and style (literal or metonymic).

In the LEDR task, each document is processed separately and entities that are mentioned in different documents are treated as different entities.

For a detailed description of entities, mentions and their attributes, and for the modifications needed to adapt the English guidelines to the specific morphosyntactic features of Italian, refer to the annotation report downloadable at: http://evalita.fbk.eu/doc/Annotation_Report_LEDR.pdf.

The Corpus (I-CAB)

Both training data and test data are part of the Italian Content Annotation Bank (I-CAB),

---

[1] Please note updates in the Data Format section (differences from the previous version of the guidelines are highlighted in bold) and new examples in Appendix A and B.

developed by FBK-irst and CELCT.

I-CAB consists of 525 news documents taken from the local newspaper "L'Adige". The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004) and are grouped into five categories: News Stories, Cultural News, Economic News, Sports News and Local News. I-CAB is divided into a development part (335 news stories, for a total of around 113,000 words) and a test part (190 news stories, for a total of around 69,000 words).

Data Format

Training data are distributed in the following formats:

- **TXT files** contain the source text. All text files are in UTF-8 (see **Appendix A**).
- APF (ACE2 Program Format) files contain the annotation in the form of XML stand-off annotation, which means that the file as a whole conforms to XML encoding standards, and the raw data being annotated resides in a separate file. The annotations "point" to portions of the raw text via indices (see **Appendix B**).

In order to have a clear visualization of the annotated APF files, it is possible to import them with **Callisto 1.5.2** (http://callisto.mitre.org), a freely distributed annotation tool developed by the MITRE Corporation.

To import the APF files follow these indications:

1. Mouse on: "File" -> "Import". A pop-up window will appear;
2. Browse to get to the file to import;
3. Select "ACE Event Task" for "Available Importers" and "**ACE2004 APF v.5.1.5**";
4. Leave the file encoding at UTF-8;
5. Press "Import".

Test data will be distributed in the **TXT format**.

Requirements:

- The data format required for system output is the APF. For each source document in the evaluation data set, the system will have to produce a single APF file as output. The XML DTD for APF can be downloaded from the EVALITA web site: http://evalita.fbk.eu/ doc/apf.v5.2.0.zip.
- You can use an XML validator to verify that a system output file conforms to the ACE DTD for APF. This java implementation is downloadable from the NIST ACE web site: ftp://jaguar.ncsl.nist.gov/ace/resources/XMLValidator.tar.gz
- Output folders are required to be named as follows:
  evalita09_ledr_participantname_run1 (or *run2*)
- Files have to be organized maintaining the same sub-folders structure (e.g. 20040907/Cultura)
- Output files have to maintain the same names of the input files
- Output folders have to be sent by e-mail both to bartalesi@celct.it and to sprugnoli@celct.it

---

[2] Automatic Content Extraction

Evaluation Metrics

The final ranking will be based on the Value score as defined in the ACE 2008 evaluation campaign.

The Value score for a system is defined to be the sum of the values of all of the system's output tokens, normalized by the sum of the values of the reference data. The possible Value of a system output token depends on how closely it matches that of the reference token to which it is mapped. The maximum possible Value score is 100% but negative scores are possible for systems that make costly errors.

For more information, see "*Automatic Content Extraction 2008 Evaluation Plan*", (http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf).

$$LEDR\_Value_{sys} = \sum_i value\_of\_sys\_token_i \left/ \sum_j value\_of\_ref\_token_j \right.$$

Parameters for scoring LEDR performance

Parameters are the same used in the ACE evaluation:

- 0.300, minimum acceptable overlap
- 4 chars, max acceptable extent difference for names and mentions to match
- 0.750, cost for spurious entity mentions
- 1.000, cost for incorrect coreference

As far as attributes are concerned, attribute weights are defined in the "*Automatic Content Extraction 2008 Evaluation Plan*" (page 10).

Please note that, even if mentions of type NAM might be extracted from the training corpus of the Named Entities Recognition task, we strong invite participants from doing it to guarantee the correctness of the evaluation.

Having confidence in the honesty of participants, we keep the same weights of ACE scorer also for mention types, as reported in the Table 1.

AttributeValue	Attribute Weight
NAM	1.00
NOM	0.50
PRO	0.10

Table 1 Default parameters for scoring LEDR attributes

Figure 1 shows an example of the evaluation output. For a brief description of the fields in each record, see "*Automatic Content Extraction 2008 Evaluation Plan*".

```
entity _EntCount_      ___DocCount___  ____DocCount(%)_____                    _____Cost(%)_____
TYPE   Ref  Det     Ref  Det  Rec    Det     Rec   B3Unweighted     Det    Attr   _Mentions__   Val    B3Valuebased
       Tot FA Mis Tot FA Mis Err  FA  Mis   Err  Pre---Rec---F   FA  Mis Err   FA  Mis  Err  (%)   Pre---Rec---F
GPE     1  1   0   1  1   0   0 100.0  0.0   0.0 100.0 100.0 100.0 7.5   0.0 0.0  0.0  0.0 0.0  92.5  98.3 100.0  99.1
LOC     2  0   0   2  0   0   0   0.0  0.0   0.0 100.0 100.0 100.0 0.0   0.0 0.0  0.0  0.0 0.0 100.0 100.0 100.0 100.0
ORG     1  0   0   1  0   0   0   0.0  0.0   0.0 100.0  66.9  80.2 0.0   0.0 0.0  0.0  9.0 0.0  91.0 100.0  82.9  90.6
PER     4  1   1   4  1   1   1  25.0 25.0  25.0  93.8  67.7  78.6 15.0 20.0 0.0  0.0 22.3 0.0  42.7  94.0  68.5  79.2
tot     9  2   2   9  2   2   1  22.2 22.2  11.1  97.1  73.1  83.4 8.2   9.1 0.0  0.0 11.8 0.0  71.0  97.1  76.8  85.7
```

Figure 1 The evaluation output

Scorer

For the official evaluation we are going to use the ACE08 scoring script ace08-eval-v17.pl, whose formulas are described in "*Automatic Content Extraction 2008 Evaluation Plan*".

You can download the scorer from the EVALITA web site: http://evalita.fbk.eu/doc/ace08-eval-v17.zip.

Contacts

Valentina Bartalesi Lenzi	Rachele Sprugnoli
Tel. 0039 0461 314885	Tel. 0039 0461 314879
bartalesi@celct.it	sprugnoli@celct.it

Appendix A

Input in the SGML format (example available at: http://evalita.fbk.eu/doc/ledr-sample.zip).

Le donne al buonconsiglio

Penelope e le altre

TRENTO - Secondo incontro, oggi alle ore 20, per approfondire i temi della mostra « Guerrieri Principi ed Eroi » al castello del Buonconsiglio di Trento. E' il ruolo della donna l'argomento centrale. Patrizia Frontini, archeologa che opera presso il Castello Sforzesco di Milano, condurrà gli intervenuti in un percorso guidato che avrà come protagonista la figura femminile dalle più antiche comunità di agricoltori e allevatori fino all'epoca dei Celti. La posizione sociale della donna sarà illustrata attraverso il commento degli oggetti esposti in mostra. L'ingresso è libero. Il prossimo incontro è in programma il 28 ottobre.

Appendix B

Output in the APF format (example available at: http://evalita.fbk.eu/doc/ledr-sample.zip).

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE source_file PUBLIC "SYSTEM" "apf.v5.2.0.dtd">

<source_file URI="adige20041007_id413720.txt" SOURCE="unknown"
TYPE="text" VERSION="5.0" AUTHOR="Evalita" ENCODING="UTF-8">
  <document DOCID="adige20041007_id413720">
    <entity ID="adige20041007_id413720-E1" TYPE="PER"
SUBTYPE="Group" CLASS="GEN">
      <entity_mention ID="adige20041007_id413720-E1-2" TYPE="NOM"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="217" END="227">della donna</charseq>
        </extent>
        <head>
          <charseq START="223" END="227">donna</charseq>
        </head>
      </entity_mention>
      <entity_mention ID="adige20041007_id413720-E1-4" TYPE="NOM"
PRIMARY="false" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="406" END="504">la figura femminile
dalle più antiche comunità di agricoltori e allevatori fino
all'epoca dei Celti</charseq>
        </extent>
        <head>
          <charseq START="409" END="414">figura</charseq>
        </head>
      </entity_mention>
      <entity_mention ID="adige20041007_id413720-E1-6" TYPE="NOM"
PRIMARY="false" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="528" END="538">della donna</charseq>
        </extent>
        <head>
          <charseq START="534" END="538">donna</charseq>
        </head>
      </entity_mention>
      <entity_mention ID="adige20041007_id413720-E1-8" TYPE="NOM"
PRIMARY="false" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="0" END="24">Le donne al
buonconsiglio</charseq>
        </extent>
        <head>
          <charseq START="3" END="7">donne</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E2" TYPE="PER"
SUBTYPE="Individual" CLASS="SPC">
      <entity_mention ID="adige20041007_id413720-E2-2" TYPE="NAM"
PRIMARY="true" METONYMY_MENTION="FALSE">
```

```
          <extent>
            <charseq START="251" END="267">Patrizia
Frontini</charseq>
          </extent>
          <head>
            <charseq START="251" END="267">Patrizia
Frontini</charseq>
          </head>
        </entity_mention>
        <entity_mention ID="adige20041007_id413720-E2-4" TYPE="NOM"
PRIMARY="false" METONYMY_MENTION="FALSE">
          <extent>
            <charseq START="270" END="279">archeologa</charseq>
          </extent>
          <head>
            <charseq START="270" END="279">archeologa</charseq>
          </head>
        </entity_mention>
        <entity_mention ID="adige20041007_id413720-E2-6" TYPE="PRO"
PRIMARY="false" METONYMY_MENTION="FALSE">
          <extent>
            <charseq START="281" END="283">che</charseq>
          </extent>
          <head>
            <charseq START="281" END="283">che</charseq>
          </head>
        </entity_mention>
        <entity_mention ID="adige20041007_id413720-E2-8" TYPE="NAM"
PRIMARY="false" METONYMY_MENTION="FALSE">
          <extent>
            <charseq START="251" END="328">Patrizia Frontini,
archeologa che opera presso il Castello Sforzesco di
Milano</charseq>
          </extent>
          <head>
            <charseq START="251" END="328">Patrizia Frontini,
archeologa che opera presso il Castello Sforzesco di
Milano</charseq>
          </head>
        </entity_mention>
        <entity_attributes>
          <name NAME="Patrizia Frontini">
            <charseq START="251" END="267">Patrizia
Frontini</charseq>
          </name>
          <name NAME="Patrizia Frontini, archeologa che opera
presso il Castello Sforzesco di Milano">
            <charseq START="251" END="328">Patrizia Frontini,
archeologa che opera presso il Castello Sforzesco di
Milano</charseq>
          </name>
        </entity_attributes>
      </entity>
      <entity ID="adige20041007_id413720-E3" TYPE="PER"
SUBTYPE="Individual" CLASS="SPC">
        <entity_mention ID="adige20041007_id413720-E3-2" TYPE="NAM"
PRIMARY="true" METONYMY_MENTION="FALSE">
          <extent>
```

```
          <charseq START="27" END="34">Penelope</charseq>
        </extent>
        <head>
          <charseq START="27" END="34">Penelope</charseq>
        </head>
      </entity_mention>
      <entity_attributes>
        <name NAME="Penelope">
          <charseq START="27" END="34">Penelope</charseq>
        </name>
      </entity_attributes>
    </entity>
    <entity ID="adige20041007_id413720-E4" TYPE="PER"
SUBTYPE="Group" CLASS="GEN">
      <entity_mention ID="adige20041007_id413720-E4-2" TYPE="PRO"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="38" END="45">le altre</charseq>
        </extent>
        <head>
          <charseq START="41" END="45">altre</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E5" TYPE="PER"
SUBTYPE="Group" CLASS="GEN">
      <entity_mention ID="adige20041007_id413720-E5-2" TYPE="NOM"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="456" END="466">agricoltori</charseq>
        </extent>
        <head>
          <charseq START="456" END="466">agricoltori</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E6" TYPE="PER"
SUBTYPE="Group" CLASS="GEN">
      <entity_mention ID="adige20041007_id413720-E6-2" TYPE="NOM"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="470" END="479">allevatori</charseq>
        </extent>
        <head>
          <charseq START="470" END="479">allevatori</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E7" TYPE="PER"
SUBTYPE="Group" CLASS="SPC">
      <entity_mention ID="adige20041007_id413720-E7-2" TYPE="NOM"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="456" END="504">agricoltori e allevatori
fino all'epoca dei Celti</charseq>
        </extent>
        <head>
```

```xml
          <charseq START="456" END="479">agricoltori e
allevatori</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E8" TYPE="PER"
SUBTYPE="Group" CLASS="SPC">
      <entity_mention ID="adige20041007_id413720-E8-2" TYPE="NOM"
PRIMARY="true" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="426" END="504">dalle più antiche
comunità di agricoltori e allevatori fino all'epoca dei
Celti</charseq>
        </extent>
        <head>
          <charseq START="444" END="451">comunità</charseq>
        </head>
      </entity_mention>
    </entity>
    <entity ID="adige20041007_id413720-E9" TYPE="GPE"
SUBTYPE="Population-Center" CLASS="SPC">
      <entity_mention ID="adige20041007_id413720-E9-1" TYPE="NAM"
PRIMARY="true" ROLE="LOC" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="51" END="56">TRENTO</charseq>
        </extent>
        <head>
          <charseq START="51" END="56">TRENTO</charseq>
        </head>
      </entity_mention>
      <entity_mention ID="adige20041007_id413720-E9-2" TYPE="NAM"
PRIMARY="false" ROLE="GPE" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="197" END="202">Trento</charseq>
        </extent>
        <head>
          <charseq START="197" END="202">Trento</charseq>
        </head>
      </entity_mention>
      <entity_attributes>
        <name NAME="TRENTO">
          <charseq START="51" END="56">TRENTO</charseq>
        </name>
        <name NAME="Trento">
          <charseq START="197" END="202">Trento</charseq>
        </name>
      </entity_attributes>
    </entity>
    <entity ID="adige20041007_id413720-E10" TYPE="GPE"
SUBTYPE="Population-Center" CLASS="SPC">
      <entity_mention ID="adige20041007_id413720-E10-1"
TYPE="NAM" PRIMARY="true" ROLE="GPE" METONYMY_MENTION="FALSE">
        <extent>
          <charseq START="323" END="328">Milano</charseq>
        </extent>
        <head>
          <charseq START="323" END="328">Milano</charseq>
        </head>
```

```xml
        </entity_mention>
        <entity_attributes>
          <name NAME="Milano">
            <charseq START="323" END="328">Milano</charseq>
          </name>
        </entity_attributes>
      </entity>
      <entity ID="adige20041007_id413720-E11" TYPE="PER"
SUBTYPE="Group" CLASS="SPC">
        <entity_mention ID="adige20041007_id413720-E11-1"
TYPE="NOM" PRIMARY="true" METONYMY_MENTION="FALSE">
          <extent>
            <charseq START="496" END="504">dei Celti</charseq>
          </extent>
          <head>
            <charseq START="500" END="504">Celti</charseq>
          </head>
        </entity_mention>
      </entity>
    </document>
</source_file>
```

References

Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. *Local Entity Detection and Recognition Annotation for Evalita 2009*.
On-line: http://evalita.fbk.eu/ doc/Annotation_Report_LEDR.pdf

Automatic Content Extraction 2008 Evaluation Plan (ACE08) Assessment of Detection and Recognition of Entities and Relations Within and Across Documents, April 2008
On-line: http://www.nist.gov/speech/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf

Web Sites

ACE Evaluation, http://www.nist.gov/speech/tests/ace/

ACE annotation, http://www.ldc.upenn.edu/Projects/ACE/

Callisto, http://callisto.mitre.org

I-CAB, http://tcc.fbk.eu/projects/ontotext/icab.html

L'Adige, http://www.ladige.it/