EVALITA 2009: Description and Results of the Local Entity Detection and Recognition (LEDR) task

Valentina Bartalesi Lenzi and Rachele Sprugnoli

CELCT, Center for the Evaluation of Language and Communication Technologies
Via Sommarive 18, 38123 Povo, Italy
{bartalesi, sprugnoli}@celct.it

Abstract. In this paper, we describe motivations and features of the LEDR (Local Entity Detection and Recognition) task at EVALITA 2009. Our work refers to the task of the same name within the Automatic Content Extraction (ACE) program. We adopted the ACE annotation scheme adapting it to the specific morpho-syntactic features of Italian in order to create training and test data to be used in the evaluation of Information Extraction systems for Italian. In this report annotated data and evaluation measures are presented. Moreover, the results obtained by the participating system are showed.

Keywords: Local Entity Detection and Recognition, Information Extraction, EVALITA 2009, evaluation of NLP systems.

1 Introduction

This report presents the Local Entity Detection and Recognition (LEDR) task organized for the second EVALITA campaign. The task was introduced to encourage research on system capable of automatically detect and recognize entities within documents.

We took the Automatic Content Extraction (ACE) program as a reference, adopting its evaluation methodology and annotation scheme. As the ACE guidelines and datasets have been developed for English, Chinese and Arabic, the main part of our effort consisted in adapting the guidelines to the morpho-syntactic features of Italian and in annotating specific training and test data.

The paper is organized as follows. Section 2 presents the definition of the task, while Section 3 describes the data for the training and the evaluation of the systems. Evaluation measures and results are showed in Sections 4 and 5. Finally, some conclusions are drawn.

2 Task Definition

The LEDR task is designed to measure a system ability to detect a set of specified entities (i.e. persons, organizations, geo-political entities and geographical locations) mentioned in source texts, to recognize selected information about these entities (i.e.

type, subtype and class), and to cluster the mentions for each entity together into a unique entity ID.

In the *Local* EDR task, each document is processed separately and entities that are mentioned in different documents are treated as different entities. It differs from the *Global* task, performed in ACE 2008, that requires systems to process the source data across documents.

In this scenario, an entity is defined as a representation of an object in the world, while an entity mention is any textual reference 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. coreferring mentions).

In the following subsections, entity and mention attributes are briefly described; for a complete description refer to the annotation guidelines [1].

2.1 Entity attributes

Each entity has three attributes: semantic type, subtype and class.

Four semantic type were defined:

1. Person: a single individual or a group of humans.
2. Organization: corporations, agencies, and other groups of people defined by an established organizational structure.
3. Geo-Political Entity: geographical regions defined by political and/or social groups (e.g. a nation, its region, its government, or its people).
4. Location: geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.

For each semantic type, various subtypes, that provide further semantic information, were identified. Table 1 shows entity types and subtypes.

Table 1. Entity types and subtypes.

Entity Type	Subtype
PER	Individual, Group, Indefinite
ORG	Government, Commercial, Educational, Entertainment, Non Governmental, Media, Religious, Medical Science, Sports
GPE	Continent, Nation, State or Province, County or District, Population Center, Cluster, Special
LOC	Address, Boundary, Celestial, Water-Body, Land-Region-natural, Region-International, Region-General

The entity class attribute describes the kind of reference each entity makes to something in the world. The allowable entity classes are listed in Table 2.

Table 2. Entity classes.

Class	Description
SPC (Specific Referential)	An entity that refers to a particular object (or set of objects) in the real world
GEN (Generic Referential)	An entity that does not refer to a particular object (or set of objects) but to a general type of objects
USP (Under-specified Referential)	A non-specific, non-generic entity
NEG (Negatively Quantified)	A negatively quantified entity

2.2 Entity Mention Detection

LEDR systems have also been scored for Entity Mention Detection (EMD) accuracy. The goal of this evaluation was to assess the system's ability to detect entity mentions and output them along with their attributes. More precisely, the output for each entity mention includes the mention type (see Table 3), its extent, the location of its syntactical head within the extent, and optionally the mention role (i.e. the role of a geo-political entity invoked by its context) and style (i.e. literal or metonymic).

Table 3. Syntactic categories of entity mentions: types.

Mention Type	Subtype
NAM (Names)	proper nouns and nicknames
NOM (Quantified Nominal Constructions)	nouns quantified with determiners, quantifiers, or possessives
PRO (Pronouns)	all pronouns and headless mentions

3 Dataset

As training and test data we have used the I-CAB (Italian Content Annotation Bank) corpus, developed by CELCT and FBK-irst and distributed upon acceptance of the agreement terms for a free research license [3, 4].

I-CAB is made of 525 news documents taken from the local newspaper "L'Adige". The selected articles 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. The development part

consists of 335 articles, for a total of around 113,000 words, and the test part consists of 190 articles, for a total of around 69,000 words.

Although we have extended the original annotation scheme to include a wider range of entities, for the purpose of this evaluation campaign we simplified the I-CAB annotation in order to conform the Local Entity Detection and Recognition task in EVALITA to the one developed in the ACE program. Table 4 presents detailed data about the annotated entities and entity mentions in training and test set for the LEDR task.

Table 4. Quantitative data about training and test set.

		Training	*Test*	*TOTAL*
PER	Entities	4,493	2,014	6,507 (53%)
	Mentions	10,086	4,425	14,511 (57%)
ORG	Entities	2,219	784	3.003 (25%)
	Mentions	4,318	1,471	5,789 (23%)
GPE	Entities	1,459	667	2,126 (17%)
	Mentions	2,920	1,323	4,243 (17%)
LOC	Entities	397	167	564 (5%)
	Mentions	574	252	826 (3%)
TOT	Entities	8,568	3,632	12,200
	Mentions	17,898	7,471	25,369

Training data were distributed in the following formats:

- TXT files in UTF-8 encoding, containing the source text;
- APF (ACE Program Format) files containing the annotation in the form of XML standoff annotation.

Test data were distributed in the TXT format, while the data format required for system output was the APF.

4 Evaluation

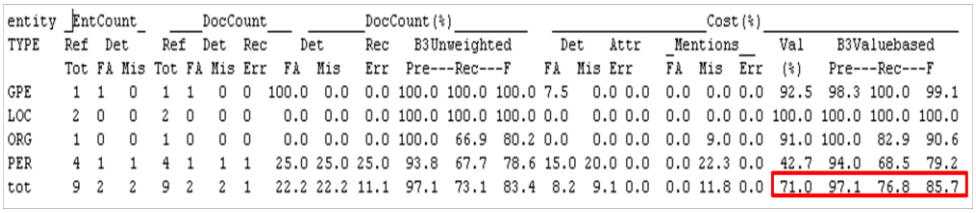
For the official evaluation we used the ACE 2008 scorer, whose formulas are described in [2]. The scorer computes the following evaluation measures:

- Value, the sum of the values of all of the system's output tokens, normalized by the sum of the values of the reference data. In the ACE campaign it is used to measure the overall performance of participating systems.
- Precision, indicates the percentage of correct positive predictions and it is computed as the ratio between the number of entities/mentions correctly identified by the system (True Positive) and the total number of entities/mentions identified by the system (True Positive plus False Positive).
- Recall, indicates the percentage of positive cases recognized by the system and it is computed as the ratio between the number of entities/mentions correctly identified

by the system (True Positive) and the number of entities/mentions that the system was expected to recognize (True Positive plus False Negative).

- F-Measure, the weighted harmonic mean of precision and recall.

Figure 1 shows an example of the scorer output.



entity	EntCount			DocCount				DocCount(%)			B3Unweighted			Cost(%)						Val	B3Valuebased		
TYPE	Ref	Det		Ref	Det		Rec	Det		Rec				Det		Attr	Mentions						
	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

Fig. 1. The evaluation output.

5 Results

Only one participant (Fondazione Bruno Kessler in conjunction with University of Trento) returns the results to the organizers. In Table 5 we provide the system evaluation in term of Value score, Precision (P), Recall (R) and F-measure (F) for both LEDR and EMD.

Table 4. Percentages for Value, Precision, Recall and F-measure of the participating system.

LEDR evaluation		**EMD evaluation**	
Value	36.7%	Value	65,7%
P	78.5%	P	78,1%
R	61.1%	R	74,1%
F	68.7%	F	76,1%

Figures 2, 3 and 4 show respectively the performance of the participating system in term of the percentage of cost (i.e. the lost Value) by entity type, entity class, and entity mention type. Much of the lost Value for entity types and entity mention types is from misses. For what concerns the entity class, the system received negative values for Generic Referential, Under-specified Referential, and Negatively Quantified entities because of the high percentage of false alarms.

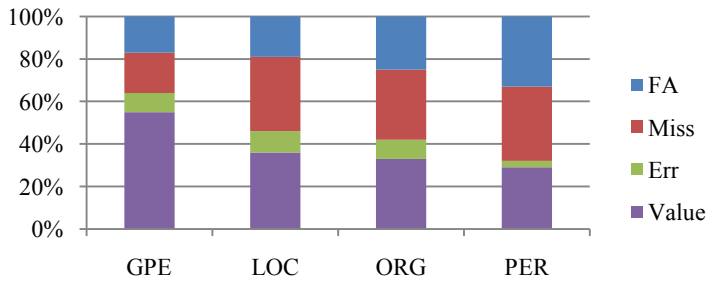


Fig. 2. Percentage of cost by entity type.

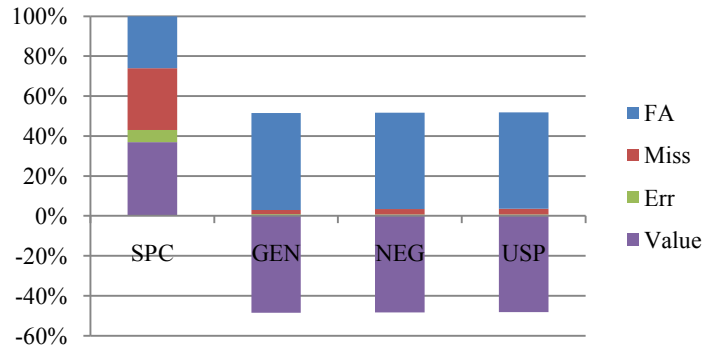


Fig. 3. Percentage of cost by entity class.

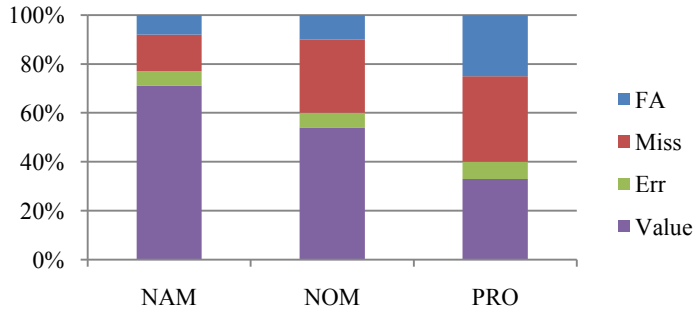


Fig. 4: Percentage of cost by entity mention type.

6 Discussion and Conclusions

At the beginning, six groups, out of which four were not Italian, registered to the task and obtained guidelines and datasets. Unfortunately, only one institution actually submitted the results. This is something that has to be discussed: the task was very

complex and given that it required a substantial effort in the pre-processing and post-processing of the data, the time left for the study of the system architecture and for the choice of algorithm and features is very limited. Moreover, we have to take into consideration that, in general, foreign groups do not have enough resources for the Italian language.

A comparison with the results of the LEDR task performed in the ACE 2008 campaign is possible [5]. The final ranking in ACE was based on the Value score and in the English task the best system achieved 52.6% and the second one 50.8%; all the other systems had a negative overall Value. In this context, the system participating in the EVALITA LEDR task obtained a good result with a Value of 36.7%. Although the ACE competition was well-established and was based on English data, in 2008 only six sites participated with their systems.

In 2009 the Knowledge Base Population task at TAC (Text Analysis Conference) took the place of the ACE evaluation [6]. This denotes a change in the approach of the scientific community to the content extraction research field. Anyway, in order to develop automatic knowledge base population systems, good Information Extraction and coreference resolution algorithms are required. Therefore, it is important to invest more in key tasks, like the LEDR one.

References

1. Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: Italian Content Annotation Bank (I-CAB): Temporal Expressions. Technical report, FBK-irst, http://evalita.fbk.eu/doc/Annotation_Report_LEDR.pdf (2007)
2. Automatic Content Extraction 2008 Evaluation Plan (ACE08), http://www.itl.nist.gov/iad/mig/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf
3. Magnini, B., Cappelli, A., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R., Romano, L., Girardi, C., Negri, M.: Annotazione di contenuti concettuali in un corpus italiano: I-CAB. In: Proceedings of SILFI 2006, X Congresso Internazionale (2006)
4. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: I-CAB: the Italian Content Annotation Bank. In: Proceedings of LREC 2006. Genoa, Italy (2006)
5. NIST 2008 Automatic Content Extraction Evaluation (ACE08) - Official Results, http://www.itl.nist.gov/iad/mig/tests/ace/2008/doc/ace08_eval_official_results_20080929.html
6. TAC 2009 Knowledge Base Population Track, http://apl.jhu.edu/~paulmac/kbp.html