EVALITA 2007: IL TASK RICONOSCIMENTO DI ENTITÀ NOMINATE

EVALITA 2007: THE NAMED ENTITY RECOGNITION TASK

MANUELA SPERANZA

SOMMARIO/ *ABSTRACT*

In questo articolo si descrive il Task di *Named Entity Recognition* organizzato nell'ambito della campagna di valutazione EVALITA 2007. In particolare, si riportano informazioni relative al dataset utilizzato e alle metriche di valutazione adottate e si presentano i risultati ottenuti dai sistemi che vi hanno partecipato.

In this paper we describe the Named Entity Recognition Task organized in the context of the EVALITA 2007 evaluation campaign. In particular, we report information about the dataset and the evaluation metrics we used, and we discuss the results obtained by participant systems.

Keywords: Named Entity Recognition, evaluation.

1. Introduction

The Named Entity Recognition (NER) task evaluated the recognition of four different types of Named Entities, i.e. Person (PER), Organization (ORG), Geo-Political Entity (GPE) and Location (LOC). The task is based on the ACE-LDC guidelines, for the ACE Entity Recognition and Normalization Task [4], with all the adaptations needed to limit the task to the recognition of Named Entities [6].

The NER Task at EVALITA 2007 had six participants from four different countries: University of Alicante (UniALi) and Yahoo! (Barcelona) from Spain, University of Dortmund (UniDort) and University of Duisburg-Essen (UniDuE) from Germany, LDC (University of Pennsylvania) from the USA, and Fondazione Bruno Kessler–irst (FBKirst) from Trento, Italy.

2. Dataset

As a dataset for the NER Task we have used the Italian Content Annotation Bank (I-CAB), developed in the context of the ONTOTEXT Project [7]. I-CAB [5] consists of 525 news stories taken from different sections (e.g. Cultural, Economic, Sports and Local News) of the local newspaper "L'Adige" [3], for a total of around 180,000 words. The selected news stories belong to four different days (September, 7th and 8th 2004 and October, 7th and 8th 2004).

Training data consist of 335 news stories, for a total of around 113,000 words, and test data consist of 190 news stories, for a total of around 69,000 words. Table 1 presents more detailed data about the size of the corpus and about the Named Entities annotated in it.

Table 1: Quantitative data about I-CAB

	Training	Test	Total
# News stories	335	190	525
# Sentences	7,227	4,002	11,229
# Words	113,634	68,930	182,564
# Tokens	132,587	79,889	212,476
# GPE	1,740	1,073	2,813
# LOC	240	122	362
# ORG	2,518	1,140	3,658
# PER	2,936	1,641	4,577

Development data made available to participants are annotated with Named Entities in the IOB2 format, i.e.with tags consisting of two parts:

- the IOB2 tag: "B" denotes the first token of a Named Entity, "I" is used for all other tokens in a Named Entity, and "O" is used for all other words;
- the Named Entity type tag (only for tokens belonging to Named Entities): PER (for Person), ORG (for Organization), GPE (for Geo-Political Entity), or LOC (for Location).

In order to make the data more accessible, we also provided some pre-processing both for the training data and the test data, i.e. sentence splitting and Part of Speech tagging (using the ELSNET tagset for Italian).

3. Evaluation Metrics

For the official evaluation of system results we have used the scorer made available by CONLL for the 2002 Shared Task, which can be freely downloaded from the CONLL website [2].

With respect to the results submitted by the participants (each participant was allowed to submit up to two runs), the CONLL scorer computes the following evaluation measures: Precision, Recall, and F-Measure (FB1).

Precision indicates the percentage of correct positive predictions and is computed as the ratio between the number of Named Entities correctly identified by the system (True Positive) and the total number of Named Entities identified by the system (True Positive plus False Positive), as shown in (1).

$$(1)\ Pr. = \frac{TP}{TP + FP} \qquad (2)\ Re. = \frac{TP}{TP + FN}$$

Recall indicates the percentage of positive cases recognized by the system and is computed as the ratio between the number of Named Entities correctly identified by the system (True Positive) and the number of Named Entities that the system was expected to recognize (True Positive plus False Negative), as shown in (2).

F-Measure, the weighted harmonic mean of Precision and Recall computed as shown in (3), has been used for the official ranking.

$$(3)\ FB1 = \frac{2 \times [precision \times recall]}{precision + recall}$$

4. Results

The F-Measure achieved by participants systems ranges from 82.14 to 63.10 (considering the best run of each group). Most systems obtained values of F-Measure between 63 and 69, while only two submissions are above 70, i.e. UniDuE_Roessler_NER which obtained FB1=72.27 (best run), and FBKirst_Zanoli_NER, which stands out as about 10 points higher than the other systems (best run FB1=82.14).

Results obtained by participant systems have been compared with two different baseline rates computed by identifying in the test data only the Named Entities that appear in the training data. In one case (baseline-u), only entities which had a unique class in the training data were taken in consideration (FB1=36.85). In the other case (baseline), entities which had more than one class in the training data were also considered, and annotated according to the most frequent class (FB1=41.11).

If we compare the results in terms of Precision and Recall (forth and fifth columns), we can see that most systems obtained higher values for Precision than for Recall, with the exception of UniDuE_Roessler_NER_r1, which slightly favors Recall against Precision (72.94% vs. 71.62), and UniAli_Kozareva_NER, whose values for Recall and Precision differ by more than eight pecentage points (70.95% vs. 62.73).

Table 2: System results in terms of F-Measure, Precision and Recall (overall and for different types of Named Entities)

Rank	Participant	Over. FB1.	Over. Prec.	Over. Recall	FB1			
					GPE	LOC	ORG	PER
1	FBKirst_Zanoli_NER_r2	**82.14**	83.41%	80.91%	85.54	73.04	64.27	92.12
2	FBKirst_Zanoli_NER_r1	**81.28**	82.97%	79.65%	85.52	73.04	64.06	90.40
3	UniDuE_Roessler_NER_r1	**72.27**	71.62%	72.94%	78.39	53.92	49.89	84.42
4	UniDuE_Roessler_NER_r2	**71.93**	73.28%	70.62%	78.75	54.73	49.01	83.64
5	Yahoo_Ciaramita_NER_r1	**68.99**	71.28%	66.85%	75.38	52.83	49.08	78.89
6	Yahoo_Ciaramita_NER_r2	**68.15**	70.44%	66.00%	75.08	52.31	46.85	78.36
7	UniDort_Jungermann_NER_r2	**67.90**	70.93%	65.12%	73.18	46.07	45.85	79.78
8	UniDort_Jungermann_NER_r1	**67.79**	70.93%	64.91%	73.18	46.07	45.74	79.58
9	UniAli_Kozareva_NER	**66.59**	62.73%	70.95%	72.60	47.26	47.81	78.66
10	LDC_Walker_NER_r1	**63.10**	83.05%	50.88%	65.25	52.94	40.70	75.39
11	LDC_Walker_NER_r2	**62.70**	82.12%	50.70%	65.13	50.56	36.26	76.44
-	BASELINE	**41.11**	42.44%	39.86%	69.67	27.63	40.32	25.48
-	BASELINE -u	**36.85**	40.29%	33.95%	57.64	26.32	39.43	25.55

However, it is worth pointing out that the system with the most striking difference between Precision and Recall is LDC_walker_NER_r1, which obtained 50.88% in terms of Recall and 83.05% (the second best score) in terms of Precision.

As far as the different types of Named Entitites are concerned (last columns), the results of the NER Task at EVALITA 2007 allow us to draw the conclusion that the recognition of Named Entities of type PER is quite an easy subtask. In fact, all participant systems obtained their highest values in terms of F-Measure in this subtask, ranging from 75.39 to 92.12.

The recognition of Geo-Political Entities does not constitute a problem for most participant systems either; in fact, F-Measure values for GPE Entities are slightly lower but still satisfactory, ranging between 65.13 and 85.54.

System results drop significantly as far as the recognition of Named Entities of type LOC are concerned, ranging between 46.07 and 73.04. The effect of relatively low results in this subtask on the overall performance of the system, however, is limited by the fact that LOC Entities constitute less than 4% of the total number of Named Entities in the corpus (see Table 1).

The most problematic subtask in NER seems to be the recognition of Named Entities of type ORG. All systems except one, in fact, obtained their lowest result in the recognition of this type of Entities, none of them being able to perform better than 65% in terms of F-Measure.

On the other hand, the complexity of recognizing the four types of Entities from the point of view of the baseline is quite different. It obtained surprizingly high values for ORG Entities (FB1=69.67), average results for ORG Entities (FB1=40.32) and low results for LOC and PER Entities (respectively FB1=27.53 and 25.48).

5. Conclusions

With the submissions of results by six different teams, we feel that we have achieved our initial goal of fostering research on Named Entity Recognition for Italian, although we had only one Italian institution among our participants. We hope that the outcome of EVALITA 2007 will stimulate the organization of other evaluations campaign for Italian in the future, where it might be interesting to propose the more complex task of detecting also entity co-reference and entity subtypes.

The approaches taken by participant systems have been described in individual papers; we look forward to discussing them at the final workshop in Rome.

Acknowledgments

This work has been partially supported the three-year project ONTOTEXT [7] funded by the Provincia Autonoma di Trento. We would like to thank the local newspaper “L’adige” and to acknowledge the people who collaborated with FBK-irst in annotating the data, i.e. Valentina Bartalesi Lenzi and Rachele Sprugnoli.

REFERENCES

[1] ACE. http://www.nist.gov/speech/tests/ace/index.htm

[2] CONLL. http://www.cnts.ua.ac.be/conll2002/ner/

[3] L’Adige. http://www.ladige.it/

[4] Linguistic Data Consortium (LDC), Automatic Content Extraction English Annotation Guidelines for Entities, version 5.6.1 2005.05.23. http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf

[5] Magnini, Cappelli, Pianta, Speranza, Bartalesi Lenzi, Sprugnoli, Romano, Girardi, Negri. Annotazione di contenuti concettuali in un corpus italiano: I-CAB. In Proceedings of SILFI 2006, X Congresso Internazionale della Società di Linguistica e Filologia Italiana, Firenze 14-17 giugno 2006.

[6] Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. Italian Content Annotation Bank (I-CAB): Named Entities, *Technical report, ITC-irst, 2007*. http://evalita.itc.it/tasks/I-CAB-Report-Named-Entities.pdf

[7] ONTOTEXT. http://tcc.itc.it/projects/ontotext

CONTACT

MANUELA SPERANZA
IRST - Centro per le Ricerca Scientifica e Tecnologica
Fondazione Bruno Kessler
Via Sommarive 18
38050 Povo (Trento)
Email: manspera@itc.it

MANUELA SPERANZA is a Junior Researcher at the Cognitive and Communication Technologies Division at FBK-irst. She holds a Master Degree in Foreign Languages and Literatures. Her main research interests are in NLP for Knowledge Management and in corpus annotation. She is currently working on the creation of evaluation benchmarks for various tasks, such as Entity Recognition and Topic Detection.