View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Archivio della ricerca - Fondazione Bruno Kessler

The Named Entity Recognition Task at EVALITA 2009

Manuela Speranza

FBK irst
38123 Povo, Trento, Italy
manspera@fbk.eu

Abstract. The submissions of results to the Named Entity Recognition task at EVALITA 2009 by seven different teams (five working in Italy and two abroad) confirms the interest displayed in the 2007 evaluation campaign. Using the same guidelines and evaluation metrics as in the previous edition, there has been a significant improvement in the average performance of the systems, with an average F-measure of the systems' best run close to 76% (in comparison to a 70% average for the 2007 evaluation) and three systems scoring above 80%.

Keywords: Evaluation, Named Entity Recognition.

1 Introduction

Following upon the success obtained by its previous edition, the Named Entity Recognition (NER) task at EVALITA 2009 was organised according to the same guidelines, with the aim of evaluating systems' performance on the recognition of four different types of Named Entities, namely 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 [3], with certain adaptations needed to limit the task to the recognition of Named Entities [5].

The NER Task at EVALITA 2009 had seven participating systems (one more than the previous edition); five of these submitted two runs and two submitted only one for a total of twelve runs to be officially evaluated.

Seven different institutions were involved in various degrees in the development of the systems which participated in the NER Task; among them, there were four Italian academic/research institutions, i.e. University of Trento (UniTN), Fondazione Bruno Kessler (FBK), University of Pisa (UniPI) and ILC-CNR, one private company, i.e. RGB s.r.l. (Milan) and two non-Italian Universities, i.e. University of Geneva (UniGen) and East China Normal University (ECNU). Two systems were the result of a collaboration between different institutions, i.e. UniTN-FBK-RGB and UniPi-ILC-CNR.

2 Dataset

As a dataset for the NER Task at EVALITA 2009 we used I-CAB, the Italian Content Annotation Bank [4]. The training data consist of the union of the training and the test data used for the 2007 evaluation, i.e. 525 news stories for a total of 212,478 tokens [6]. The test data consist of 180 news stories, for a total of 86,419 tokens (see Table 1 for more details about the size of the corpus and the news stories). The Named Entities contained in the corpus amount to 11,410 and 4,966 for training and test data respectively, with a higher percentage of PER Entities, followed by ORG and GPE Entities and a small number of LOC Entities. Table 2 reports on the distribution of the Named Entities in detail.

Table 1. Quantitative data about the training and test data.

	Training	**Test**
News stories	525	180
Sentences	11,227	4,136
Tokens	212,478	86,419
Tokens/news story	404.7	480.1

Table 2. Quantitative data about the Named Entities in the training and in the test data.

	Training		**Test**	
GPE	2,813	(24.66%)	1,143	(23.02%)
LOC	362	(3.17%)	156	(3.14%)
ORG	3,658	(32.06%)	1,289	(25.96%)
PER	4,577	(40.11%)	2,378	(47.88%)
Total	11,410		4,966	

Development data made available to the participants were 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 for both 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).

$$\text{Precision} = TP / (TP + FP) \quad (\mathbf{1})$$

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).

$$\text{Recall} = TP / (TP + FN) \quad (\mathbf{2})$$

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

$$FB1 = 2 (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (\mathbf{3})$$

4 Results

The results obtained by participating systems in the official evaluation are quite satisfactory, with values for F-Measure ranging from 82.00% to 61.03%, and four systems (out of seven) scoring above 79% (considering their best run). The four best-scoring systems, in fact, obtained very close scores in terms of F-Measure, with the best-scoring system (i.e. FBK_ZanoliPianta) achieving 82%, the second (i.e. UniGen_Gesmundo) 81.46%, the third (i.e. UniTN-FBK-RGB) 81.09% and the forth (i.e. UniTN_Nguyen) 79.77% (see Table 3).

If we compare the results in terms of Precision and Recall, we can see that all the systems obtained higher values for Precision than for Recall in all submitted runs. More precisely (Figure 1), the best-scoring system (i.e. FBK_ZanoliPianta) obtained the highest Recall (80.02%), but UniGen_Gesmundo obtained the highest Precision score (86.06%).

Table 3. Systems' results in terms of F-Measure, Precision and Recall.

	Participant	Over. FB1	Over. Prec.	Over. Recall	FB1			
					GPE	LOC	ORG	PER
1	FBK_ZanoliPianta	82.00	84.07	80.02	85.13	51.24	70.56	88.31
2	UniGen_Gesmundo_r2	81.46	86.06	77.33	83.36	50.81	71.08	87.41
3	UniTN-FBK-RGB_r2	81.09	83.20	79.08	85.25	52.24	69.61	86.69
4	UniTN-FBK-RGB_r1	80.90	83.05	78.86	85.19	54.62	69.41	86.30
5	UniTN_Nguyen_r1	79.77	82.26	77.43	82.85	42.34	67.89	86.44
6	UniTN_Nguyen_r2	79.61	81.65	77.67	82.49	50.85	67.38	86.25
7	UniGen_Gesmundo_r1	76.21	83.92	69.79	79.07	47.06	64.67	82.04
8	UniTN_Rigo_r2	74.98	81.08	69.73	75.96	38.32	60.36	83.18
9	UniTN_Rigo_r1	74.34	80.71	68.91	75.77	31.16	59.87	82.38
10	UniPI-ILC-CNR_r2	69.67	75.42	64.74	71.42	38.91	58.37	76.38
11	UniPI-ILC-CNR_r1	67.98	73.65	63.11	71.66	27.45	57.02	73.85
12	ECNU_Cai	61.03	65.55	57.09	69.25	28.72	51.49	63.49
-	BASELINE	43.99	42.80	45.25	69.00	37.07	45.54	32.06
-	BASELINE –u	39.14	40.58	37.80	52.75	28.57	44.23	32.10

As far as the different types of Named Entitites are concerned (Figure 2), the results of the NER Task at EVALITA 2009 confirm those obtained in the 2007 evaluation, according to which the Named Entities of type PER and GPE were the easiest to recognize. In fact, all participant systems obtained their highest values in terms of F-Measure in one of these subtasks. As for PER Entities, we have nine submissions out of twelve scoring above 80% in terms of F-Measure and values ranging from 63.49% to 88.31% (FBK_ZanoliPianta obtained the highest score). Similarly, for Geo-Political Entities, we have six submissions with F-Measure values above 80% and values ranging between 69.25% and 85.25% (UniTN-FBK-RGB obtained the highest score).

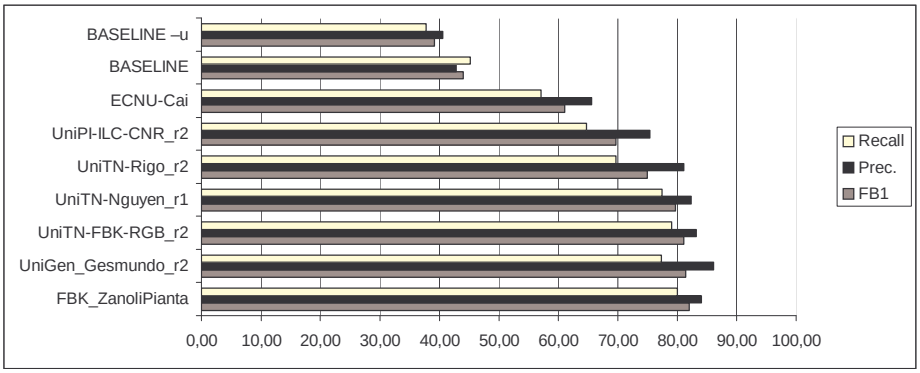

Fig 1. A chart of the overall results of the systems.

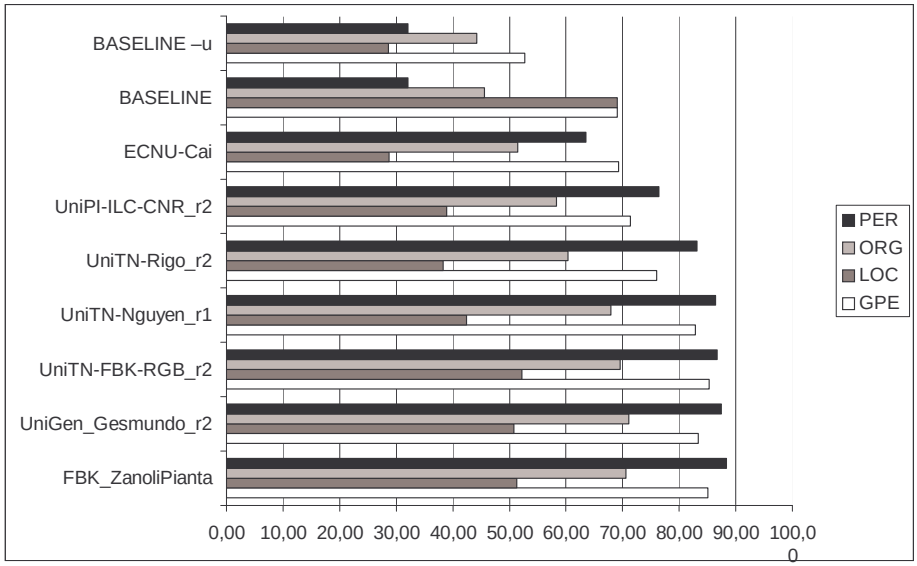

Fig. 2. A chart of the results for the different Entity types (in terms of F-Measure).

System results drop significantly as far as the recognition of Named Entities of type ORG are concerned, with only two submissions above 70% and results ranging between 51.49% and 70.56%.

The most problematic subtask in NER, however, turned out to be the recognition of Named Entities of type LOC. All systems, in fact, obtained their lowest result in the recognition of this type of Entity, none of them being able to perform better than 55% in terms of F-Measure (UniTN-FBK-RGB obtained 54.62%, the highest score). The effect of the low results in this subtask on the overall performance of the system is limited by the fact that LOC Entities constitute less than 4% of the total number of Named Entities in the corpus (see Table 2), but this is in contradiction to the findings of the 2007 evaluation when all systems except one obtained their lowest result for the recognition on ORG Entities.

As in the 2007 evaluation, results obtained by participating 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 into consideration (FB1=39.14). 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=43.99).

All systems obtained results well above the baseline rates, in terms of Precision, Recall and F-Measure. It is interesting to point out, however, that the most difficult subtask, for a simple algorithm such as the suggested baseline, is the recognition of PER Entities, where systems obtained their highest scores. The baseline obtains low results on the recognition of ORG and LOC Entities as well, while it obtains an F-Measure value of 69% on GPEs.

5 Conclusions

With the submission of results by seven different teams, for the second time the Named Entity Recognition task at EVALITA has become the reference for NER evaluation for Italian. We feel that we have satisfactorily achieved our goal of fostering research in the field. In fact, we went from having one Italian institution among our participants in 2007 to having five in 2009, and continued having the participation of institutions from different countries as well.

In addition, the results showed that the group of participating systems performed impressively as a whole, with an average score of 75.71%, going up significantly with respect to 2007 (more than five percentage points from 70.16%).

Finally, the number of systems that were competing for the best score has increased, as four systems scored very close to each other at the top, whereas in 2007 we had a single strong performer at the top with a large gap to the second best-scoring system.

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

References

1. ACE, http://www.nist.gov/speech/tests/ace/index.htm
2. CONLL, http://www.cnts.ua.ac.be/conll2002/ner/
3. Linguistic Data Consortium (LDC): ACE (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 (2005)
4. 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)
5. Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: Italian Content Annotation Bank (ICAB): Named Entities, Technical report, ITC-irst, http://evalita.fbk.eu/doc/I-CAB-Report-Named-Entities.pdf (2007)
6. Speranza, M.: The Named Entity Recognition Task at EVALITA 2007. In: Proceedings of EVALITA 2007. Intelligenza Artificiale, vol. 4, issue 2 (2007)