

EVALITA 2011: Description and Results of the Named Entity Recognition on Transcribed Broadcast News Task

Valentina Bartalesi Lenzi¹, Manuela Speranza² and Rachele Sprugnoli¹

¹ CELCT, Via alla Cascata 56/C, 38123 Povo (TN), Italy
{bartalesi, sprugnoli}@celct.it

² FBK-irst, Via Sommarive 18, 38123 Povo (TN), Italy
manspera@fbk.eu

Abstract. This report describes features and outcomes of the Named Entity Recognition on Transcribed Broadcast News task at EVALITA 2011. This task represented a change with respect to previous editions of the NER task within the EVALITA evaluation campaign because it was based on automatic transcription of broadcast news. Four participants took part in the task and submitted a total of 9 runs. In this paper, annotated data and evaluation measures are presented together with the results obtained by the participating systems.

Keywords: Evaluation, Named Entity Recognition, Broadcast News, Italian.

1 Introduction

In the Named Entity Recognition (NER) task at EVALITA 2011, systems were required to recognize different types of Named Entities (NEs) in Italian texts. As in the previous editions of EVALITA, four NE types were distinguished: Person (PER), Organization (ORG), Location (LOC) and Geo-Political Entities (GPE). Participant systems had to identify both the correct extension and type of each NE. The output of participant systems was evaluated against a manually created gold standard.

The annotation of the data was based on the ACE-LDC standard for the Entity Recognition and Normalization Task [1] adapted to Italian [2] and limited to the recognition of Named Entities [3].

The main novelty introduced for the 2011 edition is the fact that the task was based on broadcast news and consisted of two subtasks:

- **Full task:** participants had to perform both automatic transcription of the news - using an Automatic Speech Recognition (ASR) system of their choice - and recognition of the Named Entities within that transcript;
- **NER only:** organizers provided participants with the automatically created transcript of the news produced using a state-of-the-art ASR system [4] and participants had to perform Named Entity Recognition on this transcript.

For each subtask, participants were required to submit at least one run produced according to the ‘closed’ modality: only the data distributed by the organizers and no additional resources (i.e. gazetteers, NE dictionaries, ontologies, Wikipedia and complex NLP toolkits such as TextPro¹, GATE² and OpenNLP³) were allowed for training and tuning the system. Other runs could be produced according to the ‘open’ modality, i.e. using any type of supplementary data.

The NER Task at EVALITA 2011 had four participants, all for the NER only subtask, who submitted a total of nine runs to be officially evaluated. One participant submitted four runs (two in closed modality and two open), another one three runs (two in closed modality and one open) and two submitted only one closed run for a total of nine runs to be officially evaluated.

2 Dataset

The dataset consisted of 20 news broadcast provided by RTTR⁴, a Trentino service broadcaster, for a total of ten hours of transmission. These data have been manually and automatically transcribed and manual transcriptions have been annotated with NEs by three expert annotators in order to create the gold standard. Five hours of data were devoted to training while the remaining five hours composed the test set. Table 1 gives some details about the size of the dataset while Table 2 reports on the distribution of the NEs.

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

	Training	Test
News broadcast	10	10
Hours of transmission	5	5
Tokens	42,595	36,643

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

	Training	Test
GPE	747 (38,82%)	672 (39,46%)
LOC	105 (5,46%)	88 (5,17%)
ORG	618 (32,12%)	527 (30,94%)
PER	454 (23,60%)	416 (24,43%)
Total	1,924	1,703

Participants received the same training data for both subtasks, namely:

- news manually transcribed and annotated in the IOB2 format [5] where every token is annotated with a tag (‘B’ for ‘begin’ for the first token of each NE and

¹ <http://textpro.fbk.eu/>

² <http://gate.ac.uk/>

³ <http://incubator.apache.org/opennlp/>

⁴ Radio Tele Trentino Regionale, <http://www.rttr.it/>

‘I’ for ‘inside’ for other tokens of the NE) followed by the NE type; ‘O’ (‘outside’) is used for tokens that do not belong to any NE;

- automatic transcription of that same news in one-token-per-line format;
- recording of that same news in wav format.

In addition, participants could freely obtain the I-CAB corpus [2] as part of the training data.

For the test data, audio files and automatic transcriptions of the news were made available to the participants of the NER-only subtask, while none of the participants has requested the wav files for the Full subtask.

The performance of the ASR system on the two data sets (without any specific training or tuning) is the following: Word Error Rate is 16.39 on the training data and 17.91 on the test data.

3 Evaluation Procedure

The evaluation procedure, comparing systems’ results against a gold standard, consisted of three phases: transcription alignment, NER error detection, and score computation [6].

The first phase, transcription alignment, consisted of aligning the reference transcription (gold standard) with the system’s transcription by determining the best possible alignment at the word level in order to minimize the number of edit operations needed to transform the first into the second (the allowed edit operations are word Insertion, word Deletion and word Substitution).

In the next phase, we detected NER errors by comparing the reference NEs annotated in the gold standard to the hypothesis NEs annotated in the system’s transcription [7]. A hypothesis NE is correct (correct NE match) if all of the following conditions are met:

1. It has a corresponding reference NE, i.e. at least one of the words it contains is aligned to a word that is part of a reference NE;
2. Its extension is correct, i.e. an “exact match” is required in the sense that each word in the hypothesis NE must be aligned with a word in the corresponding reference NE and vice versa (one-to-one mapping between the words);
3. Its NE type is correct, i.e. it has the same NE type as the corresponding reference NE.

Hypothesis NEs which do not have a corresponding reference NE count as False Positives (FP), whereas reference NEs that do not have a corresponding hypothesis NE count as False Negatives (FN).

In the third phase, a final score was computed using the following measures: Precision, Recall and FB1-Measure.

The above described evaluation was performed automatically by means of two scripts that were available to participants before they submitted their outputs: the alignment script and the CoNLL scorer.

The alignment script takes as input the gold standard and the systems’ output, determines the best possible alignment between reference and hypothesis word

transcriptions and inserts an “O” tag in the case of word Insertion or Deletion. It produces as output a file with five columns in which the different columns contain respectively the gold standard token, the gold standard Entity tag, the type of transcription error if any, the system output token, and the system output Entity tag. It should be noticed that the alignment procedure has the effect of producing a slightly modified version of the gold standard and the system output (the insertion of an “O” tag between two tokens of an entity, for example, has the effect of splitting the entity into two⁵).

The CoNLL scorer, made available by CoNLL for the 2002 Shared Task, takes as input the gold standard and the systems’ output aligned in the previous phase and compares the two sets of NE tags. It reports the final scoring in terms of precision, recall and FB1-Measure.

4 Results

The results obtained by participant systems in the official evaluation (see Table 3), with values for FB1 measure ranging from 63.56% to 52.15% for the open modality and from 60.98% to 42.42% for the closed modality⁶ show that there is space for improvement in this task.

Table 3. Systems’ results for the closed and open modality in terms of FB1-Measure, Precision and Recall (overall and for different types of NEs).

Participant	Over. FB1	Over. Prec.	Over. Recall	FB1				
				GPE	LOC	ORG	PER	
CLOSED	1 FBK_Alam_rc1	60.98	61.76	60.23	80.12	55.21	46.82	50.96
	2 FBK_Alam_rc2	60.67	63.97	57.68	78.89	56.25	47.77	49.36
	3 FBK_Chowdhury_rc	57.02	63.31	51.86	74.61	51.01	42.14	46.10
	4 UniPi_SimiDeiRossi_rc2	56.61	64.48	50.45	76.18	44.30	38.35	46.81
	5 UniPi_SimiDeiRossi_rc1	53.47	61.61	47.23	73.60	43.97	28.79	46.77
	6 ISI_GhoshKozareva_rc	42.42	63.86	31.75	64.21	32.52	15.01	28.74
OPEN	1 FBK_Alam_ro1	63.56	65.55	61.69	80.38	56.38	53.24	51.51
	2 UniPi_SimiDeiRossi_ro1	58.19	65.90	52.09	76.25	48.78	40.60	48.75
	3 UniPi_SimiDeiRossi_ro2	52.15	54.83	49.72	72.25	30.62	33.77	46.19
- BASELINE	44.93	38.84	53.28	69.00	36.49	43.37	18.10	
- BASELINE-u	31.11	28.80	32.54	40.12	22.56	42.05	18.25	

Only the best scoring system, i.e. FBK_Alam, achieved results in terms of FB1 above 60 (63.56 in the open modality and 60.98-60.67 in the closed modality). Two

⁵ As a consequence of this, the evaluation of the performance of participant systems was computed on a total of 1,770 reference NEs.

⁶ Unless otherwise specified, in this report we consider the best run of each system.

other systems obtained very close scores: FBK_Chowdhury obtained slightly higher results than UniPi_SimiDeiRossi in the closed modality (57.02 and 56.61 respectively), but did not participate in the open modality where UniPi_SimiDeiRossi obtained a score of 58.19. It should be noted that the differences between the open modality and the closed modality are not very significant, with an improvement of around 2.5 points for FBK_Alam's best runs and 1.5 points for UniPi_SimiDeiRossi.

As in the previous editions of the NER task at EVALITA, 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 (i.e. I-CAB plus the RTTR data). In one case (baseline), entities which had more than one class in the training data were annotated according to the most frequent class (FB1=44.93). In the other case (baseline-u), only entities which had a unique class in the training data were taken into consideration (FB1=31.11). Most systems obtained results well above the baselines, with only one exception.

As far as the different types of Named Entities are concerned, it should be noted that systems in general obtained their highest scores in the recognition of Geo-Political Entities. This is in line with the results registered on written news stories in the previous editions of EVALITA, where they were reported to be among the easiest NE types to recognize [8]. The most striking difference with the previous editions is that on PER Entities, which were also reported to be among the easiest entities to recognize, all systems obtained low FB1 values in comparison not only to GPE but also to LOC Entities. The drop in the recognition of PER Entities can probably be explained by the fact the ASR system had more difficulties with the transcription of person names (which belong to a very open class) than with geo-political and purely geographic names. In line with this explanation, the fact that ORG Entities also belong to a rather open class might have contributed to confirm the position of ORG Entities as the most challenging type of NE for NER systems.

5 Conclusions

With four participants and nine runs submitted, we were satisfied overall with the outcomes of the task. On the other hand, we had a noticeable decrease in the number of participants with respect to the previous editions. A possible explanation for this, is the fact that the introduction of some novelties made the task more complex: first, the choice of evaluating NER systems on (transcribed) spoken data which made expected performance drop significantly with respect to evaluation on written news stories; second, the introduction of the "full" subtask next to the more traditional "NER only" subtask and the subsequent introduction of a new evaluation procedure applicable indistinctively to both subtasks so as not to lose comparability between the respective results. This explanation is supported by the fact that ten groups had initially registered to participate. The fact that we had no submissions to the full task (although four groups had registered) highlights that the interactions between the two research areas of automatic speech recognition and information extraction are still poor.

The outcomes of the task in terms of participant systems' performance are also satisfactory. A comparison of the results obtained with those of previous EVALITA evaluations show that for the best performance we have a decrease in performance of less than 20 percentage points, which is in line with the fact that the selected ASR system obtained a Word Error Rate of 17.91 (on the test data, without any specific training). In order to better evaluate the impact of transcription errors on NER performance, as future work, we will evaluate the results of the same systems on the manual transcription of the test data.

Acknowledgments. Special thanks to Diego Giuliani and Roberto Gretter for their helpful collaboration and advice as ASR experts.

References

1. 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)
2. 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)
3. Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: Italian Content Annotation Bank (I-CAB): Named Entities. Technical report, FBK, <http://www.evalita.it/2011/tasks/NER> (2011)
4. Falavigna, D., Giuliani, D., Gretter, R., Löff, J., Gollan, C., Schlüter, R., Ney, H.: Automatic transcription of courtroom recordings in the JUMAS project. In: 2nd International Conference on ICT Solutions for Justice. Skopje, Macedonia, pp. 65--72 (2009)
5. Sang, T. K. and Veenstra, J.: Representing Text Chunks. In: Proceedings of EACL'99, pp. 173--179 (1999)
6. Bartalesi Lenzi, V., Speranza, M., Sprugnoli, R.: Named Entity Recognition on Transcribed Broadcast News - Guidelines for Participants, <http://www.evalita.it/2011/tasks/NER> (2011)
7. Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., and Quintard, L.: Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. In: Proceedings of IJCNLP, Chiang Mai (Thailand), pp. 8--13 (2011)
8. Speranza, M.: The Named Entity Recognition Task at EVALITA 2009. In: Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, Italy (2009)