EVALITA 2011

http://www.evalita.it/2011

Named Entity Recognition on Transcribed Broadcast News

Guidelines for Participants

Valentina Bartalesi Lenzi
CELCT, Trento
bartalesi@celct.it

Manuela Speranza
FBK, Trento
manspera@fbk.eu

Rachele Sprugnoli
CELCT, Trento
sprugnoli@celct.it

1. Task description

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

The task will be based on the ACE-LDC standards for the Entity Recognition and Normalization Task (LDC 2005), with all the adaptations needed to limit the task to the recognition of Named Entities ([Magnini et al. 2011], available for download from the NER task webpage at http://www.evalita.it/2011/tasks/NER).

The main novelty introduced for the 2011 edition is the fact that the task is based on broadcast news (graciously provided by the local broadcaster RTTR) and consists of two separate subtasks:

- **Full task**: participants will 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 will provide participants with the automatically created transcript of the news produced using a state-of-the-art ASR system and participants will perform Named Entity Recognition on this transcript.

Participants can chose to participate in either one or both subtasks.

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

2. “NER only” Subtask

The NER only subtask will follow the same guidelines as the Named Entity Recognition task at EVALITA 2007 and EVALITA 2009. The data to be annotated, however, will not consist of written newspaper articles, but of automatic transcriptions of broadcast news. As a consequence of this:

- the data contain transcription errors (both in terms of word recognition and segmentation and in terms of word capitalization);
- the data contain no sentence boundaries and no punctuation.

2.1 Training and Test Data (Broadcast News)

Training data consist of spoken news (about five hours of transmission, for a total of about 40,000 words) recorded, transcribed and annotated with Named Entities, and are provided for free upon acceptance of a license agreement.

More specifically, the following data will be provided:

- news manually transcribed and annotated with Named Entities (one text file)
- automatic transcription of that same news (one text file)
- recording of that same news (one audio file for each news program)

Training data also include I-CAB, i.e. the corpus of (written) news stories annotated with Named Entities used for the Named Entity Recognition tasks at EVALITA 2007 and EVALITA 2009, which is provided upon acceptance of a separate license agreement.

Test data consist of spoken news (about five hours of transmission, for a total of about 40,000 words) recorded and automatically transcribed.

In particular, we will provide:

- recordings of the news (ten audio files, one for each news program)
- automatic transcription of that same news (one text file)

Please notice that the ASR system used for the automatic transcription has not been specifically trained on this data.

2.2 Data Format

Development data and test data consist of two separate text files, with one token per line and an empty line between different news programs (each news program corresponds to one audio file).

Each written file consists of a number of columns separated by a blank, containing respectively:

- the token (first column);
- the RTTR news program to which the token belongs (second column).

Training data are also annotated with Named Entities in the IOB2 format: the Entity tag is contained in the third column and consists of two parts:

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

Example of training data format:

```
il 06041609-rttr-16k O
capitano 06041609-rttr-16k O
della 06041609-rttr-16k O
Gerolsteiner 06041609-rttr-16k B-ORG
Davide 06041609-rttr-16k B-PER
Rebellin 06041609-rttr-16k I-PER
```

To sum up, systems participating in the ‘NER only’ subtask will have to take as input a two-column file (see Appendix A) and produce as output a three-column file annotated with Named Entities in the IOB2 format (see Appendix B).

2.3 Final Ranking

The final ranking will be based on the F-measure score obtained in Named Entity Recognition. As in the previous editions, results will be compared with a baseline rate computed by identifying in the test data only the NEs that appear in the training data. In addition, an upper bound for the specific task of NER on automatic transcriptions will be computed in terms of Precision, Recall, F-Measure and Accuracy by shifting the Entity tags of the gold standard onto the aligned automatic transcriptions (in this way, the only NER errors will be those produced as consequences of ASR errors).

After submitting systems results, participants will be provided with the manual transcription of the test data and will be asked to run the same exact participant NER systems to these data; these results will be used to analyze the impact of transcription errors and NOT to rank the systems.

3. “Full task” Subtask

3.1 Training and Test Data (Broadcast News)

Training data consist of spoken news (about five hours of transmission, for a total of about 40,000 words) recorded, transcribed and annotated with Named Entities, and are provided for free upon acceptance of a license agreement.

More specifically, the following data will be provided:

- news manually transcribed and annotated with Named Entities (one text file)
- automatic transcription of that same news (one text file)
- recordings of that same news (one audio file for each news program)

Training data also include I-CAB, i.e. the corpus of (written) news stories annotated with Named Entities used for the Named Entity Recognition tasks at EVALITA 2007 and EVALITA 2009, which is provided upon acceptance of a separate license agreement.

Test data consist of spoken news (about five hours of transmission, for a total of about 40,000 words).

In particular, we will provide:

- recording of the news (ten audio files, one for each news program)

3.2 Data Format

Audio data consist of wav audio files (one for each news program).

Written data (provided as training data) consist of a single text file, with one token per line and an empty line between different news programs (each news program corresponds to one audio file).

Each file consists of three columns separated by a blank, containing respectively:

- the token (first column);
- the RTTR news program to which the token belongs (second column)
- the Entity tag with which Named Entities are annotated in the IOB2 format (third column); the Entity tag consists of two parts:
  - the IOB2 tag: “B” (for “begin”) denotes the first token of a Named Entity, I (for “inside”) is used for all other tokens in a Named Entity, and O (for “outside”) is used for all other words;
  - the Entity type tag: PER (for Person), ORG (for Organization), GPE (for Geo-Political Entity), or LOC (for Location).

Example of data format:

```
il 06041609-rttr-16k O
capitano 06041609-rttr-16k O
della 06041609-rttr-16k O
Gerolsteiner 06041609-rttr-16k B-ORG
Davide 06041609-rttr-16k B-PER
Rebellin 06041609-rttr-16k I-PER
```

To sum up, systems participating to full task will have to take as input an audio file (.wav) and produce as output a three-column file (one token per line) annotated with Named Entities in the IOB2 format (see Appendix B).

3.3 Final Ranking

The final ranking will be based on the F-measure score obtained in Named Entity Recognition, regardless of the performance of the ASR system. The performance of ASR systems used by participants in the full task will be computed, but it will not be considered for the official ranking.

After submitting systems results, participants will be provided with the manual transcription of the test data and will be asked to run on them the same exact NER systems; these results will be used to analyze the impact of transcription errors and NOT to rank the systems.

4. Evaluation procedure

The evaluation procedure, comparing systems' results against a gold standard, consists of three phases: transcription alignment, NER error detection, and score computation.

4.1 Transcription alignment

The first phase consists of comparing the reference transcription (gold standard) with the system's transcription: both are converted to lower-case and then aligned. In particular, the best possible alignment at the word level is determined in order to minimize the number of edit operations needed to transform the reference transcription into the hypothesis transcription, where the allowed edit operations are:

- word Insertion (I);
- word Deletion (D);
- word Substitution (S).

As a result, we obtain the system transcription aligned to the reference transcription, both of which contain NE annotations. The next phase is to compare the reference NEs annotated in the gold standard to the hypothesis NEs annotated in the system's transcription.

4.2 Detection of NER errors

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 that 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 Ex. 1[1] some examples of incorrect hypothesis NEs are provided. We count it as an error when the extension (see Hyp. 1 and 2) or the NE type (see Hyp. 3) of the hypothesis NE is not correct. Hyp. 4 gives an example of a FP ("presso" is annotated as a NE), while Hyp. 5 gives an example of a FN ("FIAT SPA" is not annotated as a NE).

Ex. 1. Examples of incorrect NE matches in the case of transcription alignments with no recognition errors

Ref:	lavorare	presso		la	<ORG>	FIAT	SPA</ORG>
Hyp 1:	lavorare	presso		la	<ORG>	FIAT</ORG>	SPA
Hyp 2:	lavorare	presso	<ORG>	la		FIAT	SPA</ORG>
Hyp 3:	lavorare	presso		la	<PER>	FIAT	SPA</PER>
Hyp 4:	lavorare<LOC>	presso</LOC>		la	<ORG>	FIAT	SPA</ORG>
Hyp 5:	lavorare	presso		la		FIAT	SPA

[1] In Examples 1-3 we use <type> and </type> tags to indicate the beginning and the end of a NE.

In the following, we discuss NE evaluation in the presence of recognition errors.

Insertion/Deletion. If one or more words of a hypothesis NE correspond to an Insertion there is no possibility of a correct Entity match, because Condition 2 above can not be satisfied (the extension of the hypothesis will necessarily be incorrect). Similarly, if one or more words of a reference NE correspond to a Deletion there is no possibility of correct Entity match.

Examples of incorrect matches are presented in Ex. 2, Hyp. 1-4: in some cases all the words contained in a reference NE correspond to a Deletion (see Ex. 2, Hyp. 1 and 2, where we have FNs), in other cases only one of the words correspond to a Deletion (see Ex. 2, Hyp. 3), and in still other cases all the words contained in a hypothesis NE correspond to an Insertion (see Ex. 2, Hyp. 4, where we have a FP).

Insertions or Deletions corresponding to words which are not part of any NE (see Ex. 2, Hyp. 5 and 6) do not influence the evaluation of NEs.

Ex. 2. Examples of transcription alignments with word Deletions and Insertions

Ref:		<ORG> serie	A</ORG>	o	<ORG> B	</ORG>	di	calcio
Hyp 1:		<ORG> serie	A</ORG>	o			di	calcio
Hyp 2:				o	<ORG> B	</ORG>	di	calcio
Hyp 3:		<ORG> serie	</ORG>	o	<ORG> B	</ORG>	di	calcio
Hyp 4:	<ORG> E </ORG>	<ORG> serie	A</ORG>	o	<ORG> B	</ORG>	di	calcio
Hyp 5:		<ORG> serie	A</ORG>	o	<ORG> B	</ORG>	di i	calcio
Hyp 6:		<ORG> serie	A</ORG>	o	<ORG> B	</ORG>		calcio

Substitution. If one or more words of a hypothesis NE correspond to a Substitution there can be a correct NE match anyway, if all three conditions above are satisfied.

Ex. 3. Examples of transcription alignments with word Substitutions

Ref:	lavorare		presso	la	<ORG>	FIAT	SPA</ORG>
Hyp 1:	lavorare		presso	la	<ORG>	FIA	SPA</ORG>
Hyp 2:	lavora		presso	la	<ORG>	FIAT	SPA</ORG>
Hyp 3:	lavorare		presso	la		FIA	SPA
Hyp 4:	lavorare	<ORG>	ESSO <ORG>	la	<ORG>	FIAT	SPA</ORG>

In Ex. 3, Hyp. 1, for example, the hypothesis NE "<ORG>FIA SPA</ORG>" (where the word "FIAT" is erroneously transcribed as "FIA"), meets all three conditions and is therefore evaluated as a correct NE match for the purposes of the official scoring [Galibert et al. 2011].

NEs however can also be compared along another component besides type, i.e. content, where the focus is on the correctness of NE content transcription [Burger et al. 1998]. Even if the final ranking is based on the performance obtained on the basis of type (see the above mentioned conditions according to which the hypothesis NE in Ex. 3, Hyp. 1 is correct), the organizers will also compute, and provide the participants with, the scores achieved by considering the more restrictive condition type+content (according to which Ex. 3, Hyp. 1 would be considered incorrect).

Substitutions corresponding to words which are not part of any NE (see Ex. 3, Hyp. 2, where the word "lavorare" is transcribed as "lavora") do not influence the evaluation of

NEs. On the other hand, if we have a word Substitution in non-corresponding NEs, we have a NER error by definition (see for example Ex. 3, Hyp. 3 where we have a FN and Ex. 3, Hyp. 4, where we have a FP).

4.3 Score computation

In the third phase, a final score is computed using the following measures: Precision, Recall, F-Measure and Accuracy.

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).

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).

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

Table 1 reports three sub-sections of the example presented in Appendix A. The first column contains the tokens, the second and the third contain respectively the reference tags (from Appendix B) and tags assigned by a hypothetical system.

Token	**reference tag**	**system tag**
Il	O	O
capitano	O	O
della	O	O
Gerolsteiner	B-ORG	B-ORG
Davide	B-PER	B-PER
Rebellin	I-PER	O
ha	O	O
allungato	O	O
frazionando	O	O
il	O	O
gruppo	O	B-PER
sul	O	O
traguardo	O	O
di	O	O
Bourges	B-GPE	B-LOC

Table1. Example of system results aligned with gold standard annotations

From Table 1 we have that:

- the number of Named Entities that the system has identified correctly (True Positive) is 1, as the only correctly identified Named Entity is “Gerolsteiner” (ORG)[2];

[2] “Davide” (PER) is not correct because the correct extension was “Davide Rebellin” (PER), “gruppo”

- the total number of Named Entities that the system was expected to recognize (i.e. True Positive + False Negative) is 3: “Gerolsteiner” (ORG), “Davide Rebellin” (PER), and “Bourges” (GPE);
- the number of Named Entities that the system has identified (True Positive + False Positive), is 4: “Gerolsteiner” (ORG), “Davide” (PER), “gruppo” (PER), and “Bourges” (LOC).

As a result, the system obtains a Precision of 25% (given by 1/4) and a Recall of 33.33% (given by 1/3).

F-Measure, the weighted harmonic mean of precision and recall (see 3), will be used for the official ranking of systems.

$$(3)\ F = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

In the example above, the system would obtain a value of F-Measure equal to 28.57%.

Accuracy indicates the percentage of correct predictions and is computed with respect to tokens as shown in (4).

$$(4)\ Acc. = \frac{CorrectTokens}{TotalTokens}$$

In the example presented in Appendix A and B (see also Table 1), we have a total number of 43 tokens, 3 of which have wrong tags:

1. `Rebellin` (“O” instead of “I-PER”);
2. `gruppo` (“B-PER” instead of “O”);
3. `Bourges` (“B-LOC” instead of “B-GPE”).

The Accuracy obtained by the system would be equal to 93% (given by 40/43).

5. Scorer

The above described evaluation will be performed on the output of the participating systems against a gold standard created as follows:

- The audio data have been transcribed automatically;
- The automatic transcription has been double-checked by humans (all audio data have been listened, and the corresponding transcription checked, at least twice, each time by a different person);
- The correct transcription has been manually annotated with NEs.

The evaluation described above is performed automatically by means of two scripts that are made available to participants from the NER Task webpage at http://www.evalita.it/2011/tasks/NER: the alignment script and the CoNLL Scorer[3].

5.1 Alignment script

The alignment script takes as input the gold standard and the systems’ transcription (both with NE annotation), determines the best possible alignment between reference and

(PER) is not correct because it is not a NE, and “Bourges” (LOC) is not correct because its type is GPE.

[3] The CoNLL scorer is also available from the CONLL website: http://www.cnts.ua.ac.be/conll2002/ner.tgz

hypothesis word transcriptions and produces as output a file with five columns (see Tables 2 and 3) in which the different columns contain respectively:

- Gold standard token
- Gold standard Entity tag
- Type of transcription error (empty in case of correct transcription)
- System output token
- System output Entity tag

GOLD			SYSTEM	
lavorare	O		lavorare	O
presso	O		presso	O
la	O		la	O
FIAT	B-ORG	**S**	FIA	O
SPA	I-ORG		SPA	O

Table 2. Alignment example (see Ex. 3.3)

GOLD			SYSTEM	
Lavorare	O		lavorare	O
Presso	O	**S**	ESSO	B-ORG
La	O		la	O
FIAT	B-ORG		FIAT	B-ORG
SPA	I-ORG		SPA	I-ORG

Table 3. Alignment example (see Ex. 3.4)

In the case of token Insertion or Deletion, the script adds an "O" Entity tag. More precisely, the "O" Entity tag is added in the system output column in the case of a Deletion (see Table 4), whereas it is added in the gold standard in the case of an Insertion (see Table 5).

GOLD			SYSTEM	
serie	B-ORG		serie	B-ORG
A	I-ORG		A	I-ORG
o	O		o	O
B	B-ORG	**D**		O
di	O		di	O
calcio	O		calcio	O

Table 4. Alignment example (see Ex. 2.1)

GOLD			SYSTEM	
	O	**I**	E	B-ORG
Serie	B-ORG		serie	B-ORG
A	I-ORG		A	I-ORG
O	O		O	O
B	B-ORG		B	B-ORG
Di	O		di	O

Table 5. Alignment example (see Ex. 2.4)

5.2 CoNLL Scorer

For the official evaluation of system results, we will use the scorer made available by CoNLL for the 2002 Shared Task (http://www.cnts.ua.ac.be/conll2002/).

It takes as input the gold standard and the systems' output aligned in the previous phase (in particular the second and the fifth columns containing the Entity tags) and compares the two sets of NE tags. With respect to Table 2 the CoNLL scorer will count 1 FN, wrt Table 3 it will count 1 FP, wrt Table 4 it will count 1 FN and wrt to Table 4 it will count 1 FP.

The CoNLL scorer reports the final scoring in terms of accuracy, precision, recall and F-Measure and provides the following counts: processed tokens with the total number of NEs in the gold standard (i.e. TP+FN), total number of NEs recognized by the system ("found phrases", i.e. TP+FP) and number of correct NEs ("correct", i.e. TP).

6. Data Distribution

Training data are available for research purposes upon acceptance of a license agreement:

- If you work for a non-profit research organization, you can obtain an unlimited Research License (http://hlt.fbk.eu/en/download_ner2011dataset);
- Otherwise, you can obtain a License limited to the EVALITA 2011 evaluation (http://hlt.fbk.eu/en/download_ner2011dataset_limited).

I-CAB (the Italian Content Annotation Bank) is also available as part of the training data to all NER task participants:

- If you work for a non-profit research organization, you can obtain an unlimited Research License (http://ontotext.fbk.eu/i-cab/download-icab.html);
- Otherwise, you can obtain a License limited to the EVALITA 2011 evaluation (http://hlt.fbk.eu/en/download_icab_limited).

7. Submission of system results

- Deadline: **October 19th**, 2011, midnight (GMT + 1 hour) **- NEW**
- System results must be sent by e-mail to Manuela Speranza (manspera@fbk.eu)
- For each run, system results will consist of a single data file and will have to be named as follows: evalita11_NER_participant_run

Each participant can submit a maximum of two runs for each subtask (for a maximum total of four runs).
The first run will be produced according to the 'closed' modality: only the data we distribute and no additional resources are allowed for training and tuning the system. For external resources we intend resources used to acquire knowledge such as gazetteers, NE dictionaries, ontologies or Wikipedia. The only source of knowledge you can use is the material that the organisers provide (i.e. I-CAB plus the RTTR data). For what concerns tools, complex NLP toolkits (e.g. TextPro, GATE, OpenNLP) are forbidden for the closed run because they use NE dictionaries, while simple POS taggers or tools for lemmatization are allowed. The second run will be produced according to the 'open' modality: any type of data can be used , provided it is described in the final report. The 'closed' run is compulsory, while the 'open' run is optional.
Systems with embedded resources for which it is impossible to produce a run according to the closed modality are allowed to submit only the 'open' run provided that participants contact us in advance explaining the motivation for their request.

8. Contact Person

Manuela Speranza (manspera@fbk.eu)

Appendix A

Example of input format for the NER only subtask

```
a 06041609-rttr-16k
circa 06041609-rttr-16k
novanta 06041609-rttr-16k
chilometri 06041609-rttr-16k
dall' 06041609-rttr-16k
arrivo 06041609-rttr-16k
il 06041609-rttr-16k
capitano 06041609-rttr-16k
della 06041609-rttr-16k
Gerolsteiner 06041609-rttr-16k
Davide 06041609-rttr-16k
Rebellin 06041609-rttr-16k
ha 06041609-rttr-16k
allungato 06041609-rttr-16k
su 06041609-rttr-16k
uno 06041609-rttr-16k
dei 06041609-rttr-16k
pochi 06041609-rttr-16k
tratti 06041609-rttr-16k
in 06041609-rttr-16k
salita 06041609-rttr-16k
frazionando 06041609-rttr-16k
il 06041609-rttr-16k
gruppo 06041609-rttr-16k
alla 06041609-rttr-16k
sua 06041609-rttr-16k
ruota 06041609-rttr-16k
si 06041609-rttr-16k
sono 06041609-rttr-16k
portati 06041609-rttr-16k
altri 06041609-rttr-16k
sei 06041609-rttr-16k
corridori 06041609-rttr-16k
che 06041609-rttr-16k
hanno 06041609-rttr-16k
poi 06041609-rttr-16k
disputato 06041609-rttr-16k
lo 06041609-rttr-16k
sprint 06041609-rttr-16k
sul 06041609-rttr-16k
traguardo 06041609-rttr-16k
di 06041609-rttr-16k
Bourges 06041609-rttr-16k
```

Appendix B

Example of output format:

```
a 06041609-rttr-16k O
circa 06041609-rttr-16k O
novanta 06041609-rttr-16k O
chilometri 06041609-rttr-16k O
dall' 06041609-rttr-16k O
arrivo 06041609-rttr-16k O
il 06041609-rttr-16k O
capitano 06041609-rttr-16k O
della 06041609-rttr-16k O
Gerolsteiner 06041609-rttr-16k B-ORG
Davide 06041609-rttr-16k B-PER
Rebellin 06041609-rttr-16k I-PER
ha 06041609-rttr-16k O
allungato 06041609-rttr-16k O
su 06041609-rttr-16k O
uno 06041609-rttr-16k O
dei 06041609-rttr-16k O
pochi 06041609-rttr-16k O
tratti 06041609-rttr-16k O
in 06041609-rttr-16k O
salita 06041609-rttr-16k O
frazionando 06041609-rttr-16k O
il 06041609-rttr-16k O
gruppo 06041609-rttr-16k O
alla 06041609-rttr-16k O
sua 06041609-rttr-16k O
ruota 06041609-rttr-16k O
si 06041609-rttr-16k O
sono 06041609-rttr-16k O
portati 06041609-rttr-16k O
altri 06041609-rttr-16k O
sei 06041609-rttr-16k O
corridori 06041609-rttr-16k O
che 06041609-rttr-16k O
hanno 06041609-rttr-16k O
poi 06041609-rttr-16k O
disputato 06041609-rttr-16k O
lo 06041609-rttr-16k O
sprint 06041609-rttr-16k O
sul 06041609-rttr-16k O
traguardo 06041609-rttr-16k O
di 06041609-rttr-16k O
Bourges 06041609-rttr-16k B-GPE
```

Appendix C

The following guidelines were followed in producing the manual transcription of the news:

- Hesitations (e.g. “que- que- questo”) and autocorrections (“cre- sono convinto”) are not reflected in the transcription (the transcription for the previous examples would simply be “questo” and “sono convinto” respectively)
- Repetition of whole words is reflected in the translation (e.g. “penso che che che sia utile”)
- No punctuation is transcribed
- Proper nouns are capitalized (other words may also be capitalized, according to common usage)
- Numbers are transcribed as one word (e.g. “trentamila”) with the exception of numbers containing “milioni” and “miliardi”, which are written as separate words (e.g. “due milioni” “tre miliardi”); conjunctions within one-word numbers are not transcribed (e.g. “2009” is transcribed as “duemilanove” even if the speaker adds the conjunction “e” between “duemila” and “nove”)
- Acronyms are capitalized without any punctuation between the letters (e.g. SPA); all acronyms are spelled as such, independently on how they are pronounced (e.g. “SPA” is spelled “SPA” if is pronounced [spa] but also if it is pronounced [essepia])
- Ellipsis is not reflected in the transcription of numbers (e.g. “vent’anni” is transcribed as “venti anni” even if the speaker elides the word “venti”) and verbs (e.g. “andare” and “facciamo” even if the speaker elides the last vowel)
- Pronounciation errors are not reflected in the transcription

Appendix D

In Table 6 we present a NE match where the hypothesis NE contains both token Deletion and token Substitution. This NE match is incorrect because Condition 2[4] is not satisfied (the extension of the hypothesis NE is incorrect); in fact, there is no one-to-one mapping between the tokens it contains and the tokens in the reference NE (the token "FIAT" in the reference NE is not mapped to any token in the hypothesis NE). With reference to scoring computation this will increase the number of FPs by one and will also increase the number of FNs by one.

GOLD			**SYSTEM**	
lavorare	O		lavorare	O
presso	O		presso	O
la	O		la	O
FIAT	B-ORG	**D**		O
SPA	I-ORG	**S**	FIASPA	B-ORG

Table 6. Example of transcription alignment (Substitution and Deletion in the same NE)

Similarly, the NE match presented in Table 7 is incorrect because Condition 2 is not satisfied.

GOLD			**SYSTEM**	
lavorare	O		lavorare	O
presso	O		presso	O
la	O		la	O
	O	**I**	FI	B-ORG
FIAT	B-ORG	**S**	AT	I-ORG
SPA	I-ORG		SPA	I-ORG

Table 7. Example of transcription alignment (Substitution and Insertion in the same NE)

[4] The conditions for NE match correctness are described on p. 6.

References

Burger, J. D., Palmer, D., and Hirschman, L. Named Entity Scoring for Speech Input. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98) and the 17th international conference on Computational linguistics (COLING ’98), August 10-14, 1998, Université de Montréal, Montreal, Quebec, Canada.

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., and Quintard, L. Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. To appear in Proceedings of IJCNLP, Chiang Mai (Thaïlande), 8-13 November 2011.

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

Magnini, B., Pianta, E., Speranza, M., Bartalesi Lenzi, V., and Sprugnoli, R. Italian Content Annotation Bank (I-CAB): Named Entities, Technical report, FBK, 2011. http://www.evalita.it/sites/evalita.fbk.eu/files/doc2011/I-CAB-Report-Named-Entities.pdf

Web Sites

ACE, http://www.nist.gov/speech/tests/ace/index.htm
http://www.ldc.upenn.edu/Projects/ACE/

CoNLL 2002, http://www.cnts.ua.ac.be/conll2002/ner/

ESTER http://www.afcp-parole.org/ester/

L’Adige, http://www.ladige.it/

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

RTTR, http://www.rttr.it/