EVALITA 2007: DESCRIZIONE E RISULTATI DEL TASK TERN

EVALITA 2007: DESCRIPTION AND RESULTS OF THE TERN TASK

VALENTINA BARTALESI LENZI · RACHELE SPRUGNOLI

SOMMARIO/ *ABSTRACT*

In questo articolo si descrivono le motivazioni e le caratteristiche del task TERN (*Temporal Expression Recognition and Normalization*) di EVALITA 2007. Vengono, inoltre, presentati i dati distribuiti per l'addestramento e la valutazione dei sistemi, le metriche di valutazione utilizzate e i risultati ottenuti dai partecipanti.

In this paper, we describe motivations and features of the TERN (Temporal Expression Recognition and Normalization) task at EVALITA 2007. We also present the training and test data used in this task, evaluation measures and participants'results.

Keywords: Temporal Expression Recognition and Normalization, information extraction, EVALITA 2007.

1. Motivation

EVALITA intends to promote the development of language technologies for Italian, providing a shared framework where different systems and approaches can be evaluated in a consistent manner.

In this scenario, the Temporal Expression Recognition and Normalization task (TERN) encourages research on systems capable of automatically detect and normalize Temporal Expressions (TEs) present in Italian texts.

Our work refers to the Automatic Content Extraction (ACE) program that in 2004 adopted the TERN task with respect to the "TIDES 2005 Standard for the Annotation of Temporal Expressions" [2].

2. Task definition

The task is based on the TIMEX2 standard, with some adaptations to Italian [1]. In the Temporal Expression Recognition and Normalization task, systems are required to recognize the Temporal Expressions occurring in the source data by identifying their extension, and to normalize them, i.e. give a representation of their meaning by assigning values to a pre-defined set of attributes. The TIMEX2 attributes evaluated in the Normalization phase are listed in Table 1.

Table 1: TIMEX2 attributes

Attribute	Description
VAL	Normalized value of a TE
ANCHOR_VAL	Normalized anchoring date/time
ANCHOR_ DIR	Normalized time directionality
MOD	A temporal modifier
SET	Identifies that VAL is a set of time

Temporal Expressions to be marked include both absolute expressions ("17 luglio 2007", "12:00", "l'estate del 1980") and relative expressions ("ieri", "la prossima settimana"). Also durations ("un'ora", "due settimane"), event-anchored expressions ("due giorni prima della partenza"), and sets of times ("ogni settimana") are to be annotated. Temporal Expressions whose interpretation requires cultural or domain-specific knowledge ("anno accademico") and underspecified TEs ("per lungo tempo") are markable but receive no VAL.

The TERN task at EVALITA 2007 consisted of two subtasks:

- Temporal Expression **Recognition only**
- Temporal Expression **Recognition + Normalization**

Participants chose to participate in any of the two subtasks.

3. Dataset

Both training data and test data are part of the Italian Content Annotation Bank (I-CAB), developed by FBK-irst and CELCT, and they were distributed upon acceptance of the agreement terms for a free research licence.

I-CAB consists of 525 news documents taken from the local newspaper "L'Adige". The selected news stories 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. I-CAB is divided into a development part (335 news stories, for a total of around 113,000 words) and a test part (190 news stories, for a total of around 69,000 words).

The total number of annotated Temporal Expressions is 4,603: 2,931 and 1,672 in the training and test sections respectively.

Training data were distributed in the following formats:

- SGML files containing the source text. All SGML files are in UTF-8.
- APF (ACE Program Format) files containing the annotation in the form of XML standoff annotation, which means that the file as a whole conforms to XML encoding standards, and the raw data being annotated reside in a separate file. The annotations "point" to portions of the raw text via indices.

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

4. Evaluation

For the official evaluation we used the TERN scoring part of *ace07-eval-v01-EVALITA.pl* scorer[1], an adapted version of *ace07-eval-v01.pl* described in [3]. We introduced some modifications concerning the attribute weights for the **Recognition + Normalization** subtask as illustrated in section 4.2.

The final ranking is based on the TERN value score which is defined to be the sum of the values of all of the system's output TIMEX2 tokens, normalized by the sum of the values of all of the reference TIMEX2 tokens [3].

$$TERN\_Value_{sys} = \frac{\sum_i value\_of\_sys\_token_i}{\sum_j value\_of\_sys\_token_j}$$

The maximum possible timex2 value score is 100 percent.

We also provided the following measures:

- Precision: indicates the percentage of correct positive predictions and it is computed as the ratio between the number of Temporal Expressions correctly identified by the system (True Positive) and the total number of Temporal Expressions 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 Temporal Expressions correctly identified by the system (True Positive) and the number of Temporal Expressions that the system was expected to recognize (True Positive plus False Negative).
- F-measure: the weighted harmonic mean of precision and recall.

[1] http://evalita.itc.it/tasks/ace07-eval-v01-EVALITA.zip

Figure 1 shows an example of the evaluation output. Note that we have modified the original Perl program: the abbreviations *Rec* (Recognition) and *Norm* (Normalization) are visualized instead of *Detection* and *Rec* (Recognition) displayed in the original output.

```
Timex2 Recognition and Normalization statistics:

 ref        ______Count______    ___________Cost_(%)______________
 source   Ent   Rec      Norm     Rec     Norm  Value   Value-based
 type     Tot  FA  Miss  Err     FA  Miss  Err   (%)    Pre--Rec--F
unknown  1672  91  171   695    4.4  7.6  26.1  61.9   68.5 66.3 67.4
  total  1672  91  171   695    4.4  7.6  26.1  61.9   68.5 66.3 67.4
```

Figure 1: The evaluation output

4.1 Recognition only subtask

As far as the **Recognition only** subtask is concerned, a minimal overlap in the extent of the reference and the system output tags is required. Overlaps are measured in terms of number of characters for text input and the minimum acceptable overlap of matching is 0.300. The cost (weight) for spurious Temporal Expressions is 0.750 and the cost for missed Temporal Expressions is 1.000.

4.2 Recognition + Normalization subtask

Parameters for scoring the Recognition in this subtask are the same as above:

- 0.300, minimum acceptable overlap
- 0.750, cost for spurious TEs
- 1.000, cost for missed TEs

As far as Normalization is concerned, attribute value assignment measures the ability of the system to correctly assign the normalization attribute values of the Temporal Expressions.

The weights assigned to each TIMEX2 attribute are given in Table 2. Notice that we have modified the default weight of the ANCHOR_DIR attribute (we have assigned to ANCHOR_DIR the same value as ANCHOR_VAL). This change seemed reasonable as these two anchoring attributes are always used together.

Table 2: Default parameters for scoring TERN attributes

Attribute	Attribute Weight
VAL	1.000
ANCHOR_DIR	0.500
ANCHOR_VAL	0.500
MOD	0.100
SET	0.100

5. Results

Four teams participated in the challenge: three to the **Recognition + Normalization** subtask and one to the **Recognition only** subtask.

Table 3 presents the results for the **Recognition only** subtask in term of TERN-Value score, Precision (P), Recall (R) and F-measure (F).

Table 3: Results for the **Recognition only** subtask, percentages for Value, Precision, Recall and F-measure

Team	Value	P	R	F
FBKirst_Negri_TIME	85.7	95.7	89.8	92.6
UniPg_Faina_TIME	50.1	77.7	70.3	73.8
UniAli_Puchol_TIME	48.8	78.4	67.4	72.5
UniAli_Saquete_TIME	41.9	82.5	53.2	64.7

Table 4 lists, for each submitted run, the results obtained for the **Recognition + Normalization** subtask in term of TERN-Value score, Precision (P), Recall (R) and F-measure (F).

Table 4: Results for the **Recognition + Normalization** subtask, percentages for Value, Precision, Recall and F-measure

Team	Value	P	R	F
FBKirst_Negri_TIME	61.9	68.5	63.3	67.4
UniAli_Saquete_TIME	22.1	51.5	35.6	42.1
UniPg_Faina_TIME	11.9	24.9	19.6	21.9

The Value score achieved by the participating systems ranges from 41.9% to 85.7% in the **Recognition only** subtask, while, for the **Recognition + Normalization** subtask, the systems obtained from 11.9% to 61.9%. The submissions of FBKirst_Negri_TIME stand out as more than 35% higher than the other systems in both the task.

6. Conclusion

We received the expected attention in terms of participation considering this was a new and relatively difficult task for the Italian language. Actually, eight groups registered and were interested into participating, but four of them could not adjust their system on time.

We can be pleased of the outcome of the TERN task at EVALITA 2007, and we hope that the resources we developed and the results we obtained will encourage other teams to participate in future evaluation exercises.

Acknowledgments

We would like to thank all the people from FBK-irst who collaborated with us in processing and annotating the data, e.g. Lorenza Romano and Manuela Speranza.

REFERENCES

[1] B. Magnini, M. Negri, E. Pianta, M. Speranza, V. Bartalesi Lenzi and R. Sprugnoli. Italian Content Annotation Bank (I-CAB): Temporal Expressions. Technical report, FBK-irst, 2007. On-line: http://evalita.itc.it/tasks/timex.html.

[2] L. Ferro, L. Gerber, I. Mani, B. Sundheim and G. Wilson. TIDES 2005 Standard for the Annotation of Temporal Expressions. September 2005. On- line: http://timex2.mitre.org/annotation_guidelines/2005_timex2_standard_v1.1.pdf.

[3] The ACE 2007 (ACE07) Evaluation Plan. On-line: http://www.nist.gov/speech/tests/ace/ace07/doc/ace07-evalplan.v1.3a.pdf.

VALENTINA BARTALESI LENZI, RACHELE SPRUGNOLI
CELCT, Center for the Evaluation of Language and Communication Technologies
Via Sommarive, 18
38050 Povo (TN) - Italy
Email: {bartalesi | sprugnoli}@celct.it

VALENTINA BARTALESI LENZI graduated in Humanistic Information Technology at the University of Pisa in 2004, with a thesis on usability evaluation for museum web sites. She worked on the development of web access to cultural heritage and historical archives and she collaborated as web mistress with the University of Pisa. Since September 2005, she has been a project consultant at CELCT working on semantic annotation.

RACHELE SPRUGNOLI graduated in Humanistic Information Technology at the University of Pisa in 2004, with a thesis about the digitization of Renaissance Italian text. She did an internship at Philips in Eindhoven (NL) for the development of an Italian TTS system. Since June 2005, she has been a project consultant at CELCT working on semantic annotation.