TempEval 2
Spanish Data Release

Roser Saurí
Barcelona Media – Innovation Center
Barcelona, Catalonia

March 7, 2010

1 Corpus description

The data released for the TempEval-2 Spanish edition is a fragment of the Spanish TimeBank tailored to some of the tasks designed for this competition. Overall, it contains 210 news articles annotated with the TimeML specification language (Pustejovsky et al., 2005).

1.1 Annotated data

The data is marked up with the following information:

Time expressions. Time expressions (or timexes) are annotated with the TimeML tag `TIMEX3`, which captures dates, times, durations, as well as sets of dates or times. Timexes may denote a precise value (e.g., *April 20, 2008*, *4 days ago*), or a fuzzy value (e.g., *in the past*, *few days ago*). Each `TIMEX3` element include a `type` and a `value` attribute:

- The `type` attribute classifies the time expression into one of the following: date, time of day, duration, and set.
- The `value` attribute represents the denotation of the expression in the extended ISO 8601 format assumed by TimeML.

Time expressions in Spanish text have been annotated according to the annotation guidelines created specifically for that language (Saurí et al., 2009a).

Events. The `EVENT` tag captures expressions denoting situations that hold or take place. In TimeML, it includes both dynamic events as well as stative ones. Furthermore, each event expression provides information regarding the following attributes:

- mood: Relevant for verbal elements. It encodes the grammatical mood of these expressions in Spanish, distinguishing among: indicative, subjunctive, conditional, imperative, and none (for expressions other than verbs).
- tense: Encoding the grammatical tense of expressions in Spanish. Distinguishing among: present, past, future, and none.
- aspect: Encoding the grammatical aspect of expressions. Distinguishing among: imperfective, perfective, imperfective-progressive, perfective-progressive, or none.
- polarity: Encoding whether the expression denotes a positive or a negative event.

For a detailed description of the event annotation, refer to Saurí et al. (2009), the guidelines created specifically for Spanish data.

Temporal relations. Temporal relations hold between an event and another event, a timex and another timex, or an event and a timex. In the case of Spanish data, the following relations are annotated:

- The set of temporal relations holding between an event and the document creation time (DCT).
- The set of temporal relations holding between a timex and an event if it is the case that they both belong to the same sentence and:
  - The event expression immediately dominates the time expression, or
  - The event and the timex belong to the same NP.

Similarly, the set of 13 temporal relations types distinguished in TimeML has been simplified to 6, in accordance with the general design of the TempEval-2 evaluation. In particular, the following relation types have been used:

- before: Entity A (an event or timex) is placed before entity B on the temporal axis. This relation type includes the TimeML types of before and immediately_before.
- after: Entity A (an event or timex) is placed after entity B on the temporal axis. This relation type includes the TimeML types of after and immediately_after.
- overlap: Either entity A and B are simultaneous, or one includes the other. Hence, this relation subsumes TimeML relations of: simultaneous, identity, measure, includes, and is_included.
- before_or_overlap: This value is chosen in two different cases:
  1. Entity A begins before entity B, but at some point both entities overlap. This relation mainly subsumes the TimeML relations of: begins and is_ended.

2. Underspecified relation: entity A may have begun before B, or may overlap B. What it is certain is that entity A does not continue after B is finished.

- `overlap_or_after`: As in the previous relation, this value is chosen in two different cases:

  1. Entity A overlaps with entity , but at some point in time entity B stops whereas entity A continues. This relation subsumes the TimeML relations of: `is_begun_by` and `ends`.
  2. Underspecified relation: entity A may have begun before B, or may overlap B. What it is certain is that entity A is not before B.

- `vague`: For completely underspecified, vague cases about which we are not able to select any of the relation types above.

For further details, refer to the annotation guidelines for marking up temporal relations in Spanish (Saurí, 2010).

1.2 Statistics

The current corpus contains 210 documents, with over 68,000 tokens (including punctuation marks). It has a total of 12,385 elements tagged as events and 2,776 expressions annotated as timexes. The whole corpus has been double-annotated, and cases of disagreement have been adjudicated by a third person.

1.3 Data sources

The documents integrating this corpus are extracted from the Spanish part of the AnCora corpus (Taulé et al., 2008), a remarkable resource in that it provides annotation for a number of linguistic levels, including constituent structure, syntactic functions, dependencies, verb semantic class, argument structure, and thematic roles. This information, however, is not included in the current release.

2 TempEval-2 specifics

2.1 Targeted tasks

The current corpus has been marked up to be used in the following TempEval-2 tasks:

A. Determining the extent of time expressions as defined by the TimeML `TIMEX3` tag, as well as the value of their features `type` and `value`.

B. Determining the extent of events as defined by the TimeML `EVENT` tag, as well as the value of the features: `mood`, `tense`, `aspect`, and `polarity`.

C. Determining the relation between an event and a time expression in the same sentence. The event must either immediately dominate the time expression, or the event and the time expression must occur in the same noun phrase.

D. Determining the relation between an event and the document creation time.

2.2 Release batches

Training data. Containing a total of 175 documents. Training data will be released in two batches:

- First batch: 140 documents, to be released by March 8th.
- Second batch: 35 additional documents, to be released by March 12th.

Testing data. Containing a total of 35 documents. Released to be announced.

3 Acknowledgments

The current data have been annotated by the following contributors (alphabetically sorted): Joan Banach, Judith Domingo, Pau Giménez, Jimena del Solar, and Teresa Suñol. The adjudication has been in charge of Roser Saurí and Jimena del Solar.

This work has been supported by a EU Marie Curie International Reintegration Grant (PIRG04-GA-2008-239414) to Roser Saurí.

References

Pustejovsky, J., Knippen, B., Littman, J., & Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation*, *39*(2), 123–164.

Saurí, R. (2010). *Annotating Temporal Relations in Spanish. TimeML Annotation Guidelines.* Barcelona Media – Innovation Center. Version TempEval-2010.

Saurí, R., Batiukova, O., & Pustejovsky, J. (2009). *Annotating Events in Spanish. TimeML Annotation Guidelines.* Barcelona Media – Innovation Center. Version TempEval-2010.

Saurí, R., Saquete, E., & Pustejovsky, J. (2009). *Annotating Time Expressions in Spanish. TimeML Annotation Guidelines.* Barcelona Media – Innovation Center. Version TempEval-2010.

Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the LREC 2008.*