The It-TimeML Experience for the Ita-TimeBank

Rachele Sprugnoli
sprugnoli@celct.it

Outline

- Motivations

- Peculiarities of TimeML

- It-TimeML

- Ita-TimeBank

- Conclusions and future works

Motivations

- Many systems require temporal analysis and event-based reasoning

- Question Answering
  - *When is Ramadan this year?*
  - *What is the largest U.S. military operation since Vietnam*

- Multi-Document Summarization

14 Oct 2000 07:27:13 -0400 (EDT)
GULU: **on 13 Oct 2000**, it was reported that at least 30 people in this northern Uganda town **have died in recent weeks** of a hemorrhagic **fever** that authorities fear may be caused by the Ebola or Marburg virus.
14 Oct 2000 23:25:01 -0400
So far 10 people have died in hospital, including 3 nurses treating the

- Information Extraction
  - **LiveMemories Project** (*Active Digital Memories of Collective Life*) http://www.livememories.org

TimeML

- Specification language for the annotation of:

  - **temporal expressions**

  - **events**

  - **temporal anchoring** of events
    (relations between events and temporal expressions)

  - **temporal ordering** of events
    (relations between events and other events)

Peculiarities of TimeML (1)

1) Extends the annotation of **temporal expressions**, i.e. durations, calendar dates, times of Day and sets of time, with respect to TIMEX and TIMEX2 → **<TIMEX3> tag**

    e.g. Inserts start/end points for durations: *from 1992 to 1995*

        Inserts freq/quant attributes to addresses problem with Granularity/Periodicity: *three days every month*

2) Identifies **signals** that are tokens which point the existence of relations → **<SIGNAL> tag**

    e.g. Temporal Prepositions: *for*, *during*, *on*, *at*

        Temporal Adverbs: *before*, *after*, *while*

        Subordinators: *if*

Peculiarities of TimeML (2)

3) Has a broad notion of **event**: *something that can be said to obtain or hold true, to happen or to occur, including all kinds of actions, states, and processes* → <**EVENT**> **tag**

   e.g. Tensed and Untensed verbs; *was captured*, *to thank*
   Adjectives: *he was sick*
   Nouns and nominalizations: *strike*, *Gulf War*
   Prepositional phrases: *the man was on board*

4) Creates various kind of dependencies between events and/or temporal expressions with **3 types of relations**:

   e.g. Temporal Link <**TLINK**>: *Last week on Monday...*
   John left on Monday
   John left after the explosion
   Aspectual Link <**ALINK**>: *She finished assembling the table*
   Subordinating Link <**SLINK**>: *She came to teach Italian*

Short history of TimeML

- A short history:
    - TimeML 1.1, April 2004
    - TempEval 1, 2007
    - TimeML 1.3, February 2010
    - TempEval 2, 2010

- **ISO-TimeML** at the *Inquire Stage,* July 2009
    - → **ISO/DIS 24617-1** Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events

It-TimeML and Ita-TimeBank

- Joint work by CELCT (Valentina Bartalesi Lenzi, Rachele Sprugnoli and Emanuele Pianta) and ILC-CNR (Tommaso Caselli)

- Similar adaptations of TimeML to natural languages, such as Spanish, Catalan, French, Korean and Chinese

- Study of the language specific issues related to the linguistic phenomena taken into account

It-TimeML vs. En-TimeML

- TimeML annotation principles preserved:

  - tag extents as small as possible
  - attention to the superficial form of the annotated token(s)

- Main differences:

  - Event attributes values
  - Event attribute list
  - Annotation of modal verbs
  - Annotation of verbal periphrases

It-TimeML: MOOD and VFORM

- MOOD: captures the mood of event realizations

  – NONE: indicative mood and non verbal events
  – CONDITIONAL: *Mangerei del pesce* [I would eat fish]
  – SUBJUNCTIVE: *Voglio che tu te ne vada* [I want you to go]
  – IMPERATIVE: *Taci!* [Shut down!]

- VFORM: captures the distinctions between finite and non-finite verbs

  – NONE: all finite verb forms and non verbal events
  – INFINITIVE: *Non è possibile viaggiare* [It's not possible to travel]
  – PARTICIPLE: *Vista Maria, se ne andò* [Having seen Maria, he left]
  – GERUND: *Ha evitato l'incidente andando piano* [Driving slowly, he avoided the incident]

It-TimeML: ASPECT

- The ASPECT attribute captures standard distinctions in the grammatical category of verbal aspect

English TimeML

- NONE
- PROGRESSIVE
- PERFECTIVE
- PERFECTIVE_PROGRESSIVE

It-TimeML

- NONE
- PROGRESSIVE
- PERFECTIVE
- IMPERFECTIVE

→ Presence of the IMPERFECTIVE value

PROGRESSIVE: *is teaching* / *lei sta insegnando*
IMPERFECTIVE: *lei insegna* / *lei insegnava*

→ Absence of the PERFECTIVE_PROGRESSIVE value

It–TimeML: modal verbs

- EN: modal verbs are not annotated as Events

  You must sleep → 1 Event
  You must <EVENT…modality="MUST">*sleep*</EVENT>

- IT: modal verbs are to be considered similar to other lexical verbs

  Tu devi dormire → 2 Events
  <EVENT…modality="DOVERE">*Devi*</EVENT>
  <EVENT…>*dormire*</EVENT>

It–TimeML: verbal periphrases

- 3 kinds od verbal periphrases (Bertinetto, 1991)

– habitual periphrases

Maria stava <EVENT … ASPECT="PROGRESSIVE">*mangiando* </EVENT>
[Maria was eating]

– modal periphrases

Gli esercizi <EVENT … MODALITY="ANDARE"> *vanno* </EVENT> <EVENT … > *svolti* </EVENT> *per domani*
[Exercises must be done for tomorrow]

– phasal periphrases

Hanno <EVENT … CLASS="ASPECTUAL"> *iniziato* </EVENT> *a* <EVENT> *lanciare* </EVENT> *pietre*
[They started to throw stones.]

Annotation Guidelines VS. Annotation Specifications

- Annotation Guidelines = abstract, formal level of description

- Annotation Specifications = actual realization of the guidelines

  - facilitate the reduplication of annotations
  - justify the annotated items
  - reduce the risk of disagreement

  → Decision trees for each annotation tasks
  → FrameNet to support annotators

Specifications: Example (1)

- Decision trees to classify **TLINK**

  – "EVENT + *tra (in)* + DURATION"

  Il pacco arriverà in due giorni [the package will arrive in two days]

  <TLINK lid="l1" eventInstanceID="e1" relatedToTime="t1" signalID="s1" relType="AFTER"/>

  – "EVENT + *per (for)* + DURATION"

  Sono stati sposati per 10 anni [they have been married for ten years]

  <TLINK lid="l1" eventInstanceID="e1" relatedToTime="t1" signalID="s1" relType="MEASURE"/>

Specifications: Example (2)

▪ FrameNet to identify **Reporting Events**: "describe the action of a person or an organization declaring something, narrating an event, informing about an event, etc. "

→ Communication frame and frames using and inherited from the Communication frame having the Message as a core element and conveying a giving information speech act

– *urlare* [to scream]: from the Communication_noise frame
– *sottolineare* [to stress]: from the Convey_importance frame
– *dichiarare* [to declare]: from the Statement frame

Ita-TimeBank

- CELCT corpus + ILC corpus

    → Both corpora are created on the basis of the It-TimeML guidelines and in a **collaborative way** for the development of the annotation specifications

    → The largest resource annotated with temporal and event information for Italian

CELCT corpus: description

- 525 news stories taken from **I-CAB**
  - about 180,000 tokens
  - cultural, sports, economic, local and international news
  - already annotated with temporal expressions (TIMEX2 standard), entities and mentions (ACE formalism)

- What we have done
  - semi-automatic conversion from TIMEX2 to TIMEX3 tags on 180k tokens
  - event extents and attributes annotation on 90,000 tokens
  - signal annotation on 90,000 tokens
  - LINKs annotation on 90,000 tokens

- 2 annotators for TIMEXs, signals and events and 3 for LINKs

- CELCT Annotation Tool

ILC corpus: description

- Developed as part of a Ph.D. project
  - 171 news stories: economic, international, current affairs, politics, newswire
  - about 68.000 tokens
  - comparable in content and size to the English TimeBank

- What we have done
  - event extent annotation
  - temporal expression and normalization annotation
  - signal annotation
  - event attributes and LINKs annotation →IN PROGRESS

- Different annotators (8 volunteers) and 2 judges

- Brandeis Annotation Tool: http://www.timeml.org/site/bat/

Quantitative data: annotations

Markable	# in the CELCT corpus	# in the ILC corpus
TIMEX3	4,852 (on 180k tokens)	2,314 (on 68k tokens)
EVENT	17,554 (on 90k tokens)	10,633 (on 68k tokens)
SIGNAL	2,045 (on 90k tokens)	1,704 (on 68k tokens)
TLINK	3,373 (on 90k tokens)	1,879 (on 68k tokens)
SLINK	3,985 (on 90k tokens)	in progress
ALINK	238 (on 90k tokens)	in progress

Quantitative data: IAA on extents

Markable	CELCT corpus	ILC corpus
TIMEX3	P&R=0.94	P&R=0.95
EVENT	P&R=0.94	P&R= 0.86
SIGNAL	P&R=0.88	P&R=0.84
TLINK	P&R=0.86	P&R=0.91
SLINK	P&R=0.93	in progress
ALINK	P&R=0.90	in progress

P&R = Average Precision and Recall

Quantitative data: IAA on extents

Markable	CELCT corpus	ILC corpus	TimeBank 1.2
TIMEX3	Dice=0.94	K=0.95 P&R=0.95	P&R=0.83
EVENT	K=0.93 P&R=0.94	K=0.87 P&R= 0.86	P&R=0.78
SIGNAL	K=0.88 P&R=0.88	K=0.83 P&R=0.84	P&R=0.77
TLINK	Dice=0.86	K=0.91	P&R=0.55
SLINK	Dice=0.93	in progress	P&R=0.85
ALINK	Dice=0.90	in progress	P&R=0.81

K = Cohen' Kappa   P&R = Average Precision and Recall

Quantitative data: IAA on attributes

Tag and attribute	Fleiss' kappa CELCT corpus	Fleiss' kappa ILC corpus
TIMEX3.type	1.00	0.96
TIMEX3.value	0.92	0.96
TIMEX3.mod	0.89	0.97
EVENT.aspect	0.96	0.93
EVENT.class	0.87	0.82
EVENT.modality	1.00	0.92
EVENT.mood	0.90	0.89
EVENT.polarity	1.00	0.75
EVENT.pos	1.00	0.95
EVENT.tense	0.94	0.97
EVENT.vform	0.98	0.94
TLINK.relType	0.88	in progress
SLINK.relType	0.93	in progress
ALINK.relType	1.00	in progress

Quantitative data: IAA on attributes

Tag and attribute	Fleiss' kappa CELCT corpus	Fleiss' kappa ILC corpus	Fleiss' kappa TimeBank 1.2
TIMEX3.type	1.00	0.96	1.00
TIMEX3.value	0.92	0.96	0.89
TIMEX3.mod	0.89	0.97	0.73
EVENT.aspect	0.96	0.93	1.00
EVENT.class	0.87	0.82	0.67
EVENT.modality	1.00	0.92	1.00
EVENT.mood	0.90	0.89	
EVENT.polarity	1.00	0.75	1.00
EVENT.pos	1.00	0.95	0.96
EVENT.tense	0.94	0.97	0.93
EVENT.vform	0.98	0.94	
TLINK.relType	0.88	in progress	0.63
SLINK.relType	0.93	in progress	0.96
ALINK.relType	1.00	in progress	0.71

Conclusions and ...

- **Methodology** followed for the development of the corpora: annotation guidelines + annotation specifications

- Significant efforts for the annotation of **corpora** to be used as **benchmarks** for the task of automatic Knowledge Extraction

- Good inter annotator agreement demonstrate the **reliability** of annotation guidelines and specifications of the It-TimeML

Future Works

- **Complete** the annotation and **distribute** the two corpora

- Annotation of **relations** between I–CAB **entities and events** in which they are involved

- Promote the development of NLP tools for Italian through **evaluation campaigns**

- Feasibility study on the use of **crowdsourcing** for TimeML annotation tasks

THANK YOU!

→ T. Caselli, V. Bartalesi Lenzi, R. Sprugnoli, E. Pianta, I. Prodanof, "Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank". To appear in *Proceedings of LAW V*