GMT to +2 or
How Can TimeML Be Used in Romanian

Corina Forăscu

Research Institute for Artificial Intelligence of the Romanian Academy
&
Faculty of Computer Science, Al.I. Cuza University of Iasi, Romania

corinfor@info.uaic.ro
Outline

1. Basic concepts
2. Standard & initial corpus
3. Corpus creation & processing
4. Analysis
5. Conclusions
Temporal information in NL

1. **Time-denoting** expressions – references to a calendar or clock system (NPs, PPs, or AdvPs)
   
   - the 28th of May, 2008; Wednesday; tomorrow; the third month

2. **Event-denoting** expressions - reference to an event (sentences, NPs, Adjs, PPs)

   - Jerry is watching the talks.
   - The presenter is prepared for a possible attack.
   - A student, dormant for half of the session, suddenly started to ask questions.
Benefits from TIP for NLP

1. CL: lexicon induction, linguistic investigation
2. QA: when?, how often? or how long?
3. IE & IR
4. MT:
   - translated and normalized temporal references
   - mappings between different behavior of tenses from language to language
5. DP: temporal structure of discourse and summarization
Standard: TimeML

A metadata standard developed especially for news articles, for marking

• events: EVENT, MAKEINSTANCE
• temporal anchoring of events: TIMEX3, SIGNAL
• links between events and/or timexes: TLINK, ALINK, SLINK
McDonald's is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
TimeML: EVENTS

McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious\textsuperscript{e206} to turn\textsuperscript{e32} around KFC sales that it soon will begin\textsuperscript{e33} selling\textsuperscript{e34} hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

10/30/09

TimeML: TIMEX3

<TIMEX3 tid="t207" type="DATE" temporalFunction="true" functionInDocument="NONE" value="FUTURE_REF" anchorTimeID="t192"/>
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES" />
McDonalds is so anxious to turn around KFC sales that it will soon begin selling hamburgers for 99 cents.

<TLINK relatedToTime="t192" eventInstanceId="ei2019" relType="INCLUDES" />
<TLINK relatedToEventInstance="ei2021" eventInstanceId="ei2019" relType="BEFORE" />
McDonalds is so anxious to turn around KFC sales that it will soon begin selling hamburgers for 99 cents.

<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES" />
<TLINK relatedToEventInstance="ei2021" eventInstanceID="ei2019" relType="BEFORE" />
<TLINK relatedToTime="t207" eventInstanceID="ei2021" relType="IS_INCLUDED" />
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

<TLINK relatedToTime="t192" eventInstanceID="ei2019" relType="INCLUDES"/>
<TLINK relatedToEventInstance="ei2021" eventInstanceID="ei2019" relType="BEFORE"/>
<TLINK relatedToTime="t207" eventInstanceID="ei2021" relType="IS_INCLUDED"/>
<TLINK relatedToTime="t192" eventInstanceID="ei2021" relType="AFTER"/>
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

<SLINK signalID="s31" subordinatedEventInstance="ei2020" eventInstanceID="ei2019" relType="MODAL" />

<SLINK signalID="s31" subordinatedEventInstance="ei2020" eventInstanceID="ei2021" relType="MODAL" />
McDonalds is so anxious to turn around KFC sales that it soon will begin selling hamburgers for 99 cents.

<ALINK relatedToEventInstance="ei2022" eventInstanceID="ei2021" relType="INITIATES" />
Corpus: TimeBank

- 183 English news report documents TimeML annotated, freely distributed through LDC
- 4715 sentences with
  - 10586 unique lexical units, from
  - a total of 61042 lexical units

Non-TimeML Markup in Time Bank 1.1:
- structure information: header
- named entity recognition: `<ENAMEX>`, `<NUMEX>`, `<CARDINAL>`
- sentence boundary information: `<s>`
Corpus: TimeBank – stats

- EVENTS 7935
- INSTANCES 7940
- TIMEX3es 1414
- SIGNALS 688
- TLINKS 6418
- SLINKs 2932
- ALINKs 265
- TOTAL 27592
Parallel corpus creation & processing

1. translation
2. pre-processing
3. alignment
4. annotation import
Corpus translation

1. Translation
   • 2 “trained translators”; one final correction
   • translation criteria
   • 4715 sentences (translation units)
     • 65375 lexical tokens (61042 in English)
     • 12640 lexical types (10586 in English)

2. pre-processing

3. alignment

4. annotation import
Pre-processing the parallel corpus

1. Translation

2. Pre-processing (RACAI web services)
   1. Tokenisation – MtSeg, with idiomatic expressions, clitic splitting
   2. POS-tagging – TnT adapted & improved to determine the POS of unknown words
   3. Lemmatisation – probabilistic, based on a lexicon
   4. Chunking – REs over POS tags to determine non-recursive NPs, APs, AdvPs, PPs

3. alignment

4. annotation import
Aligning the parallel corpus

1. Translation
2. Pre-processing
3. Alignment (RACAI YAWA aligner)
   1. Content words alignment
   2. Inside-Chunks alignment
   3. Alignment in contiguous sequences of unaligned words
   4. Correction phase
      • 91714 alignments, manually checked
4. annotation import
Aligning the parallel corpus
Parallel corpus: annotation import

1. Translation
2. Pre-processing
3. Alignment (RACAI YAWA aligner)
4. Annotation import
   1. Inline markup (EVENT, TIMEX3, SIGNAL): sentence level import of XML tags from English to Romanian
   2. Offline markup (MAKEINSTANCE, ALINK, TLINK, SLINK): the transfer kept only those XML tags whose IDs belong to XML structures that have been transferred to Romanian
### Parallel corpus: annotation import

<table>
<thead>
<tr>
<th>TimeML tags</th>
<th>#</th>
<th>% transferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVENTS</td>
<td>7703</td>
<td>97.07</td>
</tr>
<tr>
<td>INSTANCES</td>
<td>7706</td>
<td>97.05</td>
</tr>
<tr>
<td>TIMEX3s</td>
<td>1356</td>
<td>95.89</td>
</tr>
<tr>
<td>SIGNALs</td>
<td>668</td>
<td>97.09</td>
</tr>
<tr>
<td>TLINKs</td>
<td>6122</td>
<td>95.38</td>
</tr>
<tr>
<td>SLINKs</td>
<td>2831</td>
<td>96.55</td>
</tr>
<tr>
<td>ALINKs</td>
<td>249</td>
<td>93.96</td>
</tr>
<tr>
<td>TOTAL</td>
<td>26635</td>
<td>96.53</td>
</tr>
</tbody>
</table>
Analysis of the annotation import

A preliminary study using 10% of the parallel corpus in order to identify:

1. Types of temporal annotation import
   1. Perfect transfer
   2. Transfer with some amendments due to TimeML specifications
   3. Transfer with amendments imposed by language specific phenomena
   4. Impossible transfer

2. Temporal elements not (yet) marked
898 inline markups (EVENT, TIMEX3, SIGNAL)

1. Types of temporal annotation import
   1. Perfect transfer: 847 (91.41%) situations
   2. Transfer with some amendments due to TimeML specifications: 40 (6.4%) situations
   3. Transfer with amendments imposed by language specific phenomena: 3 (0.36%) situations
   4. Impossible transfer: 8 (6.3%) situations

2. Temporal elements not (yet) marked in English:
   104 EVENTs, 2 TIMEX3s, 19 SIGNALs
## Annotation import: EVENTS

<table>
<thead>
<tr>
<th>Type</th>
<th>#</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect</td>
<td>785</td>
<td>😊</td>
</tr>
</tbody>
</table>
| Amendment             | 37 | TimeML rule: in cases of phrases, the EVENT tag should mark only the head of the construction:  
  • reflexive verbs: (să) se retragă – (to) withdraw  
  • verbal collocations: avea permisiunea – permit  
  • compound verb phrases: să se îndoiască – doubt |
| Language specific 3    | 3  | intercalation of an adverb/conjunction between the verbs forming a verb phrase: also said – (au) mai spus; (he) also criticised – (a) şi criticat |
| Impossible            | 4  | • missing translations:  
  forces that harbor ill intentions – forţe cu intenţii rele  
  • non-lexicalisations: give\textsuperscript{1} the view\textsuperscript{2} – arată\textsuperscript{1}  
  • missing alignments (situations corrected) |
### Annotation import: TIMEX3s

<table>
<thead>
<tr>
<th>Type</th>
<th>#</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect</td>
<td>33</td>
<td>☺</td>
</tr>
<tr>
<td>Amendment</td>
<td>3</td>
<td>• wrong marking of the Romanian prepositions as part of TIMEX3:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eight years (war) – (războiul) de opt ani</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• missing alignment: some time - un timp mai lung</td>
</tr>
<tr>
<td>Language specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impossible</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*trecc 2008 Marrakech*
## Annotation import: SIGNALs

<table>
<thead>
<tr>
<th>Type</th>
<th>#</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect</td>
<td>29</td>
<td>☺</td>
</tr>
<tr>
<td>Amendment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Impossible</td>
<td>4</td>
<td>• non-lexicalisations: <em>on Thursday – joi</em></td>
</tr>
</tbody>
</table>
New temporal elements

• 104 **EVENTs**: 70 OCCURRENCEs (**nouns**: missions, training, fight, demarcation, **verbs**: supervising, leading, include), 5 REPORTING (**say**, **said**), 21 STATEs (**belongs**, look, staying, war, policies), 1 **I_ACTION** (**include**), 7 **I_STATE** (**like**, think)

**Rationale**: each sentence expresses an event, even if not so well temporally-anchored

• 2 **TIMEX3s**: once, not that long ago

**Rationale**: non-specific value but possible to normalize according to ISO 8601 extended

• 19 **SIGNALs**: several, **when**, meanwhile, **time and again**, after, **on**

**Rationale**: identify multiple instantiations for some EVENTs (inevitable manual annotation mistakes)
Conclusions

1. The automatic import of the temporal annotations from English to Romanian is a worth doing enterprise (96.53% success rate).
2. Human introspection shows few modifications are needed.
3. The automatic transfer of (temporal) annotations represents a solution for having a (temporally) annotated corpus, if a parallel corpus & adequate processing tools exist.
4. Improvements can be done in TimeBank – consistent with TimeML developers (Boguraev, Ando, 2006).
Future work

Immediate:
- improvement & evaluation of the annotation transfer
- adequacy of temporal theories to Romanian
- translated and normalized temporal references
- mappings between different behavior of tenses from language to language

Long-term:
- (semi) automatically mark-up of the temporal information in Romanian texts (news + literature, legislation)
Acknowledgements

The author is grateful to:

- Dan Tufiş and the RACAI NLP group (especially Radu Ion) for the support and helpful discussions and advices w.r.t. this research
- Dan Cristea, Jerry Hobbs, James Pustejovsky, Marc Verhagen, and Georgiana Puşcaşu for useful research outcomes coming from discussions
- All tree 2008 organizers and reviewers
Thank you!

(Temporal) Questions???