Time Out of Joint in Temporal Annotations of Texts: Challenges for Artificial Intelligence and Human Computer Interaction

Rosella Gennari[1] and Pierpaolo Vittorini[2]

[1] Free University of Bozen-Bolzano,
gennari@inf.unibz.it
[2] University of L'Aquila,
pierpaolo.vittorini@univaq.it

Abstract. Starting from the experience of the TERENCE European project, the paper shows challenges that require a combined effort of natural language processing, automated temporal reasoning and, finally, human computer interaction. The paper starts introducing the problem of producing high quality temporal annotations for texts, and argues for a combined automated temporal reasoning and natural processing approach to tackle it. The paper then speculates that the approach would benefit from knowledge of the specific domain and of how humans interact with the annotation process, which triggers two further challenges explored in the remainder of the paper, at the intersection of natural language processing, automated reasoning and human computer interaction.

Keywords: constraint satisfaction, temporal reasoning, natural language processing, human computer interaction

1 Stories Will Teach Us Something

Reading is an important means for language acquisition, communication, sharing information and ideas. Reading transforms print to speech and print to meaning through a negotiation of meaning between the text and its reader, as a problem solving activity. Developing the capabilities of children to comprehend written texts is key to their development as young adults. More than 10% of 7–10 year old children are poor comprehenders: they have difficulties in comprehending texts, e.g., making inferences concerning the temporal flow of a story. TERENCE (10.2010-09.2013) was an FP7 European project that developed the first adaptive learning system with learning material for primary-school poor comprehenders, made of stories and quiz-like games for reasoning about stories, in English and in Italian. The material is immersed in a game world and delivered in an adaptive fashion according to children's learning needs, investigated through contextual inquiries with text comprehension experts and activities with children [1, 2], so as to promote a personalised experience. The repositories of annotated TERENCE stories and of TERENCE games are available as project deliverables [3] and [4], respectively, at www.terenceproject.eu.

Training children to reason about the temporal flow of stories as in TERENCE required to have c. 12 games of different complexity per story, for a total of more than c. 200 games per language.

Such figures led the TERENCE researchers to tackle an ambitious goal: to design a semi-automated process for generating inference-making games, of different levels, starting from stories, with the aim of significantly reducing human interventions in the generation process. In order to meet their goal, TERENCE researchers chose Artificial Intelligence (AI) for automatically extracting from stories data for semi-automatically generating inference-making games, progressively training children to text comprehension. AI took the form of Natural Language Processing (NLP) for the language-dependent aspects of games, and Automated Reasoning (AR) with temporal constraints, for completing the NLP work [5]. Such a choice meant for the TERENCE researchers to face a number of challenges, the most crucial being:

> *How to use NLP and AR to extract temporal information from stories that is critical for their comprehension?*

The TERENCE Consortium tackled that and related challenges by developing an annotation schema based on the TimeML markup language [6, 7], which already covers events and qualitative temporal information, relevant for stories, and is the de-facto standard markup language for temporal information in the NLP community. Moreover, the Consortium developed an AI-based process for generating games from stories, using NLP and AR. Notice that the process automatically generates all text-related parts of games, referred to as *textual games*, and automatically assembles them with graphical elements of games.

The automatically generated textual games were evaluated by education experts with a qualitative evaluation design, similar to Amazon Mechanical Turk, through an interface designed for them [8, 9]. Independent judges were asked to assess the textual games and revise them in case of generation errors, tracking the revision process in a structured format. Results were then revised by a further expert and differences resolved through written discussions, documented in text. Afterwards, the TERENCE system, its stories and games were used in a large-scale study at school for evaluating improvements in text comprehension and, above all, what games turned out the be most difficult for children. The evaluation of the games pointed to areas for potential improvements for the collaboration of NLP and AR.

In particular, of relevance for URANIA, the evaluation results suggested that errors in the generation process were also due to the quality of the TimeML annotation of texts for extracting temporal data, which is in turn affected by the quality of corpora over which NLP systems are trained for recognising and annotating temporal information in texts, e.g., see [10].

That led us to embark on a novel journey and tackle a new ambitious goal: to analyse errors that recur through TimeML corpora, setting “time out of joint”.

2 Analysis of Human Errors and Novel Challenges

TimeML is used in resources such as the TimeBank corpus [11], the Ita-TimeBank [12] and the data annotated for TempEval shared tasks, used for training and assessing NLP

systems [13, 14]. Specifically, the TimeML language and guidelines are used by human annotators for marking events and their temporal relations in such resources.

In TimeML, time is assumed to be linearly ordered over the real line, and relations between events are interpreted by relying on the standard order between interval endpoints. TimeML defines a qualitative time entity (e.g., action verbs) called EVENT, and a quantitative time entity (e.g., dates) called TIMEX. In particular, in TimeML events can be expressed by tensed and untensed verbs, but also by nominalizations (e.g. *invasion*, *discussion*, *speech*), predicative clauses (e.g., *to be the President of something*), adjectives (e.g. *dormant*) or prepositional phrases (e.g., *on board*).

TIMEX temporal expressions include specific dates (e.g. *June 11, 1989*), times (*twenty to ten*), durations (*three months*) and sets (*twice a week*). TIMEXs are also assigned a value that makes explicit, in ISO 8301 format, to which specific time the expression is anchored.

Time entities (EVENT and TIMEX) are linked through a TLINK relations. Several TLINK relations have been introduced in TimeML, intuitively interpreted as *Allen interval algebra* basic relations [15] in TimeML guidelines.

When human annotators manually add TimeML annotations, they can introduce different mistakes, especially in connection with temporal relations; each such mistake has potentially an impact on the quality of NLP systems that recognise and annotate temporal data. Such mistakes range from simple ones to subtle ones.

Examples of simple errors occur when annotators add two relations, such as "before" and its inverse "after", between the same two EVENT or TIMEX expressions. Such a situation is inconsistent with the assumption that time is linearly ordered—an event cannot be simultaneously before and after another. Preprocessing techniques can help in finding and fixing such errors.

Other mistakes creep into the manual annotation process, which are much harder to detect for humans: those are the cases of inconsistencies due to a chain of temporal relations between events, possibly distant in a text. An example of such error is in Table 1, found in the AQUAINT TimeML corpus. The specific error is the incorrect annotation of a "before" TLINK relation, which is inconsistent with other TLINK relations.

The way in which annotators work on corpora and the tools they have currently at their disposal all have an impact on the quality of their annotation work; see [16] and the results of contextual inquiries with NLP annotators documented in [17].

The journey into the manual annotation work led us to face a novel general **challenge** that require again a combined effort, specifically, of NLP and AR, and related to the TimeML manual annotation process:

How can AR and NLP improve on the quality of TimeML annotation work?

The question has been recently tackled in a novel manner by [16]. Chambers and colleagues devised new annotation guidelines and an AR-based support system for helping annotators in their work. The guidelines require annotators to add annotations for each pair of EVENT or TIMEX3 expressions, all interpreted as intervals over the real-line, for creating a dense temporal graph of relations. In case of doubts, annotators have to use the VAGUE relation. The system, progressively, suggests annotators TLINK relations that are consistent (once interpreted as Allen relations) with TLINK annotations

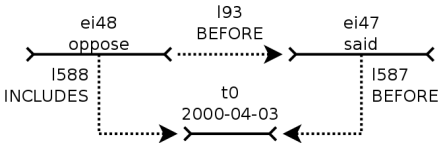
Document excerpt
Castro said^{e47} that if those who oppose^{e48} returning Elian to Cuba are worried about turning the child over to what is considered Cuban territory, then our Interests Section is willing to renounce diplomatic immunity of the residence of the chief of this section in Washington.

TLINKs

Link ID	Source	Target	Relation
l93	ei48	ei47	BEFORE
l587	ei47	t0	BEFORE
l588	ei48	t0	INCLUDES

Other temporal entities
t_0 = 2000-04-03

Interpretation over the real-line



Error
If ei48 is before ei47, and includes t_0, then ei47 cannot be before t_0 as well.

Table 1. Summary of inconsistency detection in APW20000403.0057.tml

already existing in the document. In other words, the system aid annotators in avoiding future annotation errors due to inconsistent TLINKs. However, the system does not support annotators if annotation errors sneaked early into the manual annotation, and potentially affect future correct choices of annotators. For instance, reconsider the excerpt reported in Table 1. The wrong TLINK annotation, with identifier l587, is BEFORE between ei47 and t0. This is inconsistent with the other reported TLINK annotations. If annotators introduce l587 as first, then the annotation guidelines and tool may not detect it as error and may instead mark the other annotations as errors.

In such cases, it seems beneficial to consider a complementary approach and system, which aids annotators in finding annotation errors introduced in annotated documents and due to inconsistent TLINKs (once interpreted as Allen relations or as relations of other temporal calculi). The interpretation of TLINK as Allen relations, or as relations of a different qualitative calculus, should be flexible and domain-dependent, as advanced in [18]. Moreover, we believe that the knowledge of how annotators work in annotating is pivotal for devising AR and NLP solutions that can efficiently spot such errors.

Therefore another specific **challenge** is as follows, which derives from the previous one.

How can knowledge of human processing of texts help in devising combined NLP and AR solutions for setting time in joint in document analysis?

For tackling the challenge, we are currently implementing and testing strategies for rapidly identifying possible inconsistencies, as well as designing how to highlight them in currently available tools for manual annotation and suggesting how to fix them (e.g., CAT [19]). The strategies exploit how annotators tend to work on annotating texts, e.g., sentence by sentence.

Last but not least, no support system for humans annotating texts can avoid the human-computer interaction aspects of the work, which will again require to cope with how annotators work. That is the final **challenge** we believe relevant for URANIA.

How can knowledge of human processing of texts help in devising visual metaphors that help humans in their text annotation work?

As for the visual tool, a preliminary support only for consistency checking was presented in [18], whereas improved visual metaphors are currently under investigation with HCC researchers.

References

1. Di Mascio, T., Gennari, R., Melonio, A., Tarantino, L.: Engaging "new users" into design activities: The terence experience with children. In: Smart Organizations and Smart Artifacts: Fostering Interaction Between People, Technologies and Processes. Volume 7., Cham, Springer International Publishing (2014) 241–250
2. Di Mascio, T., Gennari, R., Melonio, A., Vittorini, P.: The User Classes Building Process in a TEL Project. Advances in Intelligent and Soft Computing, year=2012, volume=152 AISC, pages=107–114
3. Moens, M.F., Kolomiyets, O.: Repository of Annotated Stories. Technical Report D3.3.3, TERENCE Project. (2013)
4. Gennari, R.: Generation Service and Repository of Textual Smart Games. Technical Report D4.3.3, TERENCE Project. (2013)
5. Gennari, R., Tonelli, S., Vittorini, P.: Challenges in Quality of Temporal Data — Starting with Gold Standards. Journal of Data and Information Quality (2015)
6. Pustejovsky, J., Castano, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Proc. IWCS-5. (2003)
7. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: ISO-TimeML: An International Standard for Semantic Annotation. In Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
8. Cofini, V., Gennari, R., Vittorini, P.: The Manual Revision of the TERENCE Italian Smart Games. In Vittorini P. et al, ed.: Proc. of the 2nd evidence-based TEL workshop (ebTEL 2013), Springer (2013)
9. Cofini, V., Di Mascio, T., Gennari, R., Vittorini, P.: The TERENCE smart games revision guidelines and software tool. Advances in Intelligent Systems and Computing **218** (2013) 17–24

10. Gennari, R., Tonelli, S., Vittorini, P.: An AI-Based Process for Generating Games from Flat Stories. In: Research and Development in Intelligent Systems XXX, Incorporating Applications and Innovations in Intelligent Systems XXI Proceedings of AI-2013, The Thirty-third SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, England, UK, December 10-12, 2013. (2013) 337–350
11. Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., others: The TimeBank corpus. In: Corpus Linguistics. Volume 2003. (2003) 40
12. Caselli, T., Lenzi, V.B., Sprugnoli, R., Pianta, E., Prodanof, I.: Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In: Proceedings of LAW V, Portland, Oregon, USA (2011)
13. Verhagen, M., Sauri, R., Caselli, T., Pustejovsky, J.: SemEval-2010 Task 13: TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, Association for Computational Linguistics (July 2010) 57–62
14. UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., Pustejovsky, J.: SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, Association for Computational Linguistics (June 2013) 1–9 bibtex: uzzaman-EtAl:2013:SemEval-2013 bibtex[bdsk-url-1=http://www.aclweb.org/anthology/S13-2001].
15. Allen, J.F.: Maintaining Knowledge about Temporal Intervals. ACM Comm **26** (1983) 832–843
16. Chambers, N., Cassidy, T., McDowell, B., Bethard, S.: Dense event ordering with a multi-pass architecture. Transactions of the Association for Computational Linguistics **2** (2014) 273–284
17. Di Mascio, T.: First user classification, user identification, user needs, and usability goals. Technical Report D1.2.1, TERENCE Project. (April 2012)
18. Gennari, R., Vittorini, P.: Qualitative Temporal Reasoning Can Improve on Temporal Annotation Quality: How and Why. Applied Artificial Intelligence **30**(7) (2016) 690–719
19. Lenzi, V.B., Moretti, G., Sprugnoli, R.: CAT: the CELCT Annotation Tool. In: LREC. (2012) 333–338