Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

FacTA: Evaluation of Event Factuality and Temporal Anchoring

Anne-Lyse Minard, Manuela Speranza, Rachele Sprugnoli and Tommaso Caselli

DOI: 10.4000/books.aaccademia.1509
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

FacTA: Evaluation of Event Factuality and Temporal Anchoring

Anne-Lyse Minard[1], Manuela Speranza[1], Rachele Sprugnoli[1-2], Tommaso Caselli[3]
[1]Fondazione Bruno Kessler, Trento
[2]Università di Trento
[3]VU Amsterdam
{minard,manspera,sprugnoli}@fbk.eu
t.caselli@vu.nl

Abstract

English. In this paper we describe FacTA, a new task connecting the evaluation of factuality profiling and temporal anchoring, two strictly related aspects in event processing. The proposed task aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making available benchmark data for high-level semantic tasks.

Italiano. *Questo articolo descrive FacTA, un nuovo esercizio di valutazione su fattualità ed ancoraggio temporale, due aspetti dell'analisi degli eventi strettamente connessi tra loro. Il compito proposto mira a fornire una cornice completa di valutazione per la fattualità, a muovere i primi passi nella direzione della valutazione dei contenitori narrativi per l'italiano e a rendere disponibili dati di riferimento per compiti semantici di alto livello.*

1 Introduction

Reasoning about events plays a fundamental role in text understanding; it involves different aspects, such as event identification and classification, temporal anchoring of events, temporal ordering, and event factuality profiling. In view of the next EVALITA edition (Attardi et al., 2015),[1] we propose FacTA (*Factuality and Temporal Anchoring*), the first task comprising the evaluation of both factuality profiling and temporal anchoring, two strictly interrelated aspects of event interpretation.

Event factuality is defined in the literature as the level of committed belief expressed by relevant sources towards the factual status of events mentioned in texts (Saurí and Pustejovsky, 2012). The notion of factuality is closely connected to other notions thoroughly explored by previous research conducted in the NLP field, such as subjectivity, belief, hedging and modality; see, among others, (Wiebe et al., 2004; Prabhakaran et al., 2010; Medlock and Briscoe, 2007; Saurı et al., 2006). More specifically, the factuality status of events is related to their degree of certainty (from absolutely certain to uncertain) and to their polarity (affirmed *vs.* negated). These two aspects are taken into consideration in the factuality annotation frameworks proposed by Saurí and Pustejovsky (2012) and van Son et al. (2014), which inspired the definition of factuality profiling in FacTA.

Temporal anchoring consists of associating all temporally grounded events to time anchors, i.e. temporal expressions, through a set of temporal links. The TimeML annotation framework (Pustejovsky et al., 2005) addresses this issue through the specifications for temporal relation (TLINK) annotation, which also implies the ordering of events and temporal expressions with respect to one another. Far from being a trivial task (see systems performance in English (UzZaman et al., 2013) and in Italian (Mirza and Minard, 2014)), TLINK annotation requires the comprehension of complex temporal structures; moreover, the number of possible TLINKs grows together with the number of annotated events and temporal expressions. Pustejovsky and Stubbs (2011) introduced the notion of *narrative container* with the aim of reducing the number of TLINKs to be identified in a text while improving informativeness and accuracy.

A narrative container is a temporal expression or an event explicitly mentioned in the text into which other events temporally fall (Styler IV et al., 2014). The use of narrative containers proved to be useful to accurately place events on timelines in the domain of clinical narratives (Miller et al., 2013). Temporal anchoring in FacTA moves in the direction of this notion of narrative container by fo-

[1] http://www.evalita.it/

cusing on specific types of temporal relations that link an event to the temporal expression to which it is anchored. However, anchoring events in time is strictly dependent of their factuality profiling. For instance, counterfactual events will never have a temporal anchor or be part of a temporal relation (i.e. they never occurred); this may not hold for speculated events, whose association with a temporal anchor or participation in a temporal relation is important to monitor future event outcomes.

2 Related Evaluation Tasks

Factuality profiling and temporal anchoring of events are crucial for many NLP applications (Wiebe et al., 2005; Karttunen and Zaenen, 2005; Caselli et al., 2015) and therefore have been the focus, either direct or indirect, of several evaluation exercises, especially for English.

The ACE Event Detection and Recognition tasks of 2005 and 2007 (LDC, 2005) took into consideration factuality-related information by requiring systems to assign the value of the *modality* attribute to extracted events so as to distinguish between asserted and non-asserted (e.g. hypothetical, desired, and promised) events. Following the ACE evaluation, a new task has recently been defined in the context of the TAC KBP 2015 Event Track.[2] The Event Nugget Detection task aims at assessing the performance of systems in identifying events and their *realis* value, which can be ACTUAL, GENERIC or OTHER (Mitamura et al., 2015). Other tasks focused on the evaluation of speculated and negated events in different domains such as biomedical data and literary texts (Nédellec et al., 2013; Morante and Blanco, 2012).

The evaluation of event modality was part of the Clinical TempEval task at SemEval 2015 (Bethard et al., 2015),[3] which also proposed for the first time the evaluation of narrative container relations between events and/or temporal expressions.

Temporal anchoring has been evaluated in the more general context of temporal relation annotation in the 2007, 2011 and 2013 TempEval evaluation exercises (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) as well as in the EVENTI task (Caselli et al., 2014) on Italian at EVALITA 2014. The TimeLine task at SemEval 2015 (Minard et al., 2015) was the first evaluation exercise focusing on cross-document event ordering; in view of the creation of timelines, it requires temporal anchoring and ordering of certain and non-negated events.

With respect to the aforementioned tasks, FacTA aims at providing a complete evaluation framework for factuality profiling, at taking the first steps in the direction of narrative container evaluation for Italian, and at making new datasets available to the research community.

3 Task Description

The FacTA task consists of two subtasks: factuality profiling and temporal anchoring of given gold event mentions. Participants may decide to take part to both or only one of the proposed subtasks.

3.1 Subtask 1: Factuality Profiling

Tonelli et al. (2014) propose an annotation schema of factuality for English based on the annotation framework by van Son et al. (2014).[4] This schema was then adapted to Italian by Minard et al. (2014). Following this, we represent factuality by means of a combination of three attributes associated with event mentions: certainty, time, and polarity. For each given gold event mention, participant systems are required to assign values for three factuality attributes.

The *certainty* attribute relates to how sure the main source is about the mentioned event[5] and admits the following four values: `certain`, `possible`, `probable`, and `underspecified`.

The *time* attribute specifies the time when an event is reported to have taken place or to be going to take place. Its values are `non_future` (for present and past events), `future` (for events that will take place), and `underspecified`.

The *polarity* attribute captures if an event is affirmed or negated and, consequently, it can be either `positive` or `negative`; when there is not enough information available to detect the polarity of an event mention, it is `underspecified`.

[2] `http://www.nist.gov/tac/2015/KBP/Event/index.html`

[3] Systems were required to distinguish actual, hedged, hypothetical and generic events.

[4] van Son et al.'s annotation framework, inspired by FactBank (Saurí and Pustejovsky, 2009), enriches it with the distinction between future and non-future events.

[5] The main source is either the utterer (in direct speech, indirect speech or reported speech) or the author of the news (in all other cases). In this framework, where factuality depends strictly on the source, factuality annotation is also referred to as attribution annotation.

Factuality value. The combination of the attributes described above determines the value of an event: `factual`, `counterfactual` or `non_factual`. More specifically, the overall factuality value is factual if its values are certain, non_future, and positive (e.g. 'rassegnato' in [1]), while it is counterfactual (i.e. the event is reported as not having taken place) if its values are certain, non_future, and negative (e.g. 'nominato' in [2]). In any other combination, the event is non factual, either because it is non certain, or future (e.g. 'nomineranno' in [1]).

(1) *Smith ha rassegnato ieri le dimissioni; nomineranno il suo successore entro un mese.* ("Smith resigned yesterday; they will appoint his replacement within a month.")

(2) *Non ha nominato un amministratore delegato.* ("He did not appoint a CEO.")

No factuality annotation. Language is used to describe events that do not correlate with a real situation in the world (e.g. 'parlare' in [3]). For these event mentions participant systems are required to leave the value of all three attributes empty.

(3) *Guardate, penso che sia prematuro parlare del nuovo preside* ("Well, I think it is too early to talk about the new dean")

3.2 Subtask 2: Temporal Anchoring

Given a set of gold events, participant systems are required to detect those events for which it is possible to identify a time anchor. Our definition of time anchor includes two different types of elements: the temporal expressions occurring in the text, as well as the Document Creation Time (DCT), which is part of the metadata associated with each document. The subtask thus includes temporal expression (or TIMEX3) detection and normalization,[6] as well as identification of temporal relations (or TLINKs) between events and temporal expressions.

TIMEX3 detection and normalization. Based on the annotation guidelines produced within the NewsReader project (Tonelli et al., 2014), which in turn are based on the ISO-TimeML guidelines (ISO TimeML Working Group, 2008), this consists of:

- TIMEX3 detection: identification and classification of temporal expressions of type date and time (durations and sets of times, on the other hand, are excluded from the task).
- TIMEX3 normalization: identification of the *value* attribute for each temporal expression.

For instance, in [1], *ieri* is a TIMEX3 of type date with value *2015-07-28* considering *2015-07-29* as DCT.

TLINK identification. This consists of detecting TLINKs of types IS_INCLUDED and SIMULTANEOUS holding between an event and a TIMEX3 (i.e. the anchor of the event), as defined in (Tonelli et al., 2014). The event (the source of the TLINK) and the TIMEX3 (the target) can either appear in the same sentence or in different sentences. For instance, in [1], *rassegnato* is anchored to *ieri* (*rassegnato*, IS_INCLUDED, *ieri*).

4 Dataset Description

4.1 Subtask 1: Factuality Profiling

As a training dataset, participants can use Fact-Ita Bank (Minard et al., 2014), which consists of 170 documents selected from the Ita-TimeBank (Caselli et al., 2011), which was first released for the EVENTI task at EVALITA 2014.[7] Fact-Ita Bank contains annotations for 10,205 event mentions and is already distributed with a CC-BY-NC license.[8]

System evaluation will be performed on the "first five sentences" section of WItaC, the NewsReader Wikinews Italian Corpus (Speranza and Minard, 2015).[9] It consists of 15,676 tokens and has already been annotated with event factuality (as this annotation has been projected from English, it will need some minor revision).

4.2 Subtask 2: Temporal Anchoring

For temporal expression detection and normalization, participant systems can be trained on the dataset used for the EVENTI Task at Evalita 2014 (Caselli et al., 2014). It also contains TLINKs between events and TIMEX3s in the same sentence but not in different sentences. To make it usable as a training corpus for temporal anchoring, we would have to add the TLINKs between events and

[6] Here, and in the remainder of the paper, we are not distinguishing between the two types of elements and we refer to them simply as temporal expressions or TIMEX3s.

[7] https://sites.google.com/site/eventievalita2014/home

[8] http://hlt-nlp.fbk.eu/technologies/fact-ita-bank

[9] The reason for selecting the first sentences was to maximise the number of articles in the corpus, while at the same time including the most salient information.

TIMEX3s in different sentences and the TLINKs between events and the DCT, which would require a big effort. Thus, we are instead planning to add the needed relations to only a subset of the corpus, namely the same 170 documents that compose Fact-ItaBank.

As test data we will use the "first five sentences" section of WItaC (Speranza and Minard, 2015), which is already annotated with TIMEX3s and with TLINKs between events and TIMEX3s in the same sentences;[10] the test set thus needs to be completed through the addition of TLINKs between events and TIMEX3s in different sentences.

5 Evaluation

Each subtask will be evaluated independently. No global score will be computed as the task aims to isolate the two phenomena.

5.1 Subtask 1: Factuality Profiling

Participant systems will be evaluated in terms of precision, recall and their harmonic mean (i.e. F1 score). We will perform the evaluation of:

- values of the factuality attributes (polarity, certainty and time);
- detection of events to which factuality values should not be assigned (i.e. "no factuality annotation" events);
- assignment of the overall factuality value (combination of the three attributes), including also the non-assignment of factuality attributes.

The official ranking of the systems will be based on the evaluation of the overall factuality value.

5.2 Subtask 2: Temporal Anchoring

For the temporal anchoring subtask, we will evaluate the number of event-TIMEX3 relations correctly identified in terms of precision, recall and F1 score. Two relations in the reference and the system prediction match if their sources and their targets match. Two sources (i.e. events) are considered as equivalent if they have the same extent, whereas two targets (i.e. TIMEX3s) match if their values are the same. Participant systems will be ranked according to the F1 score.

We will not apply the metric for evaluating temporal awareness based on temporal closure graphs proposed by UzZaman and Allen (2011), which is unnecessarily complex as we have reduced the relations to only IS_INCLUDED and SIMULTANEOUS.

6 Discussion and Conclusions

The FacTA task connects two related aspects of events: factuality and temporal anchoring. The availability of this information for Italian will both promote research in these areas and fill a gap with respect to other languages, such as English, for a variety of semantic tasks.

Factuality profiling is a challenging task aimed at identifying the speaker/writers degree of commitment to the events being referred to in a text. Having access to this type of information plays a crucial role for distinguishing relevant and non-relevant information for more complex tasks such as textual entailment, question answering, and temporal processing.

On the other hand, anchoring events in time requires to interpret temporal information which is not often explicitly provided in texts. The identification of the correct temporal anchor facilitates the organization of events in groups of narrative containers which could be further used to improve the identification and classification of in-document and cross-document temporal relations.

The new annotation layers will be added on top of an existing dataset, the EVENTI corpus, thus allowing to re-use existing resources and to promote the development of multi-layered annotated corpora; moreover a new linguistic resource, WItaC, will be provided. The availability of these data is to be considered strategic as it will help the study the interactions of different language phenomena and enhance the development of more robust systems for automatic access to the content of texts. The use of well structured annotation guidelines grounded both on official and *de facto* standards is a stimulus for the development of multilingual approaches and promote discussions and reflections in the NLP community at large.

Considering the success of evaluation campaigns such as Clinical TempEval at SemEval 2015 and given the presence of an active community focused on extra-propositional aspects of meanings (e.g. attribution[11]), making available new annotated data in the framework of an evaluation campaign for a language other than English can have a large impact in the NLP community.

[10] This also includes TLINKs between events and the DCT.

[11] Ex-Prom Workshop at NAACL 2015 `http://www.cse.unt.edu/exprom2015/`

Acknowledgements

This work has been partially supported by the EU NewsReader Project (FP7-ICT-2011-8 grant 316404) and the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3).

References

Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. 2015. State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective. *Intelligenza Artificiale*, 9(1):43–61.

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics*.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Tommaso Caselli, Antske Fokkens, Roser Morante, and Piek Vossen. 2015. SPINOZA VU: An NLP Pipeline for Cross Document TimeLines. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

ISO TimeML Working Group. 2008. ISO TC37 draft international standard DIS 24617-1, August 14. `http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf`.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, extracting and reasoning about time and events*, Dagstuhl, Germany.

LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. In *Technical Report*.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *ACL*, volume 2007, pages 992–999. Citeseer.

Timothy A Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana K Savova. 2013. Discovering narrative containers in clinical text. *ACL 2013*, page 18.

Anne-Lyse Minard, Alessandro Marchetti, and Manuela Speranza. 2014. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of CLiC-it 2014, First Italian Conference on Computational Linguistic*.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Rubén Urizar, and Fondazione Bruno Kessler. 2015. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics.

Paramita Mirza and Anne-Lyse Minard. 2014. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 66–76.

Roser Morante and Eduardo Blanco. 2012. * SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 265–274. Association for Computational Linguistics.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1014–1022. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. *The specification language TimeML*, pages 545–557.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Manuela Speranza and Anne-Lyse Minard. 2015. Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC). In *Proceedings of CLiC-it 2015, Second Italian Conference on Computational Linguistic*.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf`.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of SemEval 2013*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Chantal van Son, Marieke van Erp, Antske Fokkens, and Piek Vossen. 2014. Hope and Fear: Interpreting Perspectives by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: TempEval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.