10.12871/CLICIT2014150

Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank

Anne-Lyse Minard
FBK, Trento, Italy
minard@fbk.eu

Alessandro Marchetti
FBK, Trento, Italy
alessandro.marchetti777@gmail.com

Manuela Speranza
FBK, Trento, Italy
manspera@fbk.eu

Abstract

English. In this paper we present ongoing work devoted to the extension of the Ita-TimeBank (Caselli et al., 2011) with event factuality annotation on top of TimeML annotation, where event factuality is represented on three main axes: time, polarity and certainty. We describe the annotation schema proposed for Italian and report on the results of our corpus analysis.

Italiano. *In questo articolo viene presentata un'estensione di Ita-TimeBank (Caselli et al., 2011), con l'annotazione della fattualità delle menzioni eventive già individuate secondo le specifiche di TimeML. La fattualità degli eventi è rappresentata attraverso tre dimensioni: tempo, polarità e certezza. Lo schema di annotazione proposto per l'italiano e l'analisi del corpus sono riportati e descritti.*

1 Introduction

In this work, we propose an annotation schema for factuality in Italian adapted from the schema for English developed in the NewsReader project[1] (Tonelli et al., 2014) and describe the annotation performed on top of event annotation in the Ita-TimeBank (Caselli et al., 2011). We aim at the creation of a reference corpus for training and testing a factuality recognizer for Italian.

The knowledge of the factual or non-factual nature of an event mentioned in a text is crucial for many applications (such as question answering, information extraction and temporal reasoning) because it allows us to recognize if an event refers to a real or to hypothetical situation, and enables us to assign it to its time of occurrence. In particular we are interested in the representation of information about a specific entity on a timeline, which enables easier access to related knowledge. The automatic creation of timelines requires the detection of situations and events in which target entities participate. To be able to place an event on a timeline, a system has to be able to select the events which happen or that are true at a certain point in time or in a time span. In a real context (such as the context of a newspaper article), the situations and events mentioned in texts can refer to real situations in the world, have no real counterpart, or have an uncertain nature.

The FactBank guidelines are the reference guidelines for factuality in English and FactBank is the reference corpus (Sauri and Pustejovsky, 2009). More recently other guidelines and resources have been developed (Wonsever et al., 2012; van Son et al., 2014), but, to the best of our knowledge, no resources exist for event factuality in Italian.

2 Related work

Several studies have been carried out on the representation of factuality information. In addition to the definition of annotation frameworks, these studies have been leading to the development of annotated corpora.

Our notion of event factuality is based on the notion of event as defined in the TimeML specifications (Pustejovsky et al., 2003a) and annotated in TimeBank (Pustejovsky et al., 2003b). *Event* is a cover term for situations that happen or occur, including predicates describing states or circumstances in which something obtains or holds true (Pustejovsky et al., 2003a).

Our main reference for factuality is FactBank (Sauri and Pustejovsky, 2009), where event factuality is defined as the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in a given

[1] http://www.newsreader-project.eu/

discourse.

van Son et al. (2014) propose an annotation schema inspired by FactBank. They add the distinction between past or present events and future events (temporality) to the FactBank schema. They then use three features (polarity, certainty and temporality) to annotate event factuality on top of the sentiment annotation in the MPQA corpus (Wiebe et al., 2005).

Wonsever et al. (2012) propose an event annotation schema based on TimeML for event factuality in Spanish texts. Factuality is annotated as a property of events that can have the following values: YES (factual), NO (non-factual), PROGRAMMED_FUTURE, NEGATED_FUTURE, POSSIBLE or INDEFINITE. Besides the factuality attribute they introduce an attribute to represent the semantic time of events, which can be different from the syntactic tense. In this way they duplicate both temporal information and polarity, as the factuality values include temporal and polarity information.

For Italian, to the best of our knowledge, there are no resources for factuality. The closest work to event factuality annotation that has been done is the annotation of attribution relations in a portion of the ISST corpus (Pareti and Prodanof, 2010). An attribution relation is the link between a source and what it expresses, and contains features providing information about the type of attitude and the factuality of the attribution. The focus of this annotation is on sources and their relations with events, while our work aims at describing factuality of events without explicitly annotating the relations between events and sources.

3 Annotation of factuality

As part of the NewsReader project, Tonelli et al. (2014) have defined guidelines for intra-document annotation at the semantic level, which provide an annotation schema of factuality for English based on TimeML annotation and the annotation framework proposed by van Son et al. (2014).

Following this annotation schema, we propose guidelines for event factuality annotation in Italian where we represent factuality by means of three attributes associated to event mentions: certainty, time, and polarity.

Certainty. We define the certainty attribute as how certain the source is about an event, with the following three values: `certain`, `possible`, `probable`. Modals and modal adverbs are typical markers of both probable (e.g. *essere probabile - be likely*) and possible (e.g. *potere - may*, *can*) events. The `underspecified` value is used for events for which it is not possible to assign a certainty value. In example (1) the event *portare* is `possible` due to the presence of *potere*. Certainty is determined according to the main source, which can be the utterer (in cases of direct speech, indirect speech or reported speech) or the author of the news. In (2) the source used to determine the certainty of *detto* is the writer and for *giocato* it is *Gianluca Nuzzo*. In both cases the source is certain about the event.

(1) *L'aumento delle tasse potrebbe* ***portare*** *nelle casse più di 500.000 euro.* [The tax increase could **bring** in more than 500,000 euros.]

(2) *"Durante l'ultimo mese ho* ***giocato*** *pochissimo", ha* ***detto*** *Gianluca Nuzzo.* ["During the last month I **played** very little, said Gian Luca Nuzzo".]

Time. The time attribute specifies the time an event took place or will take place. Its values are `non future` (for present and past events), `future` (for events that will take place), and `underspecified` (used for general events and when the time of an event cannot be determined). In the case o reported speech, the value of the time attribute is related to the time of utterance and not to the time of writing (i.e. when the utterance is reported).

Polarity. 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, it is `underspecified`.

Special cases. The *special_cases* layer is needed in order to make a distinction between hypothetical events in conditionals that do not refer to the real world and general statements that are not anchored in time, among others. This annotation can have the attribute COND_ID_CLAUSE if the event is in the "if clause" of the condition, COND_MAIN_CLAUSE if it is in the main clause, GEN for a general statement or NONE otherwise.

Factuality value. Combining the three attributes certainty, time and polarity, and taking into account the special case layer, we can determine whether the term considered refers to a fac-

tual, a counterfactual or a non factual event.

We can say that an expression refers to a FACTUAL event if it is annotated as certainty `certain`, time `non future`, and polarity `positive`, while it refers to a COUNTERFACTUAL event (i.e. an event which did not take place) if it annotated as certainty `certain`, time `non future`, and polarity `negative`. In any other combination of annotation, the event referred by the term can be considered NON FACTUAL, either because it refers to a future event, or because it is not certain (`possible` or `probable`) if the event will happen or not.

The *special cases* layer changes the status of the factuality value FACTUAL to a NON FACTUAL value, i.e. an event annotated as FACTUAL will be considered as NON FACTUAL when part of a conditional construction or of a general statement.

4 The corpus

The Ita-TimeBank is a language resource manually annotated with temporal and event information (Caselli et al., 2011). It consists of two corpora, the CELCT corpus and the ILC corpus, that have been developed in parallel following the It-TimeML annotation scheme, an adaptation to Italian of the TimeML annotation scheme (Pustejovsky et al., 2003a). The CELCT corpus, created within the LiveMemories project[2], consists of news stories taken from the Italian Content Annotation Bank (I-CAB)[3] (Magnini et al., 2006), which in turn consists of 525 news articles from the local newspaper "L'Adige"[4]. The ILC corpus is composed of 171 newspaper stories collected from the Italian Syntactic-Semantic Treebank, the PAROLE corpus, and the web.

From the Ita-TimeBank, which was first released for the EVENTI task at EVALITA 2014[5], we selected a subset of news stories to be annotated with factuality. The subset consists of 170 documents taken from the CELCT corpus and contains 10,205 events.

We annotated factuality values on top of the TimeML annotation. The TimeML specifications consider as *events* predicates describing situations that happen or occur, together with predicates describing states and circumstances. Each event is classified into one of the following TimeML classes: REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE and STATE.

In the corpus, within the 10,205 event mentions, there are 6,300 verbs, 3,526 nouns, 352 adjectives and 27 prepositions. The distribution among TimeML classes is the following: 5,292 OCCURRENCE, 2,352 STATE, 900 I_ACTION, 864 I_STATE, 439 REPORTING, 258 ASPECTUAL and 100 PERCEPTION.

With respect to the TimeML annotation, we do not annotate factuality for events of the class STATE because we do not consider it relevant for "circumstances in which something obtains or holds true" (Pustejovsky et al., 2003a). Likewise we do not annotate factuality for events of the class I_STATE because we use them to determine the certainty of their eventive argument (e.g. *sperare - hope*).

The annotation of factuality has been done for 6,989 events from 170 articles by using the CELCT Annotation Tool (Lenzi et al., 2012).

5 Results

In the following section, we report on the inter-annotator agreement and then we present a first analysis of the annotated corpus.

5.1 Inter-Annotator agreement

We have computed the agreement between two annotators on the four factuality attributes assigned to 92 events. For the agreement score we used accuracy and we computed it as the number of matching attribute values divided by the number of events. For each of the four attributes we obtained good agreement, with accuracy values over 0.91.

A study of the annotations on which we found disagreement shows that the problem stems from the `underspecified` values for time, polarity and certainty attributes. The `underspecified` value is used when it is not possible to assign another value to an attribute by using information available in the text. More precise rules should be defined in order to help annotators decide if they can use the `underspecified` value or not.

5.2 Corpus analysis

Factuality attributes have been annotated on top of 4,114 verbal events and 2,870 nominal events, for a total of 6,989 events.

[2] `http://www.livememories.org`
[3] `http://ontotext.fbk.eu/icab.html`
[4] `http://www.ladige.it/`
[5] `http://www.evalita.it/2014/tasks/eventi`

	event classes					*news topics*				
	IACT	REP	PER	OCC	ASP	Trento	Sport	Economy	Culture	News
# events	900	439	100	5,292	258	3,084	886	735	684	1,600
Factual (%)	65.2	**84.5**	66.0	69.0	65.5	68.2	**71.1**	66.4	62.9	**74.6**
Counterfactual (%)	3.8	2.7	8.0	3.8	1.6	4.5	4.4	1.4	2.5	3.5
Future - certain (%)	9.0	2.5	6.0	10.9	**21.3**	9.5	14.0	**16.9**	**16.5**	4.8
Future - uncertain (%)	**14.2**	6.6	12.0	8.9	6.6	11.6	8.5	2.4	**13.6**	7.1
Non future - uncertain (%)	2.6	0.9	2	1.8	1.9	2.7	0.8	0.5	0.3	2.1

Table 1: Corpus statistics: correlation of event factuality with event classes and news topics.

We combined the values of certainty, polarity and relative time attributes of events in order to obtain their factuality value. The factuality values were then studied in comparison with event parts-of-speech, TimeML event classes and news topics. In Table 1, we report the statistics on event factuality in the corpus.

As expected, in newspaper articles the majority of events mentioned are FACTUAL. We observed that there is a higher proportion of nominal FACTUAL events (73.8%) than verbal FACTUAL events (66.1%). On the contrary, `uncertain` events are mainly verbs.

The relation between TimeML event classes and factuality values was studied in order to determine their correlation. Some expected phenomena were observed, in particular that REPORTING events[6] are mainly FACTUAL (84.5%) because they are often used to introduce reported speech and that events of the class ASPECTUAL[7] contain a high proportion of `future` events, mainly `certain`. Considering the events of the class I_ACTION[8] it can be noted that the proportion of `uncertain` events (17%) is higher than in other classes.

The distribution of the factuality value of events in the Ita-TimeBank was also studied according to the topic of each news article considered. The news of the CELCT corpus are categorized in 5 topics: news stories, local news, economy, culture and sport.

The main distinction we observed is between cultural news and all the other kinds of news. Cultural news contains a lower proportion of FACTUAL events (62.9%) and a higher proportion of `future` events (30.1%) than the other categories of news articles, while around 14% of the event mentions in cultural news were annotated as `uncertain`. Indeed cultural news contains both reports about past cultural events and announcement of future events. On the contrary, in news stories there is a high proportion of factual events and very few future events.

6 Conclusion

In this paper we have presented an annotation schema of event factuality in Italian and the annotation task done on the Ita-TimeBank. In our schema, factuality information is represented by three attributes: time of the event, polarity of the statement and certainty of the source about the event.

We have selected from the Ita-TimeBank 170 documents containing 10,205 events and we have annotated them following the proposed annotation schema. The annotated corpus is freely available for non commercial purposes from `https://hlt.fbk.eu/technologies/fact-ita-bank`.

The resource has been used to develop a system based on machine learning for the automatic identification of factuality in Italian. The tool has been evaluated on a test dataset and obtained 76.6% accuracy, i.e. the system identified the right value of the three attributes in 76.6% of the events. This system will be integrated in the TextPro tool suite (Pianta et al., 2008).

Acknowledgments

This research was funded by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project.

[6] "REPORTING events describe the action of a person or an organization declaring something, narrating an event, informing about an event, etc." (Pustejovsky et al., 2003a)

[7] ASPECTUAL events "code information on a particular phase or aspect in the description of another event" (Caselli et al., 2011)

[8] "I_ACTION events describe an action or situation which introduces another event as its argument" (Pustejovsky et al., 2003a)

References

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.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *LREC*, pages 333–338.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation*.

Silvia Pareti and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation, LREC10*.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, pages 28–34.

James Pustejovsky, Patrick Hanks, Roser Saur, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March.

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

Sara Tonelli, Rachele Sprugnoli, and Manuela Speranza. 2014. NewsReader Guidelines for Annotation at Document Level, Extension of Deliverable D3.1. In *Technical Report NWR-2014-2*.

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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, pages 162–210.

Dina Wonsever, Aiala Ros, Marisa Malcuori, Guillermo Moncecchi, and Alan Descoins. 2012. Event Annotation Schemes and Event Recognition in Spanish Texts. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 7182 of *Lecture Notes in Computer Science*, pages 206–218. Springer.