Factuality Annotation and Learning in Spanish Texts

Dina Wonsever, Aiala Rosá, Marisa Malcuori

Instituto de Computación, Facultad de Ingeniería
Universidad de la República
Julio Herrera y Reissig 565
Montevideo, Uruguay
wonsever@fing.edu.uy, aialar@fing.edu.uy, marisamalcuori@gmail.com

Abstract

We present a proposal for the annotation of factuality of event mentions in Spanish texts and a free available annotated corpus. Our factuality model aims to capture a pragmatic notion of factuality, trying to reflect a casual reader judgements about the *realis / irrealis* status of mentioned events. Also, some learning experiments (SVM and CRF) have been held, showing encouraging results.

Keywords: Event Factuality, Annotation Scheme, Corpus Annotation, Machine Learning

1. Introduction

In automated content extraction from text data, words are a primary source of information. But the presence of a word that denotes an event, for instance a noun such as *war* or a verb such as *eat*, does not imply the actual occurrence of a war or an eating act. Factuality is the status of an event mention with regards to its effective occurrence, as perceived by a reader. In this document we describe a model for representing event factuality and a Spanish corpus annotated according to this model. As previously said, the determination of events mention (or event, for short) factuality is necessary for different purposes, like Information Extraction, Information Retrieval, Question Answering or others. Our model of factuality is included in an event annotation scheme (Wonsever et al, 2012), under the form of an attribute with six possible values.

The representation and automatic detection of event factuality has been addressed by different authors, the model proposed by Roser Saurí (2008) being the main reference in the area. In Saurí's model factuality is determined as the combination of two elements: modality (CT: certain, PR: probable, PS: possible, U: underspecified) and polarity (+: positive, -: negative, u: underspecified). The combination of these two dimensions results in the following values: CT+, CT-, CTu, PR+, PR-, PS+, PS- and Uu. Another element that Saurí includes in factuality determination is the specification of the source that presents the event. The sources are represented as nested sources, following the model for opinions and emotions from (Wiebe et al, 2005). The scheme from Saurí was applied for the automatic determination of event factuality through an algorithm that uses some lexical ressources like modality and negation markers and source introducing predicates. It was also used for annotating the FactBank corpus (Saurí & Pustejovsky, 2009, 2012).

Based on the model defined by Saurí, several proposals have emerged. Van Son et al (2014) incorporated some modifications to the English scheme proposed by Saurí, including a new dimension for temporality to distinguish future from non future events. This new version of the scheme was used to annotate the corpus of the NewsReader Project (Tonelli et al, 2014). Narita et al (2013) and Matsuyoshi et al (2010) worked on the adaptation of the scheme for Japanese. They have annotated a corpus and have worked on the automatic detection of factuality. Glavas et al (2012) presented a similar work for Croatian. Minard et al (2014) annotated an Italian corpus, using Van Son's version of the Saurí schema.

The main difference between our model and the previously mentioned proposals is that we try to capture a pragmatic notion of factuality, while Saurí's and other related models are mainly logically oriented. In our model, factuality expresses the global judgment about the occurrence of the mentioned events that a casual reader would extract from texts. For instance, Saurí explicitly states that all sources are equally credible and well informed, while this is not an assumption in our case. This fact explains that Saurí has the capability of designing an algorithmic solution combining special kinds of predicates and polarity markers, while we preferred to follow a learning approach.

2. The Values for Factuality

The values we proposed for factuality are shown in Table 1.

Value	*Description*	*Example*
R	The event has happened or is happening.	*1. El tren llegó con una hora de retraso.* *The train arrived one hour late.*
NR	The event has not happened and is not happening.	*2. El tren no logró llegar a tiempo.* *The train failed to arrive on time.*
FP	Scheduled future	*3. El tren llega a las 12 del lunes próximo.* *The train will arrive at 12:00 next Monday.*
FN	Denied future	*4. El tren no llegará en hora.* *The train will not arrive on time.*
POS	Possible	*5. El tren llegará en hora si no llueve.* *The train will arrive on time if it doesn't rain.*
IND	Undefined	*6. El tren puede haber llegado en hora.* *The train may have arrived on time.*

Table 1. Factuality values

The R and NR values indicate that the occurrence (R) or non occurrence (NR) of the referred event is for sure determined. The remaining values are associated to event mentions where the occurrence or non occurrence is not determined. One can speak of certainty, focusing on the perception of a reader / observer, about the occurrence of the event in one of its two polarities, or *realis*, as opposed to irrealis, focusing here on what happens in the real world in opposition to hypothetical statements.

In the case of future events, which clearly did not happen, we decided to mark some very evident differences. On the one hand, we define the category "scheduled future" (FP), as shown in Example (3) for scheduled events; and "denied future" (FN), as the example (4), otherwise. This treatment for events with future orientation but presented with a high degree of certainty marks a difference between our proposal and the one of Sauri, where they would be listed as Ct + (or Ct-)[1].

Notice that the factuality value associated with arrive in the example (5) is not FP, as the arrival of the train in time is not stipulated as a plan but is conditioned to another event (the absence of rain). The future conjugation of the verb can be an indicator of FP and FN values, but it is not a sufficient condition.

Furthermore, the future orientation for events also determines the distinction between the values *possible (POS)* and *undefined (IND)*. In 5 the arrival event is annotated as POS as it may still occur (conditioned on the accomplishment of the *if* clause). However in 6 it is annotated as IND because the event has eventually happened (or not happened), but that information is not deductible from the text.

3. Corpus Annotation

A corpus annotation task was carried out by two annotators, in order to generate training and testing data for the development of a factuality classifier. Table 2 shows the number of verb events from each factuality value[2].

	Annotator 1	*Annotator 2*
R	540	537
NR	47	48
FP	21	8
FN	9	12
POS	159	207
IND	99	63
Total	875	875

Table 2. Annotations *per* class

Concerning future factuality values, we can observe that the number of FP and FN events is very low, and there is a significant difference in annotations for the FP value. On the other hand, for events with certain factuality values (R and NR), which are more than half of the corpus, the annotators agreement is really high.

Regarding uncertain values, the number of POS events is high (159 / 207 events in 875), compared to Possible and Probable events in FactBank (60 Pr+, Pr-, Ps+ and Ps- events in a corpus of 2192 events). In the case of events with undefined factuality, our annotators detected 63 / 99 IND events, while in FactBank the amount of Underspecified events is much higher (804). This great number of Underspecified events can be explained by the fact that, in this corpus, event references in reported speech are assigned a factuality value from the primary source perspective and a different value (generally Underspecified) for the text writer factuality. Marneffe et al (2012) carried out some annotation experiments showing that readers perceive event factuality as absolute values, regardless of who mentions them.

In both corpora, FactBank and ours, the number of events with certain negative factuality is far lower than the number of positive events. In our case, they are fewer than POS and IND events too.

Other authors mention similar results concerning the values distribution. Narita et al (2013) report a 16.8% of

1 Other schemes [4, 2, 7] also include a time attribute to distinguish futur events from other cases.

2 In this paper we do not analyze noun or adjective events.

uncertain and possible events; Matsuyoshi et al (2010) conclude that factuality classes are very skewed, uncertain and nonfactual events being the least frequent.

Table 3 shows the confusion matrix for annotations. The global inter-annotator agreement is 90.4%. To get this value we considered one of the annotators as the *gold standard* and we calculated the accuracy of the other one.

	R	*NR*	*FP*	*FN*	*POS*	*IND*
R	501	5	0	0	24	10
NR	1	40	0	4	0	2
FP	1	0	8	0	12	0
FN	0	2	0	7	0	0
POS	13	0	0	0	140	6
IND	21	1	0	1	31	45

Table 3. Confusion matrix for annotations

As Table 3 shows, IND and POS are problematic values, it seems it is difficult to distinguish them. They are also often annotted as R (factual events).

In order to apply machine learning methods on the corpus, we decided to unify some of the values of our model, to have classes with a relevant number of elements. The classes we used for training were: R (factual events), NR (nonfactual events) and IND, which includes all events with an undetermined factuality (FP, FN, POS and IND events).

We carried out a second annotation process, performed by a new annotator, according to this new scheme with three values. The extended corpus has 2080 annotated events (see Table 4).

R	1392
NR	121
IND	567
Total	2080

Table 4. 3-values annotation

4. Automatic Determination of Factuality

We have applied two different learning methods to generate classifiers: *Conditional Random Fields* (CRF) and *Support Vector Machine* (SVM). For each method, several experiments based on different attribute sets were performed. A full description of the experiments carried out can be read in (Fernández & Fernández, 2012).

4.1 Attribute Sets

a. Standard Morpho-Syntactic Information

The basic set of attributes (included in all the experiments) consists of standard morpho-syntactic information: word; lemma; part-of-speech (POS); morphological information depending on the POS, such as gender, number, person, mood, and tense. This information is provided by the FreeLing POS-tagger (Padró & Stanilovsky). An additional attribute for the dependency relation between each word and its head, obtained from the Spanish Malt-Parser (Ref), is included.

b. Verbal Morphology

We added boolean attributes for verb mood and tense, which are especially relevant features for determining factuality.

c. Lexical Resources

We developed some lists with lexical items related to the factuality status of events:

- modal markers (*suppose, impossible, may, ...*)
- negation markers (*no*, *never, fail, ...*)
- implicative verbs (100 verbs)

Each implicative verb belongs to one of four possible classes: (+ +), (+ -), (- +), (- -). This notation indicates which factuality value (R or NR) corresponds to events under the scope of the implicative verb, depending on the polarity of the verb. For example, an implicative verb (+ +), such as *lograr/succed*, with a positive polarity implies that events under its scope are factual: (*Juan logró abrir la puerta / John succeded opening the door: lograr* has positive polarity, so the factuality for *abrir* is R). On the other hand, an implicative verb (- +), such as *dudar/hesitate*, must have negative polarity to imply that events in its scope are factual (*Juan no dudó en abrir la puerta / John didn`t hesitate to open the door*: *dudar* has negative polarity, so *abrir* is R).

For each word, some boolean attributes indicate if they belong to some of the lexical items lists. For events, additional boolean attributes indicate if there are words belonging to the lists in their dependency trees, in relevant positions.

4.2 Machine Learning Experiments

We performed several experiments for different attribute sets. Final results are showed in Table 5.

Global results are slightly better for SVM than for CRF. Major differences between the two models are found for the IND value, where SVM outperforms CRF by 6.6 points. This factuality value is the one that gets the worst results.

	Precision		*Recall*		*F-Measure*		*Total Accuracy*	
	CRF	SVM	CRF	SVM	CRF	SVM	CRF	SVM
R	89.9	**91**	92.2	**94.3**	91	**92.6**	85.1	**87.4**
NR	**86.7**	85.7	**81.2**	75	**83.9**	80		
IND	65.3	**72.9**	60.4	**66**	62.7	**69.3**		

Table 5. CRF and SVM results

In general, the base-line, that reaches an accuracy of 68.5%, is largely outperformed by the two models. The base-line classifies events following this simple algorithm:

- if the event is in a future tense, then the factuality value is **IND**
- else, if the event is preceeded in the sentence by some negative word, then the factuality value is **NR**
- else, the factuality value is **R**

In particular, for IND events the base-line reaches just a 33.6% of F-Measure. As we can see in table 5, IND is also the most difficult case for both learning algorithms and the F-Measure for IND is 62.7 in CRF and 69.3 in SVM. In table 6 we show one example where both algorithms classify an R event as IND, and another one where they assign an R value instead of IND.

*1. La encuesta nacional de Factum del otro día **viene a demostrar** que se confirma una tendencia.* *The Factum national poll from the other day **shows** that a trend is confirmed.*
*2. Javier de Haedo convocó a los votantes del Partido Colorado a que se **sumen** a su proyecto.* *Javier de Haedo called the Partido Colorado voters **to join** his project.*

Table 6. Classification problems with the IND value

5. Conclusions

The determination of the factuality status for event mentions is necessary for automatic content extraction, in tasks like Information Extraction, Information Retrieval, Question Answering or others. We focused on a pragmatic reader-oriented notion of factuality, distinguishing six different values. After annotating and conducting some learning experiments, these values were conflated in three different cases. Although some subtle distinctions have been lost (distinctions in the *irrealis* area), we believe that the obtained results could be usefully integrated in a text processing pipeline.

6. Bibliographical References

Fernández, V. and E. Fernández. (2012). Determinación de factividad de los eventos mencionados en textos. Internal Report. https://www.fing.edu.uy/inco/grupos/pln/prygrado/InformeFactividad.pdf

Glavas, Snajder, Basic. (2012). Are You for Real? Learning Event Factuality in Croatian Texts. In *Conference on Data Mining and Data Warehouses* (SiKDD 2012).

Marneffe, M.C., C. Manning, C. Potts. (2012). Did It Happen? The Pragmatic Complexity of Veridicality Assessment. In *Computational Linguistics*, Volume 38, Issue 2, June 2012, pp 301-333. MIT Press, Cambridge, MA, USA.

Matsuyoshi, S., M. Eguchi, C. Sao, K. Murakami, K. Inui and Y. Matsumoto. (2010). Annotating event mentions in text with modality, focus, and source information.. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC) 2010*, pp 1456-1463.

Minard, A-L., A. Marchetti, M. Speranza. (2014). Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. In *Proceedings of CLiC-it 2014*. Pisa, Italy.

Narita, K., J. Mizuno and K. Inui. (2013). A Lexicon-based Investigation of Research Issues in Japanese Factuality Analysis. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp.587-595.

Padró, L. and Evgeny Stanilovsky. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA. Istanbul, Turkey.

Saurí, R. (2008). A Factuality Profiler for Eventualities in Text. Doctoral Dissertation. Brandeis University Waltham, MA, USA.

Saurí, R., J. Pustejovsky. (2009). FactBank: a corpus annotated with event factuality. In *Language Resources and, Evaluation*, Volume 43, Issue 3, September 2009, pp 227-268, Springer Netherlands.

Saurí, R., J. Pustejovsky. (2012). Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. In *Computational Linguistics,* Volume 38, Issue 2, June 2012, pp 261-299. MIT Press Cambridge, MA, USA.

Tonelli, S., R. Sprugnoli, M. Speranza and A-L. Minard. (2014). NewsReader Guidelines for Annotation at Document Level. Technical Report NWR-2014-2-2. August 2014. Fondazione Bruno Kessler.

Van Son, C., M. van Erp, A. Fokkens and P. Vossen. (2014). Hope and Fear: How Opinions Influence Factuality. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Wiebe, J., T. Wilson and C. Cardie. (2005). Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation*, Volume 39, Issue 2-3, pp. 165-210. Springer Netherlands.

Wonsever, D., M. Malcuori, A. Rosá. (2009). Factividad de los eventos referidos en textos. Technical Report. ISSN 0797-6410. https://www.fing.edu.uy/inco/pedeciba/bibliote/reptec/TR0912.pdf.

Wonsever, D., A. Rosá, M. Malcuori, G. Moncecchi and A. Descoins. (2012). Event Annotation Schemes and Event Recognition in Spanish Texts. In *Computational Linguistics and Intelligent Text Processing*, 13th

International Conference, CICLing 2012. Springer, p. 206-18. (Lecture Notes in Computer Science).