Overview of FACT at IberLEF 2020: Events Detection and Classification

Aiala Rosá[a], Laura Alonso[b], Irene Castellón[c], Luis Chiruzzo[a], Hortensia Curell[d], Ana Fernandez Montraveta[d], Santiago Góngora[a], Marisa Malcuori[a], Glòria Vàzquez[e] and Dina Wonsever[a]

[a]*Universidad de la República, Montevideo, Uruguay*
[b]*Universidad Nacional de Córdoba, Córdoba, Argentina*
[c]*Universitat de Barcelona, Barcelona, España*
[d]*Universitat Autonoma de Barcelona, Barcelona, España*
[e]*Universitat de Lleida, Lleida, España*

Abstract

In this paper we present the second edition of the FACT shared task (Factuality Annotation and Classification Task), included in IberLEF2020. The main objective of this task is to advance in the study of the factuality of the events mentioned in texts. This year, the FACT task includes a subtask on event identification in addition to the factuality classification subtask. We describe the submitted systems as well as the corpus used, which is the same used in FACT 2019 but extended by adding annotations for nominal events.

Keywords

Factuality Classification, Events Identification, Corpus Annotation

1. Introduction

In this paper we describe the second edition of the FACT shared task (Factuality Annotation and Classification Task), included in the 2nd Iberian Languages Evaluation Forum (IberLEF). The main objective of this task is to advance in the dastudy of the factuality of the events mentioned in texts, seeking to contrast different approaches.

Factuality is understood, following [1], as the category that determines the factual status of events, that is, "whether event-denoting expressions are presented as corresponding to real situations in the world (i.e., as facts), to situations that have not happened or hold (counterfacts), or to situations of uncertain status (possibilities)".

In order to analyze event references in texts, it is crucial to determine whether they are presented as actually having taken place or as potential or not accomplished events. This information can be used for different applications like Question Answering, Information Extraction, or Incremental Timeline Construction.

Despite its centrality for Natural Language Understanding, event factuality has been an underresearched topic, with the work by [2] as a reference for English and [3] for Spanish. For Italian, a task

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: aialar@fing.edu.uy (A. Rosá); alemany@famaf.un c.edu.ar (L. Alonso); icastellon@ub.edu (I. Castellón); luischir@fing.edu.uy (L. Chiruzzo); Ana.Fernandez@uab.cat (H. Curell); Ana.Fernandez@uab.cat (A.F. Montraveta); sgongora@fing.edu.uy (S. Góngora); marisa.malcuori@gmail.com (M. Malcuori); gloria.vazquez@udl.cat (G. Vàzquez); wonsever@fing.edu.uy (D. Wonsever)
ORCID:

similar to FACT has been proposed in the past [4]. The bottleneck to advance on this task has usually been the lack of annotated resources, together with its inherent difficulty.

For FACT2019 [5] a corpus annotated with verbal events was available for experimentation. For FACT2020 edition, we enriched the corpus by including nominal events annotations. This year we include a second task on event identification, in addition to the factuality classification task.

In this overview we describe the new version of the corpus, we present the submitted systems and their results, and we sketch some conclusions about event and factuality identification.

2. Background

A number of categories have been proposed to classify different modes of (non-)accomplishment of events. For Spanish factuality, [6] proposes a six value scheme: Accomplished, Not Accomplished, Scheduled Future, Denied Future, Possible, and Undefined. The first four categories represent a high degree of certainty, but only Accomplished and Not Accomplished categories represent events that actually happened or not. On the other hand, Possible and Undefined categories are used for events whose occurrence is uncertain (Possible for uncertain future events and Undefined for uncertain past events).

Even though this scheme provides a detailed model for factuality, the categories are too fine-grained and some of them are underrepresented in texts, making automatic recognition difficult. For this reason a simplified scheme has been used for a corpus annotation task, reducing the categories to three values: Accomplished, Not Accomplished, and Undefined [6]. This corpus is made up of Uruguayan texts and contains 2,080 verbal events (1392 Accomplished events, 121 Not Accomplished events, and 567 Undefined events), and it is the starting point for the construction of the FACT task corpus.

Starting from the Uruguayan corpus mentioned above, prior to the start of the FACT2019 shared task, an annotation process was carried out in order to extend the corpus, and to include texts from Spain and more documents from Uruguay. An annotation guideline was provided in order to explain the meaning of the tags and the scope of the annotation.

The resulting corpus contains Spanish texts with more than 5,000 verbal events classified as F (Fact), CF (Counterfact), U (Undefined). The corpus was divided in two subcorpora: the training corpus (80%), and the testing corpus (20%). The texts belong to the journalistic register and most of them are from the political sections from Spanish and Uruguayan newspapers.

3. Corpus

For FACT2020, the corpus used in the previous edition was extended by adding nominal events. We also fixed a few errors that we found in the verbal events. An excerpt of the corpus is shown below:

> *De acuerdo con el Instituto Nacional de Sismología, Vulcanología, Meteorología e Hidrología (Insivumeh), el volcán de Fuego* <event factuality="F">*ha*</event>
> <event factuality="F">*vuelto*</event> *a la normalidad, aunque*
> <event factuality="F">*mantiene*</event> <event factuality="F">*explosiones*</event>
> *moderadas, por lo que no* <event factuality="CF">*descarta*</event>">
> *una nueva* <event factuality="U">*erupción*</event>.
>
> According to the National Institute of Seismology, Volcanology, Meteorology and Hydrology (Insivumeh), the Fire volcano has returned to normal, although it maintains moderate explosions, so it does not rule out a new eruption.

Table 1
Categories distribution and corpora sizes.

Category	Train	Test	Total
Factual-Verbs	2914	698	3612
Counterfactual-Verbs	252	66	318
Undefined-Verbs	1168	310	1478
Factual-Nouns	557	82	639
Counterfactual-Nouns	26	2	28
Undefined-Nouns	276	87	363
Total	5193	1245	6438

Categories distribution and the sizes of train and test corpora are shown in Table 1.

As can be seen, the categories are highly unbalanced in the corpus, which can difficult the recognition of the least represented classes, in particular, counterfactual nominal events.

For Task 2 evaluation, we delivered a specific corpus where events were not annotated, so the participants had to identify them. This corpus contains 391 events, 326 are verbal events and 65 are nominal events.

In the following section we describe the criteria defined for the verb and noun annotation process.

3.1. Verbal Events Annotation

In order to carry out the manual annotation, general criteria for determining the object and scope of the annotation were established, together with specific criteria to solve difficult, doubtful or conflicting cases. The annotation input was Freeling's morphological analysis, which implied taking its segmentation and categorization as the starting point. All the verbs marked as predicates by the analyzer were annotated. In the case of complex tense forms and verb periphrases, each verb was tagged individually since the analyzer marks them as two predicates (even though it recognizes the main verb). The factual status of events does not relate to the status of the event in the real world but to its perception by the human annotator. For this reason, in order to determine the factivity of an event, the annotator's world knowledge and beliefs were not taken into account, only the linguistic expression.

As mentioned above, complex verb forms were annotated individually. The values within each complex verb form may coincide (había 'had' Factual visto 'seen' Factual) or not (intentó 'tried to' Factual llegar 'arrive' Undefined). Non-finite forms of complex tense forms were always annotated with the same tag as the auxiliary (habia 'had' Factual sido 'been' Factual comprado 'bought' Factual). Specific criteria were developed for structures containing subordinate clauses, non-finite forms, interrogative sentences, complex verb forms and verb periphrases. In most cases, rules were provided, such as:

> Adverbial clauses with non-finite forms (present o past participle) can be F/CF or U, depending on the tense expressed in the main clause.
>
> - *Una vez hechos (U) los deberes comeré (U) algo*
>
>   Once my homework is done (U) I'll eat (U) something

- *Una vez hechos (F) los deberes comí (F) algo*

  Once my homework was done (F) I ate (F) something

- *Come (F) viendo (F) la televisión*

  He eats (F) watching (F) TV

- *Comerá (U) viendo (U) la televisión*

  He will eat (U) watching (U) TV

The criteria for tagging verbs in interrogative sentences takes into consideration the type of interrogative: direct or indirect and total or partial.

> Interrogative sentences, both direct and indirect, if they are total interrogatives, they are always undefined (U). If they are partial interrogatives, the general criteria for verb tense forms are applied.

- *¿Vino (U) María?*

  Did Maria come (U)?

- *¿Cómo lo ha (F) hecho (F) Juan?*

  How did John do (F) it?

- *¿Dónde lo compraste (U)?*

  Where did you buy it (U)?

- *¿Qué harás (U) mañana?*

  What are you doing (U) tomorrow?

- *Juan se pregunta (F) si María lo ha hecho (U)*

  John wonders (F) if Mary has done (U) it

- *Juan no sabe (CF) a qué hora llegó (F) María*

  John doesn't know (CF) at what time María arrived (U)

- *1. Juan no sabe (CF) a qué hora llegará (U) María*

  John doesn't know (CF) at what time Mary will arrive (U)

3.2. Nominal Events Annotation

Nominal events can be expressed in a single lexical item or a multi-word expression (always denoting one entity), e.g., 'press conference,' in which case the whole phrase had to be identified. One of the most important problems was sense disambiguation. Some lexical items have more than one sense, and not all of them allow an eventive interpretation; in (1a) the process is described (event), whereas in (1b) the final product (non-event) is:

1.a. *Esta página está en construcción.*
This page is under construction.

1.b. *¿Qué es esa construcción de madera?*
What's that wooden construction?

When nominal events modify another noun, then they were only considered if the noun they modified, the head, was also a nominal event (2a); otherwise they were not annotated (2b):

2.a. *una investigación sobre la inmigración*
an investigation on immigration

2.b. *el documento sobre inmigración*
a document on immigration

The tag used to identify nominal events is the same as for verbs, but it specifies that it is a noun: <event type="noun" factuality=" "> *text* </event>. The factuality values used for nominal events were the same proposed for verbs: F (fact), CF (counterfact) and U (undefined) and the interpretation was contextual as was the case with verbs:

- *Desde enero de este año coordinaba el* <event type="noun" factivity=”F”> *proyecto* </event>.

  Since January of this year he/she coordinated the <event type="noun" factivity=”F”> project </event>.

- *El próximo año empezará una* <event type="noun" factivity=”U”> *investigación* </event> *federal sobre la trama rusa.*

  Next year a new federal <event type="noun" factivity=”U”> investigation </event> of the Russian plot will begin.

4. FACT 2020: Factuality Analysis and Classification Task

4.1. Task 1: Factuality Classification

In this task facts are not verified in regard to the real world, just assessed with respect to how they are presented by the source (in this case the writer), that is, the commitment of the source to the truth-value of the event. In this sense, the task could be conceived as a core procedure for other tasks such as fact-checking and fake-news detection, making it possible, in future tasks, to compare what is narrated in the text (fact tagging) to what is happening in the world (fact-checking and fake-news).

We established three possible categories:

- Facts: current and past situations in the world that are presented as real.
- Counterfacts: current and past situations that the writer presents as not having happened.
- Possibilities, future situations, predictions, hypothesis and other options: situations presented as uncertain since the writer does not commit openly to the truth-value either because they have not happened yet or because the author does not know.

And their respective tags:

- F: Factual
- CF: Counterfactual
- U: Undefined

The participating systems had to automatically propose a factual tag for each event in the text. Since event identification is not the scope of the task 1, the events are already annotated in the texts. The structure of the tags used in the annotation is the following:
`<event factuality="F">`*verb*`</event>`
For example, in a sentence such as:

> *El fin de semana* `<event factuality="">`*llegó*`</event>` *a Uruguay el segundo avión de la aerolínea.*
>
> The second plane of the airline arrived in Uruguay on the weekend.

The systems outcome should be:

> *El fin de semana* `<event factuality="F">`*llegó*`</event>` *a Uruguay el segundo avión de la aerolínea.*

The performance of this task was measured against the evaluation corpus using these metrics:

- Precision, Recall and F1 score for each category.
- Macro-F1.
- Global accuracy.

The main score for evaluating the submissions is Macro-F1.

4.2. Task 2: Events Identification

The second subtask proposed in FACT2020 was an experiment on event identification. In this year's edition, the systems had to recognize events expressed in verbal and nominal expressions, that is, those events expressed in nouns, such as 'destruction' or 'meeting', and in verbs, such as 'destroy' or 'meet'.

The performance of this task was measured against the evaluation corpus using Precision, Recall and F1 score.

4.3. Participating Teams and Systems Description

The different approaches submitted for each task are described in the subsections below.

4.3.1. Task 1

For Task 1 there were six participants whose systems are described below.

The first 4 systems share the same data preprocessing step, which consists of generating a sentence for each sentence containing an event. Therefore, a sentence with n events will be represented as n different sentences.

1. **t.romani** [7]) proposes a system based on word embeddings and RNN. Each word is represented as a 300-dimensional vector using word2vec; after generating the vector space, a 301th bit is added to indicate if the word represented was an event. After normalizing the length of each sentence, a 200 neurons GRU layer with a 3-dimensional output layer is used for classifying each event-word in the three possible categories.

2. **guster** [7]) proposes a system based on word embeddings, morphological analysis and a SVM classifier. The morphological analysis is performed as follows. Given an event, after POS-tagging the sentence where it is located, a vector is built regarding the grammatical category of the words surrounding it. The window size and the scores to build the vector were selected empirically in the parameter-tuning phase. Then for each word representation, its word2vec vector was concatenated. For performing the event classification into 3 classes, a SVC classifier was used, also previously configurated in the parameter-tuning phase.
3. **aaccg14** [7]) also proposes a system based on RNN but this time at char level. Each word representing an event is split into a char list, as well as the surrounding words in order to keep context information. Then, using one-hot-encoding, vectors are built marking with a flag the chars corresponding to the event. In order to perform the event classification a sequential model is used, consisting of two LSTM layers with a 3-dimensional output layer, as in the first approach.
4. **trinidadg** [7]) proposes a system based on a Random Forest classifier and morphological analysis. For performing the event classification, the input for the Random forest classifier are the resulting vectors of the same morphological analysis used in the second approach.
5. **premjithb** uses a Random Forest classifier with word2vec features of dimension 300. They also employed a cost-sensitive learning approach for avoiding any sort of imbalance in the data.
6. **garain** [8]) proposes a method for determining the factuality value for previously recognized events (nominal and verbal) using an SVM classifier, fed with BERT sentence embeddings, word2vec embeddings, sentiment words classes, subjectivity status of the sentence, TF-IDF for frequent words, and normalized auxiliary counts.

4.3.2. Task 2

Only one participant submitted results for Task 2. The system proposed by **trinidadg** uses a simple rule approach for tagging as events the following words:

- All verbs detected by Stanford tagger.
- All nouns that appear as events at least once in the training corpus. This was inspired by the baseline system.

5. Global Results

Table 2 shows the results in terms of Macro-F1, Macro-Precision, Macro-Recall and Accuracy for the teams that participated in Task 1. The baseline approach is a simple heuristic which assigns random factuality values with the following probabilities: F-70%, U-20%, CF-10%.

We can observe that the best results were obtained by as system that uses RNN and pretrained word embeddings, followed by a system based on pretrained word embeddings, morphological analysis and SVM, which achieved fairly close results. Unlike what was observed in FACT 2019 [5, 9], in this edition the experiments carried out with Random Forest did not achieve good results.

Table 3 shows the results for Task 2 in terms of F1, Precision and Recall. The baseline assigns the class ‘event’ to the words (either verb or noun) tagged as ‘event’ at least once in the training corpus.

The only system submitted for this task clearly outperforms the baseline, specially in terms of precision. This can be explained in part because all verbs are tagged as events, so the performance of the system for verbs must be almost as good as the POS tagger performance.

Table 2
Results for Task 1.

Method	Macro-F1	Macro-P	Macro-R	Accuracy
t.romani	**60.7**	61.2	**60.4**	**84.8**
guster	59.3	**62.1**	57.4	83.1
accg14	55.0	55.6	54.5	79.8
trinidadg	53.6	55.8	52.0	80.6
premjithb	39.3	45.5	37.6	71.6
garain	36.6	35.7	39.4	59.9
baseline	24.6	25.4	25.1	52.4

Table 3
Results for Task 2.

Method	F1	Precision	Recall
trinidadg	86.5	95.1	79.3
baseline	59.7	60.3	59.1

6. Conclusions

The FACT shared task (first and second editions) is an important opportunity to work on the extension and revision of an existing factuality corpus, and to perform some experiments on factuality recognition.

In this second edition we generated a corpus with verbal and nominal events annotated with their factuality category on the basis of a three valued scheme: Factual, Counterfactual, Undefined. It should be noted that the corpus is very unbalanced, both in terms of the category of events (1030 nouns and 5408 verbs), and in terms of the three factuality classes (4251 Factual, 346 Counterfactual, 1841 Undefined). This fact poses an extra challenge for experimentation with supervised methods.

The systems presented at FACT 2020 outperformed the results obtained in the 2019 edition. It is important to note, however, that the test corpus is not the same for both editions: FACT2019 corpus had only verbal events while FACT2020 corpus has nominal and verbal events.

The best results for task 1 were obtained by a system based on a RNN architecture, fed by 300 dimension word embeddings, extended by an extra value indicating if the word is an event. For task 2, only one system was submitted, which significantly outperformed the baseline despite its simplicity, showing promising results on events identification.

Some research directions we would like to pursue in the future include using the more complex six-valued annotation schema, and extending the corpus in order to have a more representative number of nominal events.

References

[1] R. Saurí, A Factuality Profiler for Eventualities in Text, Brandeis University, 2008.

[2] R. Saurí, J. Pustejovsky, Factbank: a corpus annotated with event factuality, Language resources and evaluation 43 (2009) 227.

[3] D. Wonsever, M. Malcuori, A. Rosá, Factividad de los eventos referidos en textos, Reportes Técnicos 09-12 (2009).

[4] A.-L. Minard, M. Speranza, T. Caselli, The EVALITA 2016 Event Factuality Annotation Task (FactA), in: Proceedings CLiC-it 2016 and EVALITA 2016, CEUR Workshop Proceedings, CEUR-WS, Napoli, Italy, 2016.

[5] A. Rosá, I. Castellón, L. Chiruzzo, H. Curell, M. Etcheverry, A. Fernández, G. Vázquez, D. Wonsever, Overview of FACT at IberLEF 2019 (2019).

[6] D. Wonsever, A. Rosá, M. Malcuori, Factuality annotation and learning in spanish texts., in: LREC, 2016.

[7] A. Collazo, A. Rieppi, T. Romani, G. Trinidad, FACT2020: Factuality Identification in Spanish Text (2020).

[8] B. Ray, A. Garain, Factuality Classification Using BERT Embeddings and Support Vector Machines (2020).

[9] B. Premjith, K. P. Soman, P. Poornachandran, Amrita CEN@FACT: Factuality Identification in Spanish Text, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain, 2019.