Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 111-121 recibido 19-12-2021 revisado 20-01-2022 aceptado 25-01-2022

A Methodology for the Automatic Annotation of Factuality in Spanish

Una metodología para la anotación automática de la factualidad en español

Irene Castellón Masalles,[1] **Ana Fernández Montraveta**[2] **Laura Alonso Alemany**[2]
[1]Universitat de Barcelona
icastellon@ub.edu
[2]Universitat Autònoma de Barcelona
Ana.Fernandez@uab.cat
[3]Universidad Nacional de Córdoba, Argentina
lauraalonsoalemany@unc.edu.ar

Abstract: In the last decade, factuality has undeniably been an area of growing interest in Natural Language Processing. This paper describes a rule-based tool to automatically identify the factual status of events in Spanish text, understood with respect to the degree of commitment with which a narrator presents situations. Factuality is represented compositionally, considering the following semantic categories: commitment, polarity, event structure, and time. In contrast with neural machine learning approaches, this tool is entirely based on manually created lexico-syntactic rules that systematize semantic and syntactic patterns of factuality. Thus, it is able to provide explanations for automatic decisions, which are very valuable to guarantee accountability of the system. We evaluate the performance of the system by comparison with a manually annotated Gold Standard, obtaining results that are comparable, if not better, to machine learning approaches for a related task, the FACT 2019 challenge at the IBERLEF evaluation forum.
Keywords: Factuality, event annotation, lexico-syntactic patterns, rule-based systems.

Resumen: La información factual es un área de investigación de creciente interés en el Procesamiento del Lenguaje Natural. Este artículo describe una herramienta basada en reglas para la identificación automática en español de la clase factual de los eventos en un texto, entendida con respecto al grado de compromiso con el que un narrador presenta las situaciones. En esta aproximación la información factual se representa compositivamente, considerando las siguientes categorías semánticas: compromiso, polaridad, estructura del evento y tiempo. A diferencia de los enfoques de Machine Learning, esta herramienta se basa por completo en reglas léxico-sintácticas y semánticas creadas manualmente que sistematizan los patrones semánticos y sintácticos de la información factual. Así, este sistema es capaz de proporcionar explicaciones para las decisiones automáticas, que son muy valiosas para garantía de la responsabilidad del sistema. Evaluamos el rendimiento del sistema mediante la comparación con un Gold Standard anotado manualmente, obteniendo resultados que son comparables, si no mejores, a los enfoques de aprendizaje automático para una tarea relacionada: el reto FACT 2019 del foro de evaluación IBERLEF.
Palabras clave: Factualidad, anotación de eventos, patrones lexico-sintácticos, sistemas basados en reglas.

1 Introduction and Motivation

The identification of factuality in corpora, i.e., recognizing the factual status of propositions, has been a research area of growing interest in Natural Language Processing (NLP) (Saurí, 2008; Saurí and Pustejovsky, 2009; Diab et al., 2009; Narita, Mizuno, and Inui, 2013; Soni et al., 2014), among others). In our project, following Saurí (2008)'s proposal, events' factual status is understood as the degree of commitment with which situations are presented by the narrator of the text.

The detection of this type of semantic information is extremely relevant for the semantic interpretation of texts and constitutes the base of several more complex processes and applications, such as fact-checking, fake news detection or information retrieval, among others, that need to be able to differentiate situations described as real from utterances of opinion or belief.

The work we present in this paper aims to build an automated annotator of factuality for Spanish texts exclusively based on linguistic knowledge. Unlike other annotators (Wonsever, Rosá, and Malcuori, 2016; Diab et al., 2009; Huang et al., 2019), our methodology only uses contextual linguistic knowledge, our algorithmic solution is solely based on linguistic cues. Currently, most automatic analyses of language are approached with purely statistical methods, machine learning using word embeddings, large neural language models and classifiers to perform the task (Wonsever, Rosá, and Malcuori, 2016; Huang et al., 2019; Qian et al., 2019; Rosá et al., 2020).

However, systems based on neural networks are obscure artifacts, which do not allow practitioners to understand how a given annotation has been made. For applications involving critical decision making, like fact checking, explainability is a must (Minh et al., 2021). It is important to be able to assess why a given text might be expressing a fact or a speculation, in order for people to ground their decisions with all relevant information. Indeed, regulations around the world are beginning to require that automated decision making systems can account for how decisions were reached, as in Spanish so-called rider law, which requires that companies disclose algorithms that they use to make decisions concerning labour rights [1]. In contrast with neural-based approaches, rule-based systems built upon relevant linguistic concepts provide adequate explanations, understandable by users. Some machine learning approaches, like decision trees or logistic regressions, can also provide some interpretability with respect to their decisions, however, they are dependant on big amounts of annotated text. Such big amounts of annotated text are usually not available, and, moreover, they may contain stereotypes and biases that are subsequently reproduced and amplified by the technologies based upon them, and are very difficult to detect and mitigate. In contrast, rule-based systems allow for explicit policing of biases, which makes it easier to implement positive policies and existing regulations.

Since the method this paper presents is solely based on linguistic knowledge, a prior thorough analysis of texts has been necessary to be able to identify the relevant knowledge, formalize it and systematize it as a system of rules. These rules basically exploit lexico-syntactic and morphological information that is able to capture the relevant semantic and syntactic phenomena that are related to factuality.

As will be developed below, this approach reaches good performance as evaluated on a gold standard test dataset. The domain chosen for this project, the written press, presents wide lexical diversity but is less complex in terms of the syntactic structures used, which we believe facilitates the approach of this task by means of conditions-actions.

Besides providing a tool for automated factuality analysis for Spanish, the good performance obtained by this approach allows us to automatically generate annotated corpora which will help compensate for the lack of annotated corpora at the factuality level and, in general, at the whole semantic level, for languages other than English.

The rest of the paper is organized as follows. In the next section, we describe the categories and tagset to annotate the different aspects of factuality. Then, Section 3 presents the methodology. In Section 4, we discuss some experiments to assess the performance of the system and analyze the results obtained, both quantitatively and qual-

[1] https://www.boe.es/diario_boe/txt.php?id=BOE-A-2021-7840

itatively. We then conclude with some future directions for this work.

2 Aspects of factuality to be annotated

To begin with, only declarative sentences are considered for annotation, given that interrogative or exclamatory sentences never assert facts. Only propositional content is annotated, not implications or implicatures. However, not all declarative sentences are annotated with respect to factuality, since statements describing desires or some conditional situations, for example, are not. The speaker of these types of sentences is not committing themselves to the truth of the proposition by asserting them because they do not describe 'real' situations (*realis*) and cannot be, therefore, said to be true or false (*irrealis*).

In this project, facts are understood as those present or past situations, presented by the author with commitment (that is, they are depicted as true by the speaker), and marked with positive polarity. Those situations that share these characteristics but are expressed with negative polarity are understood to depict counterfacts. All the rest of the situations belong to future or uncertain worlds and, therefore, are not considered facts. Event structure helps us to determine the eventive or stative nature of the situation described.

We use the tagset proposed within the TagFact project (Alonso Alemany et al., 2018) to describe these different aspects of facts or counterfacts, detailed in Table 1.

The combination of these four levels of linguistic description contribute to the factual interpretation of a proposition:

(1) *La presidenta madrileña que ha hecho de su hiperactividad mediática su capital político...*[2]
'The Madrid president who has made her political capital out of her media hyperactivity...'
Commitment, past, positive, event

(2) *Durante toda la jornada en la sede del Gobierno en la Puerta del Sol se aguardó por una rueda de prensa que nunca se produjo.*[3]
'Throughout the day at the Government headquarters in Puerta del Sol, they waited for a press conference that never took place.'
Commitment,past, negative, event

(3) *La Guardia Civil citará, además, como investigado para este jueves al exvicepresidente autonómico y exdirector general de la policía...*
'The Civil Guard will cite, in addition, as investigated for this Thursday the former regional vice president and former general director of the police...'
Commitment, future, positive, event

-2***Commitment**	Commitment
	Non-commitment
-2***Polarity**	Positive
	Negative
-3***Reference time**	Past
	Present
	Future
-5***Event Structure**	Event
	Mental
	Property-event
	Property-non-event
	Absolute truth

Table 1: Labels used in TagFact.

This tagset can be mapped easily to the ones used in other projects (Saurí, 2008) or (Wonsever, Rosá, and Malcuori, 2016). Below, Table 2 presents the mapping between our outcome and three other standard tagsets, Factbank, Fact Task (Rosá et al., 2019; Rosá et al., 2020) and (Qian et al., 2019), which is also based on FactBank.

This section has briefly described the tagset we used, proposed by the TagFact project, and how it relates to other similar projects. For a more detailed characterization of the tags and contexts see (Vázquez García and Montraveta, 2020).

[2] https://www.eldiario.es/politica/horas-Cifuentes-apago-focos_0_753474740.html

[3] https://www.elperiodico.com/es/politica/20180522/zaplana-detenido-por-la-guardia-civil-6832200

TagFact	**FACT task (Iberlef)**	**FactBank**	**Qian et al.**
Commitment Positive Present & Past Event - State	FACT	CT+ (certain) (incl. future situations)	CT+
Commitment Negative Present & Past Event - State	COUNTERFACT	CT-	CT-
Future Non-Commitment bv Event - State	UNDEFINED[4]	PR (probability) PS (possibility) U (undefined) PSu	PS[5]
Not applicable	UNDEFINED	U (undefined)	U

Table 2: Mapping of tagsets between TagFact and Fact Task, FactBank and Quian et al. (2019).

3 Methodology

The process leading to the creation of the automatic analyzer has been developed in the following three phases:

1. **Linguistic analysis and formalization**: The first stage consisted in the classification and characterization of the linguistic phenomena in factuality found in the TagFact corpus, a set of articles collected from several Spanish newspapers (Alonso Alemany et al., 2018; Fernández-Montraveta et al., 2020). This analysis has provided the necessary input for the implementation of the rules.

2. **Automatic processing**: This stage consisted in two steps. First, finding and evaluating the tools required for the linguistic preprocessing of the text and, second, the implementation and prioritisation of the rules. In this second step, we followed an incremental methodology: rules were developed and tested continuously, with a benchmark of representative cases for immediate assessment of the impact of each new rule, reordering or refactoring of the modules. Members of the implementation team regularly met with members of the linguistic analysis team to include new characterizations of the targeted phenomena, assess unclear cases and add further cases to the assessment benchmark.

3. **Evaluation**: Three different evaluations have been carried out, two during the development phase and the last one once the implementation was completed. The first kind of evaluation was qualitative, aimed to assess the impact of changes in the implementation: while rules were being developed, they were continuously tested against a benchmark obtained from the development corpus and some cases included ad hoc to monitor the behavior of the tool with respect to some phenomena of particular interest. The other two evaluations were quantitative and performed automatically by comparison with a manually tagged corpus. For this purpose the TagFact corpus, totalling 59.514 words, was divided into two parts, a development corpus with 82,6% of the total corpus, with 49.202 words, without manual annotations, and a test corpus, the Gold Standard, a 17,3% of the total corpus, with manual annotations. The test corpus was divided in two parts, the first consisting of three articles (1,9% of the total corpus, 1.141 words) used for the evaluation of the first implementation of the annotator, and the second (15,4% of the total corpus, 9.171 words) was reserved for the final evaluation. The evaluation was carried out quantitatively and qualitatively, the latter performed by members of the linguistic analysis team.

4 *Architecture of the automatic annotator for factuality*

The automatic annotation system detects candidate facts (situations) in text as any string tagged as a verb by the Freeling morphosyntactic annotator, and, for each situation, it assigns a value for each of the aspects of factuality described in Section 2. As a result of the combination of these aspects, the factual value of the situation is determined. The system is rule-based and works with the linguistic information available in the scope of the sentences. The rules have been implemented in Python3.6.

Within the first step of the second phase described in Section 3, the starting step of the automatic process consists in a morphosyntactic analysis of the text is carried out with Freeling (Padró et al. 2012). The output format of the analysis chosen is ConLL.As a result of this pre-process, all the lexical items that are annotated as verbs are identified as candidate facts.

The annotator of factuality consists of a sequence of sub-processes that incrementally characterise the different aspects of factuality for each of the identified situations. In the final stage, each different combination of values for the different aspects provides a standard factual value (Fact, Counterfact, Undefined).

The modules that make up the process are the following:

Module 1 selects the situations to be further annotated. In this module hypotheses, conditions and unreal worlds (irrealis) are discarded.

Module 2 assigns a polarity label to those situations selected in module 1.

Module 3 assigns a degree of commitment with which the situation is presented.

Module 4 is in charge of the analysis of referential time.

Module 5 assigns to each situation the label corresponding to the type of event denoted.

4.1 Applies

The first task is to decide whether factual analysis is applicable or not to each predicate. This is one of the most complex analyses since sentences may be describing situations in irreal worlds. In conditional constructions for example, not all hypotheses express unreal situations. A sentence, such as (8), is expressing two counterfacts, which are consistent with the real world, not situations in irreal worlds.

(4) *Si hubiera estudiado, habría aprobado.*
'Had I studied, I would have passed.'

The values assigned by this module are: Applies, Does_not_Apply and Non_pred, for those lexical items wrongly annotated as predicates in the pre-process. This assignment requires a complex analysis, which has been developed in three parts:

Local annotation: annotation of predicates in simple sentences. It takes into account the verb tense of the predicate and its context, except in the case of complex sentences with subordinate clauses, where there may be interference between the different predicates.

Conditional annotation between events: annotation inferred from the interaction of predicates between main sentences and subordinate sentences. For certain cases of non-personal forms, an inheritance mechanism has been implemented between the non-personal form and the verb that governs it.

Univocal predicate annotation: finally, some lexical items marked as predicates by the morphosyntactic pre-process are fixed forms that should not be annotated. This is the case of mira (look), used as an exclamation that formally corresponds to the imperative form of mirar (to look).

The linguistic information this module requires are: verbal tense, adverbs, conjunctions, prepositions, constructions, syntactic dependencies and syntactic functions. A total of 274 rules were developed, covering more phenomena than those appearing in the development corpus. This is because in the phase devoted to the linguistic analysis, other sources were consulted and an attempt was made to generalise the rules.

The predicates annotated with the category Applies continue the annotation process in the following modules. Events anno-

tated with the category Does_not-Apply and Non_pred are not further analyzed.

4.2 Polarity

The module dealing with the annotation of Polarity is composed of 38 rules. The label Positive is applied by default unless some triggers for Negative polarity are found in the co-text. Some examples of these kinds of triggers are adverbs of negation (5) such as no, nunca (never) or jamás (never, at no time), dependencies and syntactic functions to identify subjects or determinants (3 and 4) -nadie, ninguno, ningún + Noun (nobody, none, no + Noun) and some verb tenses (past perfect subjunctive -see (1)).

(5) *...creen que el dinero realmente nunca regresó a las arcas públicas.*[6]
'...they believe that the money never returned to the public coffers'

(6) *Nadie pone en duda la gran capacidad de trabajo que siempre ha demostrado Calvo,...*[7]
'Nobody doubts the great capacity for work that Calvo has always shown,...'

(7) *Aunque esta posibilidad en ningún momento ha sido confirmada.*[8]
'Although this possibility has never been confirmed.'

4.3 Commitment

The module in charge of assigning a Commitment value has 89 rules. It focuses on detecting expressions of doubt or uncertainty from triggers such as: creo que (I believe that), quizás (maybe), seguramente (surely), parece que (it seems that), etc. Some lexico-syntactic patterns restricted to some items have also been considered. Patterns such as the following:

(8) Existe (there exist) + det + Noun[trigger] + de (of) + que (that)
No + V0 + duda de que (there is no doubt that)
Verbs of opinion + que (that) +Verb

help us detect expressions such as sentences in (6-7):

(9) *Existe la certeza de que acudieron de noche.*
'There is a certainty that they came at night'

(10) *No le cabe la menor duda de que los empleados robaron en la sede.*
'He has no doubt that the employees robbed the headquarters.'

(11) *Considera que la solución no fue buena.*
'He considers that it was not a good solution.'

4.4 Time

In order to annotate referential time a total of 40 rules have been created. These rules deal with the recognition of referential time of simple and compound verb tenses, verb periphrases and non-personal verb forms. Besides, some rules have been developed to account for syntactic dependencies, as for example, a verb of communication in the present, if it has an animated subject refers to a past event (12):

(12) *Esa es su intención, afirma decidido, "cuando todo pase".*[9]
'That is his intention, he affirms decisively, "when everything passes".'

4.5 Event

Last, the module Event allows us to distinguish, basically, between states and events (50 rules). This module works with lists of event types. We distinguish between events, such as aprobar (pass), mental events, such as considerar (consider) and states such as tener (have). Starting from this category, the rules apply from triggers such as frequency adverbs (cada día, -every day) or specific verb forms (suele -used to or hay -there is). Some of the rules have to consult the analysis of syntactic dependencies (10) so that the category takes into consideration the co-text:

[6] https://www.elperiodico.com/es/politica/20181008/guardia-civil-desvela-gasto-32000-euros.prostibulo-fundacion-empleo.andalucia-7077756

[7] https://www.publico.es/politica/carmen-calvo-sera-vicepresidenta-del-gobierno.ministra-igualdad.html

[8] https://www.larazon.es/internacional/la-union-europea-y-reino-unido-podrian-haber.alcanzado-un-acuerdo-sobre-el-brexit.JE20174409/

[9] https://www.eldiario.es/desalambre/despues-juicio-queremos-mediterraneo-central_1_2138530.html

If a communication verbs has an inanimate subject the predicate is annotated as a state (property_non_event)

(13) *El artículo explica muy detalladamente el proceso de unificación.*
'The article explains in great detail the unification process.'

One problem that still remains to be addressed is differentiating between states (property_non_event) (14) and absolute truths or beliefs (15). Up to this moment we have not been able to formally differentiate them since, generally speaking, they share the same structure.

(14) *El presupuesto es alto.*
'The budget is high.'

(15) *La tierra es redonda.*
'The Earth is round.'

5 Gold Standard

The performance of the automatic annotator was evaluated by comparing automatic predictions against a manually annotated Gold Standard corpus.

We use a part of the Gold Standard corpus created within the TagFact project (Curell et al., 2020). It is composed of 22 press articles from Spanish generalist newspapers[10]. It contains a total of 10.272 words collected between June and September of 2020 (a mean of 553.7 words per article). The articles were mostly extracted from the Politics Section (70%) with the remaining 30% from other sections such as Economy, Sports or Technology, among others.

The corpus was first morpho-syntactically parsed and predicates were automatically identified by Freeling. Of a total of 1.696 words automatically marked as predicates only 1.319 remained after the manual phase. Then they were manually labelled (Section 2.1) by six senior linguists.

The interrater reliability was measured using Cohen's Kappa coefficient (Fernández-Montraveta Castellón in press) scoring 0,61 for the category Applies, 0,64 for Time, 0,55 for Polarity, and 0,35 for Event. The kappa value of the Commitment module could not be calculated because of lack of examples of one of the categories (Non-commitment).

[10]The articles were extracted from the following Spanish newspapers: ABC, El Diario, El Periódico and La Vanguardia.

	accuracy	**support**
Applies	81,1%	127
Time	81,35%	59
Commitment	100%	59
Polarity	98,30%	59
Event	66,1%	59

Table 3: Performance of automatic annotation in comparison with the Gold Standard.

6 Evaluation

We present here the mid-term evaluation of the project, aimed to detect how to improve the automatic annotator. This evaluation was carried out with a corpus manually annotated to calculate the inter-annotator agreement, with 127 predicates (Fernández-Montraveta et al., 2020). In what follows, we present a quantitative (6.1) and qualitative analysis (6.2) of results comparing the automatic and manual annotation.

6.1 Quantitative Analysis

Table 3 shows the general results of the comparison between the Gold Standard and the outcome of the automated process:

As can be observed, Commitment (100%) and Polarity (98,3%) are the categories showing the best behavior in terms of agreement. Second, the annotation of Applies and Time could be improved since both show around 81% accuracy, which is not a bad result but leaves room for improvement. Finally, Event is the category that shows the worst agreement rate. This fact could be explained because, first, it is the module that has more categories, some of them holding a type-subtype relation and, second, as mentioned above, the formal marks between some of them are blurred.

In a more detailed analysis, we can see the performance across the different classes for each level of analysis, as displayed in Table 3. We can appreciate that some of the proposed categories have not been evaluated because they were not found in the Gold Standard. This is the case of: non-commitment, future, property-event, mental and absolute truth.

Concerning the distinction between Applies / Does_not_Apply, we can see that the class Applies presents an F1 of 0.76, with high (0.92) recall but somehow lower 0.75 precision. Conversely, the category

'Does_not_Apply' shows a good precision (0.85) but recall drops (0.69).

Something similar happens for the Predicate / Non-Predicate distinction , with 0.90 precision but recall below 0.70. In this sense, the detection of predicates that require a factuality annotation needs to be improved.

Regarding verb tenses, present (F1 0.88) and past (F1 0.82) show good performance, with complementary distributions of precision and recall that suggest that errors in one category are confusions with the other, that is, predicates that should have been labelled as present are labelled as past and vice versa. Improvement is needed again in recall for the past tense and precision for the present. Future is underrepresented in the corpus so F1 cannot be calculated.

The annotation of Polarity reaches a very good performance, with 0,99 positive and 0,90 negative F1), as is the case with the Commitment tags (that reach 100%). Examples of non-commitment are not represented in the corpus.

Lastly, the category that shows poorer results is Event. Events have an acceptable 0.76 F1 but States perform much worse, with an F1 of 0.58, and the rest of the stative categories not even represented.

For the sake of comparison with related tasks, we have translated the annotations in the Gold Standard Corpus to the Iberlef FACT task, following the correspondence shown in Table 2. Results with this tagset can be seen in Table 5. The obtained F1 macro average is 75.6, which is better than the results obtained by machine learning approaches within the Iberlef FACT Task 1, shown in Table 6, albeit with a different corpus. We will apply the final version of this annotator to the Iberlef FACT corpus to have a more comparable assessment of performance.

Task 2 of the FACT 2019 challenge, Event Identification, is comparable to the Predicate aspect identified by our analyzer. Again, our results are comparable to those obtained by machine learning systems. We obtain 77% F1, while the only participating system for this task at FACT 2019 obtains 86.5% F1 and the baseline obtains 60%.

Therefore, our rule-based approach is competitive with machine learning approaches for a similar task, if not performing better. Nonetheless, the quantitative analysis shows that there is ample room for improvement. The categories requiring most effort to improve are Applies, Predicate and Event, and Time to a lesser extent. In what follows we carry out an analysis of errors on those categories to determine how to improve the performance of the analyzer.

6.2 Qualitative Analysis

We have carried out a systematic analysis of the cases where the automatic annotation fails, which has allowed us to create an inventory of system errors and elaborate a classification of the cases in which the analyzer fails. In order to present this classification, we describe the errors for each category.

6.2.1 Applies

The greatest number of errors in this category are predicates that should be tagged as Does_not_apply but are instead tagged as Applies. This is the case of some verb periphrases, infinitive clauses and conditional structures. Some of these problems, like the right interpretation of conditional sentences in example (17) are not an easy task to formalize in a systematic rule:

(16) *Si no se produce un acuerdo para devolver de oficio los intereses demás cobrados, el cliente bancario que esperaba la sentencia europea tiene la oportunidad de reclamar. "Primero tiene que hacerlo por vía extrajudicial, acudiendo al defensor del cliente,...".*
'If no agreement is reached to return ex officio the interest charged, the bank client who was waiting for the European judgment has the opportunity to claim. "First you have to do it extrajudicially, going to the client's ombudsman, ..

Other cases difficult to treat are infinitive clauses that do not inherit the category of the main verb because of errors in the pre-process of automatic parsing or the lack of a rule that runs through the syntactic structure.

(17) *Las entitades financieras han aprovechado la indefinición jurídica en torno a la retroactividad de las cláusulas para plantear a sus clientes cambios....*
'Financial entities have taken advantage of the legal uncertainty around the retroactivity of the

	Precision	Recall	F1
Applies			
Applies	0,753	0,920	0,828
Does_not_Apply	0,853	0,686	0,760
Predicate			
Non_pred	0,909	0,667	0,769
Time			
Present	0,795	1	0,886
Past	1	0,708	0,829
Future	na	na	na
Polarity			
Positive	1	0,981	0,990
Negative	0,833	1	0,909
Commitment			
commitment	1	1	1
non-commitment	na	na	na
Event			
event	0,851	0,696	0,766
property-non-event	0,555	0,625	0,588
property-event	na	na	na
mental event	na	na	na
absolute truth	na	na	na

Table 4: Precision, recall and F1 of the different classes for each category. When no cases were found in the corpus, "na" is reported.

	Prec	Rec	Acc	F1
Counterfact	0,833	0,555	0,954	0,667
Fact	0,924	0,710	0,812	0,803
Undefined	0,694	0,943	0,806	0,800

Table 5: Precision, recall and F1 of the automatic annotator in the Gold Standard corpus, where categories have been translated to the Iberlef Fact Task Category.

Participant	Macro-F1
t.romani	60.7
guster	59.3
accg14	55.0
trinidadg	53.6
premjithb	39.3
garain	36.6
FACT_baseline	24.6

Table 6: Results obtained in FACT 2020 Task 1, Factuality Determination.

clauses to propose changes to their customers.'

Thus, it seems difficult to address these errors in the next version of the annotator. Other errors, however, can be addressed, by incorporating additional rules to the annotator. For example, past participles premodified by a determinant ("lo cobrado" – *what is charged*) are currently tagged as Applies but a rule will be added so that they are tagged as Does_not_Apply. Additional rules will be incorporated to treat some modal periphrases that have been incorrectly labelled as Does_not_Apply.

6.2.2 Time

Most of the errors in the detection of the referential time are produced by the rule that

asserts that, in the press domain, a diction-communication verb with an animated subject [human], although a present indicative morphologically, is assigned a past value in the category referential time (18).

(18) *"Hay mucha información útil en YouTube, pero también mucha información errónea", afirma en declaraciones a The Guardian la profesora y autora del estudio Asheley LLandrum, ...* [11]
'"There is a lot of useful information on YouTube, but also a lot of misinformation," professor and study author Asheley LLandrum affirms (told) The Guardian, ...'

This temporal change does not happen when the entity is inanimate (19).

(19) *Las entrevistas realizadas a estas personas demuestran, según el estudio de la Texas_ Tech_ University, que la mayoría basan sus creencias en los vídeos que han visto en YouTube.*[12]
'The interviews carried out with these people show, according to the Texas_ Tech_ University study, that most base their beliefs on the videos they have seen on YouTube'.

In order to apply this rule, a list of animated entities was created, but still the rule fell short to account for the following cases:

- Some entities denoting collective entities were not in the list of animated entities, although they behave as such with respect to time: associations or offices, among others. These will be included in the updated list of animated entities.
- When ellided subjects were not retrievable, the rule could not apply properly.
- Errors in the syntactic pre-processing to detect the subject.
- Verbs that were not in the list of diction-communication, which will be included in the updated list for the improved version of the annotator.

[11] https://www.lavanguardia.com/tecnologia/20190219/46572983466/asi-alimenta-youtube-teorias-afirman-tierra-plana.html

[12] id. supra

6.2.3 Event

It is the category where the most errors have been detected. The annotation of this category has required creating lists of verbs lexically classified as states or events as the basis of the rules. That notwithstanding, contextual information might change the lexical event structure. In general, errors in this module come from the following factors:

- The verb of the sentence is not in the corresponding list. These have been included in the improved version.
- Lack of specific rules: for example, an inanimate object plus a communication verb produces a stative interpretation. This rule has been included in the improved version of the annotator.
- Some words were not included in the list of animated entities, and thus the relevant contextual rules could not be applied. They have now been included.

Another source of error with respect to events is that the detection of a special subkind of states, namely absolute truths, is beyond the scope of the automatic analyzer. This, however, cannot be properly addressed in the updated version of the analyzer either.

7 Conclusions and Future Work

In this paper we have presented a symbolic, rule-based system to automatically annotate factuality in Spanish text. Factuality is annotated compositionally, distinguishing different aspects of its semantics: commitment, time, eventuality and polarity. The total number of rules developed is 491, where 274 deal with searching for annotable candidates and 217 rules annotate values for the four categories.

We have shown that this approach performs comparably, if not better to machine learning approaches for the same task, but it still has room for improvement. An extensive error analysis shows where to direct efforts for future improvements, by including further rules or enhancing lists of words. Limitations of the approach have also been clearly depicted, for example, lack of accuracy due to errors in the morphosyntactic pre-processing. A future version of the analyzer will include these improvements, and will be evaluated in a holdout annotated dataset, as well as in the standard Iberlef FACT dataset.

References

Alonso Alemany, L., I. Castellón Masalles, H. Curell, A. Fernández Montraveta, S. Oliver, and G. Vázquez García. 2018. Proyecto tagfact: Del texto al conocimiento. factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural*, 61:151–154.

Curell, H., G. Vázquez, I. Castellón, A. Fernández-Montraveta, and L. Barrios. 2020. Un gold standard sobre factualidad para el español. In *III Congreso Internacional de Lingüística Computacional y de Corpus*, Colombia. Universidad de Antioquía.

Diab, M., L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Fernández-Montraveta, A., H. Curell, G. Vázquez, and I. Castellón. 2020. The tagfact annotator and editor: A versatile tool. *Research in Corpus Linguistics*, 8(1):131–146, May.

Huang, R., B. Zou, H. Wang, P. Li, and G. Zhou, 2019. *Event Factuality Detection in Discourse*, pages 404–414. 09.

Minh, D., H. Wang, Y. Li, and T. Nguyen. 2021. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 11.

Narita, K., J. Mizuno, and K. Inui. 2013. A lexicon-based investigation of research issues in japanese factuality analysis. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan*, pages 587–595.

Qian, Z., P. Li, Q. Zhu, and G. Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2799–2809, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Rosá, A., L. Alonso, I. Castellón, L. Chiruzzo, H. Curell, A. Montraveta, S. Góngora, M. Malcuori, G. Vàzquez, and D. Wonsever. 2020. Overview of fact at iberlef 2020: Events detection and classification. *CEUR Workshop Proceedings*, 2664:197–205.

Rosá, A., I. Castellón, L. Chiruzzo, H. Curell, M. Etcheverry, A. Fernández, G. Vázquez, and D. Wonsever. 2019. Overview of fact at iberlef 2019 factuality analysis and classification task. *CEUR Workshop Proceedings*, 2421:105–110.

Saurí, R. and J. Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Saurí, R. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, 01.

Soni, S., T. Mitra, E. Gilbert, and J. Eisenstein. 2014. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland, June. Association for Computational Linguistics.

Vázquez García, G. F. and A. M. Montraveta. 2020. Annotating factuality in the tagfact corpus. In M. Fuster-Márquez, C. Gregori-Signes, and J. S. Ruiz, editors, *Multiperspectives in Analysis and corpus design*. Comares, Granada. ISBN 9788413690094.

Wonsever, D., A. Rosá, and M. Malcuori. 2016. Factuality annotation and learning in Spanish texts. In *LREC'16*, pages 2076–2080, Portorož, Slovenia, May. European Language Resources Association (ELRA).