Overview of FACT at IberLEF 2019
Factuality Analysis and Classification Task

Aiala Rosá[1], Irene Castellón[2], Luis Chiruzzo[1], Hortensia Curell[3], Mathias Etcheverry[1], Ana Fernández[3], Gloria Vázquez[2], and Dina Wonsever[1]

[1] Universidad de la República, Uruguay
{aialar,luischir,mathiase,wonsever}@fing.edu.uy
[2] Universidad de Barcelona, España
icastellon@ub.edu,gvazquez@dal.udl.cat
[3] Universidad Autónoma de Barcelona, España
Ana.Fernandez,Hortensia.Curell@uab.cat

1 Introduction

In this paper we describe the FACT shared task (Factuality Annotation and Classification Task), included in the First Iberian Languages Evaluation Forum (IberLEF).

Factuality is understood, following [6], as the category that determines the factual status of events, that is, whether events are presented or not as certain. In order to analyze event references in texts, it is crucial to determine whether they are presented as 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, this task has been underresearched, with the work by [7] as a reference for English and [8] for Spanish. For Italian, a task 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. Currently PLN-InCo and GRIAL both have ongoing research projects on this topic, which are producing and will produce such annotated resources. This makes the proposal of this task even more interesting.

The main objective of this task is to advance in the study of the factuality of the events mentioned in texts, seeking to contrast different approaches. To accomplish this task a corpus annotated with factuality information is available allowing experimentation with supervised machine learning techniques.

2 Background

A number of categories have been proposed to classify different modes of (non-)accomplishment of events. For Spanish factuality, [9] 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 [9]. This corpus is made up of Uruguayan texts and contains 2,080 events (1392 Accomplished events, 121 Not Accomplished events, and 567 Undefined events).

3 Task Description

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 this task, 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.

3.1 Corpus

Starting from the Uruguayan corpus with 2,000 events mentioned above, prior to the start of this 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. An excerpt of the corpus is shown below:

Y otras generaciones que <event factuality="F">*han*</event> <event factuality="F">*vivido*</event> *más relajadamente no* <event factuality="CF">*están*</event> <event factuality="CF">*viendo*</event> *la importancia de* <event factuality="U">*luchar*</event> *para* <event factuality="U">*mantener*</event> *esa libertad de expresión...*
And other generations that have lived more relaxed are not seeing the importance of fighting to maintain that freedom of expression.

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 class (counterfactual events).

Category	Train	Test	Total
Factual	2917	702	3619
Counterfactual	255	66	321
Undefined	1171	307	1478
Total	4343	1075	5418

Table 1. Categories distribution and corpora sizes.

3.2 Participating Teams and Systems Description

There were five participating teams, one of them (garain) did not send us any description, so it is not included in this section. The systems presented by the remaining four teams are described below:

1) **Amrita CEN** (Premjith, Soman and Prabaharan [5]) proposes a system based on word embeddings using a Random Forest classifier. Taking into account the differences in the number of appearances of the different factual labels in the corpus, the implementation assigns a higher weight to the minority label (CF) and a lower one to the more frequent labels in order to improve the prediction of the less frequent categories .

2) **jimblair** (Mao and Zhong [3]) proposes the use of BERT[1], a multi-layer bidirectional transformer encoder. For this task, a BERT-Base, multilingual cased model was choosed. For the training process, the corpus was divided in two parts, Uruguayan texts and Spanish texts. The training was made for the two models independently and predicts the categories for each subcorpus. In the last step, both outputs, the Spanish and Uruguayan subcorpora, are combined in order to create the final annotation.

3) **Aspie96** (Giudice [2]) system is a character-level convolutional recurrent neural network which makes no use of pre-trained features (such as word embeddings), nor of additional knowledge or intuition about the task, but takes advantage of tokenization to classify individual words within the text. Each word is represented as a fixed-size list of vectors, each of which represents an individual character with left and right context characters added, eventually not belonging to the word. An event flag is added to indicate whether the word is an event or not. In a final step a dense layer is applied to get, for each word, its classification in one of the three classes.

4) **macro128** (Pastorini) uses SentencePiece[4], a language independent character based tokenizer, in a pre-processing phase. It is used on the training corpus to generate tokens related to the task and then to tokenize the validation / testing corpus. The pre-trained BERT language model is used, with one end layer that classifies each token among the three possible categories (F, CF or U). As not all words were initially classified, each token is randomly assigned a category. The size of the input layer (responsible for generating embeddings) is reduced to make it compatible with the amount of tokens generated in the pre-processing

[4] https://github.com/google/sentencepiece

stage. The model was first trained without the classification layer, until convergence, for a maximum of 100 epochs (using early stopping). The entire model was then trained to converge for a maximum of 100 epochs using F1 measure for early stopping.

3.3 Global Results

Table 2 shows Macro-F1 and Accuracy for the participating teams, and a simple baseline which assigns random factuality values with the following probabilities: F-70%, U-20%, CF-10%.

Team	Macro-F1	Accuracy
Amrita CEN	**0.561**	**0.721**
Aspie96	0.554	0.635
jimblair	0.489	0.622
macro128	0.362	0.579
baseline	0.340	0.524
garain	0.301	0.512

Table 2. Participants results and baseline.

The best results were obtained by Amrita CEN, whose approach is based on Random Forest and word embeddings. The remaining approaches are based on different Deep Learning models, jimblair and macro128 apply variants of BERT model, Aspie96 uses a recurrent CNN. Previous work reached higher Macro-F1 (80%) and Accuracy (87%) with a sequential version of the SVM model, training on the previous version of the corpus [9]. These previous results are not entirely representative given that the corpus used for evaluation was significantly smaller than the one used in FACT, in particular, the CF class had only 16 occurrences.

4 Conclusions

We presented the first edition of the FACT shared task, that was an important opportunity to work on the extension and revision of an existing factuality corpus, and to perform experiments on factuality recognition.

Three models based on Deep Learning were presented, but the best results were reached by a system based on Random Forest and word embeddings.

Some research directions we would like to pursue in the future include using the more complex six-valued annotation schema, and including another subtask for recognizing events (verb and noun events) together with their factuality value, instead of having the events pre-annotated in the corpus.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 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). pp. 4171–4186 (2019)
2. Giudice, V.: Aspie96 at FACT (IberLEF 2019): Factuality Classification in Spanish Texts with Character-Level Convolutional RNN and Tokenization. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
3. Mao, J., Liu, W.: Factuality Classification using the Pre-trained Language Representation Model BERT. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
4. Minard, A.L., Speranza, M., Caselli, T.: 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. Premjith, B., Soman, K.P., Poornachandran, P.: 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 (9 2019)
6. Saurí, R.: A Factuality Profiler for Eventualities in Text. Brandeis University (2008)
7. Saurí, R., Pustejovsky, J.: Factbank: a corpus annotated with event factuality. Language resources and evaluation **43**(3), 227 (2009)
8. Wonsever, D., Malcuori, M., Rosá Furman, A.: Factividad de los eventos referidos en textos. Reportes Técnicos 09-12 (2009)
9. Wonsever, D., Rosá, A., Malcuori, M.: Factuality annotation and learning in spanish texts. In: LREC (2016)