Factuality Classification Using the Pre-trained Language Representation Model BERT

Jihang Mao[1], Wanli Liu[2]

[1] Montgomery Blair High School,
51 University Blvd E, Silver Spring, MD 20901, USA
[2] TAJ Technologies, Inc.,
7910 Woodmont Ave #1214, Bethesda, MD 20814, USA

jim-blair@hotmail.com

Abstract. In this paper we report our participation in the 2019 FACT (Factuality Analysis and Classification Task) challenge task, where a corpus containing texts with verbal events is provided and systems need to automatically propose a factual tag for each event. 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. Therefore it is important to find indications of the linguistic context surrounding the events. Our approach utilizes BERT, a multi-layer bidirectional transformer encoder which can help learn deep bi-directional representations of texts, and the pre-trained model is fine-tuned on training data for FACT. The representations of an event and its sentence are fed into an output layer for classification. Our approach achieves encouraging results in evaluation, which demonstrates that it is competitive and applicable to multilingual text categorization tasks.

Keywords: BERT; Factuality Detection; Text categorization; Multilingual Model; Evaluation.

1 Introduction

With the exponential growth of user-generated content, rumors in social media platforms are widely noticed. In a Pew Research Center poll, 64% of US adults said that “made-up news” has caused a “great deal of confusion” about the facts of current events [1]. However, identifying the factual status of events early is a hard task without sufficient evidence such as responses and fact checking sites. Automating the fact-checking pipeline is rather challenging, despite the recent progress in natural language processing, databases and information retrieval [2]. Many prior studies began by manually inspecting tweet messages in the training dataset to come up with an initial human-curated list of word features. It was found that these words could be categorized into meaningful groups. Such “cue words” have been reported to be useful in identifying an author’s certainty in journalism, determining veracity of rumors and detecting disagreement in online dialogue [3-5].

It is crucial to determine whether event references are presented as having taken place or as potential or not accomplished events. Despite its centrality for Natural Language Understanding, this task has been under-researched, with [6, 7] as a reference for English and [8] for Spanish. Besides its inherent difficulty, the bottleneck to advance on this task has usually been the lack of annotated resources. Following Sauri [9], factuality is understood as the category that determines the factual status of events, i.e., whether events are presented or not as certain. Adopting the Sauri model with some changes, Wonsever et al. [10] create an annotated corpus with factuality information and an automatic annotation tool based on automatic supervised learning. Alonso et al. [11] create a tool for the annotation of factuality expressed in texts in Spanish through automatic processing, which is carried out from three different axes: multilevel, multidimensional and multitextual.

FACT (Factuality Analysis and Classification Task) is a task to classify events in Spanish texts (from Spanish and Uruguayan newspaper), according to their factuality status. The goal of FACT is the determination of the status of verb events with respect to factuality in Spanish texts. In this task, participating teams are given a text with its events already identified, and required to automatically assign a factuality category to each one of the events. Current and past situations in the world that are presented as real will be categorized into Fact, while situations that the writer presents are not having happened in real world will be categorized into Counterfacts. Situations presented as uncertain will be categorized into a class that includes a number of other values like different kinds of Future, Potential or Undefined [10]. Their tags are F (Fact), CF (CounterFact) and U (Undefined) respectively.

A brief description of our method for FACT task is presented in Section 2. In Section 3 we show the results of our method on the official FACT test datasets. In section 4 we present a discussion of the results and conclusions of our participation in this challenge.

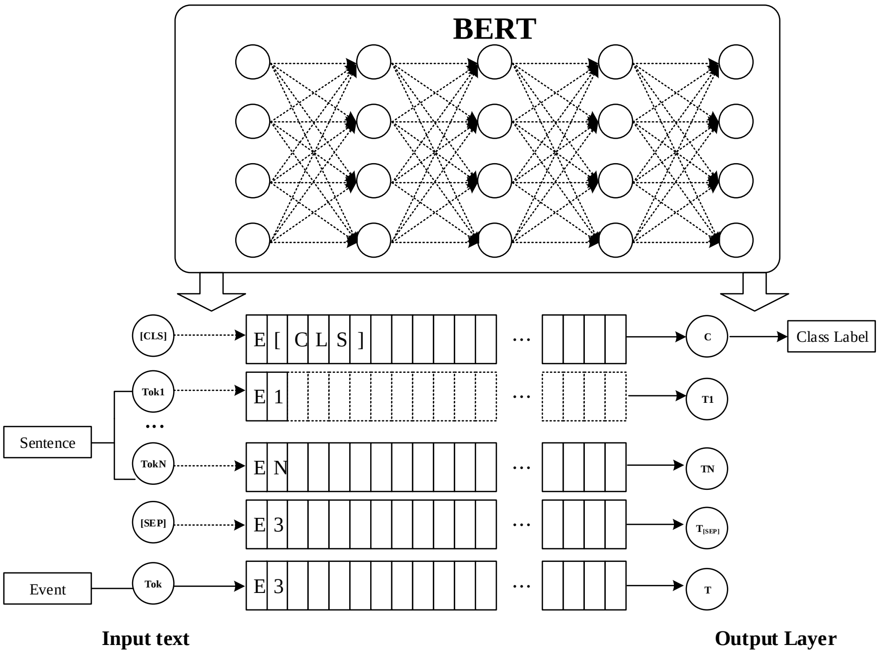
2 Methods

For FACT Task, our method builds on BERT, which has obtained state-of-the-art performance on most NLP tasks [12]. More specifically, given a sentence, our method first obtains its token representation from the pre-trained BERT model using a case-preserving WordPiece model, including the maximal document context provided by the data. Next we formulate this as a sentence-pair classification task by feeding the representations of the event and its sentence into an output layer, a multiclass classifier over the factual tags. Finally, we combine the outputs of models for Spanish and Uruguayan texts to generate the result.

BERT utilizes a multi-layer bidirectional transformer encoder which can learn deep bidirectional representations and can be later fine-tuned for a variety of tasks such as text classification. Before BERT, deep learning models, such as convolutional neural network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) have greatly improved the performance in text classification over the last few years [13]. OpenAI GPT [14] has proved the effectiveness of the generative pre-training language model.

The pre-trained BERT models are trained on a large corpus (Wikipedia + BookCorpus). There are several pre-trained models release. In FACT, we chose BERT-Base, multilingual cased model for following reasons: First, multilingual model is better for Spanish documents in FACT because the English-only model splits tokens not available in its vocabulary into sub-tokens, which will affect the accuracy of the classification task. Second, although BERT-Large generally outperforms BERT-Base in English NLP tasks, BERT-Large versions of multilingual models haven't been released. Third, the multilingual cased model fixes normalization issues in many languages, so it is recommended in languages with non-Latin alphabets (and is often better for most languages with Latin alphabets). In FACT, we use the final hidden state corresponding to a special token ([CLS]) as the aggregate sequence representation, then feed it into an output layer for classification (Figure 1).

Fig. 1. Architecture of our model for sentence pair classification. Similar to [11], we denote input embedding as E, the final hidden vector of the special [CLS] token, and the final hidden vector for the ith input token as Ti



In addition, in order to address the issue of local multilinguality, i.e. the differences of the texts from Spanish and Uruguayan newspaper, we build models for Spanish and Uruguayan texts respectively. We train the two models and predict the factual tags with corresponding training and testing texts. We then combine the outputs of the two models to generate the final results.

3 Results

The FACT corpus contains Spanish texts with approximately 5,000 verbal events classified as F (Fact), CF (Counterfact), and U (Undefined). It has been divided into two subsets: the training corpus with 4,000 events, and the testing corpus with 1,000 events.

In FACT, the performance will be measured against the evaluation corpus using the following metrics: Precision, Recall and F1 score for each category, Macro-F1, and Global accuracy. Macro-F1 is the main measure for this task. Here we present the results on the test set. In our best submission, the model was fine-tuned using the hyperparameter values suggested in [12]: learning rate (Adam) = 2e-5, number of epochs=3, max sequence length=256, and batch size=16. When fine-tuning the model for Spanish texts, we divided the training set into two subsets: 1,671 events from 20 articles for training, and 336 events from 6 articles for development. To fine-tuning the model for Uruguayan texts, 1,679 events from 22 articles is for training, and 657 events from 8 articles is for development.

As shown in table 1, our best submission significantly outperformed the baseline "fact" in both Macro-F and Accuracy, while the Macro-F score of our submission is not very far from the highest score (-0.072). We are in third place among all participants, which demonstrates a good performance of our system in automatically classifying events in Spanish texts according to their factuality status.

Table 1: Official final results for FACT

Systems	Macro-F	Accuracy
Our proposal	0.489	0.622
Baseline	0.340	0.524
Best team	0.561	0.721
Runner-up	0.554	0.635

However, although the performance of our system is reasonable on accuracy compared to other systems (0.099 and 0.013 behind the top two systems respectively), it is far from the accuracy we achieved on the development set (0.622 vs. 0.825). Table 2 shows the accuracy of the models for Spanish texts, Uruguayan texts and mixed texts on corresponding development set. The gap of performance might be caused by the differences between the training and testing sets or over-fitting of the models. We will conduct a further error analysis after the Gold Standard classifications of the test set are released.

Table 2: The Accuracy of models fine-tuned with different texts and evaluated on corresponding development set

Models	Accuracy
Spanish texts	0.840
Uruguayan texts	0.835
Mixed texts	0.825

4 Discussion & Conclusion

We described our approach that participated in the FACT: Factuality Analysis and Classification Task in IberLEF 2019. Compared to previous methods, our approach has several significant differences from system architecture to the actual implementation. It is a general and robust framework and showed competitive performance among all participating systems during the FACT evaluations. In future work, we will use a new set of random seeds each time to prevent over-fitting, and plan to explore its use in practical applications such as fact-checking and fake-news detecting.

Acknowledgements

The authors would like to thank Dr. Yutao Zhang for providing Jihang Mao the intern opportunity at George Mason University and valuable suggestions and comments on the manuscript. The authors would also like to thank the FACT task organizers for providing the data of the task.

References

1. Pew Research Center: Many Americans Believe Fake News Is Sowing Confusion. https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion (retrieved on June 21, 2019)

2. Vlachos, A. and Riedel, S. Fact checking: Task definition and dataset construction. Association for Computational Linguistics, page 18, (2014)

3. Soni, S., Mitra, T., Gilbert, E. and Eisenstein, J. Modeling factuality judgments in social media text. In ACL (2). pages 415–420. (2014)

4. Reichel, U., Lendvai, P.: Veracity Computing from Lexical Cues and Perceived Certainty Trends. Proceedings of the 2nd Workshop on Noisy User-generated Text, 4-13 (2016)

5. Misra, A. and Walker, M.A. Topic independent identification of agreement and disagreement in social media dialogue. In Conference of the Special Inte Group on Discourse and Dialogue. page 920. (2013)

6. Saurí, R., & Pustejovsky, J. FactBank: a corpus annotated with event factuality. Language resources and evaluation, 43(3), 227. (2009)

7. Gorrell, G., Aker, A., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., & Zubiaga, A. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 845-854). (2019)

8. Wonsever, D., Malcuori, M., & Rosá Furman, A. Factividad de los eventos referidos en textos. Reportes Técnicos 09-12, Pedeciba. (2009)

9. Saurí, R. A Factuality Profiler for Eventualities in Text. Ph.D. Thesis. Brandeis University. (2008)

10. Wonsever, D., Rosá, A., & Malcuori, M. (2016). Factuality Annotation and Learning in Spanish Texts. In LREC. (2016)

11. Alonso, L., I. Castellón, H, Curell, A. Fernández-Montraveta, S. Oliver, G. Vázquez. "Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español", Procesamiento del Lenguaje Natural, 61, p. 151-154. ISSN: 1135-5948. (2018)

12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, pages 4171–4186. Minneapolis, Minnesota, USA. (2019)

13. Zhang, T., Huang, M., & Zhao, L. Learning structured representation for text classification via reinforcement learning. In Thirty-Second AAAI Conference on Artificial Intelligence. (2018)

14. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)