ELiRF-UPV at IroSvA: Transformer Encoders for Spanish Irony Detection

José-Ángel González, Lluís-Felip Hurtado and Ferran Pla

VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Spain
{jogonba2,lhurtado,fpla}@dsic.upv.es

Abstract. This paper describes the participation of ELiRF-UPV team at the three subtasks proposed at IroSvA 2019 shared task. We have developed a model based on Transformer Encoders and Spanish Twitter embeddings learned from a large amount of tweets downloaded at our laboratory. Transformer Encoders are able to model long-range complex relationships among terms in a text without convolutional or recurrent layers. We addressed the three subtasks, related to three Spanish variants, using the same model. The results obtained on the validation corpus seems to confirm the adequacy of the proposed model for the irony detection task. In the final ranking, our proposal is the only system that consistently outperforms the baselines of the organizers, being the first ranked system by a considerable margin of Macro F_1 averaged on the three subtasks.

Keywords: IroSvA19 · Irony · Spanish Variants · Transformer Encoders

1 Introduction

Irony is a rhetorical device in which words are used in such a way that their intended meaning is different from the actual meaning of the words. The automatic detection of irony is an emerging topic in many natural language processing tasks. It has important implications in the final performance of some applications that need automatic processing of texts, mainly if a semantic analysis is required. For example, in tasks of sentiment analysis, polarity tends to change when irony is used. In the Semeval workshop framework, tasks such as *Task-11: Sentiment Analysis of Figurative Language in Twitter* at SemEval 2015 [3], or *Task 3: Irony Detection in English tweets* at SemEval 2018 [11] have been proposed to quantify the impact of figurative language on the Sentiment Analysis task for the English language.

In this paper, we describe the main characteristics of the system designed by the ELiRF-UPV team to address the tasks proposed at the IroSvA 2019

shared task [9]. IroSvA is focused on Spanish language from Spain, Mexico and Cuba. The task is structured into three subtasks, each one for predicting whether messages are ironic or not in one of the three Spanish variants.

2 System

In this section, we discuss the system architecture proposed to address all three IroSvA19 sub-tasks as well as the description of the resources used and the preprocessing applied to the tweets.

2.1 Resources and preprocessing

In order to learn a word embedding model for Twitter in Spanish, we downloaded 87 million tweets of several Spanish variants. To provide the embedding layer of our system with a rich semantic representation on the Twitter domain, we use 300-dimensional word embeddings extracted from a skip-gram model [8] trained with the 87 million tweets by using Word2Vec framework [4].

We have applied the same preprocessing to all the given data, both the tweets used to learn the Word2Vec embeddings model and those provided by the organization to learn the irony detection model. Firstly, a case-folding process is applied to all the tweets; Secondly, we tokenized the tweets by using TokTokTokenizer from NLTK. Thirdly, user mentions, hashtags and URLS are replaced by three generic-class tokens (*user*, *hashtag* and *url* respectively); Finally, elongated tokens are diselongated allowing the same vowel to appear only twice consecutively in a token (e.g. *jaaaa* becomes *jaa*).

2.2 Transformer Encoders

Our irony detection system is based on the Transformer [12] model. Initially proposed for machine translation, the Transformer model dispenses with convolution and recurrences to learn long-range relationships. Instead of this kind of mechanisms, it relies on multi head self-attention, where multiple attentions among the terms of a sequence are computed in parallel to take into account different relationships among them.

Concretely, we use only the encoder part in order to extract vector representations that are useful to determine the presence of irony. We denote this encoding part of the Transformer model as Transformer Encoder. Figure 1 shows a representation of the proposed architecture for irony detection.

The input of the model is a tweet $X = \{x_1, x_2, ..., x_T : x_i \in \{0, ..., V\}\}$ where T is the maximum length of the tweet and V is the vocabulary size. This tweet is sent to a d-dimensional fixed embedding layer, E, initialized with the weights of our embedding model. Moreover, to take into account positional information we also experimented with the sine and cosine functions proposed in [12]. After the combination of the word embeddings with the positional information, dropout [10] was used to drop input words with a certain probability p. On top of these

representations, Nx transformer encoders are applied which relies on multi-head scaled dot-product attention. To do this we used an architecture similar to the one described in [12]. It includes the layer normalization [1] and the residual connections.

Due to a vector representation is required to train classifiers on top of these encoders, a global average pooling mechanism was applied to the output of the last encoder, and it is used as input to a feed-forward neural network, with only one hidden layer, whose output layer computes a probability distribution over the the two classes of the task $\mathbb{C} = \{Ironic, NoIronic\}$.

We use Adam as update rule with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and Noam as learning rate schedule [12] with 15 *warmup_steps*. Weighted cross entropy is used as loss function due to the distribution of the classes is biased towards the $NoIronic$ class in a proportion of 2:1 on all the given corpora. The same proportion is used as weight terms for cross entropy loss function.

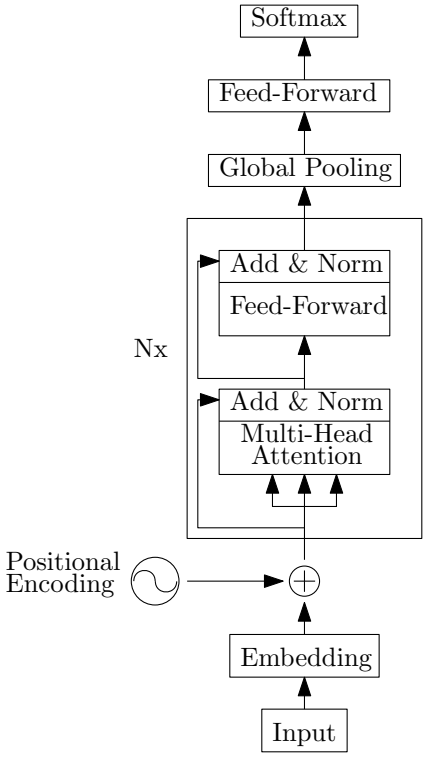

Fig. 1: The Transformer Encoders system for IroSvA19.

3 Experiments

The three subtasks proposed at IroSvA19 have the same goal: determine if a text sample is ironic or not according to a given context. The differences among them are the Spanish variant in which the text is written and the kind of text to be classified. Subtask A aims to detect irony in Spanish tweets from Spain, Subtask B aims to detect irony in Mexican Spanish tweets, and Subtask C aims to detect irony in Spanish news comments from Cuba. In this work we used only the text sample, dispensing with the context.

In order to address the three subtasks, IroSvA19 organization provided three sample sets (one per subtask). Each set is composed by 2,400 labeled documents; 1,600 of which are labeled as $NoIronic$ and the remaining 800 are labeled as $Ironic$. We divided each set provided by the organization into two subsets, a training set of 2,100 samples and a development set of 300 samples. To do this, we selected 200 $NoIronic$ and 100 $Ironic$ samples for development, maintaining the 2:1 imbalance towards the $NoIronic$ class both in the training and the development sets.

During the training phase, we fixed some hyper-parameters, concretely: $d_k = 64$, $d_{ff} = d$, $T = 50$, $h = 8$ and $batch\_size = 32$. Another hyper-parameters such as p on $warmup\_steps$ were set following some preliminary experiments to $p = 0.7$ and $warmup\_steps = 15$ epochs.

Moreover, we compare our proposal, which is based on Transformer Encoders (TE), with another deep learning systems such as Deep Averaging Networks (DAN) [7] and Attention Long Short Term Memory Networks [6] (Att-LSTM) that are commonly used in related text classification tasks obtaining very competitive results [5].

Also it is interesting to observe how some system mechanisms, like the positional encodings, or hyper-parameters like Nx affect to the results obtained in terms of macro-F_1 (MF_1), macro-recall (MR), macro-precision (MP) and class level metrics ($(F_{1i}, P_i, R_i) : i \in 0 : NoIronic, 1 : Ironic$). Concretely, we tried to remove the positional information and $1 \leq Nx \leq 2$ encoders. All these variants are applied only to the spanish subtask and the best two configurations are used also in the remaining subtasks. All these results are shown in Table 1.

As shown in Table 1, for the 1-TE-Pos and 2-TE-Pos systems, the positional encoding information harms the performance of the system. Moreover, the results obtained with 1-TE-Pos are very similar to those obtained with 1-layer Att-LSTM, that seems to indicate that the positional information, by using positional encodings or the internal memory of the LSTM, is not useful for the Spanish subtask.

It is interesting to see that when the positional information is not used, only one encoder behaves well, however, using $Nx = 2$ in this case, hurts the performance of the system in comparison to $Nx = 1$. This effect does not happen when the positional information is considered, which seems to indicate that a large number of parameters are required to take into account the positional information.

Table 1: Results on the dev set for several evaluation metrics on the three subtasks.

	F_{10}	F_{11}	MF_1	P_0	P_1	MP	R_0	R_1	MR
					ES				
DAN	85.78	69.79	77.78	84.13	**72.83**	78.48	**87.50**	67.00	77.25
Att-LSTM	81.05	66.05	73.54	84.32	61.74	73.03	78.00	71.00	74.50
1-TE-NoPos	**85.79**	**74.18**	**79.98**	**88.77**	69.91	**79.34**	83.00	**79.00**	**81.00**
1-TE-Pos	81.63	65.38	73.51	83.33	62.96	73.15	80.00	68.00	74.00
2-TE-NoPos	84.05	69.27	76.65	85.13	67.62	76.37	83.00	71.00	77.00
2-TE-Pos	82.64	62.83	72.74	80.86	65.93	73.40	84.50	60.00	72.25
					MX				
DAN	77.20	58.88	68.04	80.11	55.26	67.69	74.50	63.00	68.75
1-TE-NoPos	**79.59**	**62.91**	**71.25**	**82.35**	**59.29**	**70.82**	**77.00**	**67.00**	**72.00**
					CU				
DAN	77.86	51.85	64.85	75.83	55.06	65.44	80.00	49.00	64.50
1-TE-NoPos	**82.41**	**65.35**	**73.88**	**82.83**	**64.71**	**73.77**	**82.00**	**66.00**	**74.00**

The system 1-TE-NoPos outperforms the other systems, almost in all metrics, except on the precision over the class 1 and the recall over the class 0 with respect to DAN. Moreover, the F1 over the class 0 is very similar between both systems. However, the improvements provided by the 1-TE-NoPos system ($\sim$4.5 points of F1 on the class 1, precision and recall on the class 0 as long as the improvement of 12 points in the recall of the class 1) make this system more competitive than DAN in terms of the macro metrics.

Then, due to these two systems are the most competitive on the development set of the ES subtask, we experimented with these architectures in the other subtasks to observe their behaviour.

On the MX subtask, the results between both systems are similar, obtaining again the system 1-TE-NoPos the best results in all the metrics. However, on the CU subtask, the differences among the results of both systems are bigger, with improvements of $\sim$9 points of MF_1, MP and MR.

Finally, our best system 1-TE-NoPos (ELiRF-UPV) is used to label the test set. The results obtained are shown in Table 2. Our system outperforms the proposed baselines [2] in all the metrics, by a margin of 2 to 4 points in terms of MF_1 with respect to the best baseline. Moreover, Figure 3 shows the best five participants in the final ranking of the competition, where our proposal is the best ranked system by a margin of 2.5 points of MF_1 averaged on the three subtasks with respect to the second ranked system.

4 Conclusions

We have proposed a system based on the encoder part of the Transformer architecture in order to extract useful word representations that are discriminative to decide the presence of irony on sort texts. The results obtained by our system are very promising especially considering they have been obtained without an extensive experimentation on the hyperparameters of the model. This opens the

Table 2: Results obtained by 1-TE-NoPos on the test set for the three subtasks.

	ES	MX	CU	AVG
ELiRF-UPV	**71.67**	**68.03**	**65.27**	**68.32**
LDSE	67.95	66.08	63.35	65.79
W2V	68.23	62.71	60.33	63.76
Word nGrams	66.96	61.96	56.84	61.92
Majority	40.00	40.00	40.00	40.00

Table 3: Five best participants in the final ranking of IroSVA19.

	CU	ES	MX	AVG	RANK
ELiRF-UPV	65.27	**71.67**	**68.03**	**68.32**	1/18
CIMAT	**65.96**	64.49	67.09	65.85	2/18
LDSE	63.35	67.95	66.08	65.79	3/18
JZaragoza	61.63	66.05	67.03	64.90	4/18
W2V	60.33	68.23	62.71	63.76	5/18

door to future improvements by exploring modifications on the architecture and its hyperparameters.

Acknowledgements

This work has been partially supported by the Spanish MINECO and FEDER founds under project AMIC (TIN2017-85854-C4-2-R) and the GiSPRO project (PROMETEU/2018/176). Work of José-Ángel González is financed by Universitat Politècnica de València under grant PAID-01-17.

References

1. Ba, L.J., Kiros, R., Hinton, G.E.: Layer normalization. CoRR **abs/1607.06450** (2016)
2. Francisco Rangel, Paolo Rosso, M.F.S.: A low dimensionality representation for language variety identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), pp. 156-169 (2018)
3. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: SemEval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 470–478. Association for Computational Linguistics, Denver, Colorado (Jun 2015). https://doi.org/10.18653/v1/S15-2080
4. González, J., Hurtado, L., Pla, F.: ELiRF-UPV en TASS 2017: Análisis de Sentimientos en Twitter basado en Aprendizaje Profundo (ELiRF-UPV at TASS 2017: Sentiment Analysis in Twitter based on Deep Learning). In: Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2017, co-located with 33nd SEPLN Conference (SEPLN 2017), Murcia, Spain, September 18th, 2017. pp. 29–34 (2017)

5. González, J., Hurtado, L., Pla, F.: ELiRF-UPV en TASS 2018: Análisis de Sentimientos en Twitter basado en Aprendizaje Profundo (ELiRF-UPV at TASS 2018: Sentiment Analysis in Twitter based on Deep Learning). In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 37–44 (2018)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735
7. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1681–1691. Association for Computational Linguistics, Beijing, China (Jul 2015). https://doi.org/10.3115/v1/P15-1162
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013)
9. Ortega-Bueno, R., Rangel, F., Hernández Farías, D.I., Rosso, P., Montes-y-Gómez, M., Medina Pagola, J.E.: Overview of the Task on Irony Detection in Spanish Variants. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS.org (2019)
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)
11. Van Hee, C., Lefever, E., Hoste, V.: Semeval-2018 task 3 : irony detection in english tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 39–50. Association for Computational Linguistics (2018)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)