LaSTUS/TALN at HAHA: Humor Analysis based on Human Annotation

Lutfiye Seda Mut Altin, Àlex Bravo, and Horacio Saggion

LaSTUS-TALN Research Group
Department of Information and Communication Technologies
Universitat Pompeu Fabra
C/Tànger 122-140, 08018 Barcelona, Spain
{name.surname}@upf.edu

Abstract. In this paper we describe the participation of LaSTUS/TALN team in the shared task:“Humor Analysis based on Human Annotation” (HAHA) at the Spanish Society for Natural Language Processing (SEPLN) organized in the context of IberLEF 2019. HAHA onjective is the classification of tweets in Spanish as humorous or not, also identifying the level of funniness. This paper presents a multi-task learning approach based on bidirectional long short-term memory (biLSTM) models. The paper presents and discusses the official results achieved by our team.

Keywords: Natural Language Processing · Humor Analysis · Neural Networks · Spanish Language.

1 Introduction

Humor is a complex phenomenon in human communication that results in amusement or laughter. Although humans are very good at understanding humorous language, computers still lack this essential capability. It is therefore important to make progress in the area of humour recognition and understanding to pave the way for better human machine communication systems. Recent progress in machine learning have produced interesting results in the field of humour classification.

In this paper, we describe a neural network for humor recognition within the context of 'Humor Analysis based on Human Annotation' (HAHA) at IberLEF2019 which is based on tweets written in Spanish [6]. The task is composed of two sub-task as below:

- **Humor Detection:** Referring whether a given tweet is written by the author with an intention of humor or not.
- **Funniness Score Prediction:** Prediction of the funniness score value (average stars) for a given tweet in a 5-star ranking if it is humorous.

In Section 2 of the paper we present an overview for the related work. In Section 3 we provide information about the data and give a description of our model. In Section 4, we give the results and discuss the performance and finally in Section 5 we introduce the conclusions.

2 Related Work

Previous research for humor recognition is mainly based on taking the problem into account as a classification problem. Mihalcea et al. formulated humor recognition with a classification approach and facilitated classifiers such as SVM and Naive Bayes [7]. Purandare and Litman analyzed humorous conversations from a well-known comedy television show using standard supervised classifiers [9]. Barbieri and Saggion [1] presented a machine learning approach based on a linguistically motivated set of features which were also applied to irony detection [2].

Later on, Zhang and Liu worked on several categories of humor-related features giving input around fifty features into the Gradient Boosting Regression Tree model for automated recognition on Twitter data [12]. Radev et al. described an experiment for humor detection in cartoon captions where they compare several automatic methods for selecting the funniest caption and stated that negative sentiment, human-centeredness and lexical centrality match most strongly with the funniest captions [10]. Yang et al. constructed different computational classifiers to recognize humor, based on the designed sets of features [11]. More recently, Chen et al. presented a Convolutional Neural Network (CNN) for humor recognition focusing on lexical cues and pointed out to the advantages of CNN [4]. Chen and Soo proposed a deep learning CNN architecture that can learn to distinguish between humorous and non-humorous texts based on a large scale of balanced positive and negative dataset and reported that it outperforms the previous work [5].

On the other hand there are some researches focusing on humor ranking. The shared task:'SemEval-2017 Task 6: HashtagWars: Learning a Sense of Humor' focused on humor ranking to define the funniness level based on a dataset of funny tweets posted. The top performing system used an ensemble method of both feature based and neural network-based systems [8].

3 Data and Methodology

The corpus that was provided by the shared task organizers consist of 30,000 crowd-annotated tweets based on [3], divided in 80% (24,000 tweets) for training and 20% (6,000 tweets) tweets for testing. The annotation was made with a voting scheme in which users could select one of six options: the tweet is not humorous or, in case the tweet is humorous, an integer score between one (not funny) and five (excellent). Finally, all tweets are classified as humorous or not humorous. The humorous tweets were those which received at least three votes indicating the tweet was somehow humorous with at least five annotations. The not humorous tweets were those that received at least three votes for not humor (they might have less than five votes in total). The corpus contains tweets from every Spanish-speaking country, but the country of the user is not specified in the data-set. Most tweets are written in the Spanish language spoken in Spain, for that reason, we considered that the corpus contains tweets in Spanish.

In this work, we presented a multi-task neural network based on a bidirectional long short-term memory (biLSTM) model with two dense layers at the end. We have used data from different tasks in the context of the IberLEF 2019 evaluation which we believed can assist in humor identification (e.g. irony detection, sentiment). More specifically, we have selected three task to simultaneously train with HAHA:

- From MEX-A3T task, we used the Aggressiveness Identification track, which focuses on the detection of aggressive comments in tweets from Mexican users.
- From the TASS 2019 task, which focused on the evaluation of polarity classification systems of tweets written in Spanish, we used the data related to opinion mining. The data-set consists of tweets written in the Spanish language spoken in Spain, Peru, Costa Rica, Uruguay and Mexico, which were annotated with 4 different levels of opinion intensity (Positive, Negative, Neutral and Nothing).
- From the IroSvA task, the first shared task fully dedicated to identify the presence of irony in short messages, we also used the training dataset, which consist of 2,400 short messages annotated with irony for each Spanish variant spoken in Cuba, Mexico and Spain.

In this scenario, we defined an Embedding layer for each Spanish variant. Classification tasks with the same Spanish variant used the same Embedding layer during the training process. For instance, the embedding layer related to the Spanish from Mexico was used by the MEX-A3T task, the Mexican part of the TASS 2019 task and the tweets written in the Spanish language spoken in Mexico from IroSvA. Furthermore, all task shared the biLSTM layer during training.

In Figure 1 a simplified schema of our shared model can be seen. In the following we explain how the model works in one specific classification task. In order to train all task at the same time, we have divided each data set into the same number of batches. Then, during the training, a batch of data is randomly selected and it used to train its specific model (sharing the embedding and BiLSTM layers with other models). In this sense, we consider one epoch when all batches from all task were trained.

First, the text of the tweets were tokenized, removing punctuation marks, and keeping emoji and full hashtags since they can contribute to define the meaning of a tweet (or short message).

Second, the embedding layer transforms each element in the tokenized tweet into a low-dimension vector. The embedding layer, composed of the vocabulary of the task, was randomly initialized from a uniform distribution (between -0.8

and 0.8 values and with 100 dimensions). The initialized embedding layer was updated with the word vectors included in a pre-trained model from Regional Embeddings, which provides FastText word embeddings for Spanish language variations. After this update, words not included in the pre-trained model keep their random value.

Then, a biLSTM layer gets high-level features from previous embeddings, configured with 128 units. A disadvantage of LSTM models is that they compress all information into a fixed-length vector, causing the incapability of remembering long tweets. To overcome the limitation of fixed-length vector keeping relevant information from long tweet sequences, we added an attention layer producing a weight vector and merge word-level features from each time step into a tweet-level feature vector, by multiplying the weight vector. Finally, the tweet-level feature vector produced by the previous layers is used for classification task by two fully-connected (dense) layers. In the case of the HAHA task, the output from the classification task (humorous or not humorous) was redirected to another output layer in order to learn the funniness score value, that is, the regression task. In the test step, if a tweet is classified as humorous, the funniness score predicted was also considered, otherwise was 0.

Moreover, to be able to mitigate overfitting problem we applied dropout regularization. Dropout operation sets randomly to zero a proportion of the hidden units during forward propagation, creating more generalizable representations of data. In the model, we employ dropout on the embeddings and biLSTM layers. The dropout rate was set to 0.5 in all cases.

4 Results

In sub-task 1, we ranked 10th with an F-score of 0.759 (precision of 0.774 and recall of 0.745) and accuracy of 0.816. In sub-task 2, we ranked 7th with root mean square error of 0.919 (see Table 1).

Table 1. Scores for both subtasks: humor classification (sub-task 1) and funniness score (sub-task 2).

	sub-task 1				sub-task 2
Team	F-1	P	R	A	RMSE
LASTUS-TALN	0.759	0.774	0.745	0.816	0.919
Highest Score	0.821	0.791	0.852	0.855	0.736
Average Score	0.713	0.694	0.737	0.764	1.162
Baseline	0.440	0.394	0.497	0.505	2.455

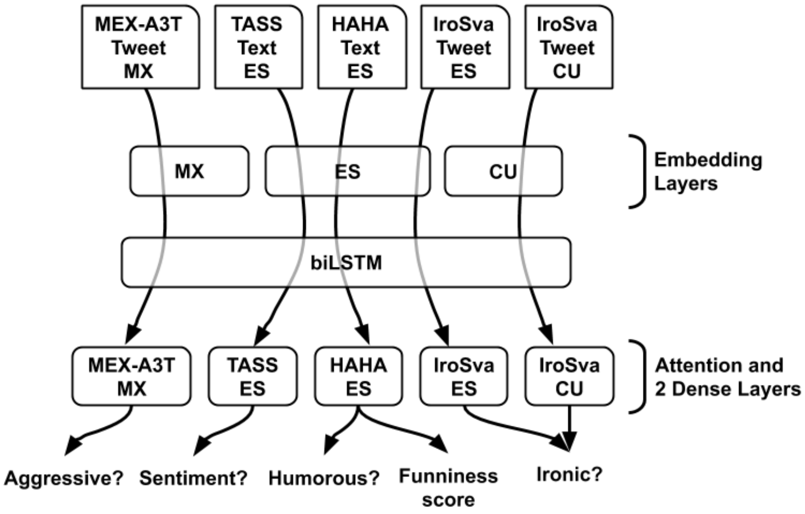

Fig. 1. Simplified schema of the multi-task model. In this example, we have only illustrated the following task: MEX-A3T, TASS (Spain), HAHA and IroSva (Spain and Cuba). Take into account, in this paper we have used all data related to the previous tasks.

5 Conclusions

In this paper, we have presented our results from the participation in the HAHA task from the IberLEF 2019. We have investigated multi-task learning on neural networks with different tasks. Our results improved the baselines presented by the organizers and also the average scores achieved by all participants. Due to time constraints, we were not able to perform an error analysis, for that reason, in future work, we will work in a detailed error analysis in order to understand the limitations of our approach. Furthermore, we want to test different types of neural networks (e.g. convolutions or combinations of convolutions and LSTM layers) and share more layers between task. Finally, we also consider that the integration of linguistic features (e.g. word frequency, POS tags and word shape) and metadata (e.g. whether a tweet is a response to another tweet) can represent useful contextual information to improve our performance.

References

1. Barbieri, F., Saggion, H.: Automatic detection of irony and humour in twitter. In: Proceedings of the Fifth International Conference on Computational Creativity, Ljubljana, Slovenia, June 10-13, 2014. pp. 155–162 (2014)

2. Barbieri, F., Saggion, H.: Modelling irony in twitter. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden. pp. 56–64 (2014)
3. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A crowd-annotated spanish corpus for humor analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media. pp. 7–11 (2018)
4. Chen, L., Lee, C.M.: Convolutional neural network for humor recognition. arXiv preprint arXiv:1702.02584 (2017)
5. Chen, P.Y., Soo, V.W.: Humor recognition using deep learning. In: NAACL-HLT (2018)
6. Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J.J., Rosá, A.: Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)
7. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 531–538. Association for Computational Linguistics (2005)
8. Potash, P., Romanov, A., Rumshisky, A.: Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 49–57 (2017)
9. Purandare, A., Litman, D.: Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 208–215. Association for Computational Linguistics (2006)
10. Radev, D., Stent, A., Tetreault, J., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., Jha, R., Mankoff, R.: Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 475–479. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1076
11. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2367–2376 (2015)
12. Zhang, R., Liu, N.: Recognizing humor on twitter. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 889–898. ACM (2014)