UO_UPV_2 at HAHA 2019: BiGRU Neural Network Informed with Linguistic Features for Humor Recognition

Reynier Ortega-Bueno[1], Paolo Rosso[2], and José E. Medina Pagola[3]

[1] Center for Pattern Recognition and Data Mining, Universidad de Oriente, Santiago de Cuba, Cuba
reynier.ortega@cerpamid.co.cu
[2] PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain
jmedinap@uci.cu
[3] University of Informatics Sciences, Havana, Cuba
prosso@dsic.upv.es

Abstract. Verbal humor is an illustrative example of how humans use creative language to produce funny content. We, as human being, access to humor or comicality with the purpose of projecting more complex meanings which, usually, represent a real challenge, not only for computers, but for humans as well. For that, understanding and recognizing humorous content automatically has been and continue being an important issue in Natural Language Processing (NLP) and even more in Cognitive Computing. In order to addressing this challenge, in this paper we describe our UO_UPV_2 system developed for participating in the second edition of the HAHA (Humor Analysis based on Human Annotation) task proposed at IberLEF 2019 Forum. Our starting point was the UO_UPV system we participated in HAHA 2018 with some modification in its architecture. This year we explored other way to inform our Attention based Recurrent Neural Network model with linguistic knowledge. Experimental results show that our system achieves positive results ranked 7th out of 18 teams.

Keywords: Spanish Humor Classification, BiGRU Neural Network, Social Media, Linguistic Features

1 Introduction

Natural language systems have to deal with many problems related with texts comprehension, but these problems become very hard when creativity and figurative devices are used in verbal and written communication. Human can easily understand the underlying meaning of such texts but, for a computer to disentangle the meaning of creative expressions such as irony and humor, it requires much additional knowledge, and complex methods of reasoning.

Humor is an illustrative example of how humans use creative language devices in social communication. Humor not only serves to interchange information or share implicit meaning, but also engages a relationship between those exposed to the funny message. It can help people see the amusing side of problems and can help them distance themselves from stressors. In the same way, it helps to regulate ours emotions. Moreover, the manners in which people produce funny content also reveal insight about their genre and personal traits.

From a computational linguistics points of view many methods have been proposed to tackle the task of recognizing humor from texts [16,15,22,20,1]. These focus the attention on investigating linguistic features which can be considered as markers and indicators of verbal humor. Also, due to the closely relation between irony and humor, other works have studied these phenomena with the goal of shed some light about what is common and what is distinct from linguistic point of view. [9,1,20]

Other methods focused on recognizing humor on messages from Twitter based on supervised learning [1,23,7,10,26]. Deep Neural Networks based methods have obtained competitive results in humor recognition on tweets. Among them, Recurrent Neural Networks (RNN) models and their bidirectional variant capture relevant information like long term dependencies. Also, attention mechanism have become to be a strong tool, that has endowed the RNN model with the capability of paying more attention to those elements that increase the effectiveness of these networks in several tasks of NLP [13,24,28,27]

Previous researches have focused on English language; however, for Spanish, the availability of corpora is scarce, which limits the amount of research done for this language. HAHA 2018 [4] became the first shared task addressing the problem of humor recognition in Spanish content in social media. Three systems were proposed to solve the task. The best ranked approach used a model based on the EvoMSA tool with uses a EvoDAG method (Evolutionary Algorithm) [21]. This is a steady-state Genetic Programming system with tournament selection. The main characteristic of EvoDAG is that the genetic operation is performed at the root. EvoDAG was inspired by the geometric semantic crossover. The second system [17] proposed a model based on Bidirectional Long Short Term Memory (BiLSTM) neural networks with attention mechanism. The authors used word2vec as input for the network and also a set of linguistically motivated features (stylistic, structural and content, and affective ones). The linguistic information was combined with the deep representation learned in the next to the last layer. The results showed that incorporating linguistic knowledge improves the overall performance. The third system presented to the shared task, trained a method based on SVM using a bag of character n-grams of sizes 1 to 8 character models.

Considering the advantages of linguistic features for capturing deep linguistics aspects of the language also the capability of RNN for learning deep representation and long term dependencies from sequential data, in this paper, we present a method that combines the linguistic features used for humor recognition and an Attention based Bidirectional Gated Recurrent Unit (BiGRU)

model. The system works with an attention layer which is applied at the top of a BiGRU to generate a context vector for each word embedding which is then fed to another BIGRU network. Finally, the learned representation is fed to a Feed Forward Network (FNN) to classify whether the tweet is humorous or not. Motivated by the results shown in [17], we explore to incorporate the linguistic information to the model through initial hidden state in the first BiGRU layer.

The paper is organized as follows. Section 2 presents a brief description of the HAHA task. Section 3 introduces our system for humor detection. Experimental results are subsequently discussed in Section 4. Finally, in Section 5 we present our conclusions and attractive directions for future work.

2 HAHA Task and Dataset

HAHA 2019 is the second edition of the first shared task that addresses the problem of recognizing humor in Spanish tweets. Similar to the first edition, in the HAHA 2019 task, two subtasks were proposed. The first one, *"Humor Detection"*, aims at predicting whether a tweet is a joke or not (intended humor by the author or not) and the second one *"Funniness Score Prediction"*, is for predicting a score value into 5-star ranking, supposing it is a joke.

Participants were provided with a human-annotated corpus of 30000 Spanish tweets [6], divided in 24000 and 6000 for training and test respectively. The training subset contains 9253 tweets with funny content and 14747 tweets considered as non humorous. As could be observed, the classes distribution are slightly unbalanced, hence a difficulty is added to learn automatically the models.

System evaluation metrics were used and reported by the organizers. They use F1 measure on humor class for the subtask of *"Humor Detection"*, moreover, precision, recall and accuracy were also reported.

3 Our UO_UPV_2 System

The motivation behind of our approach are firstly to investigate the capability of Recurrent Neural Network, specifically, the Gated Recurrent Unit (GRU) [8] to capture long-term dependencies. They showed to be able to learn the dependencies in lengths of considerably large sequences. GRU networks simplified the complexity of the LSTM networks [11], being computationally more efficient. Moreover, attention mechanisms have endowed these networks with a powerful strategy to increase their effectiveness achieving better results [27,29,24,12]. Recently, the initial hidden state of the recurrent neural network has been a successful explored way to inform the networks with contextual information [25]. Secondly, humor recognition based on features engine and supervised learning have been well studied in previous research papers. These features have proved to be good indicators and markers of humor in text. For these reasons, in this approach we propose a method that enrich the Attention-based GRUs model with linguistic knowledge which is passed to the network using the initial hidden state. In Section 3.1 we describe the tweets preprocessing phase. Following,

in Section 3.2 we present the linguistic features used for encoding humorous content. Finally, in Section 3.3 we introduce the neural network model and the way in which linguistic features are introduced. The Figure 1 shows the overall architecture of our system.
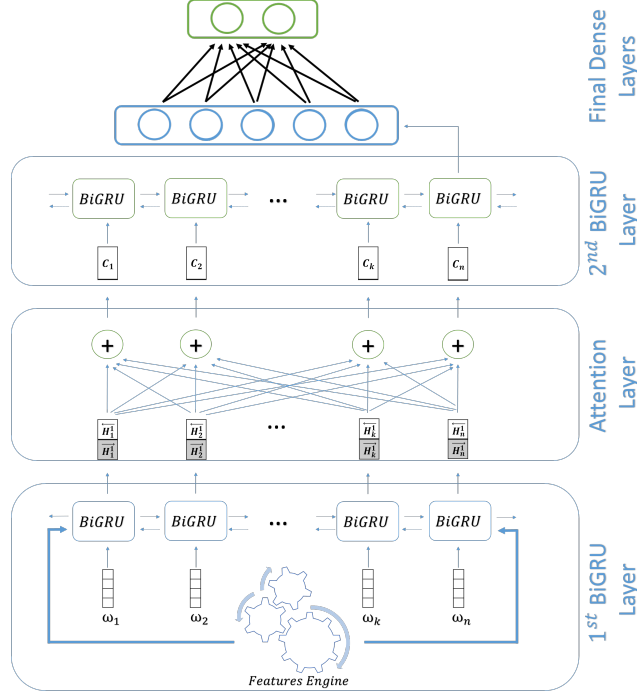

Fig. 1. Overall architecture of the system UO_UPV_2

3.1 Preprocessing

In the preprocessing step, the tweets are cleaned. Firstly, the emoticons, urls, hashtags, mentions, twitter-reserve words as RT (for retweet) and FAV (for favorite) are recognized and replaced by a corresponding wildcard which encodes the meaning of these special words. Afterwards, tweets are morphologically analyzed by FreeLing [18]. In this way, for each resulting token, its lemma is assigned. Then, the tweets are represented as vectors with a word embedding model. This embedding was generated by using the FastText algorithm [2] from the Spanish Billion Words Corpus [3] and an in-house background corpus of 9 millions of Spanish tweets.

3.2 Linguistic Features

In our work, we explored several linguistic features useful for humor recognition in texts [16,15,22,20,1,5] which can be grouped in three main categories: Stylistic, Structural and Content, and Affective. Particularly, we considered stylistic features such as: length, dialog markers, quotation, punctuation marks, emphasized words, url, emoticons, hashtag, etc. Features for capturing lexical and semantic ambiguity, sexual, obscene, animal and human-related terms, etc., were considered as Content and Structural. Finally, due to the relation of humor with expressions of sentiment and emotions we used features for capturing affects, attitudes, sentiments and emotions. For more details about the features see [17].

Notice that, our proposal did not consider the positional features used in [17]. Moreover taking into account the close relation between irony and humor and motivated by the results presented in [14] we include psycho-linguistic features extracted from the LIWC [19]. This resource contains about 4,500 entries distributed in 65 categories. Specifically, for this work we decided to use all categories as independent features.

Taking into account the previous features, we represent each message by one vector V_{T_i} with dimensionality equal to 165. Also, in order to reduce and improve this representation we applied a feature selection method. Specifically we use the Wilcoxon Rank-sum test for paired samples. By using this test all features were ranked considering their $p - value$.

3.3 Recurrent Network Architecture

We propose a model that consists in a BiGRU neural network at the word level. Each time step t the BiGRU gets as input a word vector w_t. Afterward, an attention layer is applied over each hidden state h_t. The attention weights are learned using the concatenation of the current hidden state h_t of the BiGRU and the past hidden state s_{t-1} in the second BiGRU layer. Finally, the target humor of the tweet is predicted by an FFN with one hidden layer, and an output layer with two neurons. Our overall architecture is described in the following sections.

3.4 First BiGRU Layer

In NLP problems, standard GRU receives sequentially (left to right order) at each time step a word embedding w_t and produces a hidden state h_t. Each hidden state h_t is calculated as follow:

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1} + b^{(z)}) \qquad \text{(update gate)}$$
$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1} + b^{(r)}) \qquad \text{(reset gate)}$$
$$\hat{h}_t = \tanh(W^{(\hat{h})}x_t + U^{(\hat{h})}h_{t-1} + b^{(\hat{h})}) \qquad \text{(memory cell)}$$
$$h_t = z_t \oplus h_{t-1} + (1 - z_t) \oplus \hat{h}_t$$

Where all $W^{(*)}$, $U^{(*)}$ and $b^{(*)}$ are parameters to be learned during training. Function σ is the sigmoid function and $\oplus$ stands for element-wise multiplication.

Bidirectional GRU, on the other hand, makes the same operations as standard GRU, but processing the incoming text in a left-to-right and a right-to-left order in parallel. Thus, it outputs two hidden state at each time step $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. The proposed method uses a BiGRU network which considers each new hidden state as the concatenation of these two $\hat{h}_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The idea behind this BiGRU layer is to capture long-range and backwards dependencies simultaneously. In this layer is where the linguistic information is passed throughout the model. We initialized both initial hidden state $[\overrightarrow{h_0} = g(T_i), \overleftarrow{h_0} = g(T_i)]$ where $g(.)$ receives a tweet and returns a vector which encodes contextual and linguistic knowledge $g(T_i) = V_{T_i}$.

3.5 Attention Layer

With an attention mechanism we allow the BiGRU to decide which segment of the sentence should "attend". Importantly, we let the model learn what to attend on the basic of the input sentence and what it has produced so far.
Let $H \in R^{2 \times N_h \times T}$ the matrix of hidden states $[\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_T]$ produced by the first BiGRU layer, where N_h is the size of the hidden state and T is the length of the given sequence. The goal is then to derive a context vector c_t that captures relevant information and feeds it as input to the next BiGRU layer. Each c_t is calculate as follow:

$$c_t = \sum_{t'=1}^{T} \alpha_{t,t'} \hat{h}_{t'} \quad \alpha_{t,i} = \frac{exp(\beta(s_{t-1}, \hat{h}_i))}{\sum_{j=1}^{T} exp(\beta(s_{t-1}, \hat{h}_j))}$$

$$\beta(s_i, h_j) = V_a^T * \tanh(W_a \times [s_i, \hat{h}_j])$$

Where W_a and V_a are the trainable attention parameters, s_{t-1} is the past hidden state of the second BiGRU layer and $\hat{h}_t$ is the current hidden state. The idea of the concatenation layer is to take into account not only the input sentence but also the past hidden state to produce the attention weights.

3.6 Second BiGRU Layer

The goal of this layer is to obtain a deep dense representation of the message with the intention to determine whether the tweet is humorous or not. This network at each time step receives the context vector c_t which is propagated until the final hidden state s_T. This vector is a high level representation of the tweet. Afterwards, it is passed to a feed forward network (FFN) with 3 hidden layers, and we use a softmax layer at the end as follow:

$$\hat{y} = softmax(W_0 \times dense_1 + b_0)$$
$$dense_1 = relu(W_1 \times dense_2 + b_1)$$
$$dense_2 = relu(W_2 \times dense_3 + b_2)$$
$$dense_3 = relu(W_3 \times s_T + b_3)$$

Where W_j and b_j $(j = 0,...3)$ denote the weight matrices and bias vectors for the last three layers with a softmax at the end. Finally, cross entropy is used as the loss function, which is defined as:

$$L = -\sum_i y_i * \log(\hat{y_i})$$

Where y_i is the ground true classification of the tweet (humor vs. not humor) and $\hat{y_i}$ is the predicted value by the model.

4   Experiments and Results

For tuning some parameters of the proposed model we used a stratified k-fold cross validation with 5 partitions on the training dataset. At this time, a features selection process was not performed, therefore we consider all linguistic features. During the training phase, we fixed some hyper-parameters, concretely: the batch_size =256, epochs=10, units of the GRU cell was defined as 256, optimizer="Adam" and dropout in the GRU cells=0.3. After that, we evaluated different subsets of linguistic features, particularly, five setting of features were explored. The considered subsets were: No_Fea (linguistic information is not considered), 64_Fea (the 64 best ranked features according to $p-value$), 128_Fea (the 128 best ranked features according to $p-value$), 141_Fea (all features with $p-value \geq 0.05$) and All_Fea (all linguistic features).

Table 1. Results of the UO_UPV_2 system using the feature selection strategy on the training dataset.

Features	Pr_h	Rc_h	F1_h	Pr_noh	Rc_noh	F1_noh	F1_AVG
No_Fea	0.818	0.695	0.750	0.826	0.902	0.862	0.806
All_Fea	0.795	0.751	0.767	0.851	0.871	0.859	0.813
Fea_64	0.829	0.717	0.763	0.838	0.901	0.867	0.815
Fea_128	**0.756**	**0.817**	**0.785**	**0.880**	**0.834**	**0.856**	**0.820**
Fea_141	0.779	0.769	0.771	0.858	0.859	0.857	0.814

As can be observed in the Table 1, a slight improvement is obtained when linguistic features were passed to the model. Particularly, the subset of Fea_128 achieved the best F1 score (F1_h=0.785) in the humor class. Also, when linguistic information was missing a gradual drop of 3.5%, in term of F1-score was observed in the humor class.

Regarding official evaluation and results, for the system's submission, participants were allowed to send more than one model till a maximum of 10 possible

runs. Taking into account the results showed in the Table 1 we submitted three runs. The difference among them is the number of linguistic features considered for informing the UO_UPV_2 model. In $Run1$ we use the subset of features Fea_64, for $Run2$ we used the subset of features Fea_128 while in $Run3$ the subset of features Fea_141 was considered. We achieved 0.765, 0.773, and 0.765, in terms of F1 score in the humor class, for $Run1$, $Run2$ and $Run3$ respectively. These values are consistent with the results obtained in the training phase.

Table 2. Official results for the Humor Detection subtask

Rank	Team	F1	Pr	Rc	Acc
1	adilism	0.821	0.791	0.852	0.855
2	kevinb	0.816	0.802	0.831	0.854
3	bfarzin	0.810	0.782	0.839	0.846
4	jamestjw	0.798	0.793	0.804	0.842
5	job80	0.788	0.758	0.819	0.828
6	jimblair	0.784	0.745	0.827	0.822
7	*UO_UPV_2*	*0.773*	*0.780*	*0.765*	*0.824*
8	vaduvabogdan	0.772	0.729	0.820	0.811
. . .	. . .	. . .	. . .	. . .	. . .
18	premjithb	0.495	0.478	0.514	0.591
19	hahaPLN	0.440	0.394	0.497	0.505

Regarding the official ranking, a first glance at Table 2 allows to observe that our best submissions (UO_UPV_2) was ranked as $7th$ from a total of 18 of teams.

5 Conclusion

In this paper we presented our modification of the UO_UPV system (UO_UPV_2) for the task of humor recognition (HAHA) at IberLEF 2019. We only participated in the "Humor Detection" subtask and ranked $7th$ out of 18 team. Our proposal combines linguistic features with an Attention-based BiGRU Neural Network. The model consists of a Bidirectional GRU neural network with an attention mechanism that allows to estimate the importance of each word and then, this context vector is used with another BiGRU model to estimate whether the tweet is humorous or not. Regarding the feature selection, the best result was achieved when the 128 best ranked (according to $p-value$) features were considered. The results, also shown that adding linguistic information through initial hidden state caused an improvement in the effectiveness based on F1-measure.

Acknowledgments

munication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

1. Barbieri, F., Saggion, H.: Automatic Detection of Irony and Humour in Twitter. In: Fifth International Conference on Computational Creativity (2014)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the ACL. 5, 135–146 (2017)
3. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (2016). [Online]. Available: http://crscardellino.me/SBWCE/. Retrieved May 4, 2018, `http://crscardellino.me/SBWCE/`
4. Castro, S., Chiruzzo, L., Rosá, A.: Overview of the HAHA Task : Humor Analysis based on Human Annotation at IberEval 2018. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., Carrillo De Albornoz Cuadrado, J. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). pp. 187–194. CEUR-WS.org, Sevilla, Spain (2018)
5. Castro, S., Garat, D., Moncecchi, G.: Is This a Joke? Detecting Humor in Spanish Tweets. In: Ibero-American Conference on Artificial Intelligence. pp. 139–150 (2016)
6. Castro, Santiago and Chiruzzo, Luis and Rosá, Aiala and Garat, Diego and Moncecchi, G.: A Crowd-Annotated Spanish Corpus for Humor Analysis. In: Proceedings of SocialNLP 2018, The 6th International Natural Language Processing for Social Media (2018)
7. Cattle, A., Bay, C.W., Kong, H.: SRHR at SemEval-2017 Task 6: Word Associations for Humour Recognition. In: 11th International Workshop on Semantic Evaluations (SemEval-2017). pp. 401–406 (2017)
8. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. CoRR abs/1412.3 (2014), `http://arxiv.org/abs/1412.3555`
9. Gibbs, R.W., Bryant, G.A., Colston, H.L.: Where is the humor in verbal irony ? Humor 27(4), 575–595 (2014)
10. Han, X., Toner, G.: QUB at SemEval-2017 Task 6: Cascaded Imbalanced Classification for Humor Analysis in Twitter. In: 11th International Workshop on Semantic Evaluations (SemEval-2017). pp. 380–384 (2017)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
12. Lin, K., Lin, D., Cao, D.: Sentiment Analysis Model Based on Structure Attention Mechanism. In: UK Workshop on Computational Intelligence. pp. 17–27. Springer (2017)
13. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
14. María del Pilar Salas-Zárate , Mario Andrés Paredes-Valverde , Miguel Ángel Rodriguez-García , Rafael Valencia-García, G.A.H.: Automatic Detection of Satire in Twitter: A psycholinguistic-based approach. Knowledge-Based Systems (2017), `http://dx.doi.org/10.1016/j.knosys.2017.04.009`

15. Mihalcea, R., Pulman, S.: Characterizing Humour: An Exploration of Features in Humorous Texts. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 337–347 (2007)
16. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): computational models for humor recognition. Computational Intelligence 22(2), 126–142 (2006)
17. Ortega-Bueno, R., Muñiz, C.E., Rosso, P., Medina-Pagola, J.E.: UO_UPV : Deep Linguistic Humor Detection in Spanish Social Media. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., Carrillo-de Albornoz, J. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). pp. 203–213. CEUR-WS.org, Sevilla, Spain (2018)
18. Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the (LREC 2012) (2012)
19. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71 (2001)
20. Reyes, A., Rosso, P., Buscaldi, D.: From humor recognition to irony detection: The figurative language of social media. Data and Knowledge Engineering 74, 1–12 (2012), http://dx.doi.org/10.1016/j.datak.2012.02.005
21. Salgado, V., Tellez, E.S.: INGEOTEC at IberEval 2018 Task HaHa: μTC and EvoMSA to Detect and Score Humor in Texts. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., Carrillo De Albornoz Cuadrado, J. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). pp. 195–202. CEUR-WS.org, Sevilla, Spain (2018)
22. Sjobergh, J., Araki, K.: Recognizing Humor Without Recognizing Meaning. In: International Workshop on Fuzzy Logic and Applications. pp. 469—-476 (2007)
23. Turcu, R.A., Alexa, L., Amarandei, S.M., Herciu, N., Scutaru, C., Iftene, A.: #WarTeam at SemEval-2017 Task 6: Using Neural Networks for Discovering Humorous Tweets. In: 11th International Workshop on Semantic Evaluations (SemEval-2017). pp. 407–410 (2017)
24. Wang, Y., Huang, M., Zhao, L., Others: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 606–615 (2016)
25. Wenke, S., Fleming, J.: Contextual Recurrent Neural Networks. CoRR abs/1902.0 (2019), http://arxiv.org/abs/1902.03455
26. Yan, X., Pedersen, T.: Duluth at SemEval-2017 Task 6: Language Models in Humor Detection. In: 11th International Workshop on Semantic Evaluations (SemEval-2017). pp. 385–389. No. 2 (2017)
27. Yang, M., Tu, W., Wang, J., Xu, F., Chen, X.: Attention Based LSTM for Target Dependent Sentiment Classification. In: AAAI. pp. 5013–5014 (2017)
28. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
29. Zhang, Y., Zhang, P., Yan, Y.: Attention-based LSTM with Multi-task Learning for Distant Speech Recognition. Proc. Interspeech 2017 pp. 3857–3861 (2017)