UO_UPV: Deep Linguistic Humor Detection in Spanish Social Media

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

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

Abstract. Natural Language Understanding becomes very hard when creativity and figurative language are used in social communication. Humor constitutes an illustrative example of how humans use creative language to produce funny content. Therefore, create new methods and resources for analyzing properly humorous texts is an important issue in Natural Language Processing (NLP) and even more in Human Computer Interaction (HCI). In this sense, this paper introduces our UO_UPV system developed for the Humor Analysis based on Human Annotation (HAHA) track proposed in IberEval 2018 Workshop. The task focuses on classifying tweets in Spanish as humorous or not, and predicting how funny they are. To solve this task, our proposal combines both linguistic features and an Attention-based Recurrent Neural Network, where the attention layer helps to calculate the contribution of each term towards targeted humorous classes. Experimental results show that our model achieves encourage results.

Keywords: Spanish Humor Recognition, Deep Recurrent Neural Network, Social Media, Linguistic Features

1 Introduction

Social Media have provided to modern societies easy and attractive ways for sharing their point of views on the most diverse subjects. For this reason, new challenges in information processing and management, decision support systems, and human computer interaction has been opened. Dealing with multilingualism, with multiples genres and written styles, as well as other subjectives language devices (sentiments, emotions, attitudes and opinions) has captured the focus of several research. However, these faced 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 expressions but, for a computer to disentangle the meaning of creative expressions such as irony and humor, it requires much additional knowledge.

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.

Twitter has become as a popular information source for gathering, in transparent way, spontaneous user's generated content. It enables accessing humorous messages, it is useful for recognizing and identifying how humor arises through language. From a computational linguistics point of view many methods have been proposed to tackle the fascinating task of recognizing humor from texts [15,14,24,20,1]. These focus the attention on investigating linguistic features which can be considered as markers and indicators of verbal humor. Other methods focused on recognizing humor on messages from Twitter based on supervised learning [1,25,5,8,27]. Deep Neural Networks based methods have obtained competitive results in humor recognition on tweets. Among them, Long Short Term Memory (LSTM) models and their bidirectional variant (Bi-LSTM) capture relevant information like long term dependencies. Finally, attention mechanism have been use in a wide range of application in the NLP field obtaining excellent results too [13,26,29,28]

Considering the advantages of linguistic features for capturing deep linguistics aspects of the language also the capability of Recurrent Neural Network for learning features and long term dependencies from sequential data, in this paper, we present a new method that combines the most relevant linguistic features used for humor recognition and an Attention based LSTM. The system works with an attention layer which is applied on the top of a Bidirectional LSTM to generate a context vector for each word embedding which is then fed to another LSTM network to detect whether the tweet is humorous or not. To the best of our knowledge, there has not been any other work exploring the use of attention-based architectures for humor recognition in Spanish tweets.

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

Humor recognition and generation on social media content have become interesting research areas from the computational point of view. Most studies in the field of humor recognition from textual sources have focused on English more than on other popular languages such as Spanish. In this context, HAHA can be considered as the first shared task that facing the fascinating problem of recognizing humor in Spanish tweets. In the HAHA task, two subtasks were proposed by the organizers. The first one, *"Humor Detection":* aim to predict if a tweet is

a joke or not (intended humor by the author or not) and the second one *"Funniness Score Prediction":* for predicting a score value (average stars) for a tweet into 5-star ranking, supposing it is a joke.

Participants were provided with a human-annotated corpus of 20000 Spanish tweets [4], divided in 16000 from training and 4000 for testing. The annotation was made with a voting scheme, in which annotators could choose one of six options: the tweet does not contain humor, or the tweet contains humor and a number of stars from one to five. The training subset contains 5865 tweets with funny content and 10135 tweets considered as non humorous. As could be observed, the classes in the training are unbalanced, hence a difficulty is added to automatic learning models.

System evaluation metrics were used and reported by the organizers. Their choice was to use F1 measure on humor class for the subtask of *"Humor Detection"*, moreover, precision, recall and accuracy were also reported. The Root Mean Squared Error (RMSE) was used to assess the systems effectiveness in the subtask of *"Funniness Score Prediction"*.

3 Our Approach

The motivation of our approach is twofold: firstly, the capability of Recurrent Neural Network, specifically, the LSTMs [10] to capture long-term dependencies and, therefore, their suitability for NLP. They are able to learn the dependencies in lengths of considerably large sequences. Moreover, attention mechanisms have endowed these networks with a powerful strategy to increase their effectiveness achieving better results [28,30,26,12,21]. 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 LSTMs model with linguistic knowledge. In Section 3.1 we describe the tweets preprocessing phase. Following, in Section 3.2 we present the linguistic features used in this work for encoding humorous content. Finally, in Section 3.3 we introduce the neural network model and the way in which linguistic features are introduced to it.

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 [19]. 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 of Word2Vec algorithm [16] from the Spanish Billion Words Corpus [2] and an in-house background corpus of 9 millions of Spanish tweets. We decided to merge both corpora in order to obtain a better

representation of words in context and also taking advantage of the peculiar writing style used by Twitter's users.

3.2 Linguistic Features

In our work, we explored some linguistic features useful for humor recognition in texts [15,14,24,20,1,3] which can be grouped in three main categories: Stylistic, Structural and Content, and Affective. We define a set of features distributed as follows:

Stylistic Features

- *Multiple Statement*:This feature takes into account whether the tweet is composed or not by multiple lines (single vs. many lines).
- *Length*: Three different features were considered: number of words, number of characters, and the means of the length of the words in the tweet.
- *Dialog*: Two different features were considered: tweet is a dialog and the tweet has multiple statements starting with dialog marker (-).
- *Hashtags*: The amount of hashtags in the tweet is counted.
- *Urls*: The amount of url in the tweet is counted.
- *Emoticons*: The amount of emoticons in the tweet is counted.
- *Exclamations*: The amount of exclamation marks is counted.
- *Emphasized Words*: Four different features were considered: word emphasized through repetition, capitalization, character flooding and exclamation marks.
- *Punctuation Marks*: The frequency of dots, commas, semicolons, and question marks.
- *Quotations*: The number of expressions between quotation marks.
- *Alliteration*: This feature tries to capture the occurrence of alliteration in the tweet. We used a fixed length (3) sequence of phonetic prefix.

Structural and Content Features

- *Animal Vocabulary*: This feature counts the number of words in a dictionary of animal names.
- *Toponym Vocabulary*: This feature counts the number of words in a dictionary of countries, capitals, cities and nationalities.
- *Sexual and Obscene Vocabulary*: This feature counts the number of words in a dictionary of sexual and obscene words.
- *Antonyms*: This feature considers the number of pairs of antonyms existing in it. WordNet [18] antonym relationship and Spanish language enrichment provided by the Multilingual Central Repository (MCR) [7] are used for this.
- *Lexical Ambiguity*: Thee different features were considered using MCR: the first one is the mean of the number of synsets of each word of the tweet. The second one is the greatest number of synsets that a single word has in the tweet. The last is the difference between the number of synsets of the word with major number of synsets and the average number of synsets.

- *Domain Ambiguity*: Three different features were considered using MCR: the first one is the mean of the number of domains of each word of the tweet. The second one is the greatest number of domains that a single word has in the tweet. The last one is the difference between the number of domains of the word with major number domains and the average number of domains. It is important to clarify that the resources WordNet Domains[4] and SUMO[5] were separately used.
- *Semantic Classes:* These features try to capture distinct semantic frames of the verbs in the tweet according to ADDESE[6].
- *Persons*: This feature tries to capture verbs conjugated in the first, second, third persons and nouns and adjectives which agree with such conjugations.
- *Tenses*: This feature tries to capture the different verbal tenses used in the tweet.
- *Questions-answers*: Occurrences of questions and answers structure in the tweet is counted.
- *Part of Speech*: The number of nouns, verbs, adverbs and adjectives in the tweet are quantified.
- *Negation*: The amount of negation words in the tweet is counted.

Affective Features

- *Sentiments*: These feature count the number of positive and negative words according to a sentiment resource. Notice that, for each resource two features are built. In this work we explore four distinct dictionaries: Spanish Sentiment Lexicon [6], Elhuyar Sentiment Lexicon [22], CriSol Lexicon [6], and the Lexicon presented in [9].
- *Emoti Sentiments*: These count the number of positive and negative emoticons in the resource Emoticons Sentiment [11].
- *Attitudes*: These features count the number of words according to the three distinct attitude categories (affect, judgment, and appreciation) proposed in [9].
- *Emotions*: These feature count the number of words according to the six basic emotions provided by the resource SEL [23].

We also try to analyze the occurrence of some of the previous features into specific positions in the tweets. For instance, the feature occurring at the beginning or at the ending of a tweet. Based on all the features mentioned above a vector (VF_t) is built for each tweet in the training and test datasets.

3.3 Recurrent Network Architecture

We propose a model that consists in a Bidirectional LSTM neural network (Bi-LSTM) at the word level. Each time step t the Bi-LSTM gets as input a word

[4] http://wndomains.fbk.eu/hierarchy.html
[5] http://www.adampease.org/OP/
[6] http://adesse.uvigo.es/data/clases.php

vector w_t with syntactic and semantic information, known as word embedding [17]. 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 Bi-LSTM and the past hidden state s_{t-1} of a Post-Attention LSTM (Pos-Att-LSTM). Finally, the target humor of the tweet is predicted by this final Pos-Att-LSTM network.

3.4 Pre-Attention Bi-LSTM

In NLP problems, standard LSTM 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:

$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) && \text{(input gate)}\\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) && \text{(forget gate)}\\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) && \text{(output gate)}\\
u_t &= \sigma(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) && \text{(new memory cell)}\\
c_t &= i_t \oplus + f_t \oplus c_{t-1} && \text{(final memory cell)}\\
h_t &= o_t \oplus tanh(c_t)
\end{aligned}
$$

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 LSTM, on the other hand, makes the same operations as standard LSTM but, processes 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 Bi-LSTM network which considers each new hidden state as the concatenation of these two $\hat{h_t} = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The idea of this Bi-LSTM is to capture long-range and backwards dependencies.

3.5 Attention Layer

With an attention mechanism we allow the Bi-LSTM to decide which part 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 $H \in R^{2\times N_h \times T_x}$ the matrix of hidden states $[\hat{h_1}, \hat{h_2}, \ldots, \hat{h_{T_x}}]$ produced by the Bi-LSTM model, where N_h is the size of the hidden state and T_x is the length of the given sentence. The goal is then to derive a context vector c_t that captures relevant information and feed it as input to the next level (Pos-Att-LSTM). Each c_t is calculate as follow:

$$
c_t = \sum_{t'=1}^{T_x} \alpha_{t,t'}\hat{h_{t'}} \quad \alpha_{t,i} = \frac{\beta_{t,i}}{\sum_{j=1}^{T_x}\beta_{t,j}} \qquad \beta = \tanh(W_a \times [h_t, \hat{s}_{t-1}] + b_a)
$$

Where W_a and b_a are the trainable attention parameters, s_{t-1} is the past hidden state of the Pos-Att-LSTM 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 Post-Attention LSTM

The goal of the Post-Att-LSTM is to predict 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_x}. This vector is a high level representation of the tweet and is used in the final softmax layer combined with the linguistic feature vector as follow:

$$\hat{y} = softmax(W_g \times [S_{T_x}, LF_d] + b_g)$$

$$LF_d = relu(W_d \times LF_t + b_d)$$

Where W_g and b_g denote the weight matrix and bias vector for the last layer with a softmax at the end. LF_d is the result of passing the VF_t vector associated to each tweet through a dense layer before the softmax layer. 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). For predicting the funniness score we use an architecture similar to the one described above. For predicting the funniness score we use an architecture similar to the one described above. The most salience change is at the last hidden layer and the loss function. Specifically, the last layer was changed to a dense one with just one neuron as output and the mean square error (MSE) was used as loss function for optimizing the model.

4 Experiments and Results

In this section we show the results of the proposed method in the shared task of "Humor Detection" and discuss them. For the system's submission, participants were allowed to send more than one model till a maximum of 4 possible runs. In Tables 1 we report our three best performing systems (run1, run2 and run3 sent by UO_UPV) for humor detection task on two classes (humor, not humor). The first run is based on the Attention based model mixed with linguistic features which are fed to the model in the last Dense Layer. Run2 is similar to run1, the major difference consists in a dimension reduction of linguistic features by using Random Forest[7] as strategy to rank the importance of each feature. Finally, in

[7] We use the implementation provided by the sklearn tool

the run 3 we evaluate our proposal without linguistic features. As can be shown in Table 1, run1 and run2 achieved F1=0.7851 and F1=0.7785 scores, respectively while run3 obtains a F1=0.7702. Experiments showed that introducing linguistic information to the Attention LSTM model (run1, run2) improves the performance of the model. Contrary, to our expectations, the reduction of the dimensionality of the linguistic feature vector applied in run2 obtained a drop in term of F1 measure. Our 4^{th} run (UO_UPV run4) addressed the task of *"Funiness Score Prediction"*. For that, we consider a setting similar to run1 where all linguistic features and recurrent neural network model were combined, but considering the modification of the model explained in the Section 3.6 to deal with the regression task. Our run4 obtained a value of 1.5919 in terms of RSME and positioned at last place out of two runs. Also our result do not surpass the baseline established by the task.

Table 1. Official results for the Humor Detection subtask

Team	**Run**	**Acc**	**Prec**	**Rec**	**F1**
INGEOTEC	run 2	0.8452	0.7796	0.8157	0.7972
UO_UPV	*run 1*	**0.8455**	**0.8158**	*0.7567*	**0.7851**
UO_UPV	*run 2*	*0.8448*	**0.8322**	*0.7312*	**0.7785**
ELiRF-UPV	run 1	0.8367	0.8046	0.7426	0.7724
UO_UPV	*run 3*	*0.8397*	*0.8281*	*0.7198*	**0.7702**
INGEOTEC	run 1	0.8403	0.8557	0.6877	0.7625
ELiRF-UPV	run 2	0.7552	0.6546	0.7279	0.6893

Regarding the official results, at a first glance on Table 1 it is possible to observe that our submissions (run1, run2 and run3) were ranked on 2^{nd}, 3^{rd} and 6^{th} respectively from a total of 7 runs of 3 teams. Notice that, our best result (run1) achieves slightly better accuracy that the first score obtained by the INGEOTEC Team. It is important to remark that in term of precision on humor class our run1 and run2 outperform the best ranked submissions (INGEOTEC run2).

5 Conclusion

In this paper we presented the UO_UPV system for the task of humor recognition (HAHA) at IberEval 2018. We participated in the "Humor Detection" subtask and ranked 2^{nd} out of 7 submissions. Our proposal combines linguistic features with an Attention-based Long Short-Term Memory Network. The model consists of a Bidirectional LSTM neural network with an attention mechanism that allows to estimate the importance of each word and then, this context vector is used with another LSTM model to estimate whether the tweet is humorous or not.

The results shown that the consideration of linguistic features in combination with the deep representation learned by the neural network model obtains better effectiveness based on F1-measure the in humor class. Due to encouraging results of our approach, we think that including the linguistic features of humor into the embedding layer could be a way to increase the effectiveness. We would like to explore this approach in the future work.

Acknowledgments

The work of the fourth author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

1. Barbieri, F., Saggion, H.: Automatic Detection of Irony and Humour in Twitter. In: Fifth International Conference on Computational Creativity (2014)
2. Cardellino:, C.: Spanish Billion Words Corpus and Embeddings (2016), http://crscardellino.me/SBWCE/
3. 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)
4. 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)
5. 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)
6. González, M.D.M., Cámara, E.M., Valdivia, M.T.M.: CRiSOL:Base de conocimiento de opiniones para el español. Procesamiento del Lenguaje Natural (55), 143–150 (2015)
7. Gonzalez-Agirre, A., Laparra, E., Rigau, G.: Multilingual central repository version 3.0. LREC pp. 2525–2529 (2012)
8. 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)
9. Hernández, L., López-Lopez, A., Pagola, J.E.M.: Classification of Attitude Words for Opinions Mining. International Journal of Computational Linguistics and Applications 2(1-2), 267–283 (2011)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
11. Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., Kaymak, U.: Exploiting Emoticons in Sentiment Analysis. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. pp. 703–710. SAC '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2480362.2480498
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. 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)
15. Mihalcea, R., Strapparava, C.: Learning to laugh (automatically): computational models for humor recognition. Computational Intelligence 22(2), 126–142 (2006)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Nips pp. 1–9 (2013)
17. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: International Conference on Learning Representations (ICLR 2013). pp. 1–12 (2013), http://arxiv.org/pdf/1301.3781v3.pdf
18. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
19. Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (may 2012)
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. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685 (2015)
22. Saralegi, X., Vicente, I.S.: Elhuyar at TASS 2013. In: XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). pp. 143–150 (2013)
23. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., D\'\iaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. In: Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I. pp. 1–14. MICAI'12, Springer-Verlag, Berlin, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-37807-2{_}1
24. Sjobergh, J., Araki, K.: Recognizing Humor Without Recognizing Meaning. In: International Workshop on Fuzzy Logic and Applications. pp. 469—-476 (2007)
25. 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)
26. 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)
27. 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)
28. Yang, M., Tu, W., Wang, J., Xu, F., Chen, X.: Attention Based LSTM for Target Dependent Sentiment Classification. In: AAAI. pp. 5013–5014 (2017)
29. 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)

30. Zhang, Y., Zhang, P., Yan, Y.: Attention-based LSTM with Multi-task Learning for Distant Speech Recognition. Proc. Interspeech 2017 pp. 3857–3861 (2017)