Humor recognition using deep learning

Peng-Yu Chen
National Tsing Hua University
Hsinchu, Taiwan
pengyu@nlplab.cc

Von-Wun Soo
National Tsing Hua University
Hsinchu, Taiwan
soo@cs.nthu.edu.tw

Abstract

Humor is an essential but most fascinating element in personal communication. How to build computational models to discover the structures of humor, recognize humor and even generate humor remains a challenge and there have been yet few attempts on it. In this paper, we construct and collect four datasets with distinct joke types in both English and Chinese and conduct learning experiments on humor recognition. We implement a Convolutional Neural Network (CNN) with extensive filter size, number and Highway Networks to increase the depth of networks. Results show that our model outperforms in recognition of different types of humor with benchmarks collected in both English and Chinese languages on accuracy, precision, and recall in comparison to previous works.

1 Introduction

Humor, a highly intelligent communicative activity, provokes laughter or provides amusement. The role that humor plays in life can be viewed as a sociological phenomenon and function. Proper use of it can help eliminate embarrassment, establish social relationships, create positive affection in human social interactions. If computers can understand humor to some extent, it would facilitate predicting human's intention in human conversation, and thereby enhance the proficiency of many machine-human interaction systems.

However, to automate the humor recognition is also a very challenging research topic in natural language understanding. The extent to which a person may sense humor depends on his/her personal background. For example, young children may favor cartoons while the grownups may feel the humor in cartoons boring. Also, many types of humor require substantial such external knowledge as irony, wordplay, metaphor and sarcasm. These factors make the task of automated humor recognition difficult.

Recently, with the advance of deep learning that allows end-to-end training with big data without human intervention of feature selection, humor recognition becomes promising. In this work, we propose a convolutional neural network (CNN) with augmentation of both the filter sizes and filter numbers. We use the architecture called highway network to implement a much more proficient model for humor recognition. The performance on many benchmarks shows a significant improvement in detecting different humor context genre.

2 Related Work

The task of automatic humor recognition refers to deciding whether a given sentence expresses a certain degree of humor. In early studies, most of them are formulated as a binary classification, based on selection on linguistic features. Purandare and Litman analyzed humorous spoken conversations from a classic comedy television show. They used standard supervised learning classifiers to identify humorous speech (Purandare and Litman, 2006). Taylor and Marlack focused on a specific type of humor, wordplays. Their algorithm of the study was based on the extraction of structural patterns and peculiar structure of jokes (Taylor and Mazlack, 2004). Later, Yang et al. (2015) formulated a classifier to distinguish between humorous and non-humorous instances, and also created computational models to discover the latent semantic structure behind humor from four perspectives: incongruity, ambiguity, interpersonal effect and phonetic style.

Recently, with the rise of artificial neural networks, many studies utilize the methods for humor recognition. Luke and Alfredo applied recurrent neural network (RNN) to humor detec-

Proceedings of NAACL-HLT 2018, pages 113–117
New Orleans, Louisiana, June 1 - 6, 2018.

tion from reviews in Yelp dataset. In addition, they also applied convolutional neural networks (CNNs) to train a model and the work shows that the model trained with CNNs has more accurate humor recognition (de Oliveira and Rodrigo, 2015). In other research (Bertero and Fung, 2016), CNNs were found to be a better sentence encoder for humor recognition as well. In a recent work, Chen and Lee predicted audience's laughter also using convolutional neural network. Their work gets higher detection accuracy and is able to learn essential feature automatically (Chen and Lee, 2017). However, there are still some limitations: (a) they focused on only a specific humor type in TED data, that is puns. (b) the datasets in most studies are English corpus. (c) the evaluations are isolated from other research.

In our work, we build the humor recognizer by using CNNs with extensive filter size and number, and the result shows higher accuracy from previous CNNs models. We conducted experiments on two different dataset, which were used in the previous studies. One is Pun of the Day (Yang et al., 2015), and the other is 16000 One-Liners (Mihalcea and Strapparava, 2005). In addition, we constructed a Chinese dataset to evaluate the generality of the method performance on humor recognition against different languages.

3 Data

To fairly evaluate the performance on humor recognition, we need the dataset to consist of both humorous (positive) and non-humorous (negative) samples. The datasets we use to construct humor recognition experiments includes four parts: Pun of the Day (Yang et al., 2015), 16000 One-Liners (Mihalcea and Strapparava, 2005), Short Jokes dataset and PTT jokes. The four datasets have different joke types, sentence lengths, data sizes and languages that allow us to conduct more comprehensive and comparative experiments. We would like to thank Yang and Mihalcea for their kindly provision of two former datasets. And we depict how we collect the latter two datasets in the following subsections. Table 1 shows the statistics of four datasets.

Dataset	#Pos	#Neg	Type	Lang
16000 One-Liners	16000	16002	One-liner	EN
Pun of the Day	2423	2403	Pun	EN
Short Jokes	231657	231657	All	EN
PTT Jokes	1425	2551	Political	CH

Table 1: Statistics of four datasets

3.1 16000 One-Liners

16000 One-Liners dataset collected humorous samples from daily joke websites while using formal writing resources (e.g., news titles) to obtain non-humorous samples. A one-liner is a joke that usually has very few words in a single sentence with comic effects and interesting linguistic structure. While longer jokes can have a relatively complex linguistic structure, a one-liner must produce the humorous effect with very few words.

3.2 Pun of the Day

Pun of the Day dataset was constructed from the Pun of the Day website. The pun, also called paronomasia, is a form of wordplay that exploits multiple meanings of a term, or of similar-sounding words, for an intended humorous or rhetorical effect. The negative samples of this dataset are sampled from news website.

3.3 Short Jokes Dataset

Short Jokes dataset, which collected the most amount of jokes among four datasets, are from an open database on a Kaggle project[1]. It contains 231,657 short jokes with no restriction on joke types scraped from various joke websites and length ranging from 10 to 200 characters. We use it as our positive samples. For the negative samples, we choose WMT16[2] English news crawl as our non-humorous data resource. However, simply treating sentences from the resource as negative samples could result in deceptively high performance of classification due to the domain differences between positive and negative data. So we try to minimize such domain differences by selecting negative samples whose words all appear in the positive samples and whose average text length being close to the humorous ones.

3.4 PTT Jokes

PTT Bulletin Board System (PTT, Chinese: 批踢踢, telnet://ptt.cc) is the largest terminal-based bulletin board system (BBS) in Taiwan. It has more than 1.5 million registered users and over 20,000 boards covering a multitude of topics. Every day more than 20,000 articles and 500,000 comments are posted. Additionally, there is a

[1]https://www.kaggle.com/abhinavmoudgil95/short-jokes
[2]http://www.statmt.org/wmt16/translation-task.html

board called joke that we could acquire large amount of Chinese humor samples. Thus, we use some political-related words to extract political jokes from PTT and treat them as the positive samples. For the negative samples, we use Yahoo News in politics and select the samples by the same method we use in Short Jokes dataset to prevent from the problem of domain difference.

4 Method

In this section, we describe how we design our model for humor recognition.

4.1 CNN

Convolutional neural network (CNN) is a neural network architecture designed to extract local features in high dimensional data such as image or speech signal. When it comes to natural language processing (NLP), CNN also shows successes in several text categorization tasks (Johnson and Zhang, 2015). The input of most NLP tasks, such as a sentence or a document could be represented as a 2D structure with word embedding (Mikolov et al., 2013). In the input 2D matrix, each row is a vector of a word, a word segment or even a character that depends on the embedding methods. And typically we make the window width of the filters the same as the embedding dimension. Thus, the filter size varies according to a sliding window size we decide.

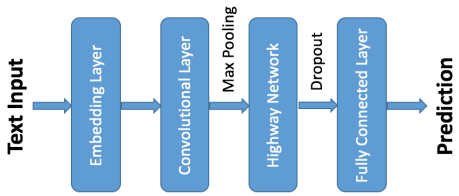

Figure 1: Network Architecture

4.2 Model Setting

In this paper, our CNN model's setup follows the Kim (2014) for the task of text classification. Figure 1 depicts the model's details. We firstly convert tokenized input sentence (length L) with word vector (dimension d) to a 2D matrix (L x d) by utilization of the GloVe embedding vectors (Pennington et al., 2014) which trained on 6B tokens and 400K vocabulary words of Wikipedia 2014 + Gigaword 5 as our embedding layer. Next, according to the average sentence length in the dataset, we tried different filter sizes with a range from 3 to 20. For each filter size, 100-200 filters are applied to the model. After convolutional layer, we exploit max pooling and then flatten the output. Assume we totally have n filters, eventually it will lead to a flatten 1D vector with dimension n at the prediction output.

4.3 Highway Layer

To improve the performance we usually can connect the flattened output with a fully connected layer and predict labels. In this paper, we would like to evaluate the performance improvement as we increase the network depth. However, the training of deeper networks becomes more difficult with increasing depth. So we use the concept of highway network (Srivastava et al., 2015) to help improve our model. The highway network allows shortcut connections with gate functions. These gates are data-dependent with parameters. It allows information unimpeded to flow through several layers in information highways. The architecture is characterized by the gate units that learn to regulate the flow of information through a network. With this architecture, we could train much deeper nets. In the end, we also use dropout and connect the results to the output layer.

5 Experiment

In this section, we describe how we formulate humor recognition as a text classification problem and conduct experiments on four datasets which we mentioned in Section 3. We validate the performance of different network structure with 10 fold cross validation and compare with the performance of previous work.

Table 2 shows the experiments on both 16000 One-Liners and Pun of the Day. We set the baseline on the previous works of Yang et al. (2015) by Random Forest with Word2Vec + Human Centric Feature (Word2Vec + HCF) and Chen and Lee (2017) by Convolutional Neural Networks. We choose a dropout rate at 0.5 and test our model's performance with two factors F and HN. F means the increase of filter size and number as we mentioned in section 4. Otherwise, the window sizes would be (5, 6, 7) and filter number is 100 that is the same with Chen and Lee (2017)'s. HN indicates that we use the highway layers to train deep networks and we set the HN layers = 3 because it has better stability and accuracy in training step. We could observe that when we use both F and

		16000 One-Liners				Pun of the Day			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Previous Work									
	Word2Vec+HCF	0.854	0.834	0.888	0.859	0.797	0.776	0.836	0.705
	CNN					0.861	0.857	0.864	0.864
Our Methods									
	CNN	0.877	**0.899**	0.856	0.877	0.867	0.880	0.859	0.869
	CNN+F	0.892	0.896	0.928	0.898	0.892	0.886	0.907	0.896
	CNN+HN	0.885	0.877	0.902	0.889	0.892	**0.889**	0.903	0.896
	CNN+F+HN	**0.897**	0.872	**0.936**	**0.903**	**0.894**	0.866	**0.940**	**0.901**

Table 2: Comparison of Different Methods of Humor Recognition

HN our model gives the best performance on both accuracy and F1-Score and this conclusion is consistent across two datasets. The results show that our model helps increase F1-Score from 0.859 to 0.903 on 16000 One-Liners and from 0.705, 0.864 to 0.901 on Pun of the Day compared to previous work

Dataset	Accuracy	Precision	Recall	F1
Short Jokes	0.906	0.902	0.946	0.924
PTT Jokes	0.957	0.927	0.959	0.943

Table 3: Result of Short Jokes and PTT Jokes datasets

Table 3 presents the result of Short Jokes and PTT Jokes datasets. As we can see, for the datasets was construed, it achieve 0.924 on Short Jokes and 0.943 on PTT Jokes in terms of F1 score respectively. It shows that the deep learning model can, to some extent learn the humorous meaning and structure embedded in the text automatically without human selection of features.

6 Discussion

In this section, we show a sample in each category (true positive, false positive, true negative and false negative) to get a sense of what kinds of sentences are predicted correctly and incorrectly. The sentences are shown in the table 4.

	Sentence
TP	when he gave his wife a necklace he got a chain reaction
TN	the barking of a dog does not disturb the man on a camel
FP	rats know the way of rats
FN	it's a fact taller people sleep longer in bed

Table 4: Example Sentences

The TP sentence "when he gave his wife a necklace he got a chain reaction" shows that our model seems to be able to catch not only the literal meaning between the "necklace" and "got a chain reaction". Besides, the TN sentence "the barking of a dog does not disturb the man on a camel" means that if you're lucky enough to own your own camel, a little thing like a barking dog won't bother you. The example is a proverb but not a joke and our model correctly recognizes it as a non-humor one. Model misclassifies certain instances such as the FP sentence "rats know the way of rats" is actually derived from a Chinese proverb and the model predict it as humor. In addition, the FN sentence "it's a fact taller people sleep longer in bed" is obviously a joke but it is not considered as a humor by the model. To deal with more subtle humor/non-humor, the model has room to be improved.

7 Conclusion

In this study, we have extended the techniques of automatic humor recognition to different types of humor as well as different languages in both English and Chinese. We proposed a deep learning CNN architecture with high way networks that can learn to distinguish between humorous and non-humorous texts based on a large scale of balanced positive and negative dataset. The performance of the CNN model outperforms the previous work. It's worth mentioning that the recognition accuracy on PTT, political jokes in Chinese, and the short jokes dataset with various types of jokes in English are both as high as above 90%. The novel deep learning model relieves the required human intervention of selection linguistic features for humor recognition task. In future work, we would conduct more rigorous comparative evaluation with human humor recognition and look into how the humorous texts can be generated using deep learning models as well.

References

Dario Bertero and Pascale Fung. 2016. A long short-term memory framework for predicting humor in dialogues. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 130–135.

Lei Chen and Chong MIn Lee. 2017. Predicting Audience's Laughter Using Convolutional Neural Network. *ArXiv e-prints:1702.02584*.

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 103–112.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 531–538.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Luke de Oliveira and Alfredo Lainez Rodrigo. 2015. Humor detection in yelp reviews.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Amruta Purandare and Diane J. Litman. 2006. Humor: Prosody analysis and automatic recognition for f*r*i*e*n*d*s*. In *EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 208–215.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2377–2385.

Julia M. Taylor and Lawrence J. Mazlack. 2004. Computationally recognizing wordplay in jokes. In *In Proceedings of CogSci 2004*.

Diyi Yang, Alon Lavie, Chris Dyer, and Eduard H. Hovy. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2367–2376.