Extended Abstract for WiNLP

Anonymous ACL submission

Abstract

Humor represents one of the most unique and intelligent activities that define humans. Our research addresses humor detection task by using a variety of methods with the focus on the continuous and subjective nature of humor based on a particular humorous Twitter dataset, instead of treating it as a binary classification task. So far we have applied language model approach to this problem. This paper presents the current work of our research along with promising results, as well as possible future work.

1 Introduction

Humor is considered to be a human-only trait and one of the most amusing and mystifying human activities. It enters the domain of philosophy, sociology, psychology, linguistics and computer science. With the increasing development of Artificial Intelligence(AI), Machine Learning(ML) and computational linguistics, *Computational humor* has found its way to numerous studies. Humor generation and humor detection have been the prevailing focus of computational humor. Researchers have developed well working systems to produce humor (e.g., (Stock and Strapparava, 2003), (Özbal and Strapparava, 2012)). However, *humor detection* remains a less explored and challenging problem.

Most studies have treated humor detection as a binary classification problem, which is to classify whether a piece of data is humorous (e.g., (Mihalcea and Strapparava, 2006), (Zhang and Liu, 2014), (Shahaf et al., 2015), (Miller and Gurevych, 2015)). Although the binary classification technique is an applicable approach, it ignores the continuous and subjective nature of humor.

In our research, we specifically take these factors into account when we choose methods and develop systems to detect humor. In order to address the subjective characteristic of humor, we focus on learning a sense of humor from a particular tweet dataset constructed by Potash et al (Potash et al., 2016), which consists of humorous tweets submitted to a comedy show @*midnight*. To represent the continuous trait of humor, we use methods that rank tweets based on their humorous level rather than classifying whether each tweet is humorous.

2 Language Models

We chose traditional N-gram Language Models (LMs) as a starting point of our research for two reasons: First, training N-gram LMs allows us to detect humor by learning a sense of humor from a word and its neighbors in a tweet. Additionally, language model approach is suitable to represent the continuous nature of humor considering its ability of ranking tweets based on the probability score it computes for each tweet. Our research is associated with the SemEval-2017 Task 6 #HashtagWars: Learning a Sense of Humor (Potash et al., 2017) since its purpose corresponds to ours. There are two subtasks involved: Pairwise Comparison (Subtask A) and Semi-ranking (Subtask B). More details about the task refers to (Potash et al., 2017).

In the SemEval task, we developed system Duluth [1] to estimate tweet probability using N-gram LMs. Our system ranked tweets based on how funny they are by training N-gram LMs on two different datasets: the funny tweet data provided by the task and the news data which is freely available for research [2]. Tweets that have a higher proba-

[1] https://xinru1414.github.io/HumorDetection-SemEval2017-Task6/

[2] http://www.statmt.org/wmt11/featured-translation-task.html

bility according to the tweet data language model are ranked as being funnier. However, tweets that are less probable according to the news language model are considered the funnier since they are the least like the (unfunny) news corpus. We relied on both bigrams and trigrams when training our models considering tweets' concise and short nature. KenLM (Heafield et al., 2013) was used as our language modeling tool, with modified Kneser-Ney smoothing and back-off technique. More details about our system refers to (Yan and Pedersen, 2017).

Table 1 shows our results based on LMs trained on both datasets using bigram and trigram. Note that the accuracy and the distance measurements listed are defined by the task organizers (Potash et al., 2017). For Subtask A, higher accuracy score represents better system performance; for Subtask B, lower distance score means the system works more effective.

DataSet	N-gram	Subtask A Accuracy	Subtask B Distance
tweets	trigram	0.397	0.967
tweets	bigram	0.406	0.944
news	trigram	0.627	0.872
news	bigram	0.624	0.853

Table 1: Results based on bigram and trigram LMs on both datasets. The trigram LM trained on the news data ranked 4th place on Subtask A and 1st place on Subtask B during the SemEval task evaluation.

From the results we can tell that: LMs trained on the news data had a significant advantage over models trained on the tweets data; Bigram LMs performed slightly better than trigram LMs on both datasets; Our system had a better performance on Subtask A over B.

We believe that the significant advantage of the news data over the tweet data is caused by the difference in quantity between corpora. The tweet data was significantly smaller than the news data. In the future we intend to collect more tweet data, especially the ones participating the #Hashtag-Wars. We also plan to experiment on cutting the amount of news data and then build the models to see how the results compare.

Our system performed much better on Subtask A comparing to B, which reveals that the semi-ranking task has a higher requirement on the system to represent the continuous nature of humor comparing to the pair-wise comparison.

3 Deep Learning

Our system performed well in the SemEval task, confirming the ability of language models to detect humor with focusing on its continuous and subjective nature. Going forward, language models certainly can characterize humor in a more comprehensive way with more complex implementations.

Recently, studies have shown that Recurrent Neural Networks (RNN) such as LSTM neural networks are exceptionally powerful in language modeling based on its ability to take into account of all preceding words over a word sequence (Sundermeyer et al., 2012) (Sundermeyer et al., 2015). Also, in their work of learning a sense of humor, (Potash et al., 2016) points out that due to the high portion of out of vocabulary (OOV) words generated by puns in the humorous tweet data, character-level Constitutional Neural Networks (CNN) model would be more suitable for capturing the single-token puns comparing to token level models. In addition, (Kim et al., 2015)'s study proves that character-level neural language models outperform state-of-art word-level neural language models. Moving forward, considering the promising future of neural networks, we would like to build a character-level neural language model (NLM) which relies on both LSTM RNN and character-level CNN to improve our current system performance.

Last but not the least, (Potash et al., 2016) states that external knowledge is crucial in order to detecting humor from the tweet dataset based on its nature. In the future we intend to make use of external knowledge in the NLM in possibly two ways: With a deeper understanding of what kinds of external knowledge are the most useful such as movie and book titles, celebrity and song names, one is to incorporate them as features in the NLM; The other is to combine the NLM with a Neural Turing Machine (NTM) (Graves et al., 2014), which is particularly designed for interacting with external interfaces, with a profound and solid study of NTM.

4 Conclusion

Humor has not been addressed broadly in current computational research area. Our research focuses on humor detection, developing systems that can

capture its continuous and subjective nature.

References

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* .

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pages 690–696.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2015. Character-aware neural language models. *arXiv preprint arXiv:1508.06615* .

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence* 22(2):126–142.

Tristan Miller and Iryna Gurevych. 2015. Automatic disambiguation of english puns. 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)*. Association for Computational Linguistics, Beijing, China, pages 719–729.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 703–711.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2016. # hashtagwars: Learning a sense of humor. *arXiv preprint arXiv:1612.03216* .

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: #HashtagWars: learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, BC.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '15, pages 1065–1074.

Oliviero Stock and Carlo Strapparava. 2003. Getting serious about the development of computational humor. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Acapulco, pages 59–64.

Martin Sundermeyer, Hermann Ney, and Ralf Schlüter. 2015. From feedforward to recurrent lstm neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 23(3):517–529.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Interspeech*. pages 194–197.

Xinru Yan and Ted Pedersen. 2017. Duluth at semeval-2017 task 6: Language models in humor detection. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, BC.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '14, pages 889–898.