Interspeech 2018
2-6 September 2018, Hyderabad

Computational modeling of conversational humor in psychotherapy

Anil Ramakrishna[1], *Timothy Greer*[1], *David Atkins*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, USA
akramakr@usc.edu, timothdg@usc.edu, datkins@uw.edu, shri@ee.usc.edu

Abstract

Humor is an important social construct that serves several roles in human communication. Though subjective, it is culturally ubiquitous and is often used to diffuse tension, specially in intense conversations such as those in psychotherapy sessions. Automatic recognition of humor has been of considerable interest in the natural language processing community thanks to its relevance in conversational agents. In this work, we present a model for humor recognition in Motivational Interviewing based psychotherapy sessions. We use a Long Short Term Memory (LSTM) based recurrent neural network sequence model trained on dyadic conversations from psychotherapy sessions and our model outperforms a standard baseline with linguistic humor features.

Keywords: Automatic Humor Recognition, Psychotherapy, Motivational Interviewing

1. Introduction

Humor is a significant aspect of language and is well studied in a variety of disciplines such as psychology [1], linguistics [2] and psychotherapy [3]. It has been associated with an individual's well-being [4], higher intimacy [5] and trust in relationships [6, 7]. It is also often suggested as a means to reduce anxiety [8]. This makes it a useful tool in psychotherapy where a trained counselor tries to address psychological conditions or provide emotional support to clients [3].

Motivational Interviewing (MI) is a well known form of psychotherapy, commonly used in addressing conditions such as substance abuse disorders [9]. In MI, the counselor tries to elicit the motivation for behavioral change from within the client by subtly steering conversations towards this goal. Humor can be especially useful in this context for establishing the relationship and trust with the client; it was previously studied in this context by [10] where they examined the association of shared laughter with desirable counselor behavior.

Recent advancements in Artificial Intelligence have resulted in development of several computer assisted psychotherapy tools including virtual reality based exposure therapy, therapeutic computer games and intelligent agents for psychotherapy [11]. Intelligent agents have been used to simulate both an artificial therapist for counseling [12] and a client for training therapists [13]. These are typically conversational agents that use a Natural Language Understanding (NLU) component at their core and are usually capable of analyzing sentiment and humor. As a result computational modeling of humor has gained considerable interest in the NLU community. A model to identify and/or generate humor automatically can be used in a variety of conversational agents including the virtual therapy systems described above. However, this is a challenging task due to the inherent ambiguities and subjectiveness in the definition of humor. Moreover, the data used to train humor recognition systems has been typically limited to short sentences. In this work, we propose a model to automatically recognize humor from arbitrarily long MI therapy conversations which can make use of contextual information to improve predictions.

Several theories have been proposed to model humor, specially in the context of linguistics [2]. Humor is often categorized in three modes: incongruity (contrasting meanings), hostility (derision) and release theories [14]. Of particular importance is the theory of incongruity where humor is associated with the presence of benign violations from situational expectations [15], which are connected with surprise and emotional transformation in the receiver [16]. Linguistic features have been developed to capture this incongruity with some success in computational humor recognition systems [17]. However these have been limited to making predictions on one line jokes or tweets and fail to capture contextual information that is often relevant in conversational humor. We address this issue here by using a recurrent neural network (RNN) based sequence model to capture the relevant contextual information for humor prediction. RNNs are designed to be able to capture context between arbitrarily far inputs and are hence well suited for sequential data such as language. We use a hierarchical model with two Long Short Term Memory (LSTM) cells used to encode input utterances and make humor predictions respectively. We present results on two variants of this model and compare it with a standard baseline that uses humor and context features.

The rest of the paper is organized as follows: we explain the related work in next section and describe the data set in section 3. We explain our model in section 4, the experimental setup in section 5 followed by results in section 6 before concluding in section 7.

2. Related work

Despite their recent popularity, computational humor recognition systems are still limited in their applicability due to a variety of factors such as personal and cultural subjectivities involved in humor along with the myriad subcategories. In contrast, humor generation has been been well studied [18] thanks to the large number of theories of humor which can be instantiated based on context to generate humorous text. Humor recognition has also been limited in the domains in which it has been deployed owing to the limited number of datasets available with humor labels. Most research works in this thread construct humorous data from tweets and one liners and non-humorous data samples from often unrelated domains such as news articles [18].

One of the early efforts in automatic humor recognition was in [17] where the authors proposed simple linear classification systems which used carefully selected stylistic features designed to capture humorous intent from text. Such features were further expanded in several subsequent works such as [19]

10.21437/Interspeech.2018-1583

Table 1: *Statistics of dataset used*

Number of sessions	96
Number of utterances	26428
Number of humorous utterances	2251 (8.5%)

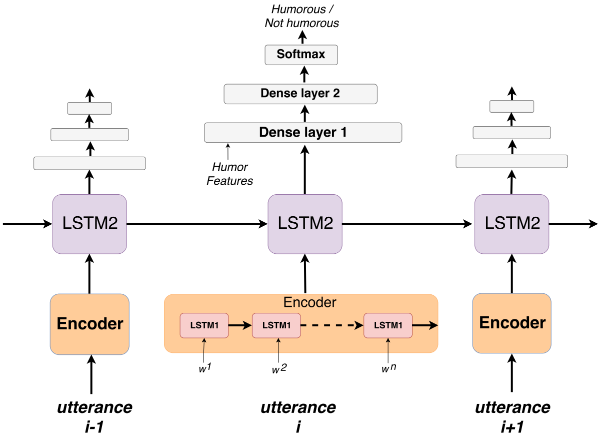

Figure 1: *Hierarchical LSTM sequence model with sentence encoder. Words from the i^{th} utterance are input sequentially to the encoder. Last state of the encoder is passed to the second LSTM cell. Humor features are concatenated with the output of the second LSTM cell and passed to the first dense layer.*

and [18].

Though viable in recognizing humor from one liners and tweets, the linguistic features fail to capture context between utterances and hence not readily usable in conversational agents. For example, an utterance may contradict a statement made several turns ago resulting in a humorous remark. Handling these using linguistic features would need an expanded window over which the features are computed. A more elegant solution would be to model the utterances sequentially, which is the main theme of this paper. Our work is similar to [20], where they use an RNN model to identify dialogs that are followed by audience/canned laughter from television sitcoms, which are tagged as humorous. However, since these are almost always laughter induced by the show writers, it may be better described as *intended humor* and not *perceived humor*. Further, since the conversations in sitcoms are not necessarily indicative of real life, it is unclear if the system is generalizable outside their setting. Our model avoids these by training on more authentic conversations from psychotherapy sessions.

3. Data

Our dataset consists of conversations from 353 psychotherapy sessions which were part of six motivational interviewing based clinical trials (ARC, BAER, ESP21, ESPSB, HMCBI, ICHAMP) [21, 22]. In all of these sessions, the counselors use MI to address various forms of substance abuse with sessions varying in duration from 8 minutes to over an hour. Each session used here was transcribed manually along with utterance level behavioral labels from the Motivational Interviewing Skill Code (MISC) [23] and non-verbal cues such as laughter.

We use these laughter tags to identify humorous utterances. However, since laughter may also indicate nervousness, we only label those instances in which both the client and the therapist share laughter with separate laughter tags within a fixed search window (of size 5 utterances) as humorous. We also filter out sessions with fewer than five such shared laughters in order to minimize the class imbalance. Statistics for the final dataset used in our experiments are shown in table 1.

4. Model

We use a hierarchical RNN model with Long Short Term Memory (LSTM) cells [24] as shown in figure 1. The LSTM cell uses a carefully designed structure to retain information between inputs for a long time. They avoid the issues of vanishing or exploding gradients found in conventional recurrent networks using a memory cell which contains the information shared between time units. The contents of the cell are regulated using three neurons or gates: an *input gate*, *output gate* and a *forget gate*.

Our model uses two LSTM cells: the first cell (*encoder*) combines words from an input utterance to create a fixed dimensional distributed representation and the second cell (*classifier*) accepts these representations as input and makes humor predictions. The second cell operates across utterances by capturing the context, leading to a two layer hierarchical structure as shown in the figure. We also experimented with attention mechanism [25] to combine the word embeddings but the performance was slightly lower than the above model.

The encoder cell takes a sequence of words as input and generates a sentence level *embedding* as the output of the cell corresponding to the last word. Embeddings are representations of words and sentences in a fixed dimensional real vector space and are useful in capturing semantic similarities. Words from an utterance can be input to the encoder cell using either a standard one hot representation, or using generic or task specific word embeddings. In our experiments, we evaluate both task specific word embeddings trained in an end to end manner as well as generic glove embeddings. The sentence embeddings from the first cell are passed as inputs to the classifier cell which predicts humor at the utterance level. As shown in the figure, outputs from the second cell are concatenated with humor features from section 4.4 before passing them through two fully connected dense layers with hyperbolic tangent activation functions and a softmax layer. The second dense layer has fewer neurons than the first. In our experiments, this multilayer structure had better performance than passing the LSTM output through one dense layer as is common in practice. The second LSTM cell makes prediction for each input utterance. We compare our model with an utterance level SVM baseline that operates on sentence level embeddings concatenated with the humor features of section 4.4.

4.1. SVM

Our baseline is a Support Vector Machine (SVM) classifier trained on glove based sentence embeddings along with task specific humor features described in section 4.4. SVMs have been used to make binary predictions of humor in several previously reported works [17, 26]. Glove [27] is an unsupervised task agnostic word representation algorithm trained using co-occurrence of words in a corpus. It has been shown to quantitatively capture various forms of semantic similarities in words.

The SVM model was trained on utterance level embedding vectors computed by averaging word level glove embeddings, concatenated with the humor features.

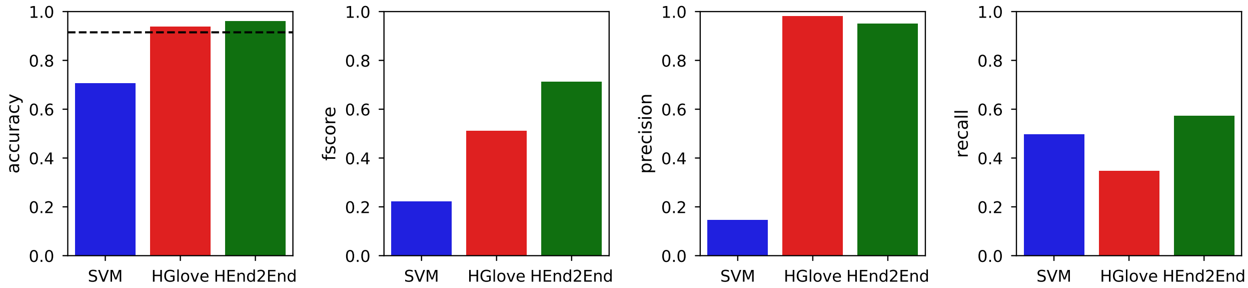

Figure 2: *10-fold CV performance of the three models; dashed line represents majority baseline*

4.2. Hierarchical Glove based LSTM (HGlove)

In this model, we train the full Hierarchical LSTM network on the input utterances but fix the input word embeddings from glove. This model was explored to evaluate the applicability of task agnostic word embeddings such as glove in humor recognition when trained with a sequence model.

4.3. Hierarchical End to End LSTM (HEnd2End)

Our final model is similar to HGlove except it was trained in an end to end fashion such that task specific word level embeddings were trained as part of the humor classification task. This model can make use of the training labels and induce a distribution in the word embeddings that can better capture the word level features required for predicting humor. Similar to the previous model, output of the classifier cell was passed through two fully connected layers before the softmax layer.

4.4. Humor features

Several linguistic features have been proposed as candidates to capture different forms of humor [2]. These include stylistic features such as rhyme chains and alliteration chains, ambiguity measures, measures of emotional content, etc. We use a subset of these in our classifiers.

4.4.1. Structural features

Our first set of features includes simple counts such as number of words, average word length and percentage of uppercase and lowercase characters as suggested by [28].

4.4.2. Stylistic features

[17] reported successes in using stylistic features such as number of rhymes or alliterations in the utterance. These are phonetic characteristics of words which are used such that their arrangement leads to humor. Rhyming words end with similar sounds (ex: *clean* and *glean*) while alliteration uses words that start similar (ex: *Peter* and *Piper*). Several studies have reported the use of rhyming words [29] and alliteration to deliver or enhance humor.

Similar to [17], we use the CMU pronunciation dictionary[1] to extract phonetic transcriptions for each word in the utterance and identify non-overlapping and longest possible chains for both rhymes and alliterations. We use counts of both types of chains as features in our experiments.

4.4.3. Ambiguity

Several works report the use of ambiguity in predicting humor from text [28]. This can be attributed to the theory of subverting expectations which is frequently associated with humor [16, 15]. To measure ambiguity in an utterance, we used Wordnet[2] similar to [28] to get the average number of *synsets* associated with each word, since higher this number, higher the apparent ambiguity in its meaning.

5. Experimental setup

The SVM model was trained using a linear kernel (selected by tuning on a held out set). We used l2 penalty and hinge loss along with class balancing to handle the skewed labels. The RNN models were implemented using PyTorch [30] and trained on a CUDA enabled machine. In all experiments, word level embeddings (including glove) were fixed to be 100 dimensional while sentence embeddings were 200 dimensional in the RNN models. The LSTM cells used a single layer with hidden state of 50 dimensions. The dense layers were of dimensionality 100 and 25 respectively in both the RNN models.

Dropout [31] was enabled ($p = 0.5$) on both the LSTMs and the fully connected layers to avoid overfitting. The neural networks were trained for 25 epochs with batch size of 5 and were optimized using RMSProp [32] with learning rate 10^{-3} and cross entropy loss. All models were trained using session level 10-fold cross validation, where a subset of the therapy sessions were held out as the test set.

6. Results

Figure 2 shows bar plots for 10 fold cross validated accuracy, f-score, precision and recall for the three models. Note that despite the majority class achieving higher accuracy than the SVM as shown, we do not include that as a baseline since it would have 0 recall and f-score. The RNN models were trained with sequence length (number of utterances) set to 5.

Both the RNN models show higher accuracy and f-score compared to the SVM model, with the end to end model outperforming the HGlove model in both metrics. The HGlove model seems to be predicting a small number of utterances as positive as evidenced by its higher precision and reduced recall. This indicates that the generic glove embeddings maybe limited in their capacity to capture the task and domain specific features maybe required for predicting humor. On the other hand, the

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] http://wordnet.princeton.edu/

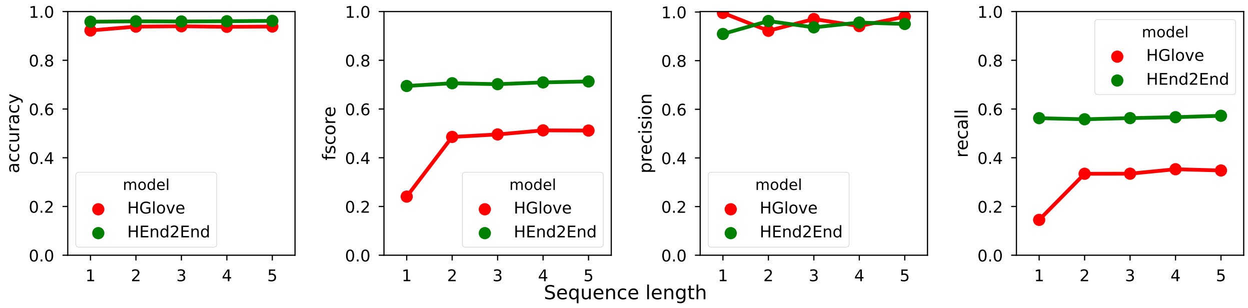

Figure 3: *Performance of the RNN models for different sequence lengths (best viewed in color).*

HEnd2End makes use of the training labels to induce a task specific distribution for the word embeddings which are able to better capture the relevant features for predicting humor. As seen in the figure, HEnd2End has the highest recall of the three models along with a high precision score comparable with HGlove.

To evaluate the effect of sequence length on the RNN models, we also ran experiments for lengths 1 through 5 and the results are shown in figure 3. Both models maintain high accuracy and precision over all sequence lengths. However, with just one utterance, the HGlove model seems to identify only a small fraction of utterances as humorous leading to perfect precision but very low recall and hence a low f-score. As we increase the this length, it seems to be able to make use of available context to achieve higher recall and f-score, suggesting the benefits of using a sequence model for humor prediction. The HEnd2End model, on the other hand only shows marginal improvement when trained with more context. It maintains high performance for all sequence lengths, suggesting that the model is able to learn task specific features which can predict humor even in the absence of context.

7. Conclusions

We presented an recurrent neural network model to predict humor from psychotherapy conversations. Our model used a hierarchical two layer structure with an LSTM based sentence encoder to learn utterance level embeddings. We evaluated two variants of the model with generic and task specific embeddings. In both cases the RNN models outperformed a standard baseline trained with linguistic humor features. The glove based model showed improved performance when trained with longer sequences indicating that context can be useful in humor prediction. The end to end model showed higher performance than the glove based model even in the absence of any context by making use of task specific embeddings.

Future work includes extending the trained model to other domains. Further analysis of the learned distribution of the word embeddings may lead to development of new linguistic features relevant in predicting humor. Error analysis on the types of mistakes made by the model may also help uncover hidden patterns in humor.

8. Acknowledgements

The authors would like to thank Karan Singla and Jimmy Gibson for their feedback and all the helpful discussions.

9. References

[1] J. E. Roeckelein, *The psychology of humor: A reference guide and annotated bibliography.* Greenwood Press/Greenwood Publishing Group, 2002.

[2] S. Attardo, *Linguistic theories of humor*. Walter de Gruyter, 2010, vol. 1.

[3] K. Rutherford, “Humor in psychotherapy.” *Individual Psychology: Journal of Adlerian Theory, Research & Practice*, 1994.

[4] N. A. Kuiper and R. A. Martin, “Humor and self-concept,” *Humor-International Journal of Humor Research*, vol. 6, no. 3, pp. 251–270, 1993.

[5] W. P. Hampes, “Relation between intimacy and humor,” *Psychological reports*, vol. 71, no. 1, pp. 127–130, 1992.

[6] B. Muthayya, “Relationship between humour and inter-personal orientations.” *Journal of Psychological Researches*, 1987.

[7] W. P. Hampes, “The relationship between humor and trust,” *Humor-International Journal of Humor Research*, vol. 12, no. 3, pp. 253–260, 1999.

[8] T. E. Ford, S. K. Lappi, E. C. OConnor, and N. C. Banos, “Manipulating humor styles: Engaging in self-enhancing humor reduces state anxiety,” *Humor*, vol. 30, no. 2, pp. 169–191, 2017.

[9] S. Rubak, A. Sandbæk, T. Lauritzen, and B. Christensen, “Motivational interviewing: a systematic review and meta-analysis,” *Br J Gen Pract*, vol. 55, no. 513, pp. 305–312, 2005.

[10] R. Gupta, T. Chaspari, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, “Analysis and modeling of the role of laughter in motivational interviewing based psychotherapy conversations,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] D. D. Luxton, “Artificial intelligence in psychological practice: Current and future applications and implications.” *Professional Psychology: Research and Practice*, vol. 45, no. 5, p. 332, 2014.

[12] A. Rizzo, G. Lucas, J. Gratch, G. Stratou, L. Morency, R. Shilling, and S. Scherer, “Clinical interviewing by a virtual human agent with automatic behavior analysis,” *The 2016 Proceedings of the ICDVRAT*, 2016.

[13] P. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo, “Virtual patients for clinical therapist skills training,” in *International Workshop on Intelligent Virtual Agents*. Springer, 2007, pp. 197–210.

[14] V. Raskin, *Semantic mechanisms of humor*. Springer Science & Business Media, 2012, vol. 24.

[15] A. P. McGraw and C. Warren, “Benign violations: Making immoral behavior funny,” *Psychological science*, vol. 21, no. 8, pp. 1141–1149, 2010.

[16] T. C. Veatch, “A theory of humor,” 1998.

[17] R. Mihalcea and C. Strapparava, “Learning to laugh (automatically): Computational models for humor recognition,” *Computational Intelligence*, vol. 22, no. 2, pp. 126–142, 2006.

[18] D. Yang, A. Lavie, C. Dyer, and E. Hovy, “Humor recognition and humor anchor extraction,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2367–2376.

[19] R. Zhang and N. Liu, “Recognizing humor on twitter,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 889–898.

[20] D. Bertero and P. Fung, “A long short-term memory framework for predicting humor in dialogues,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 130–135.

[21] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, “Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification,” *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[22] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, “Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors,” *Journal of substance abuse treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[23] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, “Manual for the motivational interviewing skill code (misc),” *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2014.

[26] B. Charalampakis, D. Spathis, E. Kouslis, and K. Kermanidis, “A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets,” *Engineering Applications of Artificial Intelligence*, vol. 51, pp. 50–57, 2016.

[27] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[28] A. Morales and C. Zhai, “Identifying humor in reviews using background text sources,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 492–501.

[29] W. Menninghaus, I. C. Bohrn, U. Altmann, O. Lubrich, and A. M. Jacobs, “Sounds funny? humor effects of phonological and prosodic figures of speech.” *Psychology of aesthetics, creativity, and the arts*, vol. 8, no. 1, p. 71, 2014.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *NIPS-W*, 2017.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.