PREDICTING HUMOR RESPONSE IN DIALOGUES FROM TV SITCOMS

Dario Bertero, Pascale Fung

Human Language Technology Center
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
dbertero@connect.ust.hk, pascale@ece.ust.hk

ABSTRACT

We propose a method to predict humor response in dialog using acoustic and language features. We use data from two popular TV sitcoms – "The Big Bang Theory" and "Seinfeld" – to predict how the audience responds to humor. Due to the sequentiality of humor response in dialogues we use a Conditional Random Field as classifier/predictor. Our method is relatively effective, with a maximum precision obtained of 72.1% in "Big Bang" and 60.2% in "Seinfeld". Experiments show that audio, speed, word and sentence length features are the most effective. This work is applicable to develop appropriate machine response empathetic to emotion in dialog, in addition to humor.

Index Terms— humor response, dialog, empathetic computing, TV sitcoms, emotion detection

1. INTRODUCTION

Humor describes a verbal or textual, sometimes physical stimuli that triggers a response such as laughter in the recipient. Humor response have been found to benefit humans in a number of ways. Researchers [1, 2, 3] report a wide range of ways where humor is helpful to promote physical well-being and to establish and maintain human relationships. Therapists and counselors report that humor facilitates problem solving and reduces anxiety and stress. Laughter helps to reduce depression, aggression and negative reactions, even boosts the immune system and improves the body's response to virus and even cancers [4, 5, 6]. Humor is without doubt an important emotion in human communications.

Recently there have been a surge in research on humor classification of individual sentences [7], Twitter data [8, 9, 10], product reviews [11] and discussion forums [12]. Our work aims to predict humor response from dialogues, and therefore departs from the previous humor detection tasks. We aim to predict where, in a typical dialog, the recipient would laugh. This analysis has great implications in the field of emotion detection and empathetic computing. It is a first step towards building machine systems, such as chat robots, that can fully understand and share a joke, and respond appropriately to humor and other emotional stimuli from human users. This has great potential for human-machine interaction systems in the future.

All the previous cases employ a supervised classification task based on language features, with some attempts to capture syntactic and semantic structures highly correlated with humor [10, 7]. Only [13] made use of acoustic and prosodic features, and none of them so far has taken into account the combination of acoustic and language features. Moreover previous methods assume humor exists in isolated sentences or utterances, or eventually grouped under a common topic but still independent. In a spoken or written dialog, however, the humorous effect is often generated by the context – psychologists observe a certain "setup" of humor where the recipient is "prepared" to receive a stimuli which then came in the form of humorous "triggers", followed by the "punchline", which are then immediately followed by laughter and other types of humor response. In themselves, specific utterances or discourse segments might or might not trigger laughter depending on where and when they are used. A short but clear example is shown in figures 1 and 2: the utterances underlined are very similar to each other, but in figure 1 it is just the setup for a subsequent sarcastic joke, while in figure 2 the utterance itself is the punchline that triggers the laugh.

We analyze dialogues from two popular TV sitcoms, namely "The Big Bang Theory" (i.e "Big Bang") and "Seinfeld". Sitcoms are interesting for several reasons: they provide a proper dialog flow, with both the audio and the transcription, to allow a multi-modal analysis. They include canned laughters which follow each punchline [14]. They provide a pretty good indication of when the audience would laugh and are easy indicators of humor response. Two example extracts are shown in figures 1 and 2. Finally the sitcom domain has never been studied computationally previously. Yet most of the work based on this domain may be transposed to build a real conversational humor generation and response in an empathetic human-machine interactive system. In this paper we propose to learn the dialog act of human response sequentially and predict when a laughter occurs from previous utterances that might include the setup and the trigger, regardless of the specific kind of humor which generates it.

This work is funded by the Hong Kong PhD Fellowship Scheme.

Fig. 1: Example from The Big Bang Theory:
LEONARD: I did a bad thing.
SHELDON: Does it affect me?
LEONARD: No.
SHELDON: ***Then suffer in silence.*** **LAUGH**

Fig. 2: Example from Seinfeld:
GEORGE: You simply must apologize.
JERRY: Must I?
GEORGE: Yes. Because it is the mature, adult thing to do.
JERRY: ***How does that affect me?*** **LAUGH**

2. METHODOLOGY

We propose to use a sequential supervised classification approach to predict humor response in a dialog. We extract a combination of acoustic features from the audio track, and language features from the scripts and train a Conditional Random Field [15] to fully take advantage of the sequential structure of our data.

2.1. Acoustic features

It is very typical in our everyday life that the same utterance, repeated with different intonations, loudness or generically in a different way generates a very different emotional effect on the listener. For this reason we extract several typical acoustic features from the audio track of each utterance.

We use the openSMILE software package [16] to extract a total of around 2500 acoustic and prosodic features from the "emobase" and "emobase2010" (made of the feature set from the INTERSPEECH 2010 paralinguistic challenge [17]) packages provided. These packages include features specifically assembled for emotion classification tasks. Among them there are MFCC, pitch, intensity, loudness, probability of voicing, F_0 envelope, Line Spectral Frequencies, Zero-Crossing Rate and their variations (delta coefficients).

In addition we also take into account each utterance rate in the form of duration in time, obtained from the subtitles files, divided by the number of words. We would expect that a deliberately fast or slow pace would sometimes imply a humorous intent.

2.2. Language features

Different aspects of language can represent humor. Therefore we include different views of features extracted from the closed caption provided. Some of our features are also based on the comparison between an utterance and its preceding ones. Our feature set includes:

1. Lexical: unigram, bigram and trigrams. We discard n-grams that appears less than 5 times.
2. Structural features [9]: number and proportion of nouns, verbs, adjectives and adverbs, sentence length, difference in sentence length with the previous utterance, and average word length. They model the structure and the syntactic content of each utterance.
3. Ambiguity [9]: we take the mean, the maximum and the difference between mean and maximum of the number of WordNet synsets [18] of each word.
4. Antonym: we identify the presence of antonyms, retrieved from WordNet, with the previous utterance. We use four binary features for nouns, verbs, adjectives and adverbs.
5. Sentiment [9]: they aim to evaluate the sentiment content and polarity of words and utterances. For each word we extract the positive and negative sentiment scores from SentiWordNet [19]. Then we take the average of all positive scores, the average of all negative scores, the average and the difference between these two scores.
6. Latent semantic features: we take the cosine similarities between latent semantic vector representations of each utterance with the four previous, and with the whole scene vector. We used the model introduced in [20], trained with the default corpus consisting of WordNet sense definitions, Wikitionary definitions and examples and the Brown corpus. These values are intended to capture shifts in the lexical and semantic content along the discourse, and utterances out of context with the scene. These elements may often trigger humor, in the form of sarcasm or nonsense.
7. Speaker turns: character who speaks the utterance and position of the utterance inside the speaker turn (beginning, middle, end, isolated). These features are aimed to capture the characters more likely to be humorous, and the fact that humor is often triggered at the end of long turns or with short isolated replies.

	The Big Bang Theory				Seinfeld			
Features	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
All positive baseline	42.8	42.8	100.0	59.9	26.6	26.6	100.0	42.0
All negative baseline	57.2	0.0	0.0	0.0	73.4	0.0	0.0	0.0
Acoustic +speed	69.2	67.1	55.1	60.5	73.8	51.4	24.4	33.1
Language	62.8	58.0	47.6	52.3	68.7	37.4	26.1	30.7
Language -ngrams	62.3	60.4	34.6	44.0	72.0	41.7	13.2	20.1
All -speed -ngrams	65.2	61.5	49.8	55.0	73.3	49.5	24.6	32.9
All -latent -ngrams	72.6	70.9	60.8	65.5	**76.6**	**60.2**	35.2	44.4
All -structural -ngrams	70.3	67.7	58.4	62.7	74.7	54.3	30.0	38.7
All -antonyms -ngrams	72.8	71.3	60.8	65.7	76.3	58.7	35.6	44.4
All -ambiguity -ngrams	**73.4**	**72.1**	61.8	**66.5**	76.5	59.4	36.5	45.2
All -sentiment -ngrams	72.7	71.0	61.2	65.7	76.1	58.5	34.9	43.7
All -turns -ngrams	72.4	71.1	60.0	65.0	76.1	59.1	33.1	42.4
All -ngrams	72.8	71.3	60.9	65.7	76.4	59.2	35.7	44.6
All -ambiguity	71.3	68.3	61.3	64.7	74.3	52.1	42.6	**46.9**
All features	72.0	69.4	**62.4**	65.6	73.8	50.9	**42.7**	46.5

Table 1: Results obtained on the two corpora, percentage, using the training set and test set of the same show.

2.3. Classifier

The Conditional Random Field (CRF) [15] is a classifier specifically developed to deal with time series. It has been successfully used in other similar sequence-tagging tasks dealing with speech and dialogues, such as disfluency detection [21] or meeting summary extraction [22]. Compared to other popular sequence-based classifiers, such as Hidden Markow Models, the CRF is a discriminative model, so it does not require independence among features. A sequence-based classifier allows to make predictions which take into account not just the individual utterance or a fixed-size sliding window, but also the relations among all the utterances in the whole sequence. To further justify the choice of a CRF we also compare it with a standard logistic regression (LR), trained on the same features.

3. EXPERIMENTS

3.1. Corpus

We used a corpus of dialog utterances from TV sitcoms. We selected all episodes in Big Bang Theory seasons 1 to 6, and most episodes in Seinfeld seasons 5 to 9, discarding clip-shows or episodes with errors in the transcripts. We also discarded all the initial and final monologues, for not being domain consistent with the other dialog utterances.

For each sitcom we retrieved the audio track, the subtitles and the scripts (from https://bigbangtrans.wordpress.com and www.seinology.com respectively for Big Bang Theory and Seinfeld). Utterances were divided according to the subtitles files, delimited by the timestamps in them. All utterances were divided into scenes and annotated with speaker identification. We retrieved the canned laughters position from the audio tracks with a two step process. In the first step we used the vocal removal tool of Audacity. In the second step we used a sound/silence detector to retrieve the timestamps of the laughters. Afterwards we compared these timestamps with the utterances timestamps of the subtitle files. Each utterance followed by a laughter immediately or within $1s$ was annotated as punchline. We assume laughter comes right after the punchline though sentences prior to the punchline are useful as setup and triggers. The laughters timestamps were also used to cut the ending of the utterance audio tracks during audio features retrieval in order to avoid a possible bias in the classification with acoustic features.

We annotated in this way a total of 135 Big Bang and 102 Seinfeld episodes, each of a duration between 20 and 22 minutes. In Big Bang there are a total of 1589 scenes, 42.8% of the utterances being punchlines, an average interval of 2.2 utterances between two laughters and 7 characters appearing for more than 500 utterances (around 1% of the total). In Seinfeld 2267 scenes, 26.6% punchlines, an average of 3.1 utterances between two punchlines and 6 characters with more than 500 utterances. We divided each sitcom into a training set of about 80% of the episodes, a development set and a test set of 10% each. Episodes were drawn with the same proportion from each season (two episodes per season in the development set and two in the test set). Overall Big Bang training set has 35865 utterances, the development set 3904 and the test set 3903. Seinfeld training set has 36692 utterances, the development set 4097 and the test set 4945. Thus the two corpora are roughly the same size, but with different proportion of punchlines.

3.2. Experimental setup

We took language features on a window of size 5, including the utterance to classify and the four previous, eventually interrupted by the scene boundary. For the acoustic features we assigned only the ones of the last utterance. The development

	The Big Bang Theory				Seinfeld			
Features	A	P	R	F1	A	P	R	F1
All -ngrams -ambiguity LR	72.0	70.3	59.9	64.7	73.8	55.3	7.7	13.5
All -ngrams -ambiguity CRF	73.4	72.1	61.8	**66.5**	76.5	59.4	36.5	**45.2**
All features LR	71.6	68.2	62.8	65.4	74.0	53.8	15.3	23.9
All features CRF	72.0	69.4	62.4	**65.6**	73.8	50.9	42.7	**46.5**

Table 2: Comparison between logistic regression and conditional random field.

	Train: Seinfeld ->Test: Big Bang				Train: Big Bang ->Test: Seinfeld			
Features	A	P	R	F1	A	P	R	F1
Acoustic +speed	64.1	66.8	31.9	43.2	70.7	43.8	36.5	39.8
Language -speaker	47.2	43.0	72.6	54.0	44.2	27.4	66.7	38.9
All -ngrams -speaker	69.4	68.4	53.0	59.7	72.5	48.1	44.4	46.2

Table 3: Results obtained training the model on one series and testing on the other one

set was used to tune the hyperparameters and the regularization. L2 regularization was more effective than L1 in both corpora, with a greater difference in Big Bang. We used CRFSuite [23] implementation for the CRF. We ran experiments training and testing on the same sitcom, training on one and testing on the other, and to compare CRF with LR.

3.3. Results

Table 1 shows the results of the CRF with different feature settings, and the comparison with an all positive/negative baseline. We obtain the best result of 72.1%/61.8%/66.5% Precision, Recall and F-score in Big Bang (over a baseline of 59.9% F-score) with all features except ngrams and ambiguity, and of 52.1%/42.6%/46.9% in Seinfeld (over a baseline of 42.0% F-score) with all features except ambiguity. In the latter we also obtained an overall maximum precision of 60.2% with all features except ngrams and latent. The results on the evaluation of different training-test set pairs are shown in table 3, with 59.7% F-score obtained on Big Bang test set with a model trained on Seinfeld data, and 46.2% training the model on Big Bang data and evaluating it on Seinfeld. Finally table 2 shows the CRF/LR comparison results. CRF clearly outperforms LR in Seinfeld obtaining more than 20% of recall, while in Big Bang the performance differences are of 2% F-score without ngrams, and almost negligible with ngrams.

4. CONCLUSION AND DISCUSSION

We have described a first-ever method that predicts laughter responses in dialogues containing humor. We obtained the best performances combining acoustic features retrieved from the audio with language features retrieved from the transcriptions. Our method yields the best result of 72.1% precision on "The Big Bang Theory" and of 60.2% on "Seinfeld".

Structural features – word length, sentence length and part of speech proportion – and the utterance rate are the most effective features. A possible explanation is that deliberately slow or long utterances are often used to generate humor, and sometimes followed by very short ones in the following turn enhancing the effect, for example:

SHELDON: *Are you pointing out that California is a community-property state and since Howard and Bernadette are married, intellectual property in the letter is jointly owned by the two spouses?* **LAUGH**
PENNY: *Yeah, obviously.* **LAUGH**

Ambiguity features are not found to be effective. N-grams are helpful in raising the recall without audio features, yet don't seem to always help with audio features.

Our results show that our method is quite effective in predicting punchlines when acoustic and language features are combined. However, there is an important challenge in using canned laughter from a sitcom for learning. Canned laughter are inserted to solicit humor response in the audience. They might even be inserted to strengthen a weak joke. The actual response of the home audience is not known. This is evident especially when the model is trained on Seinfeld series, where the recall is not very high.

Our method performs better on Big Bang data. The main reason is the more balanced combination of punchline vs non-punchline in the former. Big Bang has also a cleaner audio track, fewer characters and a simpler humor structure with clear and frequent punchlines. In figures 1 and 2: "suffer in silence" can be perceived as sarcastic on its own, while "How does it affect me?" only works when associated with the appropriate context or when said with the appropriate tone. It is also worth noting that training the classifier on Seinfeld and testing it on Big Bang still yields a comparable precision. On the contrary there is a larger performance difference between the CRF and the LR on Seinfeld. The CRF learns transition scores which are useful to model the lower chance in Seinfeld to jump to a punchline.

In future work we plan to integrate humor generation and response prediction into a dialog system with the objective for a more empathetic human-machine interaction.

5. REFERENCES

[1] Ann D Sumners, "Humor: coping in recovery from addiction," *Issues in mental health nursing*, vol. 9, no. 2, pp. 169–179, 1988.

[2] William H Martineau, "A model of the social functions of humor," *The psychology of humor: Theoretical perspectives and empirical issues*, pp. 101–125, 1972.

[3] Lawrence La Fave, Jay Haddad, and William A Maesen, "Superiority, enhanced self-esteem, and perceived incongruity humour theory," *Humor and laughter: Theory, research, and applications*, pp. 73–91, 1976.

[4] Craig A Anderson and Lynn H Arnoult, "An examination of perceived control, humor, irrational beliefs, and positive stress as moderators of the relation between negative stress and health," *Basic and Applied Social Psychology*, vol. 10, no. 2, pp. 101–117, 1989.

[5] Herbert M Lefcourt, Karina Davidson, Kenneth M Prkachin, and David E Mills, "Humor as a stress moderator in the prediction of blood pressure obtained during five stressful tasks," *Journal of Research in Personality*, vol. 31, no. 4, pp. 523–542, 1997.

[6] Harvey Mindess, "The sense in humor," *Saturday Rev*, vol. 21, no. 8, pp. 10–12, 1971.

[7] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy, "Humor recognition and humor anchor extraction," in *EMNLP*, 2015, pp. 2367–2376.

[8] Antonio Reyes, Paolo Rosso, and Tony Veale, "A multidimensional approach for detecting irony in twitter," *Language Resources and Evaluation*, vol. 47, no. 1, pp. 239–268, 2013.

[9] Francesco Barbieri and Horacio Saggion, "Modelling irony in twitter: Feature analysis and evaluation," in *Proceedings of LREC*, 2014, pp. 4258–4264.

[10] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang, "Sarcasm as contrast between a positive sentiment and negative situation.," in *EMNLP*, 2013, pp. 704–714.

[11] Antonio Reyes and Paolo Rosso, "Making objective decisions from subjective data: Detecting irony in customer reviews," *Decision Support Systems*, vol. 53, no. 4, pp. 754–760, 2012.

[12] Byron C. Wallace, Do Kook Choe, and Eugene Charniak, "Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics: Long Papers*, Beijing, China, July 2015, pp. 1035–1044.

[13] Rachel Rakov and Andrew Rosenberg, "" sure, i did the right thing": a system for sarcasm detection in speech.," in *INTERSPEECH*, 2013, pp. 842–846.

[14] Michael J Platow, S Alexander Haslam, Amanda Both, Ivanne Chew, Michelle Cuddon, Nahal Goharpey, Jacqui Maurer, Simone Rosini, Anna Tsekouras, and Diana M Grace, "its not funny if theyre laughing: Self-categorization, social influence, and responses to canned laughter," *Journal of Experimental Social Psychology*, vol. 41, no. 5, pp. 542–550, 2005.

[15] John Lafferty, Andrew McCallum, and Fernando CN Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[16] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 835–838, ACM.

[17] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[18] George A Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] Andrea Esuli and Fabrizio Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, 2006, vol. 6, pp. 417–422.

[20] Weiwei Guo and Mona Diab, "Modeling sentences in the latent space," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, 2012, pp. 864–872.

[21] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1526–1540, 2006.

[22] Justin Jian Zhang and Pascale Fung, "Automatic parliamentary meeting minute generation using rhetorical structure modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2492–2504, 2012.

[23] Naoaki Okazaki, "CRFsuite: a fast implementation of conditional random fields (CRFs)," 2007.