Predicting laughter relevance spaces in dialogue

Vladislav Maraev, Christine Howes and Jean-Philippe Bernardy

Abstract In this paper we address the task of predicting spaces in interaction where laughter can occur. We introduce the new task of predicting actual laughs in dialogue and address it with various deep learning models, namely recurrent neural network (RNN), convolution neural network (CNN) and combinations of these. We also attempt to evaluate human performance for this task via an Amazon Mechanical Turk (AMT) experiment. The main finding of the present work is that deep learning models outperform untrained humans in this task.

1 Introduction

Non-verbal vocalisations, such as laughter, are ubiquitous in our everyday interactions. In the Switchboard Dialogue Act corpus (Jurafsky et al., 1997), which we use in the current study, non-verbal dialogue acts (that are explicitly marked as non-verbal) constitute 1.7% of all dialogue acts and laughter tokens make up 0.5% of all the tokens that occur in the corpus. In order to make state-of-the-art dialogue systems more natural and cooperative, it is vital to enable them to understand non-verbal vocalisations and react to them appropriately. With regards to laughter, the most important issues are in understanding the coordination of laughter with speech, social and pragmatic functions of laughter, and the reasons for laughter.

Vladislav Maraev
Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, e-mail: vladislav.maraev@gu.se

Christine Howes
Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, e-mail: christine.howes@gu.se

Jean-Philippe Bernardy
Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, e-mail: jean-philippe.bernardy@gu.se

The social function of laughter is well documented (e.g., Mehu, 2011): laughter is associated with senses of closeness and affiliation, establishing social bonding and smoothing away discomfort. Laughter can also have a pragmatic function, such as indicating a mismatch between what was said and what was meant, for example by indicating that a speaker was 'just kidding' (for further details on classification of laughter see the work of Mazzocconi et al. (2018)).

Although laughter is closely associated with humour, and humorous and joyful remarks can be thought as a prerequisite for laughter, this is not necessarily the case: laughter can display surprise, nervousness, embarrassment, disagreement etc. (Poyatos, 1993). This suggests that laughter is not exclusively associated with positive emotions (happiness, joy, pleasure and more) — other emotional dimensions and their (perhaps contradictory) combinations should also be considered. Nevertheless, positive emotional states are an intuitive notion of where laughter occurs.

In the current study we focus on the issues of laughter relevance and predictability. We introduce the term *laughter relevance spaces* analogously with backchannel relevance spaces (Heldner et al., 2013) and transition relevance places (Sacks et al., 1978). We define a *laughter relevance space* as a position within the interaction where an interlocutor can appropriately produce a laughter (either during their own or someone else's speech). Following the approach of Heldner et al. (2013) to backhannels, we distinguish *actual laughs* and *potential laughs*. By definition, the number of potential spaces for laughter is larger than number of actually produced laughter spaces.

In this work we are guided by the following research questions: i) can laughs be predicted from the textual data either by humans or by deep learning systems, and ii) to what extent can these predictions be compared. In an attempt to address these questions we present:

- The task of predicting laughter from dialogue transcriptions.
- Human annotations of potential laughs from dialogue transcriptions.
- Automatic methods for predicting actual laughs with deep learning models.

In the rest of the paper we present details of the dataset and the task (section 2). We then describe the Amazon Mechanical Turk (AMT) experiment and its evaluation (section 3). We present our sentiment analysis baseline in section 4. In section 5 we present our deep learning models and summarise the results. We conclude with some pointers for future work (section 7).

2 Data

The Switchboard Dialogue Act Corpus (Jurafsky et al., 1997) consists of 1155 dyadic telephone conversations (221,616 utterances) between participants who were unfamiliar to each other. For the purposes of our study we make use of the disfluency annotations (Meteer et al., 1995) in the corpus.

For our experiments we split utterances into tokens using the Python library SWDA[1] and combine consecutive laughs within a turn into a single laughter token. The laughter tokens are then removed from the text and replaced by laughter annotations. That is, the data is a sequence of tuples (t_i, l_i) such that:

- $t_i \in \mathbb{N}$ is the ith speech (typically a representing a word) or turn-taking token (For either A or B).
- $l_i \in \{0, 1\}$ is a laughter marker, which indicates whether laughter follows immediately after the token t_i.

The goal of the current study is to determine whether l_i can be predicted, that is, does laughter occur after a given sequence of tokens $(t_0..t_i)$.

Exploratory task

The obvious way to tackle the goal is to predict the probability of laughter for each token. To do so we split the corpus on turn boundaries, with no overlap (Figure 1) and train an RNN model (see Section 5.1.1) on 80% of the corpus (total number of samples range from 17k examples for 10-turn split to 73k for 3-turn split). In Table 1 we report the results depending on a turn span and threshold for converting predicted probability of laughter into a binary value. We observed that adding more context leads to better predictions even if it leads to decreasing the size of training data[2]. Yet, even in the case of a 10-turn span, the recall was only 1.5%. A direct attempt to increase the recall by increasing the threshold to report a laughter lowered the precision to unacceptable levels.

```
1 sp_A {F Oh, } I know. /
1 sp_A It's really amazing. /
1 sp_B Yeah. /
2 sp_A It's, {F uh, } <LAUGHTER> -/
2 sp_B Beautiful, beautiful machine. /
2 sp_A Absolutely, /
```

Fig. 1 Example of dialogue split into two samples. The leading number shows to which of the training samples each utterance will be related to based on 3-turn span.

Balanced task

The above experiment indicates to us that this task is difficult to tackle using deep learning models. We attribute this difficulty to the corpus being unbalanced towards negative predictions, due to the sparsity of laughs. Indeed, the proportion of actual

[1] https://github.com/cgpotts/swda

[2] In all our experiments we keep 80%/10%/10% training/validation/test split.

Table 1 Predicted laughs depending on a turn span and threshold. Number of laughs to predict vary due to different splits of the data.

span	threshold	laughs to predict	precision	recall	F_1
3	0.50	1128	0.733	0.010	0.007
5	0.50	1116	0.786	0.010	0.005
10	0.50	1127	0.630	0.015	0.018
10	0.45	1127	0.407	0.020	0.132
10	0.40	1127	0.400	0.039	0.036
10	0.35	1127	0.255	0.060	0.049

laughter tokens is around 0.5% in the whole corpus. Additionally, it is also a hard and unrealistic task for humans because annotating every token is tedious.

We therefore, instead, fix the point of focus to given positions, and attempt to predict the incidence of laughter those given points. We select these points so that the frequency of laughter is equal to the frequency of non-laughter at this points. To do so, we run a sliding window through all the examples. The size of the sliding window is fixed to a given number of tokens (not turns), in our case, 50 or 100 tokens. All the laughs (except the final one for the sequence) are represented as a special token and the final laughter is removed and represented as a positive label. The resulting training set (80% of all data) contained around 17k samples and remaining 20% were left out for validation and testing. This amended task is the focus of the rest of the paper.

3 Amazon Mechanical Turk

In order to understand how well humans perform at this task, we conducted an experiment with naive annotators located in the US. They were given a task description with the following salient points:

1. An invitation to complete the task with a notice that native level of English is required.
2. A sound sample of a dialogue containing laughs (in order to help coders understand that laughter can occur in non-humorous conditions).
3. Three excerpts from test set with removed non-verbal signals, disfluency and discourse annotations. Each of the excerpts has to be annotated regarding the potential to elicit laughter as: a) very unlikely, b) not very likely, c) quite likely, and d) very likely.

The subset of 399 excerpts was annotated by at least two annotators per sample. We computed Cohen's κ chance-adjusted inter-annotator agreement both for four-class predictions and for predictions converted into binary: judgements "quite likely" or "very likely" are counted as positive and "very unlikely" or "not very likely" — as negative. The resulting κ was very low (below chance level: $\kappa = -0.125$ for four-

class predictions and $\kappa = -0.071$ for binary predictions), which indicates either that quality of AMT annotations are very low or that human judgements about laughter are very subjective.

Subjects also showed a disposition towards laughter: 66% of excerpts were annotated as "quite likely" or "very likely", and only 2% were annotated as "very unlikely" or "not very likely" by both annotators. After comparison with the distribution of actual laughs in the corpus we observe that AMT respondents are not very good in predicting whether there was actually a laughter at the end of the sequence, but they might instead be predicting *potential* laughter, which is suggested by the predominance of such predictions. This means that participants are judging whether a laugh could appropriately be produced at that point in the dialogue, not whether the dialogue participants themselves actually did produce one. We conjecture that this result extends to the general population, if asked to make laughter judgements in the same conditions (i.e. when little effort is spent for each judgement). The expert prediction of laughter, that is by subjects trained in the relevant psycholinguistic aspects, is beyond the scope of the present paper. In Table 2 we show accuracy and F_1 score for human predictions of actual laughs.

Table 2 Human annotations as compared with the test set. Scores are computed based on the valence. For all the cases examples labelled as "quite likely" or "very likely" valence is positive, and the rest — as negative.

selection principle	accuracy	precision	recall	F_1
average of 4-class annotations	0.51	0.50	0.92	0.65
average of binary annotations	0.51	0.49	0.67	0.57
agreed judgements w.r.t. the valence (in 271 cases out of 499)	0.51	0.49	0.98	0.66

4 Off-the-shelf sentiment analyser

Even though laughter can be associated with a variety of sentiments, it is often naively associated with positive sentiment. Therefore, as a baseline, we employed the VADER sentiment analyser (Gilbert, 2014) to check whether its prediction of positive sentiment correlates with laughter. VADER is designed to classify sentiment along the positive/negative scale and mainly used for sentiment classification in social media which is not specifically designed for the dialogue task but arguably should perform relatively well on "noisy" texts such as those found in the Switchboard corpus (Howes et al., 2014). VADER is built in the Python NLTK library (Bird and Loper, 2004).

The sentiment analyser showed a predominance towards positive sentiment (and hence laughter) but the accuracy was only slightly above the majority vote baseline (51.1%).

5 Deep learning

5.1 Models

We present several deep learning models to tackle our task, either recurrent neural networks (RNN), convolutional neural networks (CNN) or combinations of these.

These models are implemented using our own high-level deep-learning library, which uses TensorFlow as a backend[3,4].

5.1.1 RNN model

Our RNN-based model architecture is shown in figure 2 and consists of three layers:

1. **An Embedding layer** which is characterised by the size of token embeddings (d).
2. **An LSTM recurrent layer** characterised by state size n. Each LSTM cell additionally includes dropout (on its inputs, outputs and hidden state inputs) of a probability ε.
3. **A Dense layer** which predicts laughter relevance for each token. We have exactly two classes: relevant (1) or irrelevant (0). For the main task we only output the final prediction of the dense layer.

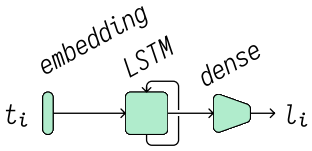

Fig. 2 Architecture of the RNN model ("rolled" view). For the main task only the final prediction (l_n) is considered.

[3] TypedFlow: https://github.com/GU-CLASP/TypedFlow

[4] Models and data are available at: https://github.com/GU-CLASP/laughter-spaces

5.1.2 CNN model

The convolution neural network model includes the following parts:

1. **An Embedding layer** which is characterised by size of token embeddings (d).
2. **A first 1-D Convolution layer** characterised by filter size h_1 and number of filters k_1. The layer is followed by a rectified linear unit (ReLU).
3. **A first max-pooling layer** with a stride $s = 2$.
4. **A second 1-D Convolution layer** characterised by filter size h_2 and number of filters k_2. The layer is followed by ReLU.
5. **A max-over-time pooling layer** which computes element-wise maximum along all the features of the second convolution layer.
6. **A Dense layer** that predicts laughter relevance for the sequence.

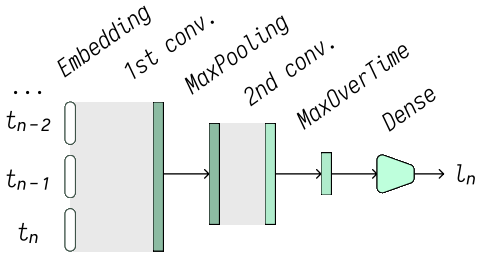

Fig. 3 Architecture of the CNN model.

5.1.3 Combinations of the models

In order to estimate whether RNN and CNN models pick up on either the same or different features, we also tested two combinations of the above models:

1. **A Fusion model** (Figure 4) where outputs of an RNN and a CNN model (both without a dense layer) are concatenated, and a dense layer operates on this concatenation.
2. **A Hybrid model** (Figure 5) similar to the fusion model, but when token embeddings are shared between RNN and CNN.

5.2 Results

We present results for the different models in Table 3. Given the results of the AMT experiment, we posited that the task of predicting actual laughters in dialogue is

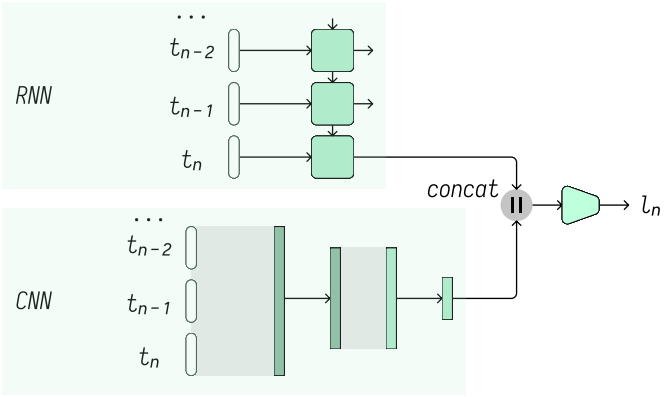

Fig. 4 Architecture of the fusion model. Outputs of the RNN's last cell and CNN's max-over-time pooling layers are concatenated and then dense layer is applied.

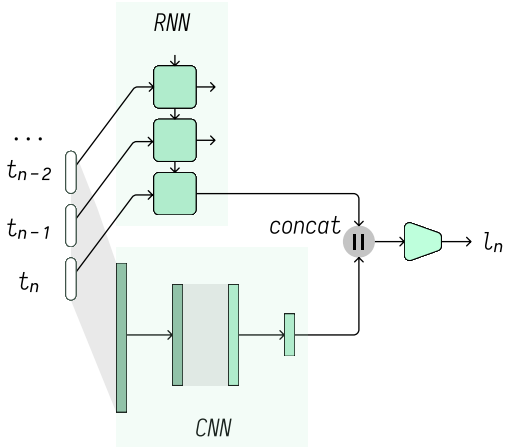

Fig. 5 Architecture of the hybrid model. Token embeddings are shared between RNN and CNN.

hard for untrained humans to perform. We also saw that the task is difficult to tackle by simple sentiment analysis. Thus we expect the task to be difficult for deep learning models as well. However, the deep learning models perform considerably better than our baselines, especially in terms of accuracy. Additionally, the CNN model consistently outperforms the RNN model. Further, combining RNN with CNN provides no significant benefit. This suggests that the RNN model does not detect more laughter-inducing patterns than the CNN model.

Table 3 Summary of the prediction results.

model	val. acc.	test. acc.	test. precision	test. recall	test. F_1
AMT	–	0.510	0.500	0.920	0.650
VADER	–	0.518	0.511	0.749	0.607
RNN (span=50)[a]	0.762	0.743	0.732	0.763	0.747
RNN (span=100)[a]	0.756	0.770	0.761	0.777	0.769
CNN (span=50)[b]	0.789	0.765	0.761	0.771	0.766
CNN (span=100)[b]	0.783	0.787	0.777	0.794	0.785
fusion (span=50)[c]	0.794	0.766	0.760	0.778	0.768
hybrid (span=50)[d]	0.793	0.776	0.775	0.774	0.774

Hyperparameters:

[a] $d = 50$; $n = 40$; $\varepsilon = 0.1$
[b] $d = 100$; $k_{1,2} = 40$; $h_{1,2} = 7$
[c] see LSTM and CNN
[d] see LSTM and CNN, shared embedding layer: $d = 100$

6 Error analysis

After analysing the results, we noted that there were a large number of examples where laughter occurs at a turn boundary. In this case the last token of the sample is a turn change (TC) token (`sp_A` or `sp_B`). A concern was that this would significantly affect the results. In order to measure this effect, we removed these results from the test set and observed the accuracy and F-measure shown in Table 4. We observe a drop of F-score (around 6 percentage points) but accuracy is almost unchanged. This indicates that system relies on turn change (possibly combined with other features — and consequently captured by neural networks) as an important predictor for laughter, for both basic models of the system. Examples where laughter is predicted to occur immediately after a turn change are shown in (1) and (2)).

(1) A A: let me ask you this.
A: How, how old are you?
B: I'm, uh, thirty-three.
A: Thirty-three?
B: Thirty-two,
B: excuse me.
A: Okay.
B: ((correct prediction: LAUGHTER))

(2) A B: when I was a freshman in college
A: Uh-huh.
B: uh, my degree was in computer, uh, technology originally

B: and it seemed like it would,
B: ((wrong prediction: LAUGHTER))

Table 4 Performance of the models before and after removing the examples where turn change token is the last token. As a result, the dataset is 22% smaller and it is missing 36% of positive examples. All deep learning models use the dataset with the span of 50 tokens.

model	accuracy	precision	recall	F_1
AMT	0.500	TBD	TBD	0.660
VADER	0.518	0.511	0.749	0.607
RNN	0.743	0.732	0.762	0.747
RNN (last TC removed)	0.738	0.673	0.705	0.689
CNN	0.765	0.761	0.771	0.766
CNN (last TC removed)	0.761	0.715	0.694	0.705

In conversational analysis studies many laughs are considered to form adjacency pairs with prior laughs (Jefferson et al., 1987), and preceding laughter by another speaker seem to be a relevant feature for our models (e.g., (3)). However, in excepts where there are a lot of laughs the system sometimes gets it wrong (e.g., (4)).

(3) A: I'm not really sure what the ((LAUGHTER))
B: Yeah,
B: really,
B: it's one of those things that you read once,
B: and then, if you're not worried about it, you just forget about it ((LAUGHTER))
A: ((correct prediction: LAUGHTER))

(4) A: (...) don't get a hot tub and
B: ((LAUGHTER)) Yes.
A: shave my legs, I'm going to die ((LAUGHTER))
A: And I had ((LAUGHTER))
B: Yes
B: I understand that ((LAUGHTER))
A: I got enough of it right ((wrong prediction: LAUGHTER))

7 Conclusions and future work

The main conclusion of our experiments is that for the given task deep learning approaches perform significantly better than untrained humans.

We are optimistic that the introduced task and the approaches that we have developed are a big step towards inferring appropriate spaces for laughter from textual data. Together with approaches based on audio components (e.g. El Haddad et al., 2017) this should enable future dialogue systems to understand when is it appropriate to laugh. Nevertheless, we are aware of the fact that this requires understanding laughter on a deeper level, including its various semantic roles and pragmatic functions (see (Maraev et al., 2018) for a discussion about integrating laughter in spoken dialogue systems).

We are planning to extend our Amazon Mechanical Turk experiments by introducing more annotators in order to get more consistent results. We are going to introduce probabilistic annotations in our future crowdsourced experiments following e.g. Passonneau and Carpenter (2014).

Regarding the task itself, we are planning to address it in a more "dialogical" way. We will consider the data not as one input to a neural network which contains speaker tokens but as two possibly overlapping streams. This will introduce the notion of coordination between speakers into the prediction model. Two streams can be also extended by additional information provided in separate inputs, such as information about disfluencies, discourse markers, fundamental frequency and other acoustic features. We are planning to see what features will make a more robust contribution to the task of predicting relevant laughs.

8 References

References

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, and Thierry Dutoit. 2017. Amused speech components analysis and classification: Towards an amusement arousal level assessment system. *Computers & Electrical Engineering*, 62:588–600.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*, pages 137–146. Peter Lang Publishing Group.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 7–16.

Gail Jefferson, Harvey Sacks, and Emanuel A Schegloff. 1987. Notes on laughter in the pursuit of intimacy.

D Jurafsky, E Shriberg, and D Biasca. 1997. Switchboard dialog act corpus. *International Computer Science Inst. Berkeley CA, Tech. Rep*.

Vladislav Maraev, Chiara Mazzocconi, Christine Howes, and Jonathan Ginzburg. 2018. https://doi.org/10.21437/AI-MHRI.2018-3 Integrating laughter into spoken dialogue systems: preliminary analysis and suggested programme. In *Proceedings of the FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction*, pages 9–14.

Chiara Mazzocconi, Vladislav Maraev, Christine Howes, and Jonathan Ginzburg. 2018. Analysis of laughables: a preliminary perception study. In *Proceedings of the Workshop on Dialogue and Perception*, pages 43–47.

Marc Mehu. 2011. Smiling and laughter in naturally occurring dyadic interactions: relationship to conversation, body contacts, and displacement activities. *Human Ethology Bulletin*, 26(1):10–28.

Marie W Meteer, Ann A Taylor, Robert MacIntyre, and Rukmini Iyer. 1995. *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania Philadelphia, PA.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Fernando Poyatos. 1993. *Paralanguage: A linguistic and interdisciplinary approach to interactive speech and sounds*, volume 92. John Benjamins Publishing.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.