Predicting laughter relevance spaces

Vladislav Maraev, Christine Howes
and Jean-Philippe Bernardy

Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg, Sweden

CLASP seminar · Feb 6, 2019

CLASP centre for linguistic theory and studies in probability

GÖTEBORGS UNIVERSITET

Why laughter?

- Non-verbal vocalisations, such as laughter, are ubiquitous in our everyday interactions.
- In Switchboard Dialogue Act Corpus (Jurafsky et al., 1997) (SWDA) 1.7% of all dialogue acts are non-verbal, and **laughter tokens** make up 0.5% of all the tokens.

Why laughter?

- Non-verbal vocalisations, such as laughter, are ubiquitous in our everyday interactions.
- In Switchboard Dialogue Act Corpus (Jurafsky et al., 1997) (SWDA) 1.7% of all dialogue acts are non-verbal, and **laughter tokens** make up 0.5% of all the tokens.

We need to make sense of laughter:

- coordination with speech
- social and pragmatic functions
- reasons for laughter

What do we know

1. Laughter has a **social function**: it is associated with senses of closeness and affiliation, establishing social bonding and smoothing away discomfort.
2. Laughter has a **pragmatic function**: e.g. indicate a mismatch in 'just kidding' sense.
3. Laughter is not exclusively associated with positive emotions, but positive emotional state is an intuitive notion of where laughter occurs.

Laughter relevance spaces

Our main focus: **laughter relevance and predictability**

Laughter relevance spaces

Our main focus: **laughter relevance and predictability**

We introduce the term laughter relevance spaces:

a position within the interaction where an interlocutor can appropriately produce a laughter (either during their own or someone else's speech)

- Analogous to backchannel relevance spaces (Heldner et al., 2013) and transition relevance spaces (Sacks et al., 1978).
- Following Heldner et al. (2013) we distinguish **actual laughs** and **potential laughs**.

Research questions

- Can laughs be predicted from the textual data either by humans or by deep learning systems?
- To what extent can these predictions be compared?

Research questions

- Can laughs be predicted from the textual data either by humans or by deep learning systems?
- To what extent can these predictions be compared?

We present:

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

Next

The task and the data

Amazon Mechanical Turk

Deep learning models

Error analysis

Conclusions

Data

- Switchboard Dialogue Act Corpus (Jurafsky et al., 1997)
- 1155 dialogues, 221616 utterances
- disfluencies (Meteer et al., 1995)
- laughter – 0.5% of all tokens

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

Data preparation

1. We split utterances into tokens using `swda.py` library
2. The laughter tokens are then removed from the text and replaced by laughter annotations, so

```
data: sequence of tuples (tᵢ, lᵢ)
  - tᵢ ∈ N      -- i-th speech or speaker token
  - lᵢ ∈ {0,1} -- laughter marker
```

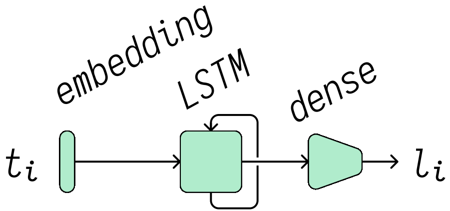
- The goal is to predict laughter token l_i after a given sequence of tokens $(t_0..t_i)$.

Exploratory task

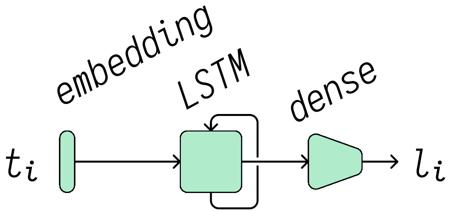
- split the corpus on turn boundaries with no overlap
- predict laughter for every token
- training data (80%) ranges from 17k samples (10-turn span) to 73k (3-turn span)

```
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, } -/
2 sp_B Beautiful, beautiful machine. /
2 sp_A Absolutely, /
```

Model and results



Model and results



span	th	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

Balanced set

- proportion of laughs is 0.5%
- instead we fix the positions of laughs to predict, such that frequency of laughs will be equal to the frequency of non-laughs
- sliding window (50 or 100 tokens)
- training set (80%) 17k samples, 10% val. and 10% test.

Next

The task and the data

Amazon Mechanical Turk

Deep learning models

Error analysis

Conclusions

Amazon Mechanical Turk

task

- 400 samples, 2 annotations per sample
- listen to the audio
- a) very unlikely, b) not very likely, c) quite likely, d) very likely

result

- very low Cohen’s kappa (below chance level: $\kappa = -0.125$ for four-class predictions and $\kappa = -0.071$ for binary predictions)
- 66% of excerpts were annotated as “quite likely” or “very likely”
- only 2% were annotated as “very unlikely” or “not very likely” by both annotators

As compared with actual laughs

Selection principle	accuracy	precision	recall	F_1
avg. of 4-class annot.	0.51	0.50	0.92	0.65
avg. of binary annot.	0.51	0.49	0.67	0.57
annot. agree on valence	0.51	0.49	0.98	0.66

Annotators might be predicting potential laughter, which is suggested by the predominance of such predictions.

Next

The task and the data

Amazon Mechanical Turk

Deep learning models

Error analysis

Conclusions

Deep learning models

Baseline

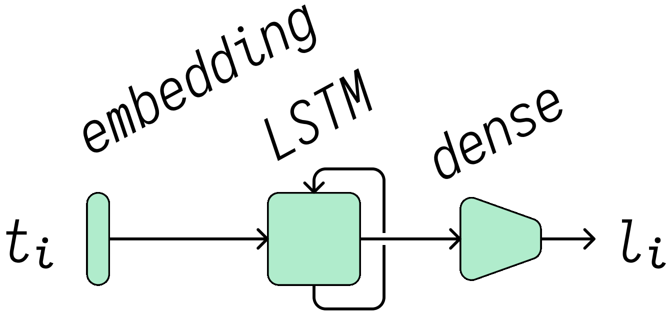
- We employed sentiment analysis baseline: VADER Gilbert (2014) designed for social media texts (part of NLTK).

Neural networks

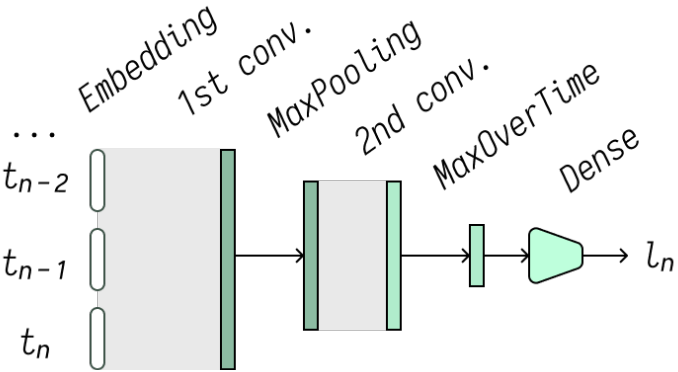
- RNN (LSTM)
- CNN
- two combinations of RNN and CNN

Implemented in TypedFlow:
`https://github.com/GU-CLASP/TypedFlow`
`logo_desperately_needed.png`

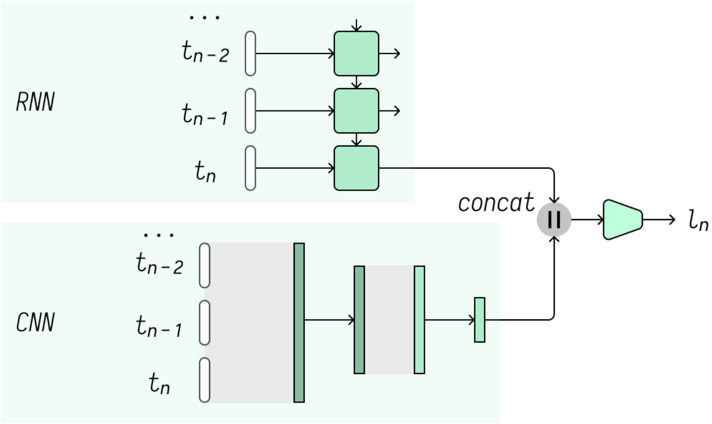
RNN



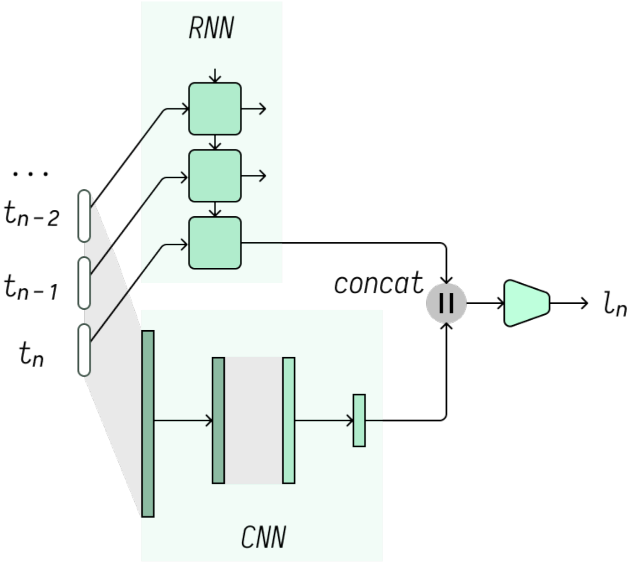
CNN


Embedding
1st conv.
MaxPooling
2nd conv.
MaxOverTime
Dense
...
t_{n-2}
t_{n-1}
t_n
l_n

Fusion



Hybrid



Results

Model	accuracy	precision	recall	F_1
AMT	0.510	0.500	0.920	0.650
VADER	0.518	0.511	0.749	0.607
RNN (span=50)	0.743	0.732	0.763	0.747
RNN (span=100)	0.770	0.761	0.777	0.769
CNN (span=50)	0.765	0.761	0.771	0.766
CNN (span=100)	0.787	0.777	0.794	0.785
fusion (span=50)	0.766	0.760	0.778	0.768
hybrid (span=50)	0.776	0.775	0.774	0.774

Next

The task and the data

Amazon Mechanical Turk

Deep learning models

Error analysis

Conclusions

Turn boundaries

Laughters tend to occur at a turn boundary

```
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: <LAUGHTER> [correct!]

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: <LAUGHTER> [wrong!]
```

We removed these samples...

Table: 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**	$\mathbf{F_1}$
AMT	0.510	0.500	0.920	0.650
VADER	0.518	0.511	0.749	0.607
RNN	0.743	0.732	0.763	0.747
RNN (removed)	0.738	0.673	0.705	0.689
CNN	0.765	0.761	0.771	0.766
CNN (removed)	0.761	0.715	0.694	0.705

Laughter as a predictor

```
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: <LAUGHTER> [correct!]

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 <LAUGHTER> [wrong!]
```

Next

Conclusions

Main conclusion

for the given task deep learning approaches perform significantly better than untrained humans

Conclusions

Main conclusion

for the given task deep learning approaches perform significantly better than untrained humans

- step towards inferring **appropriate spaces for laughter** from textual data
- this should enable future dialogue systems to understand when is it appropriate to laugh
- but…

Conclusions

Main conclusion

for the given task deep learning approaches perform significantly better than untrained humans

- step towards inferring **appropriate spaces for laughter** from textual data
- this should enable future dialogue systems to understand when is it appropriate to laugh
- but…

… we are aware that this requires understanding laughter on a deeper level, including its various semantic roles and pragmatic functions.

Future work

1. Extend our AMT experiments, introduce probabilistic annotations (Passonneau and Carpenter, 2014)
2. Address the task in a more 'dialogical' way:
    - input: two possibly overlapping streams instead of one
    - **coordination between speakers** as a predictor
    - extend the streams with features:
        - disfluencies
        - discourse markers
        - acoustic features (f0)

```
-- Thank you! <LAUGHTER?> <QUESTIONS?>
```

```
https://github.com/GU-CLASP/laughter-spaces
```

The work was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References I

Gilbert, C. H. E. (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)*.

Heldner, M., Hjalmarsson, A., and Edlund, J. (2013). Backchannel relevance spaces. In *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*, pages 137–146. Peter Lang Publishing Group.

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

Meteer, M. W., Taylor, A. A., MacIntyre, R., and Iyer, R. (1995). *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania Philadelphia, PA.

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

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.