Tickle me BERT: The effect of laughter on dialogue act recognition

Bill Noble
Vladislav Maraev

CLASP seminar
Feb. 26th

BERT

- Large-scale language models (e.g. BERT) achieve state-of-the-art results on traditional NLP tasks.
- But are they useful for dialogue?
- To answer this, we decided to see how good BERT is on a dialogue act recognition (DAR) task.

Dialogue acts (DAs)

- Theory of dialogue acts is based on the theory of speech acts.
- The idea is that utterances can convey actions *(e.g. promising or apologising)*.
- DAMSL schema for dialogue act tagging
- forward-looking *(expecting a response)* and backward-looking *(responding to a preceding utterance)* DAs

Dialogue act recognition—
assigning DA tag to every utterance

	Utterance	Dialogue act
A:	*Well, I'm the kind of cook that I don't normally measure things,*	Statement-non-opinion (sd)
A:	*I just kind of throw them in*	sd
A:	*and, you know, I don't to the point of, you know, measuring down to the exact amount that they say.*	sd
B:	*That means that you are real cook.*	Statement-opinion
A:	*<Laughter> Oh, is that what it means*	Downplayer
A:	*Uh-huh.*	Backchannel
A:	*<Laughter>*	Non-verbal

Data: corpora

Switchboard	AMI Corpus
Dyadic	Multi-party
Casual conversation	Mock business meeting
Telephone	Face-to-face
English	English
Native speakers	Native & non-native speakers
early 90's	2000's
2000 conversations	171 meetings
1115 in SWDA	*131 in AMI-DA*
400k utterances	118k utterances
3M tokens	1.2M tokens

But would it be a problem for BERT?

- **Different sequential structure of discourse** *(taking turns and switching perspectives)*
- **Internal structure is different** *(disfluencies, non-verbal vocalisations, NSUs, etc.)*
- **Syntactic structure is different** *("I don't to the point of, you know, measuring down to the exact amount that they say")*

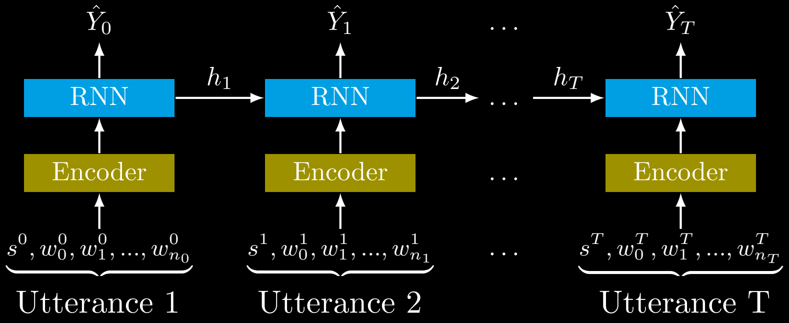
And of course: laughter

- In Switchboard it comes about every 200 tokens.
- It is related to discourse item (laughable), which can be described in the dialogue.
- Laughter can help to determine sincerity of an utterance, e.g. to detect sarcasm.*
- Laughs appear in any kind of DA.

---

* Joseph Tepperman, David Traum, and Shrikanth Narayanan (2006) "Yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth International Conference on Spoken Language Processing*.


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SWDA
AMI-DA
DA has laughter in one of its adjacent utterances
DA contains laughter



Neural dialogue act recognition sequence model

Utterance encoder: BERT

- **Multi-layer transformer** *(Base model: 768-dimension hidden, 12 layers)*
- **Trained on BookCorpus*** *(800M words)* **+ English Wikipedia** *(2,500 words)*

---

* https://www.smashwords.com/books/

Pre-training BERT

- Masked token prediction

```
[CLS] my dog is [MASK] [SEP] -> hairy
```

- Next sentence prediction

```
[CLS] the man went to [MASK] store
[SEP] he bought a gallon [MASK] milk
[SEP] -> IsNext

[CLS] the man [MASK] to the store
[SEP] penguin [MASK] are flight
##less birds [SEP] -> NotNext
```

Utterance encoder: CNN

- Kim (2014)-style encoder
  Word-level CNN
  Window sizes 3, 4, 5
  100 feature maps

- Word embeddings
  gloVe
  100 dimensions

Preprocessing

- We remove disfluencies and speech-laughs
- Laughs are normalised: `[LAUGH]`
- All utterances are lower-cased.
- We use BERT's word piece tokeniser with a vocabulary of 30,000.
- We prepend each utterance with a speaker token: `[SPKR_A], [SPKR_B]...`

Experiment 1...

Experiment 1.
Impact of laughter

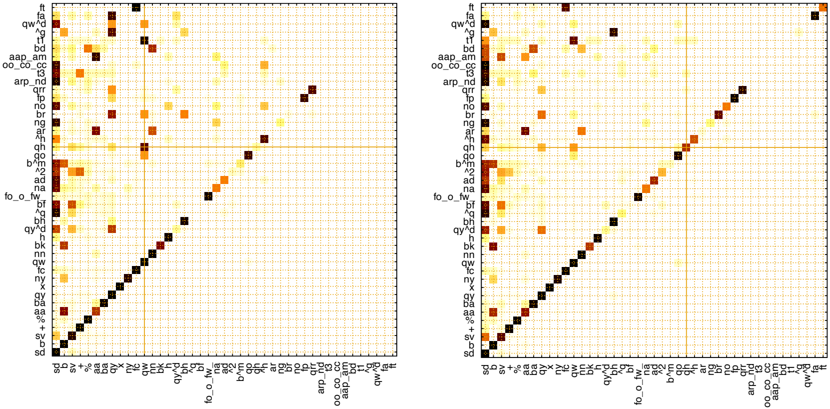
Is laughter helpful for DAR?

- We train two versions for each utterance encoder: with and without laughter and compare them.

Experiment 1: Results

	SWDA		AMI-DA	
	*F1**	*acc.*	*F1*	*acc.*
BERT–NL	38.10	77.07	49.09	67.06
BERT–L	45.99	76.93	50.17	67.12
CNN–NL	37.23	75.08	38.37	63.46
CNN–L	27.59	75.40	37.94	64.30

* Henceforth, we report macro-averaged F1 scores.



Confusion matrices:

BERT–NL *(left)* **vs** BERT–L *(right)*

The case of rhetorical questions

	Utterance	Dialogue act
B:	*Um, as far as spare time, they talked about,*	sd
B:	*I don't, + I think,*	Uninterpretable
B:	*who has any spare time <laughter>?*	Rhetorical-Q.
A:	*<laughter>.*	Non-verbal

Experiment 2. Impact of pre-training vs. fine-tuning

How does pre-training affect BERT's DAR performance?

	Pre-trained	Fine-tuned
BERT–FT	👍	👍
BERT–RI	👎	👍
BERT–FZ	👍	👎

Experiment 2: Results

	SWDA		AMI-DA	
	F1	*acc.*	*F1*	*acc.*
BERT–FT	45.99	76.93	50.03	66.94
BERT–RI	32.18	73.80	33.45	61.53
BERT–FZ	7.75	55.61	14.44	46.59

- Fine-tuning makes difference: 7.3% contain laughter (4.6% overall)
- AMI: 9.6% (8.5% overall)

Experiment 2: Fine-tuning laughs

- In `BERT-FZ` laughter token is randomly initialised and frozen.
- In SWDA fine-tuning makes difference: 7.3% contain laughter (4.6% overall).
- In AMI: 9.6% (8.5% overall)

Experiment 3.
Impact of dialogue pre-training

How does additional in-domain pre-training affect BERT's DAR performance?

- SWnDA: SWDA without DA tags
- AMI: AMI-DA + 32 dialogues without tags
- Combined corpus (SWnDA + AMI)

Experiment 3: Results

		SWDA		AMI-DA	
		F1	*acc.*	*F1*	*acc.*
Fine-tuned	BERT	45.99	76.93	50.03	66.94
	BERT-ID	45.48	77.02	46.56	68.66
	BERT-CC	47.78	77.35	48.72	66.58
Frozen	BERT	7.75	55.61	14.44	46.60
	BERT-ID	6.46	52.30	14.43	48.07
	BERT-CC	5.76	51.14	12.56	42.42

Conclusions

- Laughter is useful for dialogue act recognition, and its impact varies across different dialogue acts.
- During fine-tuning, BERT learns to represent laughter, a dialogical feature not seen in pre-training.
- Standard BERT pre-training is useful for DAR, but the model performs poorly without fine-tuning.
- Further pre-training with in-domain data shows promise for dialogue, but further investigation with larger dialogue corpora is required.