Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions

Nikolaos Flemotomos[1], Victor R. Martinez[1], James Gibson[1], David C. Atkins[2], Torrey A. Creed[3], Shrikanth Narayanan[1]

[1] Signal Analysis and Interpretation Laboratory
University of Southern California

[2] Department of Psychiatry and Behavioral Sciences
University of Washington

[3] Department of Psychiatry
University of Pennsylvania

Interspeech 2018
September 5

Quality Assessment in Psychotherapy

https://cdn-images-1.medium.com/max/1600/1*BHBHQcA30AjIpkEJFoVJxg.jpeg

- interventions based on spoken language ⇒ quality encoded in therapists' and patients' speech/language characteristics
- traditionally addressed by human raters using recorded sessions

Quality Assessment in Psychotherapy

https://cdn-images-1.medium.com/max/1600/1*BHBHQcA30AjIpkEJFoVJxg.jpeg

- interventions based on spoken language ⇒ quality encoded in therapists' and patients' speech/language characteristics
- traditionally addressed by human raters using recorded sessions
  - time consuming
  - cost prohibitive

⇓

- computational methods for automatic evaluation
  - already succesfull application in *Motivational Interviewing*

psychotherapy focused on behavior change, often used to treat addiction

D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [Perspectives]". *IEEE Signal Processing Magazine*, 34(5), pp.196-195., 2017

B. Xiao, C. Huang, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling", *PeerJ Computer Science*, vol. 2, p. e59, 2016

Cognitive Behavior Therapy

What is CBT

- the therapist works towards the modification of the patient's belief system
- based on the *cognitive model*: the link between a person's thoughts and feelings a primary factor contributing to mental illness
- original focus on depression but has expanded

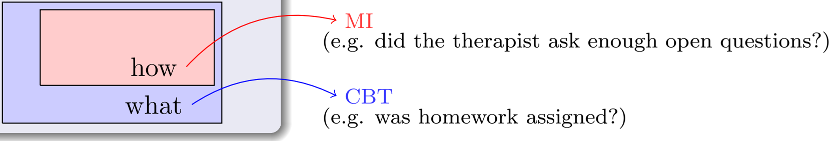
Cognitive Behavior Therapy

What is CBT

- the therapist works towards the modification of the patient's belief system
- based on the *cognitive model*: the link between a person's thoughts and feelings a primary factor contributing to mental illness
- original focus on depression but has expanded

Differences from MI

- *topics* not restricted to substance use
- *quality assessment*



how

what

MI
(e.g. did the therapist ask enough open questions?)

CBT
(e.g. was homework assigned?)

Cognitive Therapy Rating Scale

- 11 session-level codes scored on a 7-point Likert scale (0=poor, 6=excellent)
- $\sum_{i=1}^{11} \text{code}_i \geq 40 \Rightarrow$ competent delivery of CBT

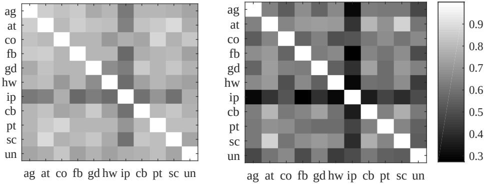
Table: CBT quality codes defined by CTRS.

abbreviation	meaning	
ag	agenda	*management and structure*
fb	feedback	
pt	pacing and efficient use of time	
hw	homework	
un	understanding	*good relationship*
ip	interpersonal effectiveness	
co	collaboration	
gd	guided discovery	*conceptualization*
cb	focusing on key cognitions and behaviors	
sc	strategy for change	
at	application of cognitive-behavioral techniques	

Dataset

- Beck Community Initiative: recorded sessions, annotated with the CTRS, used for training in CBT
- *adout* set: 386 adult outpatient sessions from 131 therapists
    - *trans* set: 92 sessions from 70 therapists
        - SNR > 7dB
        - highest/lowest total CTRS in *adout*
        - manually transcribed

Figure: Correlation matrices.



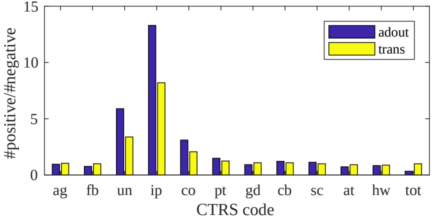
(a) *trans* (b) *adout*

T. Creed, *et al*, "Implementation of transdiagnostic cognitive therapy in community behavioral health: The Beck Community Initiative.", *Journal of consulting and clinical psychology*, vol. 84, no. 12, pp. 1116-1126, 2016

Dataset

additional training of the therapist?
alternative strategies for the patient?

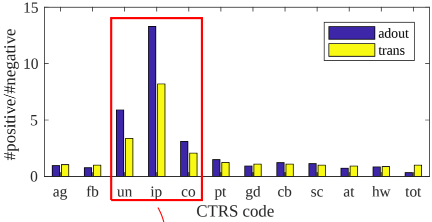
- binary classification problem:
  Is CBT delivery satisfactory or in need of improvement?
- binarization: code $\geq 3 \Rightarrow$ satisfactory (positive)
  $\sum_{i=1}^{11} \text{code}_i \geq 40 \Rightarrow$ satisfactory (positive)



Dataset

additional training of the therapist?
alternative strategies for the patient?

- binary classification problem:
  Is CBT delivery satisfactory or in need of improvement?
- binarization: code $\geq 3 \Rightarrow$ satisfactory (positive)
  $\sum_{i=1}^{11} \text{code}_i \geq 40 \Rightarrow$ satisfactory (positive)



codes related to
patient-therapist relationship

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words
- pretrained word embeddings (GloVe)
  - session embedding = mean(utterance embeddings)
  - utterance embedding = mean(word embeddings)
  - 300-dimensional

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words
- pretrained word embeddings (GloVe)
  - session embedding = mean(utterance embeddings)
  - utterance embedding = mean(word embeddings)
  - 300-dimensional
- LIWC features
  - wide use in psychology domain
  - word occurences belonging to pre-defined category dictionaries
  - 46-dimensional (psychological processes + personal concerns)

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words
- pretrained word embeddings (GloVe)
  - session embedding = mean(utterance embeddings)
  - utterance embedding = mean(word embeddings)
  - 300-dimensional
- LIWC features
  - wide use in psychology domain
  - word occurences belonging to pre-defined category dictionaries
  - 46-dimensional (psychological processes + personal concerns)
- Psycholinguistic Norm Features (PNFs)
  - lexical norms encoding aspects such as emotion, age, valence, etc.
  - session norm = mean(utterance norms)
  - utterance norm = mean(word norms)
  - 39-dimensional

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words
- pretrained word embeddings (GloVe)
  - session embedding = mean(utterance embeddings)
  - utterance embedding = mean(word embeddings)
  - 300-dimensional
- LIWC features
  - wide use in psychology domain
  - word occurences belonging to pre-defined category dictionaries
  - 46-dimensional (psychological processes + personal concerns)
- Psycholinguistic Norm Features (PNFs)
  - lexical norms encoding aspects such as emotion, age, valence, etc.
  - session norm = mean(utterance norms)
  - utterance norm = mean(word norms)
  - 39-dimensional
- Dialogue Acts
  - capture some form of dyadic interaction
  - features: total number of each DA in a session
  - 7 dimensions: question / statement / agreement / appreciation / incomplete / backchannel / other

Feature Sets

- unigrams with tf-idf
  - word occurences, downscaled for very common words
- pretrained word embeddings (GloVe)
  - session embedding = mean(utterance embeddings)
  - utterance embedding = mean(word embeddings)
  - 300-dimensional
- LIWC features
  - wide use in psychology domain
  - word occurences belonging to pre-defined category dictionaries
  - 46-dimensional (psychological processes + personal concerns)
- Psycholinguistic Norm Features (PNFs)
  - lexical norms encoding aspects such as emotion, age, valence, etc.
  - session norm = mean(utterance norms)
  - utterance norm = mean(word norms)
  - 39-dimensional

✓ interpretability

- Dialogue Acts
  - capture some form of dyadic interaction
  - features: total number of each DA in a session
  - 7 dimensions: question / statement / agreement / appreciation / incomplete / backchannel / other

Experimental Workflow

Feature Extraction

- *trans*: extract the features from the transcribed text
- *adout*: first, decode the audio session
  - VAD
  - diarization
  - role matching
  - ASR

Feature Normalization

- all the features standardized, tf-idfs l_2-normalized
- dimensionality reduction for tf-idfs:

  select K best features based on F-test and 5-fold cross-validation on total CTRS

Final Results

- 5-fold cross-validation across therapists
- classifier: linear support vector machine

B. Xiao, C. Huang, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling", *PeerJ Computer Science*, vol. 2, p. e59, 2016

Results

Therapist

Patient

majority class

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	**0.91**	0.69	0.45	0.82	0.78	0.61	0.73	0.35	0.78	0.68	0.32
fb	**0.83**	0.69	0.48	0.82	0.75	0.62	0.69	0.32	0.73	0.67	0.32
un	**0.55**	0.47	0.46	0.51	0.52	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.46	0.43	0.41	**0.62**	0.46	0.56	0.44	0.39	0.47	0.49	0.57
co	0.63	0.56	0.49	**0.65**	0.57	0.57	0.61	0.33	0.71	0.57	0.40
pt	**0.87**	0.63	0.51	0.77	0.70	0.65	0.64	0.38	0.68	0.60	0.35
gd	**0.85**	0.67	0.47	0.74	0.71	0.54	0.66	0.41	0.64	0.64	0.34
cb	**0.85**	0.70	0.52	0.76	0.75	0.57	0.64	0.35	0.59	0.62	0.32
sc	**0.86**	0.69	0.50	0.81	0.78	0.58	0.68	0.38	0.69	0.61	0.31
at	**0.86**	0.71	0.50	0.76	0.75	0.67	0.63	0.38	0.70	0.61	0.34
hw	**0.82**	0.61	0.49	0.71	0.70	0.56	0.66	0.40	0.70	0.67	0.34
tot	**0.86**	0.71	0.49	0.81	0.76	0.63	0.68	0.37	0.71	0.65	0.31

Table: Averaged F_1 score for the classification of the *trans* sessions.

Results

Therapist

Patient

majority class

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	**0.91**	0.69	0.45	0.82	0.78	0.61	0.73	0.35	0.78	0.68	0.32
fb	**0.83**	0.69	0.48	0.82	0.75	0.62	0.69	0.32	0.73	0.67	0.32
un	**0.55**	0.47	0.46	0.51	0.52	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.46	0.43	0.41	**0.62**	0.46	0.56	0.44	0.39	0.47	0.49	0.57
co	0.63	0.56	0.49	**0.65**	0.57	0.57	0.61	0.33	0.71	0.57	0.40
pt	**0.87**	0.63	0.51	0.77	0.70	0.65	0.64	0.38	0.68	0.60	0.35
gd	**0.85**	0.67	0.47	0.74	0.71	0.54	0.66	0.41	0.64	0.64	0.34
cb	**0.85**	0.70	0.52	0.76	0.75	0.57	0.64	0.35	0.59	0.62	0.32
sc	**0.86**	0.69	0.50	0.81	0.78	0.58	0.68	0.38	0.69	0.61	0.31
at	**0.86**	0.71	0.50	0.76	0.75	0.67	0.63	0.38	0.70	0.61	0.34
hw	**0.82**	0.61	0.49	0.71	0.70	0.56	0.66	0.40	0.70	0.67	0.34
tot	**0.86**	0.71	0.49	0.81	0.76	0.63	0.68	0.37	0.71	0.65	0.31

Table: Averaged F_1 score for the classification of the *trans* sessions.

poor performance

Results

Therapist

Patient

majority class

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	**0.91**	0.69	0.45	0.82	0.78	0.61	0.73	0.35	0.78	0.68	0.32
fb	**0.83**	0.69	0.48	0.82	0.75	0.62	0.69	0.32	0.73	0.67	0.32
un	**0.55**	0.47	0.46	0.51	0.52	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.46	0.43	0.41	**0.62**	0.46	0.56	0.44	0.39	0.47	0.49	0.57
co	0.63	0.56	0.49	**0.65**	0.57	0.57	0.61	0.33	0.71	0.57	0.40
pt	**0.87**	0.63	0.51	0.77	0.70	0.65	0.64	0.38	0.68	0.60	0.35
gd	**0.85**	0.67	0.47	0.74	0.71	0.54	0.66	0.41	0.64	0.64	0.34
cb	**0.85**	0.70	0.52	0.76	0.75	0.57	0.64	0.35	0.59	0.62	0.32
sc	**0.86**	0.69	0.50	0.81	0.78	0.58	0.68	0.38	0.69	0.61	0.31
at	**0.86**	0.71	0.50	0.76	0.75	0.67	0.63	0.38	0.70	0.61	0.34
hw	**0.82**	0.61	0.49	0.71	0.70	0.56	0.66	0.40	0.70	0.67	0.34
tot	**0.86**	0.71	0.49	0.81	0.76	0.63	0.68	0.37	0.71	0.65	0.31

Table: Averaged F_1 score for the classification of the *trans* sessions.

best performance

Results

Therapist

Patient

majority class

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	**0.91**	0.69	0.45	0.82	0.78	0.61	0.73	0.35	0.78	0.68	0.32
fb	**0.83**	0.69	0.48	0.82	0.75	0.62	0.69	0.32	0.73	0.67	0.32
un	**0.55**	0.47	0.46	0.51	0.52	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.46	0.43	0.41	**0.62**	0.46	0.56	0.44	0.39	0.47	0.49	0.57
co	0.63	0.56	0.49	**0.65**	0.57	0.57	0.61	0.33	0.71	0.57	0.40
pt	**0.87**	0.63	0.51	0.77	0.70	0.65	0.64	0.38	0.68	0.60	0.35
gd	**0.85**	0.67	0.47	0.74	0.71	0.54	0.66	0.41	0.64	0.64	0.34
cb	**0.85**	0.70	0.52	0.76	0.75	0.57	0.64	0.35	0.59	0.62	0.32
sc	**0.86**	0.69	0.50	0.81	0.78	0.58	0.68	0.38	0.69	0.61	0.31
at	**0.86**	0.71	0.50	0.76	0.75	0.67	0.63	0.38	0.70	0.61	0.34
hw	**0.82**	0.61	0.49	0.71	0.70	0.56	0.66	0.40	0.70	0.67	0.34
tot	**0.86**	0.71	0.49	0.81	0.76	0.63	0.68	0.37	0.71	0.65	0.31

Table: Averaged F_1 score for the classification of the *trans* sessions.

second best performance

Results

Therapist

Patient

majority class

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	**0.91**	0.69	0.45	0.82	0.78	0.61	0.73	0.35	0.78	0.68	0.32
fb	**0.83**	0.69	0.48	0.82	0.75	0.62	0.69	0.32	0.73	0.67	0.32
un	**0.55**	0.47	0.46	0.51	0.52	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.46	0.43	0.41	**0.62**	0.46	0.56	0.44	0.39	0.47	0.49	0.57
co	0.63	0.56	0.49	**0.65**	0.57	0.57	0.61	0.33	0.71	0.57	0.40
pt	**0.87**	0.63	0.51	0.77	0.70	0.65	0.64	0.38	0.68	0.60	0.35
gd	**0.85**	0.67	0.47	0.74	0.71	0.54	0.66	0.41	0.64	0.64	0.34
cb	**0.85**	0.70	0.52	0.76	0.75	0.57	0.64	0.35	0.59	0.62	0.32
sc	**0.86**	0.69	0.50	0.81	0.78	0.58	0.68	0.38	0.69	0.61	0.31
at	**0.86**	0.71	0.50	0.76	0.75	0.67	0.63	0.38	0.70	0.61	0.34
hw	**0.82**	0.61	0.49	0.71	0.70	0.56	0.66	0.40	0.70	0.67	0.34
tot	**0.86**	0.71	0.49	0.81	0.76	0.63	0.68	0.37	0.71	0.65	0.31

Table: Averaged F_1 score for the classification of the *trans* sessions.

second best performance
out of interpretable features

Most Informative Words

- backward selection to find the 5 best words (tf-idfs) in each fold
- correlation of the words (tf-idfs) with the codes

⇒ ‘homework’, ‘agenda’, ‘evidence’ constantly among the best

Most Informative Words

- backward selection to find the 5 best words (tf-idfs) in each fold
- correlation of the words (tf-idfs) with the codes

⇒ 'homework', 'agenda', 'evidence' constantly among the best

Experiment

classify the *trans* sessions after deleting all those words

before deleting

after deleting

	tf-idf_T	da_T	tf-idf_T′	da_T′	tf-idf_T′ +da_T′
ag	0.91	0.78	0.73	0.78	**0.80**
fb	0.83	0.75	0.69	0.74	**0.78**
un	0.55	0.52	0.49	0.52	**0.60**
ip	0.46	0.46	0.46	0.47	0.47
co	0.63	0.57	0.53	**0.57**	0.56
pt	0.87	0.70	0.71	0.70	**0.75**
gd	0.85	0.71	0.66	0.71	**0.74**
cb	0.85	0.75	0.74	0.75	**0.78**
sc	0.86	0.78	0.74	0.78	**0.80**
at	0.86	0.75	0.68	0.75	**0.80**
hw	0.86	0.70	0.65	0.70	**0.73**
tot	0.86	0.76	0.71	**0.76**	**0.76**

Table: Averaged F_1 score for the classification of the *trans*.

Most Informative Words

- backward selection to find the 5 best words (tf-idfs) in each fold
- correlation of the words (tf-idfs) with the codes

⇒ ‘homework’, ‘agenda’, ‘evidence’ constantly among the best

Experiment

classify the *trans* sessions after deleting all those words

before deleting

after deleting

	tf-idf_T	da_T	tf-idf_T′	da_T′	tf-idf_T′ +da_T′
ag	0.91	0.78	0.73	0.78	**0.80**
fb	0.83	0.75	0.69	0.74	**0.78**
un	0.55	0.52	0.49	0.52	**0.60**
ip	0.46	0.46	0.46	0.47	0.47
co	0.63	0.57	0.53	**0.57**	0.56
pt	0.87	0.70	0.71	0.70	**0.75**
gd	0.85	0.71	0.66	0.71	**0.74**
cb	0.85	0.75	0.74	0.75	**0.78**
sc	0.86	0.78	0.74	0.78	**0.80**
at	0.86	0.75	0.68	0.75	**0.80**
hw	0.86	0.70	0.65	0.70	**0.73**
tot	0.86	0.76	0.71	**0.76**	**0.76**

tf-idfs significantly affected

Table: Averaged F_1 score for the classification of the *trans*.

Most Informative Words

- backward selection to find the 5 best words (tf-idfs) in each fold
- correlation of the words (tf-idfs) with the codes

⇒ ‘homework’, ‘agenda’, ‘evidence’ constantly among the best

Experiment

classify the *trans* sessions after deleting all those words

before deleting

after deleting

	tf-idf_T	da_T	tf-idf_T′	da_T′	tf-idf_T′ +da_T′
ag	0.91	0.78	0.73	0.78	**0.80**
fb	0.83	0.75	0.69	0.74	**0.78**
un	0.55	0.52	0.49	0.52	**0.60**
ip	0.46	0.46	0.46	0.47	0.47
co	0.63	0.57	0.53	**0.57**	0.56
pt	0.87	0.70	0.71	0.70	**0.75**
gd	0.85	0.71	0.66	0.71	**0.74**
cb	0.85	0.75	0.74	0.75	**0.78**
sc	0.86	0.78	0.74	0.78	**0.80**
at	0.86	0.75	0.68	0.75	**0.80**
hw	0.86	0.70	0.65	0.70	**0.73**
tot	0.86	0.76	0.71	**0.76**	**0.76**

DAs not affected!

Table: Averaged F_1 score for the classification of the *trans*.

Results after decoding

	tf-idf_T	tf-idf_T +da_T	baseline
ag	0.71	0.71	0.33
fb	0.64	0.62	0.36
un	0.46	0.46	0.46
ip	0.48	0.48	0.48
co	0.45	0.43	0.43
pt	0.60	0.64	0.37
gd	0.63	0.68	0.34
cb	0.67	0.67	0.35
sc	0.61	0.66	0.35
at	0.62	0.64	0.37
hw	0.63	0.65	0.35
tot	0.56	0.58	0.42

Table: Averaged F_1 score for the classification of the *adout* sessions.

- performance drop due to
  - ASR errors
  - more imbalanced classes
- not significant differences after feature fusion

Conclusions

- early results for interpretable evaluation of CBT
- therapist-related features have greater predictive power
- unigrams under tf-idf yield the best performance, but
  - sensitive to specific words ⇒ prone to ASR errors
  - fail to capture information relevant to the imbalanced, human-centric codes (un, ip, co)

Future Work

- regression instead of classification
- examine the extent to which different annotation systems (i.e. MI vs. CBT) capture unique therapeutic content