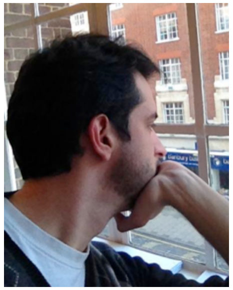
Analysing Dialogue for Diagnosis and Prediction in Mental Health

Matthew Purver
Queen Mary University of London







with Christine Howes, Rose McCabe, Claire Kelleher, Niall Gunter, Julian Hough

Mental Health & Language

- Communication is important in mental health:
  - Linguistic indicators of conditions
  - Communication *during treatment:*
    - Communication quality associated with outcomes
    - Conversation structure (how) and content (what)
  - Can NLP techniques help us analyse & understand therapy/conditions?

- PPAT project:
  - schizophrenia: face-to-face outpatient conversation
- AOTD project:
  - depression & anxiety: online text-based therapy
- SLADE project:
  - dementia: face-to-face clinical conversation

- (Howes, McCabe, Purver 2012-2014, 2018 to appear)

Questions

- Does language correlate with and/or predict symptoms & outcomes?
  - Can we use this to help diagnosis and/or treatment?
- What are the informative features?
  - Topic?
  - Sentiment/emotional content?
  - Conversation structure?
- Can we detect them automatically?
  - Accurately
  - Robustly
  - Using existing NLP techniques/tools
- How can we do better?

PPAT: Face-to-Face Dialogue

- Transcripts of therapy for schizophrenia
- Measures of symptom severity
  - *positive* (delusions, hallucinations, beliefs)
  - *negative* (withdrawal, blunted affect, alogia)
- Recorded related outcomes
  - ratings of communication quality
  - future adherence to treatment (6 months later):
    - non-adherence: risk of relapse 3.7 times higher
  - shared understanding known to be a related factor
- Manual annotation & statistical analysis
    - McCabe et al (2013)
- Automatic NLP processing & machine learning
    - Howes et al (2012; 2013)

Using Brute Force

- Classify entire dialogues (patient turns only) with SVMs, ngrams
  - Predict **non-**adherence to treatment 6 months later

Features	P (%)	R (%)	F (%)
Class of interest	28.9	100.0	**44.8**
Baseline features	27.0	51.9	**35.5**
Best ngram features	70.3	70.3	**70.3**

- Similar for symptoms, some outcomes e.g. HAS, PEQ
- Human psychiatrist given same task:

Data	P (%)	R (%)	F (%)
Text transcripts	60.3	79.6	**68.6**
Transcripts + video	69.6	88.6	**78.0**

- But how well will this generalise? And what does it **mean**?

Images: wikipedia, coursera.com

Using Brute Force

- Classify entire dialogues (patient turns only) with SVMs, ngrams
  - Predict **no**…

F…
Class…
Basel…
Best ng…

- Similar for s…
- Human psy…

Text…
Transc…

X_1

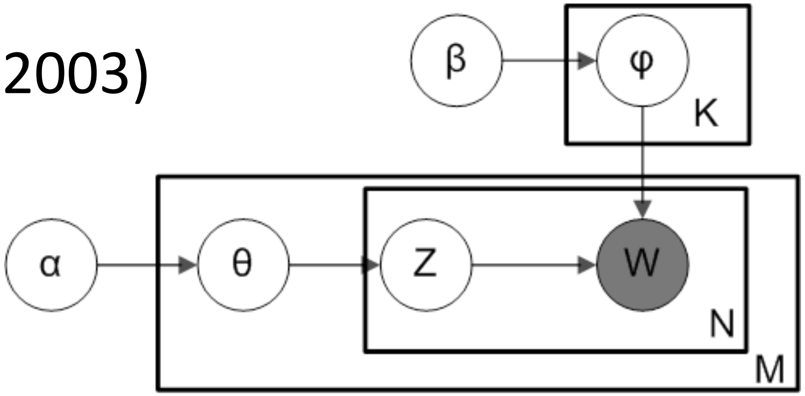
- But how well will this generalise? An…

Images: wikipedia, coursera.com

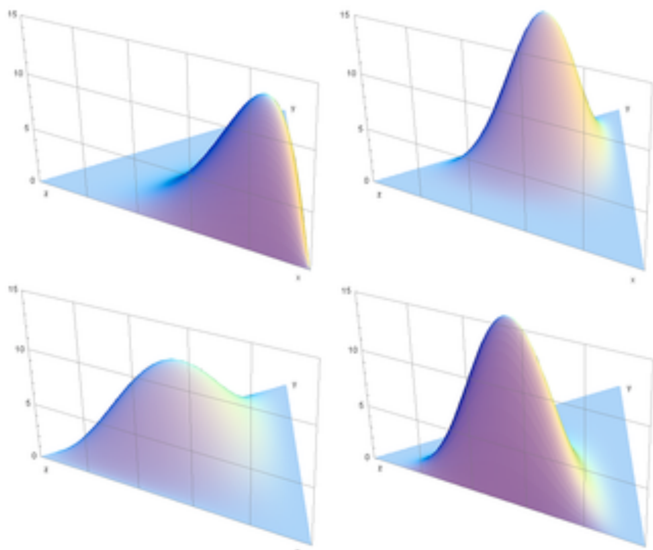
Manual topic segmentation

Topic	Name	Description
01	Medication	Any discussion of medication, excluding side effects
02	Medication side effects	Side effects of medication
03	Daily activities	Includes activities such as education, employment, household chores, dail
04	Living situation	The life situation of the patient, including housing, finances, benefits, plan
05	Psychotic symptoms	Discussion on symptoms of psychosis such as hallucinations and delusion
06	Physical health	Any discussion on general physical health, physical illnesses, operations,
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions, compulsions, phobias
08	Suicide and self harm	Intent, attempts or thoughts of self harm or suicide (past and present)
09	Alcohol, drugs & smoking	Current or past use of alcohol, drugs or cigarettes and their harmful effect
10	Past illness	Discussion of past history of psychiatric illnesses, including previous adm
11	Mental health services	Care coordinator, community psychiatric nurse, social worker or home tre
12	Other services	Primary care services, social services, DVLA, employment agencies, poli
13	General chat	Includes introductions; general topics; weather; holidays; end of appointm
14	Explanation about illness	Patients diagnosis, including doctor explanations and patients questions ab
15	Coping strategies	Discussions around coping strategies that the patient is using or the doctor
16	Relapse indicators	Relapse indicators and relapse prevention, including early warning signs
17	Treatment	General and psychological treatments, advice on managing anxiety, buildi
18	Healthy lifestyle	Any advice on healthy lifestyle such as dietary advice, exercise, sleep hyg
19	Relationships	Family members, friends, girlfriends, neighbours, colleagues and relations
20	Other	Anything else. Includes e.g. humour, positive comments and non-specific

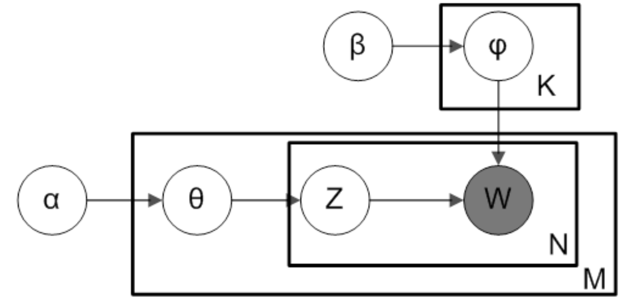
Topic Modelling



- Latent Dirichlet Allocation (Blei et al, 2003)
- Unsupervised Bayesian model:
  - texts as mixtures of "topics"
  - topics as distributions over words
- No prior knowledge of topics
  - number of topics
  - likely distribution shapes
  - (automatically optimised)
- Successful application in a wide range of domains & tasks



LDA topic modelling



- Infer 20 lexical "topics":

Topic 0	feel low alright mood long drug feeling tired time confider
Topic 4	voices pills mood cannabis telly voice shaking chris contro
Topic 5	letter health advice letters council copy send dla cpn prob
Topic 7	church voice voices hear medication sister bad hearing tak
Topic 9	school children kids back september oclock gonna phone
Topic 10	weight months medication stone risk lose eat write gp has
Topic 11	place support work centre gotta job stress feel psychologis
Topic 12	door house police thought ring knew worse wall hadnt sat
Topic 13	doctor alright years nice ill anxious write long sit eye hear
Topic 14	drug taking milligrams hundred doctor night time medicat
Topic 15	sort medication work drugs kind team issues drink alcohol
Topic 16	mum place brother tablets died dad depot house meet mo
Topic 17	people life drug make care lot friends dry camera live cop
Topic 18	alright house drink drinking money alcohol god drugs livir

LDA topic modelling

- LDA topics given manual “interpretations”:
  - (some include positive/negative sentiment aspect)

	Interpretation	Example words from top 20
0	Sectioning/crisis	hospital, police, locked
1	Physical health - side-effects of medication and other	gp, injection, operation
2	Non-medical services - liaising with other services	letter, dla, housing
3	Ranting - negative descriptions of lifestyle etc	bloody, cope, mental
4	Meaningful activities - social functioning	progress, work, friends
5	Making sense of psychosis	god, talking, reason
6	Sleep patterns	sleep, bed, night
7	Social stressors - other people stressors/helpful	home, thought, told
8	Physical symptoms - e.g. pain, hyperventilating	breathing, breathe, burning
9	Physical tests - Anxiety/stress arising from tests	blood, tests, stress
10	Psychotic symptoms - e.g. voices, etc.	voices, hearing, evil
11	Reasurrance/positive feedback/progress	sort, work, sense
12	Substance use - alcohol/drugs	drinking, alcohol, cannabis
13	Family/lifestyle	mum, brother, shopping
14	Non-psychotic symptoms - incl. mood, paranoia	feel, mood, depression

Manual vs LDA topic correlation

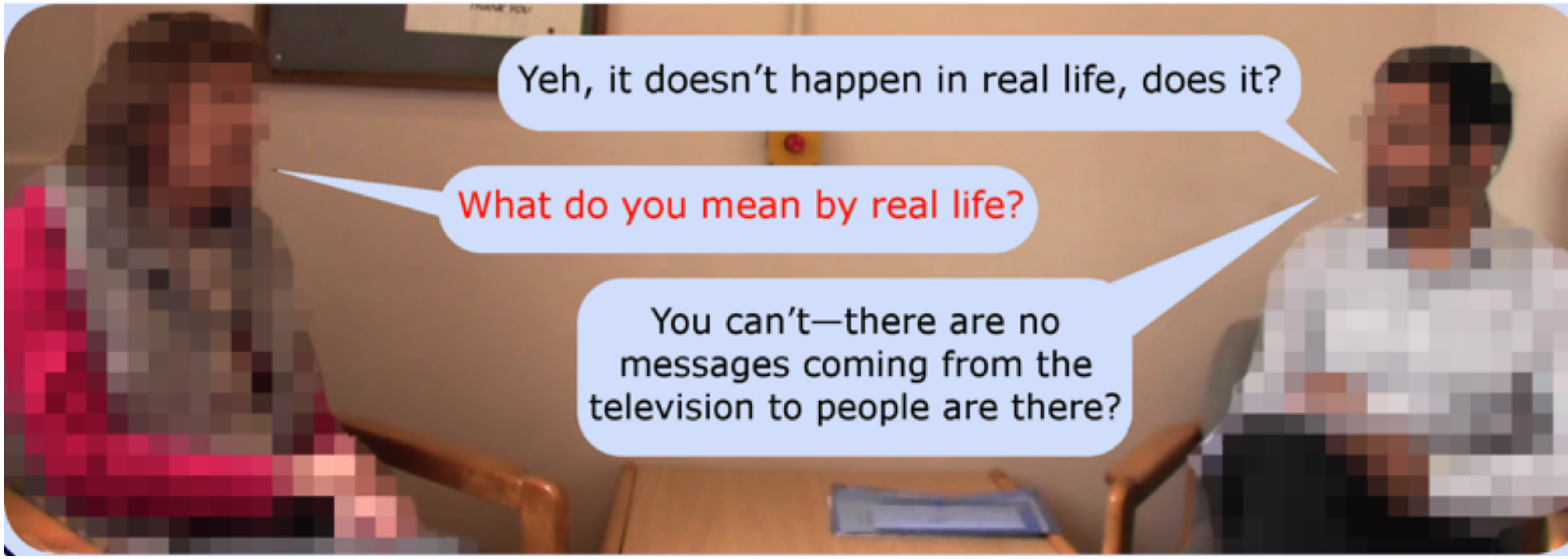
Hand-coded topic	Automatic topic	r	p
Medication	Medication regimen	0.643	<0.001
Psychotic symptoms	Making sense of psychosis	0.357	<0.001
Psychotic symptoms	Psychotic symptoms	0.503	<0.001
Physical health	Physical health	0.603	<0.001
Non-psychotic symptoms	Sleep patterns	0.376	<0.001
Suicide and self-harm	Weight management	0.386	<0.001
Alcohol, drugs and smoking	Substance use	0.651	<0.001
Mental health services	Non-medical services	0.396	<0.001
General chat	Sectioning/crisis	0.364	<0.001
Treatment	Medication issues	0.394	<0.001
Healthy lifestyle	Weight management	0.517	<0.001
Relationships	Ranting	0.391	<0.001
Relationships	Social stressors	0.418	<0.001
Relationships	Leisure	0.341	<0.001

Outcome prediction using topics

- Use topic weight per dialogue, with general Dr/P factors, as features:

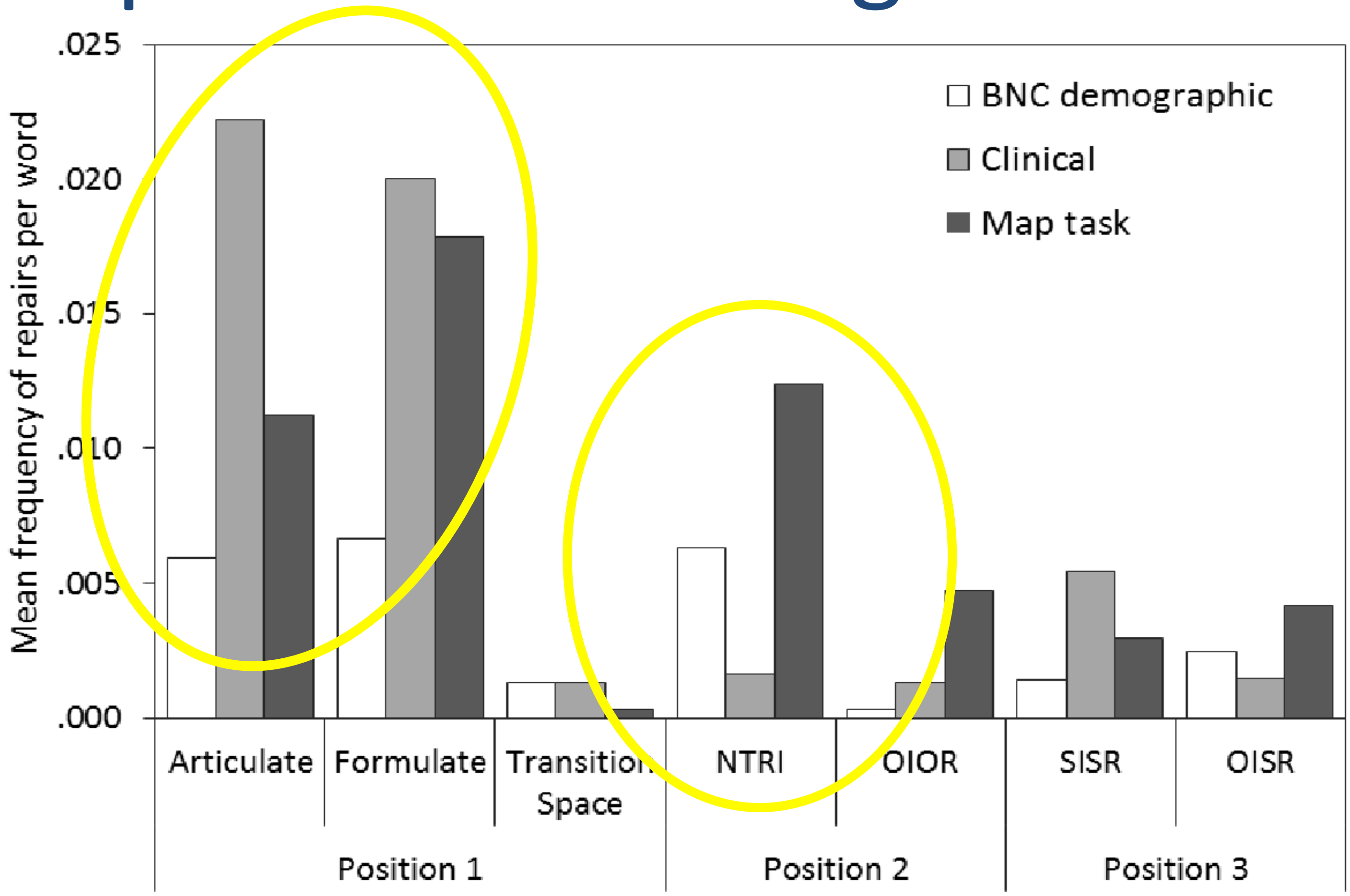
Measure	Manual Acc (%)	LDA Acc (%)
HAS Dr	**75.8**	**75.0**
HAS P	59.0	53.7
PANSS positive	**61.1**	58.8
PANSS negative	**62.1**	56.1
PANSS general	59.5	53.4
PEQ communication	59.7	56.7
PEQ comm barriers	**61.9**	**60.4**
PEQ emotion	57.5	57.5
Adherence (balanced)	**66.2**	54.1

Linguistic analysis: Repair



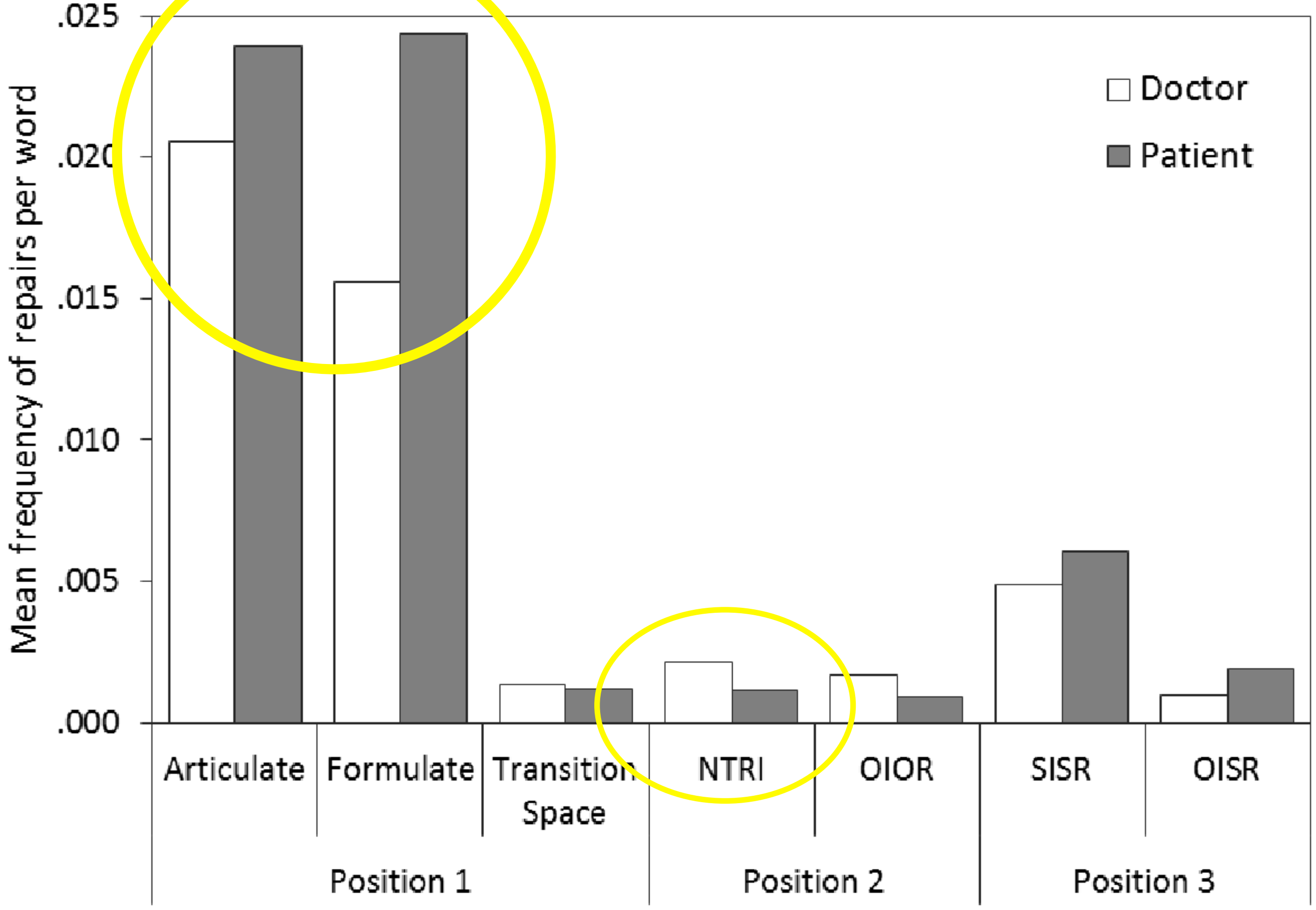
- Manual linguistic analysis
  - Significant role of *repair*
  - Patient-initiated other-repair & self-repair

Compare other dialogue contexts



- Therapy: more self-repair, less other-repair & initiation

Patient-doctor comparison



- Patients: more self-repair, less other-repair & initiation

But …

- Experiments with automatic other-repair detection didn't help:
  - A very sparse problem (e.g. <1% of turns)
  - Needs a general measure of parallelism
  - Needs vocabulary-independence
- So sparse, even perfect performance wouldn't have helped prediction
- We can do better now!
  - (see later)

Schizophrenia: Summary

- Predicting future adherence to treatment:
  - words & ngrams (phrases): 70%
  - humans: 70% (transcripts), 80% (video)
  - topics: 66% (manual), 54% (auto)
  - i.e. we can do it, but we don't really understand how …
- Topic modelling provides useful features:
  - topics correlate well with human-annotated topics
  - topics predict symptom severity
  - topics predict therapeutic relationship ratings
  - topics & emotion/stance interrelate
- Repair correlates with adherence
  - but automatic detection is difficult
  - … and it's a very sparse phenomenon

Online Text-based Therapy

- Text-based therapy for **depression & anxiety**
  - IESO Digital Health Ltd
- Cognitive Behavioural Therapy
  - 2,000 sessions, 500 patients, mean 5.65 sessions/patient
- Anonymisation using RASP
  - (Briscoe et al, 2006)
  - Non-trivial
- Outcome measure
  - Patient Health Questionnaire (PHQ-9)
  - Current severity, progress since start

Topics

- Themes include family, sleep, symptoms, progress, process:

0	time session sorry today great send next now one work thanks see thank please help make able perhaps look
1	feel life think know way things now like want make self feelings people change maybe someone much need others
2	right well great sure appointment feel thank just lol tonight please know get sorry say bye meeting last though
3	eating eat food weight sick drink meal now lunch control great chocolate absolutely day healthy dinner put use really
4	time husband mum family feel children now dad want see said friends also kids home life got school daughter
5	people say angry situation anger situations said way social others like one friends talk someone person behaviour saying know
6	get go know like need things going just think try want one something time good now make day start

Topics vs Schizophrenia

Sleep patterns	day sleep week time bed work mood night get things days	sleep day time feel bed bit things hours morning sleeping night
Family	time husband mum family feel children now dad want see said friends also kids home life	mum money dad brother shopping died enjoy tablets blood bad daughter sister
Food / weight	eating eat food weight sick drink meal now lunch control great chocolate absolutely day healthy	weight stone eat medication gain hospital twelve weigh exercise cut gym
Negative feelings	feel life think know way things now like want make self feelings	feel medication feeling thoughts time mood low head past illness
Crises	get help gp depression pain know medication health therapy sorry appointment last face moment	remember doctor hospital reason police people memory ring shaking headaches door
Social stress	work job time good stress working get school life money wife issues	things back place years thought bit ago home put day coming

Topic vs severity & progress

0	Materials, self-help, procedures	–		10	Unhelpful thinking/habits		
1	Feelings/effects of relationships on sense of self	+	+	11	Work/training/education issues/ goals		
2	Positive reactions/encouragement			12	Agenda/goal setting & review		
3	Issues around food			13	Panic attack description/explanation	–	–
4	Family/relationships & issues with (mostly negative)	+		14	Other healthcare professionals, crises, risk, interventions	++	
5	Responses to social situations			15	Sleep/daily routine	+	
6	Breaking things down into steps	+		16	Positive progress, improvements	––	–
7	Worries/fears/anxieties	–		17	Feelings, specific occasions/thoughts		
8	Managing negative thoughts/ mindfulness			18	Explaining/framing in terms of CBT model		+
9	Fears, checking, rituals, phobias	–	–	19	Techniques for taking control	–	–

Sentiment/Emotion Detection

- Detect positive & negative sentiment
  - see e.g. (DeVault et al, 2013)
- Detect anger
  - challenge & emotion elicitation in CBT process
- Compared existing tools
  - Manually annotated 85 utterances in 1 session
    - *positive / negative / neutral* (inter-annotator agreement $\kappa = 0.66$)
- Dictionary-based LIWC
  - sentiment 34-45%; anger recall = 0
- Data-based (RNNs) Stanford, trained on news text (85%)
  - sentiment 51-54% (no anger)
- Data-based (SVMs), trained on Twitter text
  - sentiment 63-80%

Sentiment/Emotion vs PHQ

	Severity (PHQ)	Progress (ΔPHQ)
Sentiment mean	––	–
Sentiment std dev		+
Anger mean/max	+	
Anger std dev	+	

- More positive sentiment ➔ better PHQ, progress
- More variable sentiment ➔ worse progress
- More/more variable anger ➔ worse PHQ

Predicting final outcomes

- Changes in levels help predicting final in/out-of-caseness:
  - using features from initial and/or final sessions:

	Final In-caseness
Baseline proportion	*26.8%*
First + last session features, incl deltas	**0.71 (0.48)**
Including early PHQ scores	**0.76 (0.51)**

- Features chosen are informative:
  - Levels of sentiment & anger, progress & crisis/risk topics
    - Deltas between sessions
  - PHQ scores at assessment and initial treatment sessions

Predicting dropout

- Can we predict dropout & non-engagement?
  - 148 of 500 did not enter or stay in treatment

	Dropout
Baseline proportion	*29.6%*
Assessment session features	0.65 (0.26)
Treatment session features	**0.70 (0.59)**
Both sessions	**0.73 (0.64)**

- >70% accuracy using initial session features
  - But only by including fine-grained word features

Predicting therapy quality

- Can we distinguish “good” from “bad” therapists?
  – Top 25% vs bottom 25% based on number of patients recovered

	Dropout
Baseline proportion	*50%*
Only high-level features	0.67 (0.63)
Incuding lexical features	**0.78 (0.74)**

- Good accuracy using initial & final session features
  – But mostly by including fine-grained word features

Depression/Anxiety: Summary

- Topic modelling provides useful features:
  - correlate well with human-annotated topics and previous study
  - topics correlate with symptom severity and progress
- Emotion detection provides useful features:
  - levels and variability predict symptoms and progress
  - needs care choosing & training tools

- Predicting useful outcome measures:
  - recovery: 71%, 76% with PHQ information
    - emotion levels & variability; talk about progress & dealing with crises
    - (are we starting to understand what's going on?)
  - dropout: 73%
  - therapist quality: 78%
    - details of content & structure: we still don't understand these …

SLADE: Dementia Diagnosis

- U. Exeter dataset
  - 148 diagnosis conversations with doctor (& carer)
    - 70 positive diagnosis of dementia
    - 78 negative diagnosis (Mild Cognitive Impairment in some cases)
  - After referral from GP, memory tests/scans
  - Given diagnosis, advice
- Relatively early stage
  - Can we aid diagnosis?

Dementia & Language

- Vocabulary reduction (e.g. Hirst & Feng, 2012)
  - Authors over long timescales
- Content reduction (e.g. Orimaye et al 2014)
  - Fewer predicates, fewer utterances, shorter sentences
  - DementiaBank: 74%
- Speech features (Jarrold et al, 2014)
  - Including lexical class features e.g. pronoun/noun/verb frequencies
  - Small set, healthy controls: 80-90%
- Combined features (Fraser et al, 2016)
  - Impairment: semantic, syntactic, information, acoustic
  - DementiaBank: c.80%

- But we have short timescales, diagnosis-dependent content …
  - Advice on driving, legal requirements, future planning
  - And many other features e.g. length
  - Need **content-independent** features

Conversation-based studies

- Many CA-like studies (Watson et al 1999 … Jones et al 2015)
- Indicative dialogue-structural features:
  - “Lack of fluency”
    - Self-repair
    - Lack of topic coherence
  - Other-repair
    - Types, appropriateness, answering behaviour, lack of corrections
  - Question-answering
    - Avoidance strategies, contentlessness
  - Pausing behaviour
    - Intra- and inter-utterance
  - Backchannel behaviour
    - More contentless utterances vs lower use of continuants?
  - Laughter

Testing Interactional Features

- Add interactivity features to Fraser et al (2016) model:
  - Self-repair indices:
    - Pauses, filled pauses, incomplete words, repetition, edit terms
    - Similarity of post-filler terms
  - Other-repair indices:
    - Question forms, inter-turn pauses
    - Backchannels, answer length
  - General indices:
    - Laughter
    - “Don’t know” answers
    - Participant turn & question frequency/ratios

- Top 3 most predictive features are interactional (Kelleher et al, in prep)
- Performance improvement:
  - Benchmark replication: F=73.8%
  - Adding interactive features: F=79.4%

Doing better: Other-Repair

- (Howes et al SIGDIAL 2012; Purver et al, 2018 to appear)
- Discriminative classification
  - Weighted classifiers to combat sparsity
  - Per-turn incrementality
  - Assume adjacent antecedent-repair pairs
    - 85% of cases (Purver et al, 2003)
- Define features manually, extract automatically
  - Linguistically/observationally informed:
    - Wh-question words, closed class repair words
    - Backchannel behaviour, fillers, pauses, overlaps
    - Lexical parallelism, syntactic (POS) parallelism
    - Semantic parallelism (Mikolov et al 2013; Turian et al 2010)
  - Brute force "ceiling": all unigrams

Other-repair results

- Results on real unbalanced data:

Target	Features	P (%)	R (%)	F (%)	PRC
PCC	(baseline)	1	100	**3**	0.01
PCC	All high-level	44	43	**44**	0.40
PCC	All features	46	47	**46**	0.48
BNC	(baseline)	4	100	**8**	0.04
BNC	All high-level	55	55	**55**	0.52
BNC	All features	57	62	**60**	0.61
SWBD	(baseline)	0.2	100	**1**	0.00
SWBD	All high-level	54	52	**53**	0.50
SWBD	All features	52	60	**56**	0.58

- Worse on some datasets e.g. MapTask F-score 38-50 (PRC 0.55)

Doing better: Self-repair

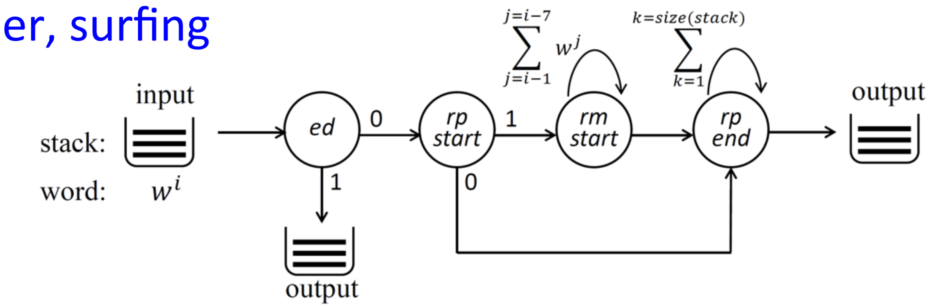
- (Hough & Purver, EMNLP 2014; Purver et al, 2018 to appear)
- "Disfluency detection" for speech recognition

  A flight to Boston – uh, I mean, to Denver

  ➔ A flight to Denver
- Per-word incremental output, maintaining semantic context:

  The interview was – it was alright

  I went swimming with Susan – or rather, surfing



- Domain-general, information-theoretic features:
  – Similarities between probability distributions
  – Changes in probability & entropy given repair hypotheses
  – Combined in random forest classifiers

Self-repair results

- Designed & trained on Switchboard corpus:
  - State-of-the-art accuracy F=0.85 (P 0.93 >> R 0.79)
  - Per-utterance correlation 0.96
  - Faster incrementality
- Transfer to mental health domain (PPAT):
  - F=0.62 (P 0.66 > R 0.59)
  - Per-utterance correlation 0.81
  - Per-dialogue correlation 0.94
- Other corpora less impressive (BNC, Colman & Healey 2011):
  - F=0.42 (P 0.40 < R 0.44)
  - Per-utterance correlation 0.58 ($p < 0.001$)

Summary

- We can predict useful outcome measures
  - (diagnosis, severity, adherence)
  - but we'd really like an interpretable model
- Topic & emotion detection is a useful step
  - needs care choosing & training tools
  - good for predicting symptoms and progress
  - not good for other outcomes (therapy quality, adherence …)
- Interaction modelling is another useful step
  - particularly the role of **repair** (self- and other-)
  - approximations help dementia diagnosis
  - general models are now in a state to be applied!

Acknowledgements

The CMSI project was supported by CreativeWorks London, a Knowledge Exchange Hub for the Creative Economy funded by the Arts and Humanities Research Council; and completed in collaboration with Chatterbox Labs Ltd & the Barbican

The PPAT & AOTD projects were supported by Queen Mary University of London's EPSRC-funded Pump-Priming and Innovation Funds, and PsychologyOnline Ltd; and completed in collaboration with PsychologyOnline Ltd and iLexIR Ltd.

The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Com- mission, under FET grant number 611733§