Language and Outcome Prediction in Patient-Clinician Dialogues

Matthew Purver, Christine Howes, Rose McCabe

School of Electronic Engineering and Computer Science
School of Medicine and Dentistry

Queen Mary, University of London
m.purver@qmul.ac.uk

Outline

1. Background
   - Therapy and doctor-patient communication
   - Repair
2. Repair in therapy dialogue
   - Trial and corpus study
   - Automatic detection of repair
3. Predicting therapy outcomes
   - Automatic prediction of adherence
4. Topic modelling for therapy dialogue
   - Manual & automatic topic identification
   - Automatic prediction of outcomes

Outline

1. Background
   - Therapy and doctor-patient communication
   - Repair
2. Repair in therapy dialogue
   - Trial and corpus study
   - Automatic detection of repair
3. Predicting therapy outcomes
   - Automatic prediction of adherence
4. Topic modelling for therapy dialogue
   - Manual & automatic topic identification
   - Automatic prediction of outcomes

Doctor-patient communication

- Mental ill health nearly half all ill health (Layard, 2012)
- Non-specific effects account for 60% of variance in patient outcome in clinical trials (Walach et al, 2005)
- One locus of non-specific effects is doctor-patient communication
- Shared understanding important (Mead et al, 2000)
  - How the patient understands the doctor
  - Common understanding of goals and implementation of treatment
- Associated with patient outcomes (Ong et al, 1995)
  - Patient satisfaction
  - Treatment adherence

Doctor-patient communication in schizophrenia

- Schizophrenia is a serious but treatable condition
  - estimated to affect 400,000 people in England
- Range of symptoms
  - Positive symptoms: delusions, hallucinations, beliefs
  - Negative symptoms: withdrawal, blunted affect, alogia
  - 30 item Positive and Negative Syndrome Scale (PANSS)
- Non-adherence to treatment a significant problem
  - About half of patients are non-adherent in the year after discharge from hospital (Weiden & Olfson, 1995)
  - Risk of relapse 3.7 times higher (Fenton et al, 1997)
- Shared understanding important (McCabe et al, 2002)
  - Patient Experience Questionnaire (PEQ)
  - Helping Alliance Scale (HAS)

Doctor-patient communication in schizophrenia

- Shared understanding important (McCabe et al, 2002)
- Recent research: features of therapy dialogue associated with outcomes, including adherence
- Dialogue content: *topics* (Hermann et al., in prep.)
  - Patients focus on symptoms, doctors on treatment
- Dialogue structure: *repair* (McCabe et al., in prep.)
  - Building shared understanding?

Repair is . . .

- A dialogue phenomenon – used by speakers to:
  - formulate understanding of one's own talk
  - clarify understanding of other's talk
  - address misunderstanding of own and other's talk
- Pervasive in dialogue (e.g. Schegloff, 1992)
  - identifying and resolving (potential) misunderstandings
  - driving the process of grounding (Clark, 1996)
  - collaboration to achieve shared understanding
- Important in dialogue systems
  - signalling (potential) misunderstandings
  - understanding corrections and confirmations

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):

Example - Turn positions

A:	How are you?	← *position 1*
B:	Fine thanks	← *position 2*
A:	That's good	← *position 3*

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):
  - P1SISR - Signal problem and resolve in own turn

Example - P1SISR - Articulation

Dr: You probably have seen so many psychiatrists **o o** **over the years**

Example - P1SISR - Formulation

Dr: **Did you feel that** **did you despair so much that** you wondered if you could carry on

Example - P1SISR - Transition Space

P: Where I go to do **some printing** **lino printing**

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):
  - P3SISR - Signal problem and resolve in subsequent own turn

Example - P3SISR

Dr: **Clorazil** or
P: Yeah
Dr: **Clozapine** **yes**

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):
  - P2OIOR - Other person signals and resolves in next turn

Example - P2OIOR

Dr: Rather than **the diazepam** which I don't think is going to do you any good
P: **the valium**

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):
  - P2NTRI - Other person signals a problem for the original speaker to resolve
  - P3OISR - Original speaker resolves a problem signalled by the other

Example - P2NTRI with P3OISR

Dr:	Yeh, it doesn't happen in real life does it?
P:	**What do you mean by real life?**

Types of repair

- Repair classified according to position (which turn), initiator (self or other) and repairer (self or other):
  - P2NTRI - Other person signals a problem for the original speaker to resolve
  - P3OISR - Original speaker resolves a problem signalled by the other

Example - P2NTRI with P3OISR

Dr:	Yeh, it doesn't happen in real life does it?
P:	**What do you mean by real life?**
Dr:	**You can't - there are no messages coming from the television to people are there?**

Outline

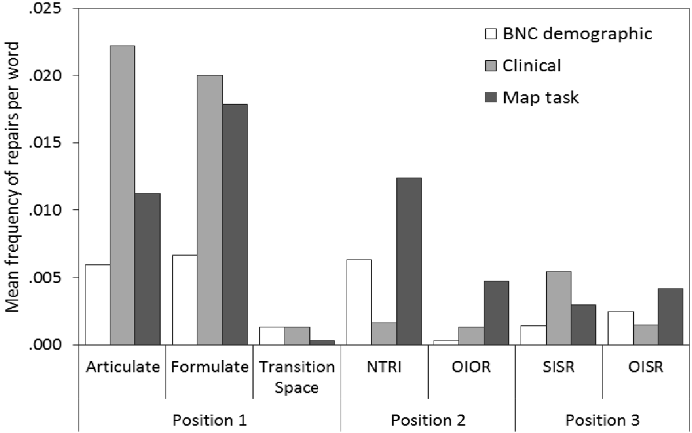
Recent study

- Observational study of psychiatrist-patient consultations
  - McCabe et al. (in prep.), Howes et al., (2012)
- Assess the use of repair in negotiating shared understanding
- Test the hypothesis that more repair is associated with better treatment adherence
- Explore which types of repair are relevant for adherence
- Compare with other dialogue data
  - BNC demographic, MapTask (Colman & Healey, 2011)

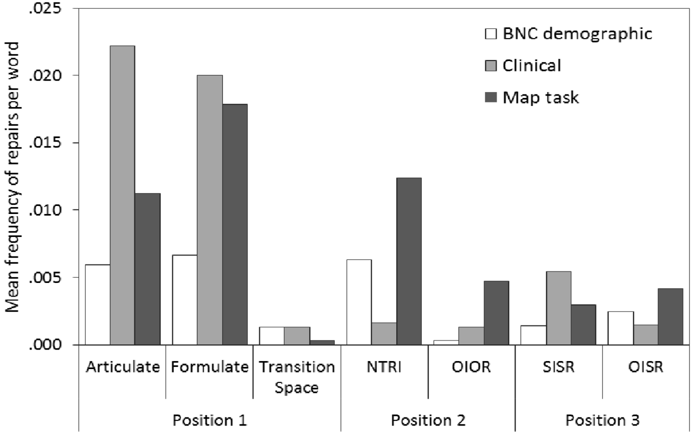
Design

- 138 consultations audio-visually recorded
  - Consultations transcribed and repair annotated
  - Inter-annotator agreement good ($\kappa = 0.73$)
- Patients interviewed to assess symptoms and evaluate communication
  - 30 item PANSS symptom scale
  - Patient Experience Questionnaire (PEQ)
- Adherence assessed after 6 months (general/medication)
  - Good $>$75%
  - Average 25-75%
  - Poor $<$25%

Comparison with other dialogue contexts

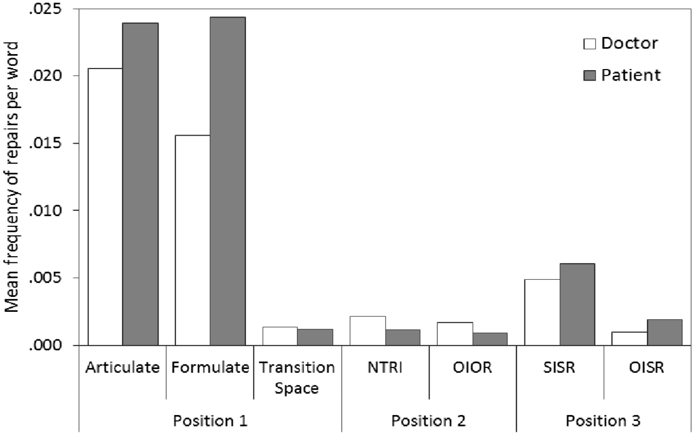


Comparison with other dialogue contexts

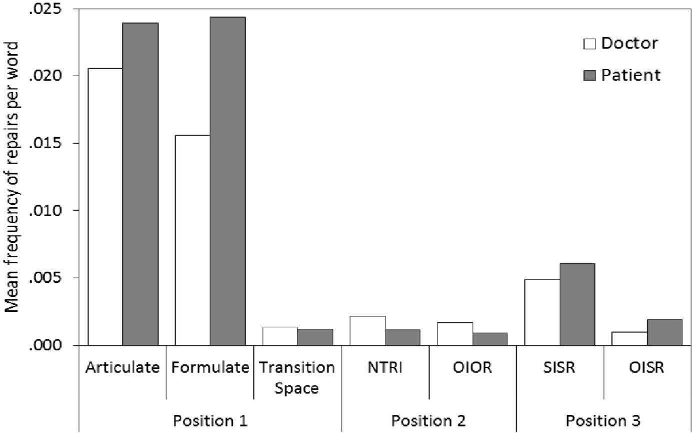


- More self-repair in clinical context
- Fewer NTRIs in clinical context

Mean frequencies of repair types per word



Mean frequencies of repair types per word



- Patients do more formulation repairs than doctors
- Patients produce fewer NTRIs than doctors

Repair and adherence

- Response variable (adherence) binary; good ($>$75%) or not
- Adherence not associated with length of illness or symptoms
- Principal component analysis to reduce number of variables
  - 4 factors, explaining 72% of the variance:
    - Psychiatrist led clarification and patient response (31%)
    - Patient led clarification and psychiatrist response (17%)
    - Patient reformulation (14%)
    - Psychiatrist reformulation (9%)
- Regression model (mixed effects)
  - One significant association with adherence after 6 months
  - Patient led clarification: odds ratio 5.82, $p = 0.02$
  - But 95% confidence margin wide: 1.3-25.8

Questions

- If repair correlates with adherence:
  - can we automatically detect repair?
  - can we automatically predict adherence?

Questions

- If repair correlates with adherence:
  - can we automatically detect repair?
  - can we automatically predict adherence?
- What does patient-led clarification measure?
  - Engagement
  - Understanding
  - Confrontation
  - Psychiatrist's communication style

Questions

- If repair correlates with adherence:
  - can we automatically detect repair?
  - can we automatically predict adherence?
- What does patient-led clarification measure?
  - Engagement
  - Understanding
  - Confrontation
  - Psychiatrist's communication style
- What else correlates with adherence (or with other outcomes)?
  - Other dialogue phenomena
  - Content of discussion

Examples

- Time for some quick examples . . . ?

Task: detecting patient-led clarification

- Identify repair phenomena of interest
  - next turn repair initiators (NTRI)
  - position 2 repairs (P2R)
  - especially in patient talk
- Automatically annotate transcripts for relevant features
  - following e.g. (Purver et al, 2001)
- Supervised discriminative classification
  - following (Fernandez et al. 2006, Schlangen 2005)

Dialogue act tagging

- Related to the general dialogue act tagging task
  - (label each turn with indication of function)
  - DA tagsets often include e.g. `check` category
- This is a sparse phenomenon
  - little attention paid in general tagging
  - (less common than question, statement etc)
- Accuracies on repair-related DA tags generally low:
  - e.g. Surendran & Levow (2006) on text
  - `check` 8% turns: 45% f-score
  - `clarify` 4% turns: 19% f-score
- Even sparser in our data: 0.8% of patient turns

Emotions in suicide notes

- Liakata, Kim, Saha et al (2012)
- Tagging sentences into a set of categories
- Some categories common (90% of data)
  - instructions, information
  - love, hopelessness, guilt, blame
- Sparse for some categories
  - happiness, hopefuless, fear, pride, abuse ...
- Union of set of supervised classifiers
  - Sequence classifiers (CRFs) as well as SVMs
  - 45.6% f-score overall
  - Sparse categories as low as 5% f-score

Specific dialogue phenomena

- Existing work for similar dialogue phenomena
- Fernández, Ginzburg and Lappin (2006)
  - Classifying *non sentential utterances* (fragments)
  - Supervised classification, lexical/dialogue features
  - Accuracy very good (>90% for clarification)
  - Only a subset of our task:
    - fragments only, fragment already identified
- Schlangen (2005)
  - Finding fragments (and antecedents)
  - Fragments relatively sparse: 5% of turns
  - Finding fragments accuracy low (30-40% f-score)

Clarification request (CR) taxonomies

- Purver, Ginzburg and Healey (2001)
- CRs are requests for the other person to provide repair of some aspect of a prior turn – subset of NTRIs
- Taxonomy of surface forms that CRs can take including
    - reprise fragments and sentences (*"a handbag?"*)
    - wh-fragments (*"who?"*)
    - wh-substituted reprises (*"have I what?"*),
    - explicit questions (*"what do you mean by that?"*)
    - general forms (*"eh?"*, *"pardon?"*)
- See also (Schlangen & Rodriguez, 2004)

Clarification request (CR) taxonomies

- Purver, Ginzburg and Healey (2001)
- CRs are requests for the other person to provide repair of some aspect of a prior turn – subset of NTRIs
- Taxonomy of surface forms that CRs can take including
  - reprise fragments and sentences (*"a handbag?"*)
  - wh-fragments (*"who?"*)
  - wh-substituted reprises (*"have I what?"*),
  - explicit questions (*"what do you mean by that?"*)
  - general forms (*"eh?"*, *"pardon?"*)
- See also (Schlangen & Rodriguez, 2004)
- No automatic detection/classification
- But can provide basis for distinctive features

Using specific lexical items

Dr: Ok you have done it before
P: **Pardon?**
Dr: If you have done it before

Dr: Presumably the ice has gone
P: **Eh?**
Dr: Presumably the ice has gone, it was quite icy this morning

Dr: Now from the psychiatric point of view because I'm not really a physical doctor that is your GP, who is your GP now
P: **What?**
Dr: Who is your GP

Using specific sentences

P: They're not negative erm but they're positive as i eh erm um it's like imagining how your life will be
Dr: Ok, ok, ok so thinking about how
P: **Do you know what I'm talking about?**
Dr: What, what you want to achieve in the future what you want to do

Dr: Well what kind things when you see yourself and you say you want to go back to to where you left of how you see yourself
P: **I'm not with you**
Dr: How do you look at yourself as in do you see positive things do you see negative things

Repeating lexical items

Dr: Yep well that is a possible side effect
P: **Side effect?**
Dr: Of the err Haliperidol

Dr: One thing that I ask you is when you were low in mood did you have suicidal thoughts
P: **Did I have . . . ?**
Dr: Suicidal thoughts

Reformulating

Dr:	Paroxitine
P:	**Fluoxitine**
Dr:	Ah Fluoxitine

Dr:	Right oh that's right so it's that it's gone back up to 130
P:	**150**
Dr:	150

Dr:	Who's your key worker there do you know
P:	**Err the person who comes to see me?**
Dr:	Yeah the person you see most often I suppose

Extending

Dr:	Yeah well as um shall we just um re re-start where we were we just commencing starting the interview when we um coz we see you was it couple of months three months
P:	**Since I saw you?**
Dr:	Yeah when was the last time I saw you

Dr:	Can you remember what you are on five
P:	No
Dr:	Or
P:	**10 milligrams?**
Dr:	10 milligrams

Combinations

Dr: Have you experienced this sensation in the past
P: **Have I what?**
Dr: If you have experienced this sensation in the past

Dr: So how are the headaches have they changed at all
P: **What do you mean changed** I got a headache now
Dr: Have they got worse or are they getting

Dr: Are you suspicious are you suspicious of people
P: **Suspicious?**
Dr: Paranoid
P: **Jealous?**
Dr: Jealous yeah

Combinations

Dr: Aaa so have you had any more thoughts about studying
P: **What music?**
Dr: Well you just you need to come up with a few ideas about what you might study

Dr: Do you do you really feel it or is it a sensation
P: **Is it what I'm thinking is that what you mean?**
Dr: No is it just err the mind playing tricks on you or is it something

Dr: That's right I don't think we've actually booked another time for the Clozapine clinic
P: **Have we already?**
Dr: No I don't think we have

Method

- Define features manually, then extract automatically:
  - Linguistically/observationally informed:
    - wh-question words, closed-class repair words
    - repeated fragments
    - ...
  - Brute force:
    - *all* the words used (unigrams)
    - patient only, to avoid doctor-specificity
- Train machine learning classifiers to detect NTRIs/P2Rs
  - Supervised classification (SVMs)
- 138 dialogues, c.44,000 turns (c.21,000 by patient)
  - 567 NTRIs (159 patient), 830 P2Rs (262 patient)
  - 5-fold cross-validation

Features

- Full feature set (one raw, one proportional):

Feature	Description
Speaker	Doctor, Patient, Other
NumWords	Number of words in turn
OpenClassRepair	Contains *pardon*, *huh* etc
WhWords	Number of wh-words (e.g. *what, who, when*)
Backchannel	Number of backchannels (e.g. *uh-huh, yeah*)
FillerWords	Number of fillers (e.g. *er, um*)
RepeatedWords	Number of words repeated from preceding turn
MarkedPauses	Number of pauses transcribed
OverlapAny	Number of portions of overlapping talk
OverlapAll	Entirely overlapping another turn

Results - balanced data

- Repair detection, balanced (i.e. small!) dataset:

Target	Features	Accuracy (%)
NTRI	Repeated proportion	65.9
NTRI	All high-level	78.1
NTRI	All unigrams	78.4
NTRI	All features	80.4
P2R	Repeated proportion	64.5
P2R	All high-level	75.7
P2R	All unigrams	77.2
P2R	All features	79.9

Results - balanced data

- Repair detection, balanced dataset, patient only:

Target	Features	Accuracy (%)
NTRI	Repeated proportion	61.2
NTRI	All high-level	83.4
NTRI	All unigrams	82.4
NTRI	All features	86.3
P2R	Repeated proportion	61.5
P2R	All high-level	78.5
P2R	All unigrams	77.1
P2R	All features	79.8

Results - balanced data

- Repair detection, balanced dataset, patient only:

Target	Features	Accuracy (%)
NTRI	Repeated proportion	61.2
NTRI	All high-level	83.4
NTRI	All unigrams	82.4
NTRI	All features	86.3
P2R	Repeated proportion	61.5
P2R	All high-level	78.5
P2R	All unigrams	77.1
P2R	All features	79.8

- But this is a sparse phenomenon (0.8% of turns)

Results - raw data

- Repair detection, raw (unbalanced) dataset:

Target	Features	F (%)	P (%)	R (%)
NTRI	High-level	27.3	36.0	22.3
NTRI	All	32.9	38.9	29.2
P2R	High-level	24.2	32.7	19.3
P2R	All	30.9	37.5	26.5

- Repair detection, raw (unbalanced) dataset, patient only:

Target	Features	F (%)	P (%)	R (%)
NTRI	OCRProportion	35.8	85.7	22.6
NTRI	High-level	41.4	42.8	40.6
NTRI	All	44.0	44.9	43.6
P2R	OCRProportion	19.6	56.4	11.8
P2R	High-level	31.6	36.2	28.4
P2R	All	35.4	43.8	30.3

Next steps

- We're ignoring non-transcript features
  - Intonation
  - Non-verbal behaviour
- We're ignoring dialogue *context*
  - Human annotators rely on subsequent turn
  - Presence of P3OISR
  - Some similar features to NTRIs (repetition etc)
- Joint problem:
  - Some similarity with decision detection
  - Fernandez et al (2007); Bui & Peters (2010)
- Does it actually help?

Outline

Method

- Apply the same approach to classifying entire dialogues
  - and therefore individual patients
- 125-138 dialogues only!
  - 37 associated with low subsequent adherence
  - 5-fold cross-validation
- Features normalised per dialogue, per word, per turn
- Lexical unigram feature space is very large ...
  - use correlation to find most predictive
  - patient only, to avoid doctor-specificity

Features

- Feature set used (one each for Doctor, Patient, Other):

Feature	Description
Turns	Total number of turns
Words	Total number of words spoken
Proportion	Proportion of talk in words
WordsPerTurn	Average length of turn in words
WhPerWord	Proportion of wh-words
OCRPerWord	Proportion of open class repair initiators
BackchannelPerWord	Proportion of backchannels
RepeatPerWord	Proportion of words repeated
OverlapAny	Proportion of overlapping talk
OverlapAll	Proportion entirely overlapping other turn
QMark	Proportion containing question intonation
TimedPause	Pause of more than c.200ms (where marked)

Results - raw data

- Adherence prediction, raw (unbalanced) dataset:

Features	F (%)	P (%)	R (%)
Baseline (all)	44.8	28.9	100
High-level	35.5	27.0	51.9
+ repair features	35.5	27.0	51.9
Best features (false!)	86.2	89.4	84.8

Results - raw data

- Adherence prediction, raw (unbalanced) dataset:

Features	F (%)	P (%)	R (%)
Baseline (all)	44.8	28.9	100
High-level	35.5	27.0	51.9
+ repair features	35.5	27.0	51.9
~~Best features (false!)~~	~~86.2~~	~~89.4~~	~~84.8~~
Best features (correct!)	70.3	70.3	70.3

Results - raw data

- Adherence prediction, raw (unbalanced) dataset:

Features	F (%)	P (%)	R (%)
Baseline (all)	44.8	28.9	100
High-level	35.5	27.0	51.9
+ repair features	35.5	27.0	51.9
~~Best features (false!)~~	~~86.2~~	~~89.4~~	~~84.8~~
Best features (correct!)	70.3	70.3	70.3

- Best (10-20) features selected over each training fold
  - Only words mentioned $>$ 40 times across set

Results - raw data

- Adherence prediction, raw (unbalanced) dataset:

Features	F (%)	P (%)	R (%)
Baseline (all)	44.8	28.9	100
High-level	35.5	27.0	51.9
+ repair features	35.5	27.0	51.9
~~Best features (false!)~~	~~86.2~~	~~89.4~~	~~84.8~~
Best features (correct!)	70.3	70.3	70.3

- Best (10-20) features selected over each training fold
  - Only words mentioned $>$ 40 times across set
- Human psychiatrist given same task:

Text transcripts	68.6	60.3	79.6

Results - raw data

- Adherence prediction, raw (unbalanced) dataset:

Features	F (%)	P (%)	R (%)
Baseline (all)	44.8	28.9	100
High-level	35.5	27.0	51.9
+ repair features	35.5	27.0	51.9
~~Best features (false!)~~	~~86.2~~	~~89.4~~	~~84.8~~
Best features (correct!)	70.3	70.3	70.3

- Best (10-20) features selected over each training fold
  - Only words mentioned $>$ 40 times across set
- Human psychiatrist given same task:

Text transcripts	68.6	60.3	79.6
Transcripts + video	78.0	69.6	88.6

Next steps

- So: we can (apparently) predict adherence as well as a human
- What next?

Next steps

- So: we can (apparently) predict adherence as well as a human
- What next?
- Add multimodal information
  - Video processing? Audio processing?
- How can we interpret what we have?
  - Why do we do this well?
  - What can we tell therapists?
- How well will it generalise?

Lexical features - predicting non-adherence

- Words chosen reflect some topical content:

air	fill	mates	simply
anyone	finished	monthly	sodium
balanced	fish	mouse	stable
bleach	flashbacks	nowhere	stock
build	grass	pains	symptoms
building	grave	possibly	talks
busy	guitar	pr	teach
challenge	h	recent	terminology
chemical	hahaha	removed	throat
complaining	lager	ri	virtually
cup	laying	schizophrenic	was
dates	lifting	sensation	wave
en	lucky	sickness	worse

Lexical features - predicting PEQ overall

- Different content with patient evaluation:

20th	electric	onto	sometime
ages	energy	overweight	son
angry	environment	oxygen	standing
anxiety	experiencing	packed	stomach
background	facilities	percent	suddenly
bladder	friendly	personally	sundays
booked	helps	picture	suppose
boy	ignore	played	table
broken	immediately	programs	team
bus	increased	progress	television
certificate	irritated	provide	thursdays
dead	kick	public	troubles
deep	later	quid	uhhm
drunk	lee	radio	upsetting
earn	loose	realised	walks
eeerrrr	low	reply	watchers

Outline

1. Background
   - Therapy and doctor-patient communication
   - Repair
2. Repair in therapy dialogue
   - Trial and corpus study
   - Automatic detection of repair
3. Predicting therapy outcomes
   - Automatic prediction of adherence
4. Topic modelling for therapy dialogue
   - Manual & automatic topic identification
   - Automatic prediction of outcomes

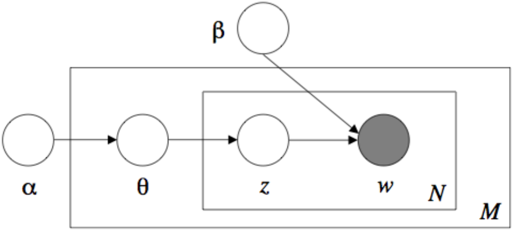
Can topics provide useful features?

- Existing manual definition of 20 *"topics"*
  - Medication, side-effects, treatment, management
  - Symptoms, health, self-harm
  - Daily activities, living situation, relationships, ...
- Higher-level aspect of content – more generalisable?
- Annotated over all 138 dialogues in same dataset

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
04	Living situation	The life situation of the patient, including housing, finances, b
05	Psychotic symptoms	Discussion on symptoms of psychosis such as hallucinations a
06	Physical health	Any discussion on general physical health, physical illnesses,
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions, compulsi
08	Suicide and self harm	Intent, attempts or thoughts of self harm or suicide (past and p
09	Alcohol, drugs & smoking	Current or past use of alcohol, drugs or cigarettes and their ha
10	Past illness	Discussion of past history of psychiatric illnesses, including p

Automatic topic modelling

- Can we learn topics from the data?
- Latent Dirichlet Allocation (Blei et al, 2003)



- Unsupervised generative approach:
    - "Document" = probability distribution over "topics"
    - "Topic" = probability distribution over words
    - Bayesian approach: integrate over possibilities
    - Hyperparameters govern sparseness of distributions

LDA topics

- LDA "topics" are lexical probability distributions:

Topic 0	feel low alright mood long drug feeling tired time confidence coming
Topic 4	voices pills mood cannabis telly voice shaking chris control inside ma
Topic 5	letter health advice letters council copy send dla cpn problems housir
Topic 7	church voice voices hear medication sister bad hearing taking felt nev
Topic 9	school children kids back september oclock gonna phone social son w
Topic 10	weight months medication stone risk lose eat write gp hasnt exercise
Topic 11	place support work centre gotta job stress feel psychologist theyll cor
Topic 12	door house police thought ring knew worse wall hadnt sat coming fea
Topic 13	doctor alright years nice ill anxious write long sit eye heart ring lovely
Topic 14	drug taking milligrams hundred doctor night time medication voices
Topic 15	sort medication work drugs kind team issues drink alcohol things sup
Topic 16	mum place brother tablets died dad depot house meet money lives da
Topic 17	people life drug make care lot friends dry camera live cope thing canc
Topic 18	alright house drink drinking money alcohol god drugs living basically
Topic 19	kind day time remember side weeks blood hospital appointment case

LDA topics

- LDA topics given a manual "interpretation":

	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

Comparing automatic and manual topics

- Cross-correlations across dialogues:

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

Do topics predict symptoms?

- Correlations between symptoms and topics across dialogues:

	Symptom scale	Topic	r	p
Hand-coded	positive	daily activities	-0.249	0.004
		psychotic symptoms	0.487	$<$0.001
	negative	daily activities	-0.211	0.015
		psychotic symptoms	0.206	0.018
	general	daily activities	-0.254	0.003
		psychotic symptoms	0.383	$<$0.001
		healthy lifestyle	-0.235	0.007
		suicide and self harm	0.230	0.008
Automatic	positive	ranting	0.265	0.002
		making sense of psychosis	0.378	$<$0.001
		physical tests	0.233	0.007
		psychotic symptoms	0.316	$<$0.001
	negative	weight management	-0.202	0.019
	general	ranting	0.234	0.007
		making sense of psychosis	0.316	$<$0.001

Classification experiments

- Include topic weight per dialogue as features
- Outcome prediction, manual topics:

	+Dr/P factors		Topics	
Measure	J48	SVM	J48	SVM
HAS Dr	**75.8**	**71.2**	50.8	56.8
HAS P	46.3	49.3	50.7	47.0
PANSS pos	58.0	59.5	**61.1**	58.0
PANSS neg	58.3	59.1	**61.4**	57.6
PANSS gen	51.9	55.0	55.7	59.5
PEQ comm	50.0	56.0	55.2	55.2
PEQ comm barr	50.7	**61.9**	52.2	52.2
PEQ emo	51.2	45.7	51.2	49.6
Adherence (balanced)	51.4	**66.2**	51.4	44.6

Classification experiments

- Include topic weight per dialogue as features
- Outcome prediction, LDA topics:

	+ Dr/P factors		Topics	
Measure	J48	SVM	J48	SVM
HAS Dr	**75.0**	**75.0**	**65.2**	**62.9**
HAS P	49.3	48.5	53.7	47.0
PANSS pos	45.0	58.8	51.1	50.4
PANSS neg	50.8	52.3	48.5	50.8
PANSS gen	47.3	50.4	53.4	48.9
PEQ comm	51.5	56.0	56.7	53.7
PEQ comm barr	56.7	**60.4**	51.5	56.0
PEQ emo	57.5	49.6	52.8	53.5
Adherence (balanced)	47.3	54.1	47.3	51.4

Conclusions

- We can detect repair quite well
  - ... but it's too sparse to predict outcomes
- We can predict some outcomes (including adherence)
  - ... but with specific & unhelpful lexical features
  - ... or (for some) with manually defined topics
- We can detect topics similar to manual topics
  - ... but they're good for some things, bad for others
- How do we separate form from content?
- What are the important features of each?