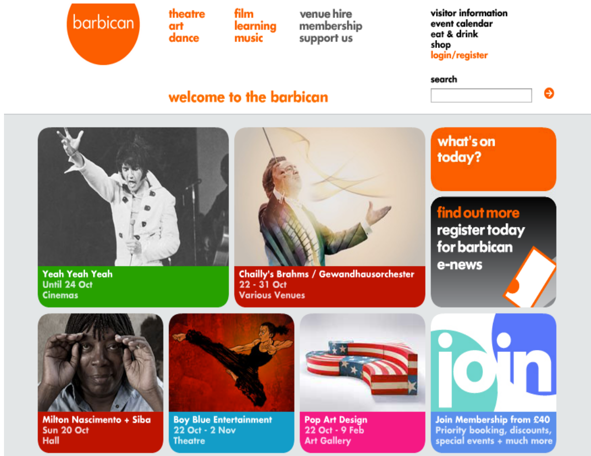
Sensing Mental States from Language

(While Learning Structure and Meaning)

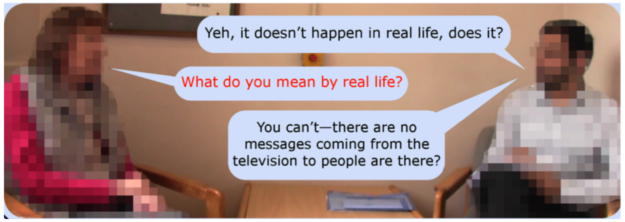
Matthew Purver
(and many others)

Kyoto University, March 2015

Language as a Sensor



- Can we use language to sense:
  - Mental states
  - Interests, emotions
- Two projects:
  - Characterising audiences via social media
  - Characterising patients from therapy



- How do we do better?
  - Sensing interaction structure
  - Learning meaning from distributions

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 sup- port 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§

CREATIVE WORKS LONDON

Arts & Humanities Research Council

Queen Mary University of London

PsychologyOnline

CONCRETE

EPSRC Pioneering research and skills

chatterboxlabs Data Science & Monetization

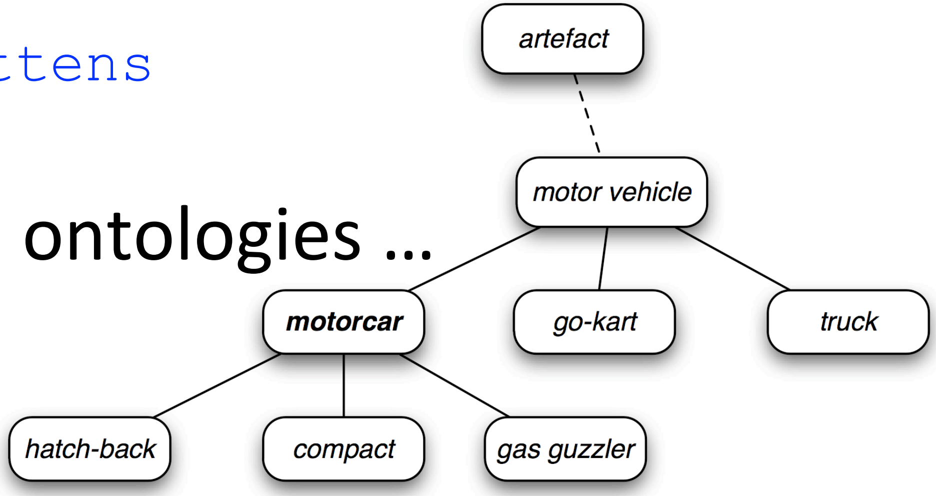
iLexIR NLP Consultancy

barbican

SEVENTH FRAMEWORK PROGRAMME

Sensing Language

- Structure at various levels: e.g. syntax, **interaction**
- The primary data (words) are **categorical**
  - How do we know when two items are "similar"?
    - `dogs chase cats`
    - `puppies pursue kittens`
- We can use dictionaries, ontologies …



- But what if the language isn't what you expect?

Social media language

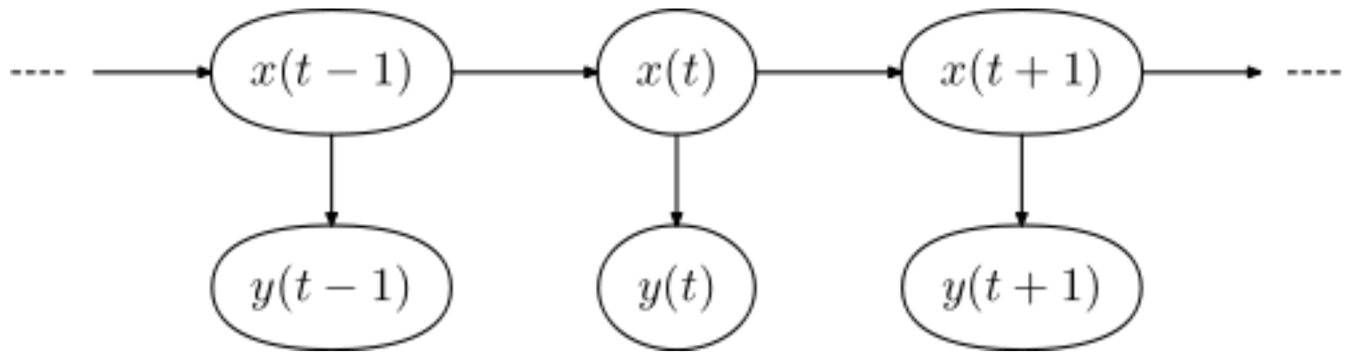
Nyt alexx tweetdreamsh RT @JDBAustralia: Goodnight everyone, i will tweet you all tomorrow <3 #loveislouder

Im Not Goin o2 Be Sad o2day Imah $MILE , Jus o4 Big Bruhh !

LOL IM BOR3D @ENYCHARM YU GOIN O2 DA M33TING?

Statistical Modelling

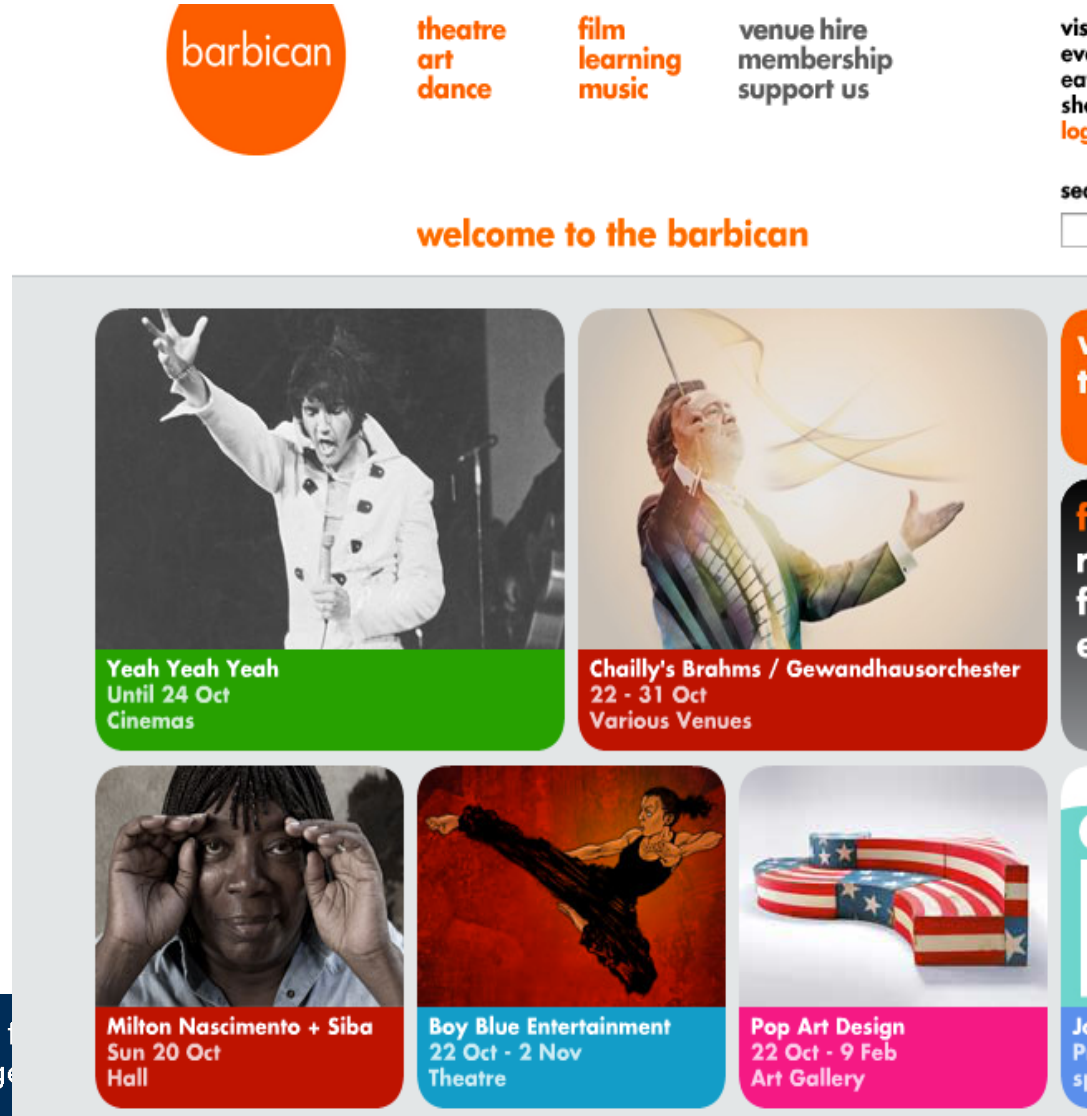
- Statistical models of word sequence
  - Generative language models



  - Discriminative classifiers with n-gram features
  - Convolutional deep neural networks
  - …

CMSI Project: the Barbican

- (Concannon & Purver, 2014)
- Understand audiences better
- Relate to offerings across artistic genres
- Locate curious audience members and learn from them
- Diversify audiences
- As automatically as possible
- Twitter: naturally occurring conversation as opposed to surveys etc



Topic

- #uburoi at the Barbican. Simply extraordinary. Awesome. #cheekbyjowl
- RT @AndreaTyrimos: First day painting at @BarbicanCentre to debut my 'BRICK'- am so excited! 1st pic of the day... @nick_that #Brick …
- @heylanikai Cool. Smoke Fairies outside Sister Ray on Berwick Street at 1600h and they're ace. I'm at Barbican in eve for Ana Moura. ☺
- Cider at the barbican in the sun #loveit
- @BrittenSinfonia @barbicancentre I want one with David Butcher jumping up
- "@BarbicanCentre @nicomuhly Certainly hope so, we'll see what we can do the week before when we play our concert with Ian Bostridge..."

Relevance

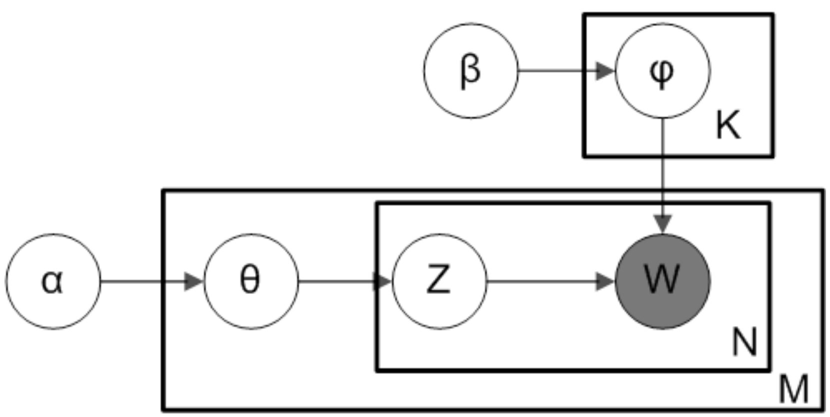
- Stand-up @MrEdByrne is in his Roaring Forties and coming to the York Barbican this October! http://t.co/qu1HpGHyn7
- I miss Bank Holidays on the Barbican #drunk #memories #Plymouth
- @Kimbletron Rub it in.. I wish I was down there on the Barbican with a big ice lolly.
- He must be baffled about what those painted lines in parking lots are for [Fontana Pharmacy, Barbican, Kingston]
- I swear @Shuhzia is so dumb, she thought I was drinking alcohol -.- it's Barbican, non alcoholic, ARABIC beverage
- Pomegranate barbican is an addiction♡
- Something going on at Barbican station about 12 emergency vehicles there mostly fire engines :-/

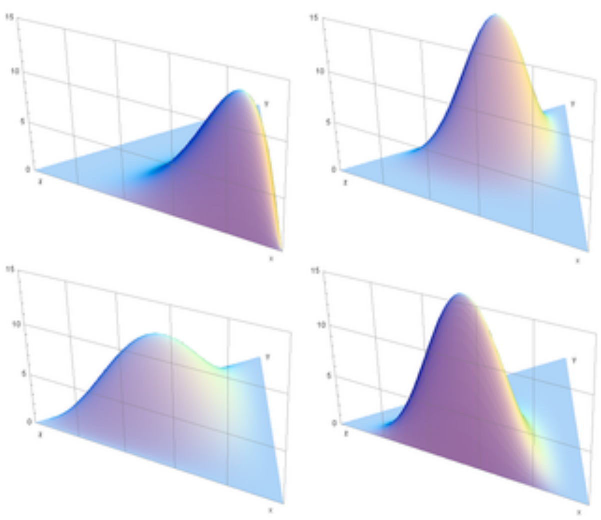
Supervised Approach

- Discriminative classifier (SVM/NBayes)
  - 1-3-gram features, 10-fold cross-validation
- Relevance: manual labels, c.3,500 examples
  - 95.2% accuracy
- Topics:
  - Barbican genre system: Art, Dance, Theatre, Music, Film
  - 84% accuracy
  - Some difficulties:
    - Film vs Music
    - Music vs Theatre
    - Festivals ("Hack the Barbican") vs all topics
  - One-vs-all classifiers: 87-97% accuracy
    - Maybe it's not that simple!

Unsupervised Approach

- Maybe tweets (like documents) are weighted **mixtures** of genres …
- Latent Dirichlet Allocation (Blei et al, 2003)



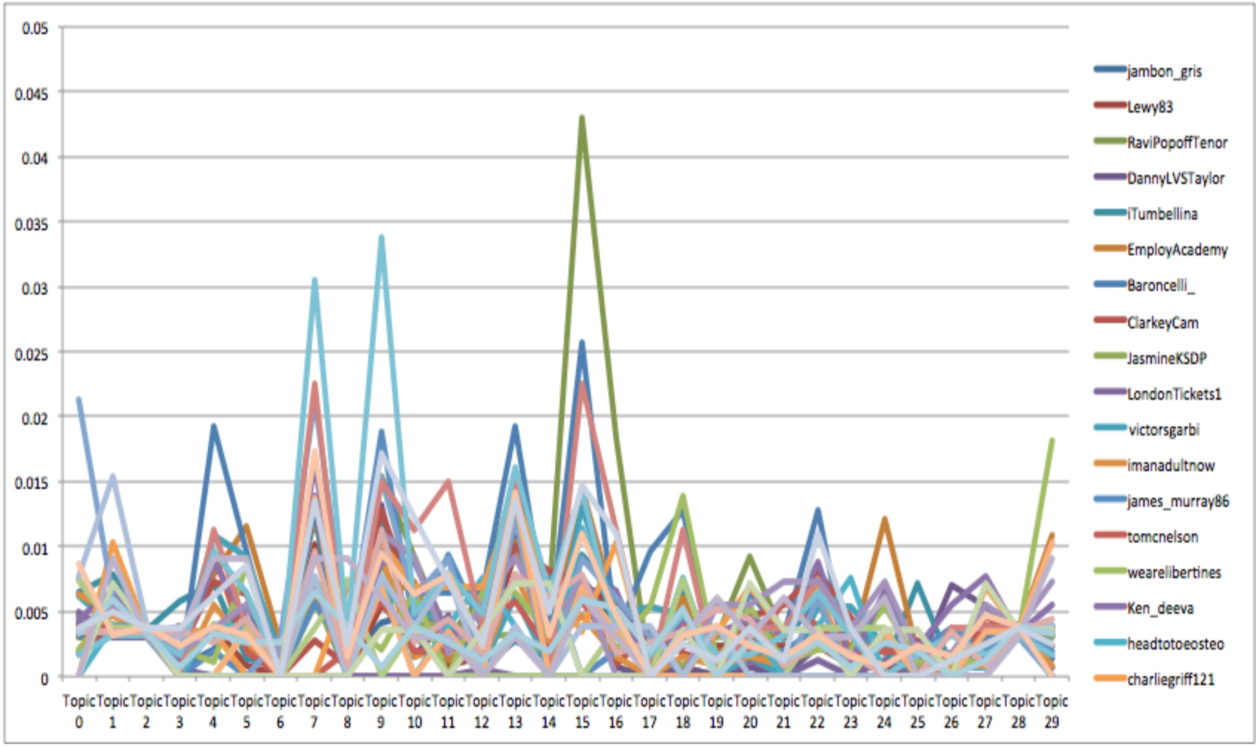


Unsupervised Approach

Enjoyment	Ludovico	HTB	Gigs	Festivals
time	einaudi	hackthebarbican	mogwai	london
good	ludovicoeinaud	uk	tickets	create
love	night	dollop	zidane	festival
lunch	tonight	hack	tonight	openeast
day	ludovico	today	live	park
back	amazing	jackmaster	mogwaiband	weekend
walk	concert	part	devendra	olympic
night	music	loefah	london	east
nice	time	htb	banhart	music
haha	evening	free	bought	open
beautiful	great	week	music	great
great	forward	installation	ticket	openeastfestival

Boundary Crossers

- And the same is true for people & their interests
  - Take last 100 tweets from timeline
  - Assign topic distributions from previous LDA experiment



NLP for Mental Health

- Communication is important in mental health:
  - Communication quality associated with outcomes
    - (Ong et al, 1995; McCabe et al, 2013)
  - Communication *during treatment:*
    - Conversation structure (how)
    - Conversation content (what)

- Can NLP techniques help us analyse & understand therapy?
- PPAT project:
  - transcripts of face-to-face therapy for schizophrenia
- AOTD project:
  - online text-based therapy for depression & anxiety

- (Howes, McCabe, Purver, SIGDIAL 2012, IWCS 2013, ACL 2014)

Face-to-Face Dialogue

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

Face-to-Face Therapy

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

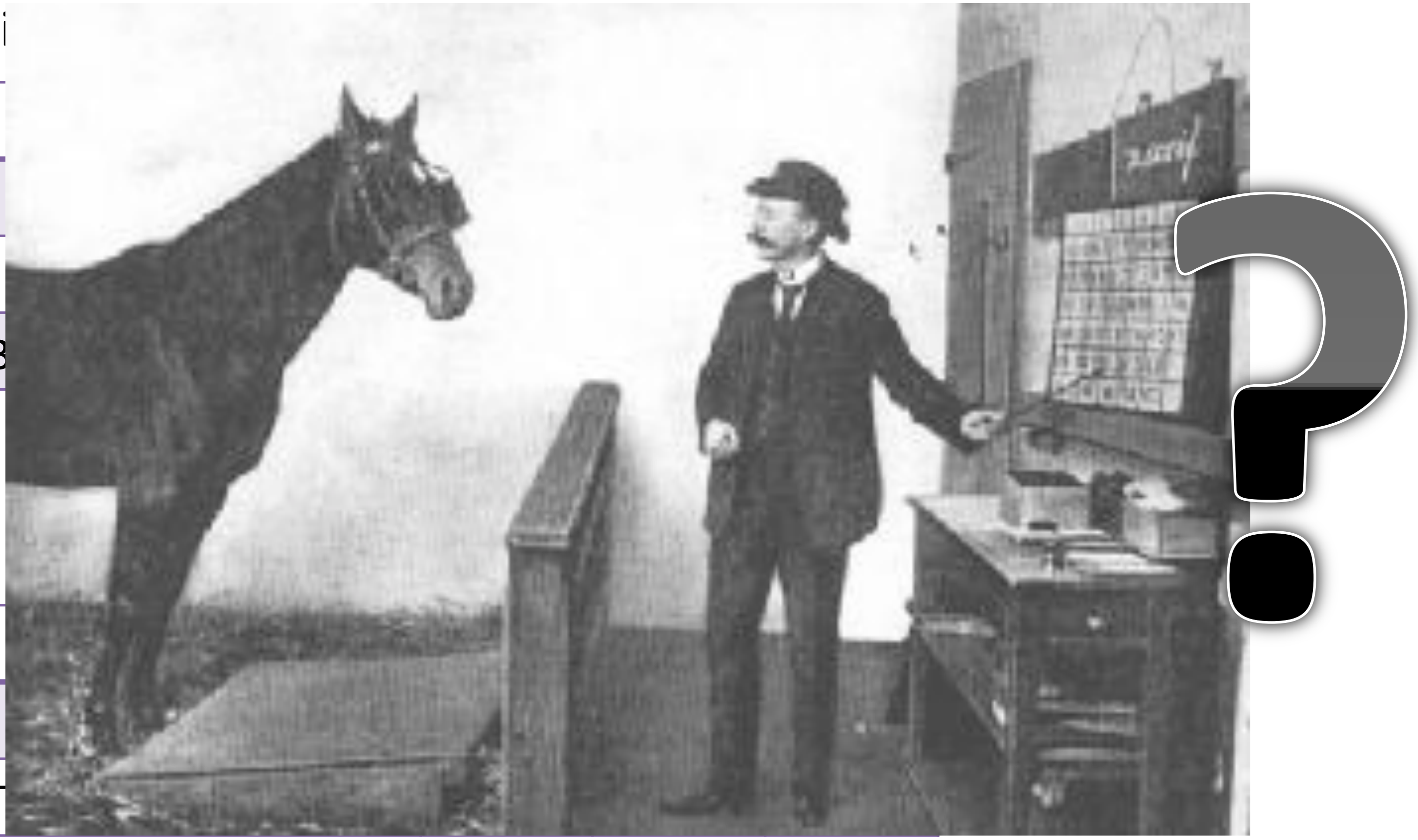
Features	P (%)	R (%)	F (%)
Class of interest	28.9	100.0	**44.8**
Baseline features	27.0	51.9	**35.5**
Best ngram features	70.3	70.3	**70.3**

- 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**

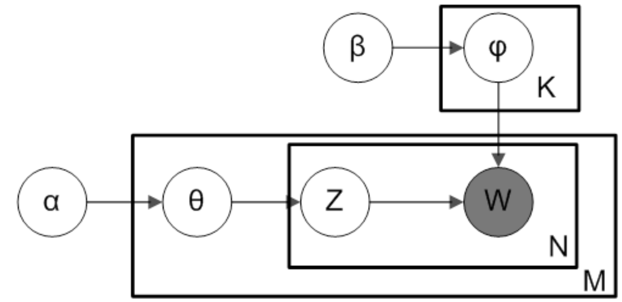
Face-to-Face Therapy

- Classify entire dialogues (patient turns only) with SVMs, ngrams
  - Predi



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

LDA topic modelling



- Infer 20 lexical "topics":

Topic 0	feel low alright mood long drug feeling tired time confide
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":
  - (including 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

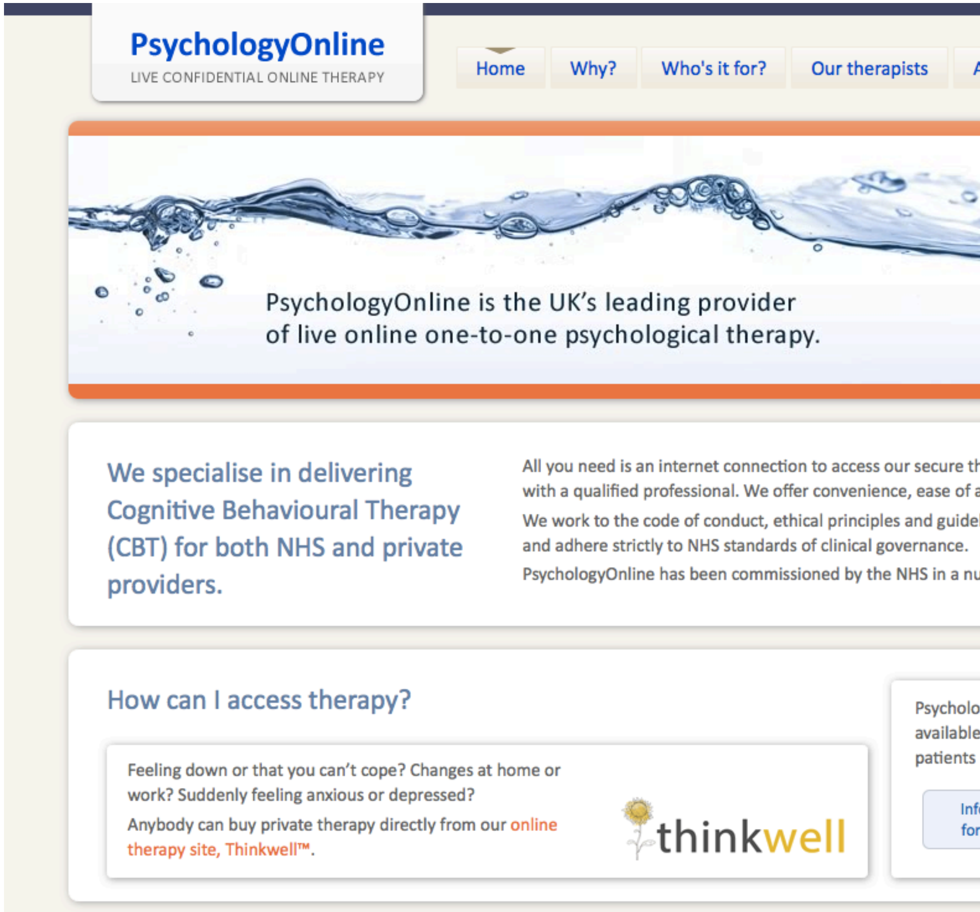
Outcome prediction using topics

- Include 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

Online Text-based Therapy

- Text-based therapy for depression & anxiety
  - PsychologyOnline Ltd
- Cognitive behavioural therapy
  - 500 patients, mean 5.65 sessions/patient
- Outcome measure
  - Patient Health Questionnaire (PHQ-9)
  - Current severity, progress since start



Online Therapy

- Online, anonymous, text-based therapy for depression and anxiety (PsychologyOnline Ltd)
  - Cognitive Behavioural Therapy (CBT)
  - 500 patients (352 female, 146 male, 2 unknown), 64 therapists
  - 2066 sessions, 1864 from ongoing or complete treatment
  - mean 5.65 sessions per patient (min 1, max 15)
- Anonymisation independently (via iLexIR Ltd)
  - Using RASP toolkit (Briscoe et al, 2006)
    - Person & organisation names, places, dates
    - Harder than standard text tasks …
    - … so some errors, manually corrected
- Outcome measures
  - Patient Health Questionnaire (PHQ-9)

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
  – sentiment 51-54% (no anger)

Distant Supervision


- A common technique for sentiment detection

```
Best day in ages! #Happy :)
just because people are celebs they dont
reply to your tweets! NOT FAIR :(
```

Distant Supervision

- A common technique for sentiment detection

  Best day in ages!

  just because people are celebs they dont reply to your tweets! NOT FAIR

Distant Supervision

- A common technique for sentiment detection

  Best day in ages!

  just because people are celebs they dont reply to your tweets! NOT FAIR

  再做个梦如果明天我中奖了该怎么支配呢每次想这个问题都 觉得很美*^_^*

  离队倒计时,期待奇迹的发生 (T_T)

Distant Supervision

- A common technique for sentiment detection

  Best day in ages!

  just because people are celebs they dont reply to your tweets! NOT FAIR

  再做个梦如果明天我中奖了该怎么支配呢每次想这个问题都 觉得很美

  离队倒计时,期待奇迹的发生

- e.g. Go et al (2009): works well ***if*** you have a reliable but (semi-)independent label to hand

Distant Supervision

- Can be applied to finer-grained emotions (Purver & Battersby, EACL 2012)
  - But quite bad for some … how reliable are these?:

`:-O` `:-@` `:-$` `:-P`

- Can also get supervision from responses:

_AggieGirl16: @captain_lizard lol yeaaaah. I'm pretty lucky! Haha!

captain_lizard: @_AggieGirl16 I'm glad you're happy, Monica! :)

Sentiment/Emotion Detection

- Detect positive & negative sentiment
  - see e.g. (DeVault et al, 2013)
- Detect anger
  - challenge & emotion elicitation in CBT process
- Compared 3 existing tools
  - 1 dictionary-based: LIWC
  - 2 data-based: Stanford (news), Sentimental (social media)
- Manually annotated 85 utterances in 1 session
  - *positive / negative / neutral*
  - Inter-annotator agreement $\kappa = 0.66$
  - LIWC 34-45%; Stanford 51-54%; Sentimental 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 seem informative:
  - Levels of anger, progress & crisis/risk topics
  - PHQ scores at assessment and initial treatment sessions

Predicting dropout

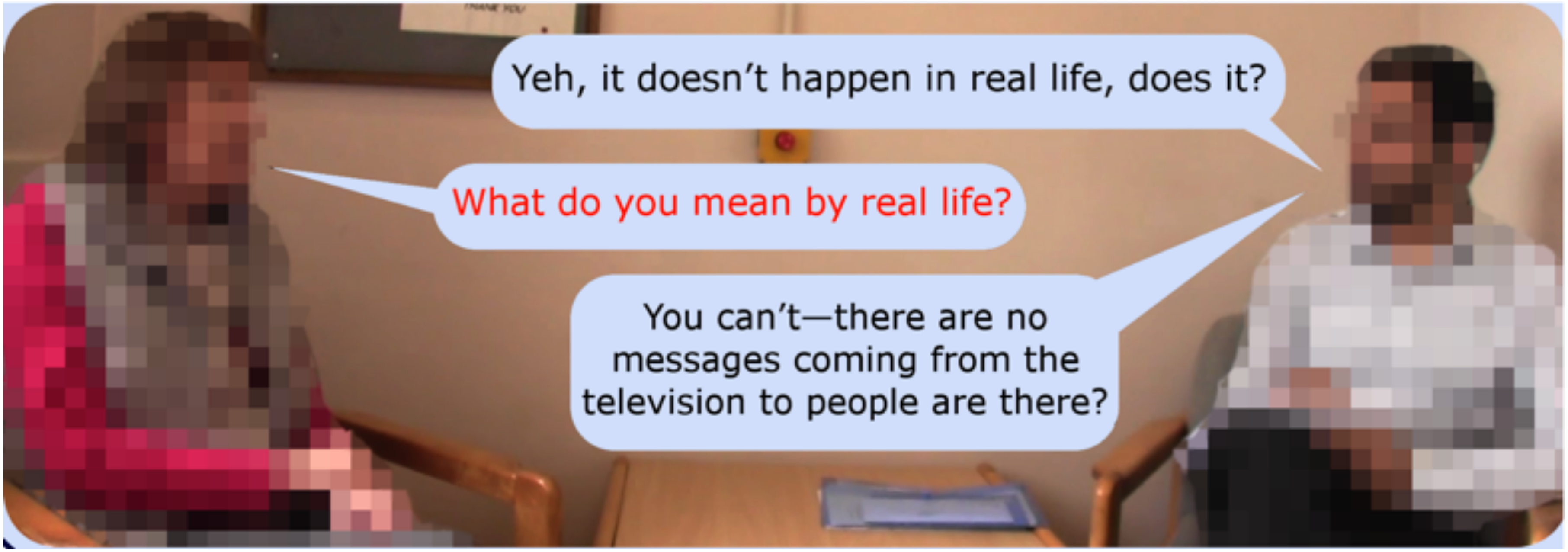
- Can we predict dropout & non-engagement?
  - 148 of 500 did not enter or stay in treatment

	Dropout
Baseline proportion	*29.6%*
Assessment session features	0.65 (0.26)
Treatment session features	**0.70 (0.48)**
Both sessions	**0.73 (0.52)**

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

- But sometimes this doesn't work as well …

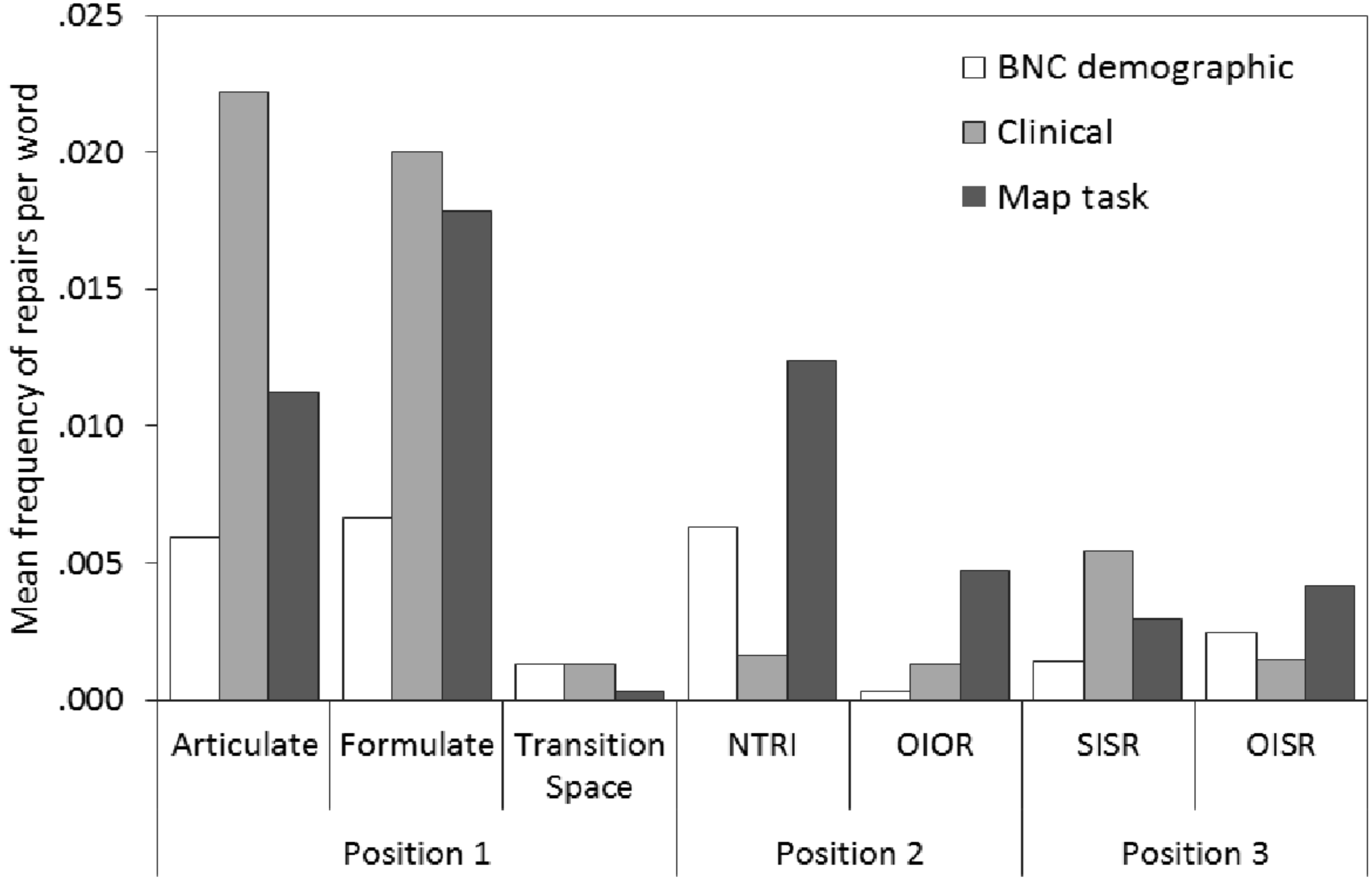
Schizophrenia & Repair



- Manual linguistic analysis
  - Significant role of *repair*
  - Patient-initiated other-repair (above)
  - And self-repair:

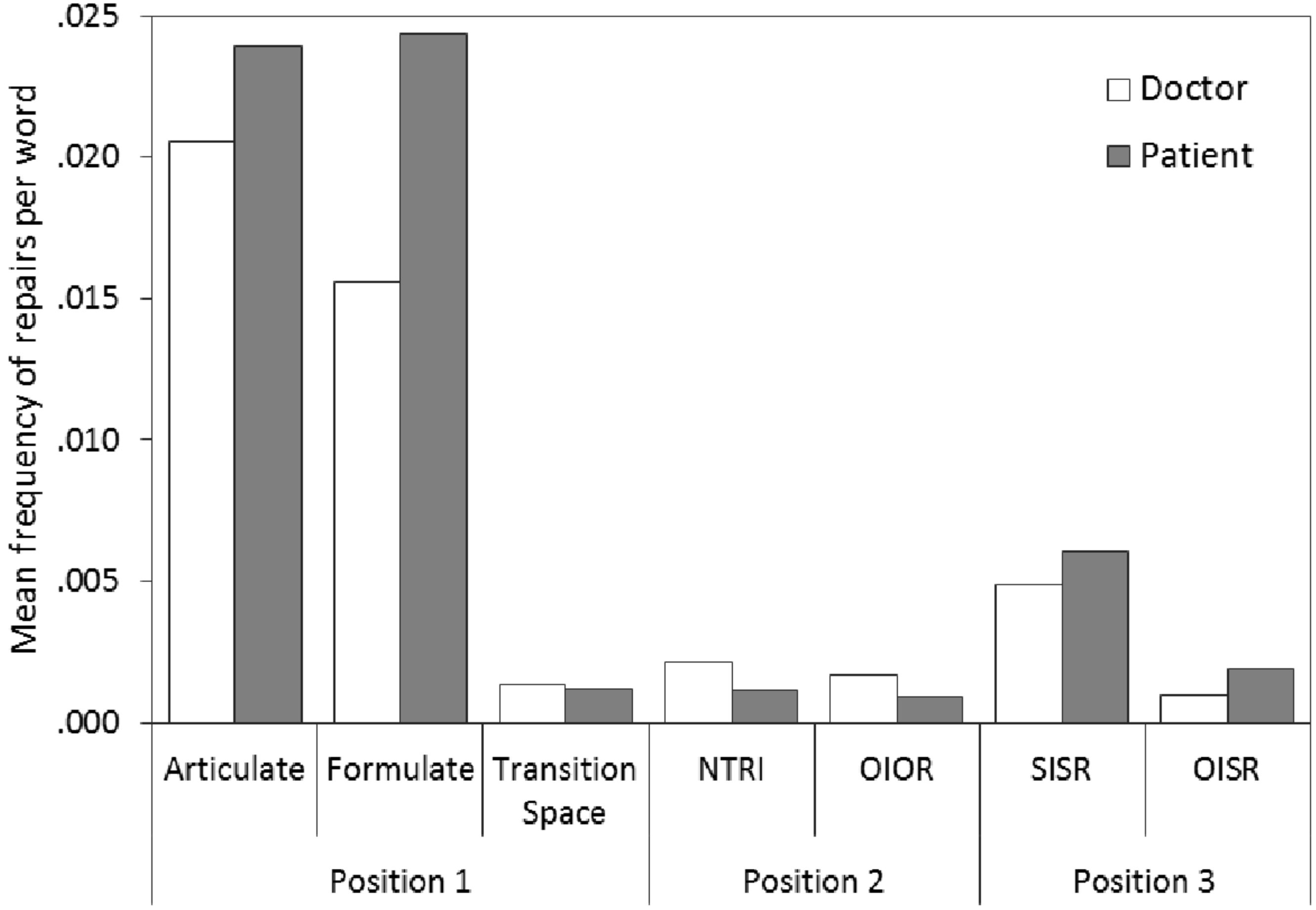
Did you feel that – *did you despair so much that* – you wondered if you could carry on

Dialogue context comparison



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

Patient-doctor comparison



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

Self-repair

- (Hough & Purver, SemDial 2012 - EMNLP 2014)
- "Disfluency detection" for speech recognition

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

  → A flight to Denver

  John likes, uh, loves Mary

  → John loves Mary

- But what about:

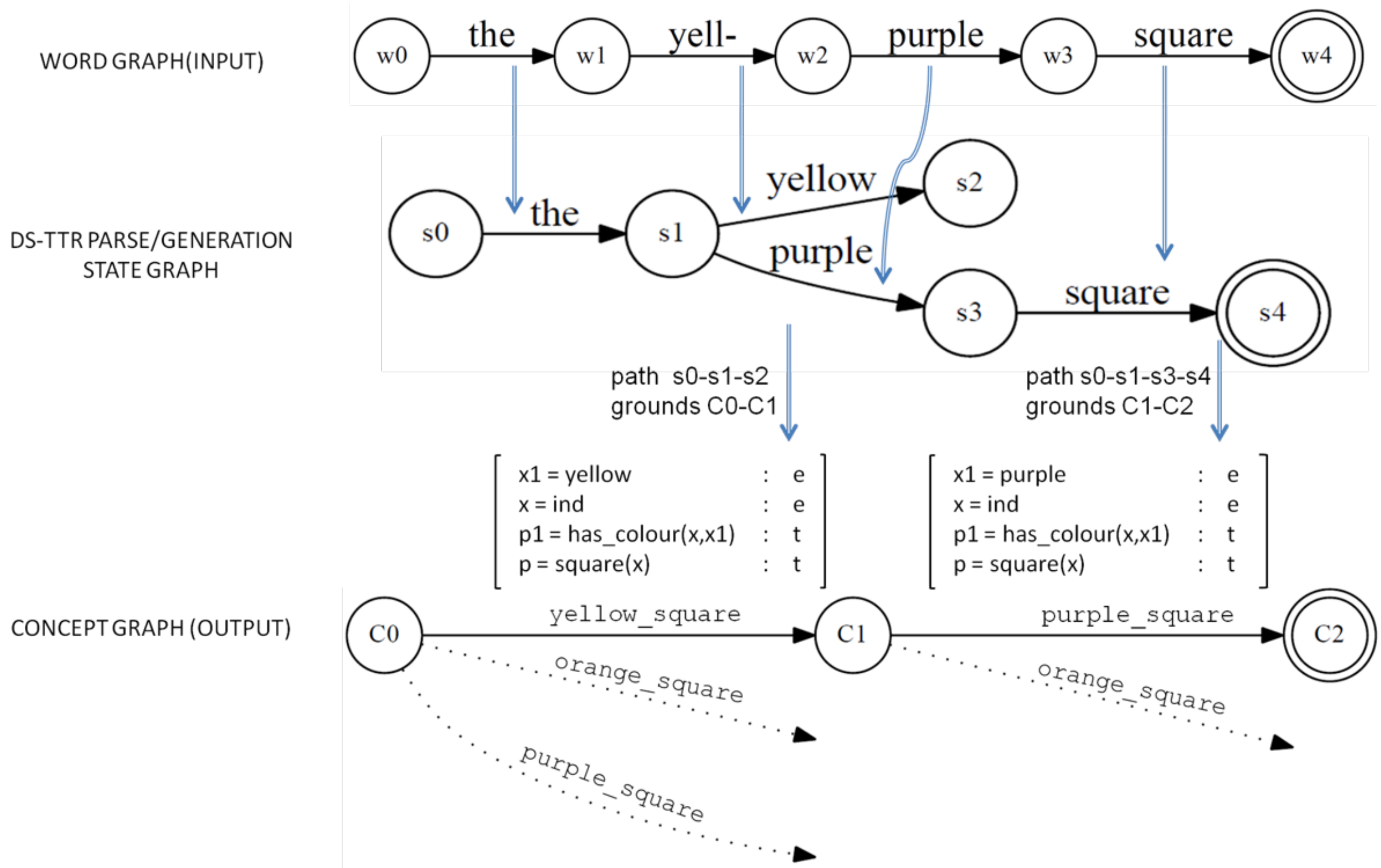
  The interview was – it was alright

  I went swimming with Susan – or rather, surfing

- Incrementality & monotonicity:
  - Maintain semantic context, but with …
  - incremental parsing & choice mechanisms
  - Using domain-general methods

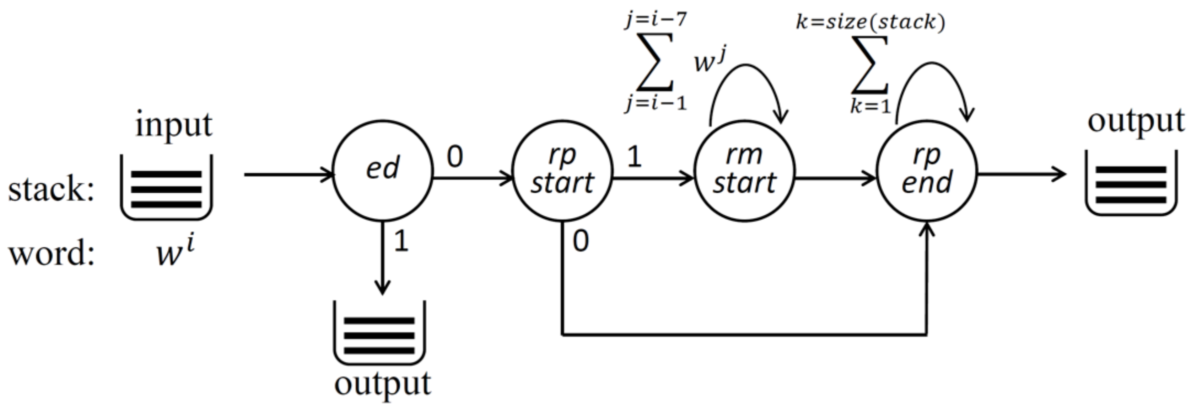
Self-repair

- Incremental, monotonic context model



Self-repair

- Incremental, information-theoretic repair point classifier



- Domain-general features:
  - Similarities between probability distributions
  - Changes in probability & entropy given repair hypotheses
  - Combined in random forest classifier
  - Near state-of-the-art F-score 0.81, with faster incremental performance
  - Transfer to mental health domain: 0.68, per-dialogue correlation 0.95

Other-Repair

- (Howes, McCabe, Purver SIGDIAL 2012)
- Define features manually, extract automatically
  – Linguistically/observationally informed:
    - Wh-question words, closed class repair words
    - Repetition, parallelism
    - Backchannel behaviour, fillers, pauses, overlaps
  – Brute force: all unigrams
- Train SVMs to detect repairs (NTRIs & P2Rs)
  – 44,000 turns, only 567 NTRIs (159 patient), 830 P2Rs (262)
  – 80-86% on balanced data
  – but only 35-44% F-scores (above 20-36% baselines)

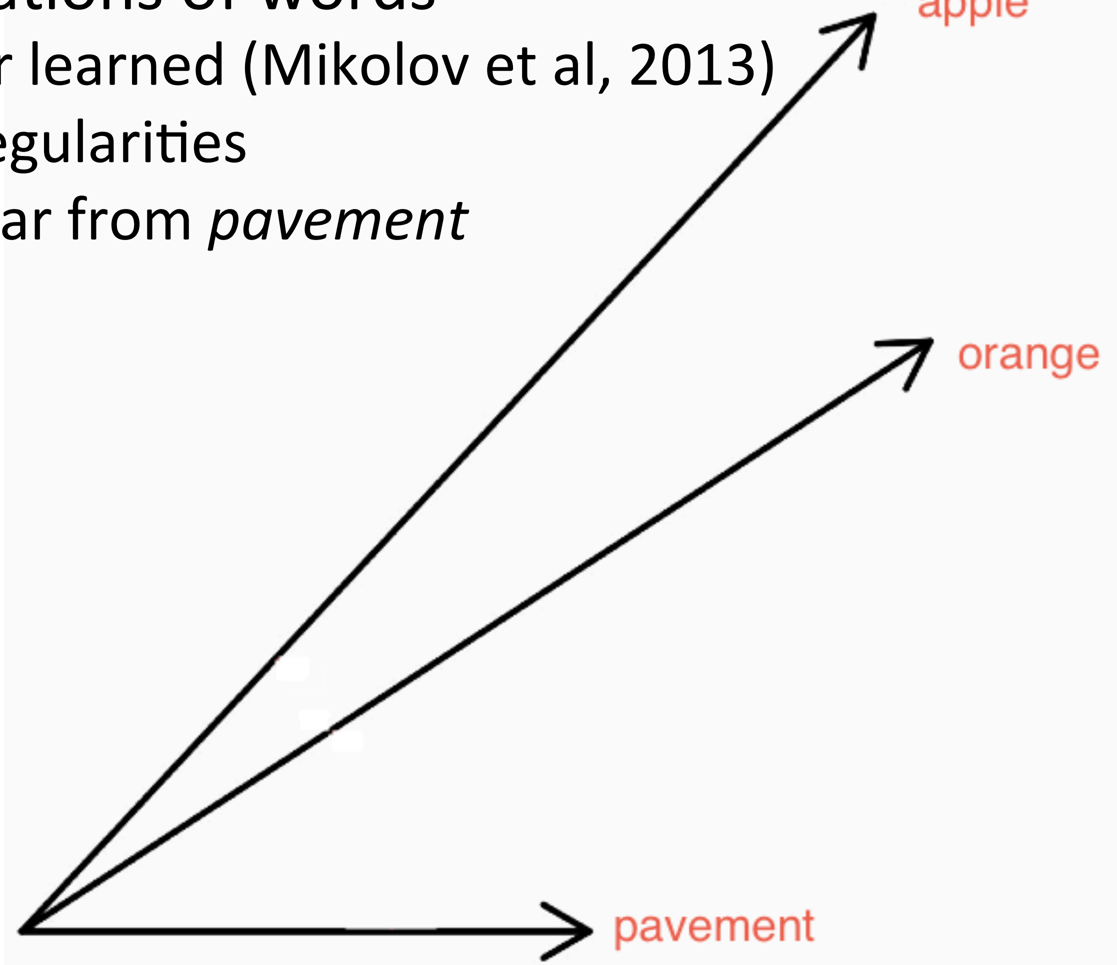
- How can we do better?
  – Repair involves parallelism: not always lexical, but semantic
  – Self-repair model: language model distributions
  – Other-repair: lexical repetition

Distributional Semantics

- Vector space representations of words
  - Co-occurrence-based or learned (Mikolov et al, 2013)
  - Semantic similarity & regularities
  - *apple* close to *orange*, far from *pavement*
  - *(king – queen)*

    ≈ *(man – woman)*

    ≈ (*uncle – aunt)*



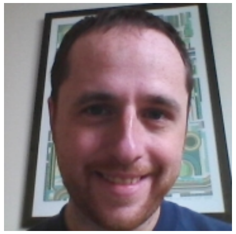
Distributional Semantics

- Standard distributional models help DA tagging …
  - (Milajevs et al, EMNLP 2014)
  - … but not much! (0.60 -> 0.63 accuracy)
- Standard models reflect within-sentence distributions:
  - word2vec (Mikolov et al, 2013) on Google News 100bn wd
  - Closest neighbours of "hello":

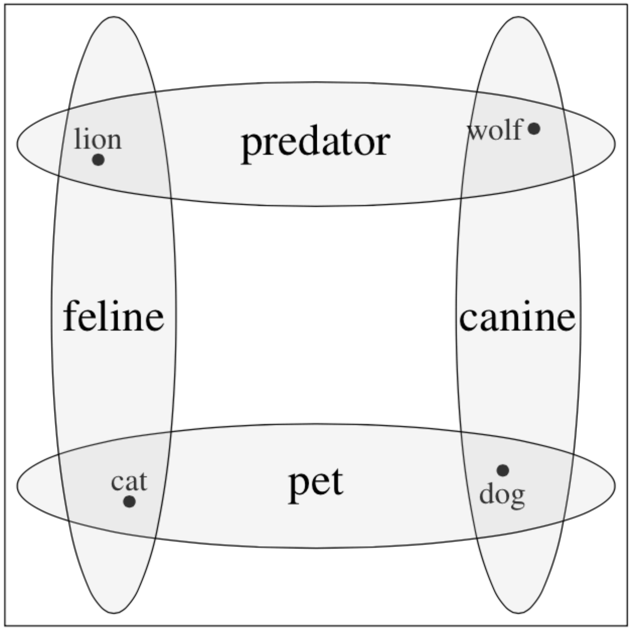
– hi	0.654899
– goodbye	0.639906
– howdy	0.631096
– goodnight	0.592058

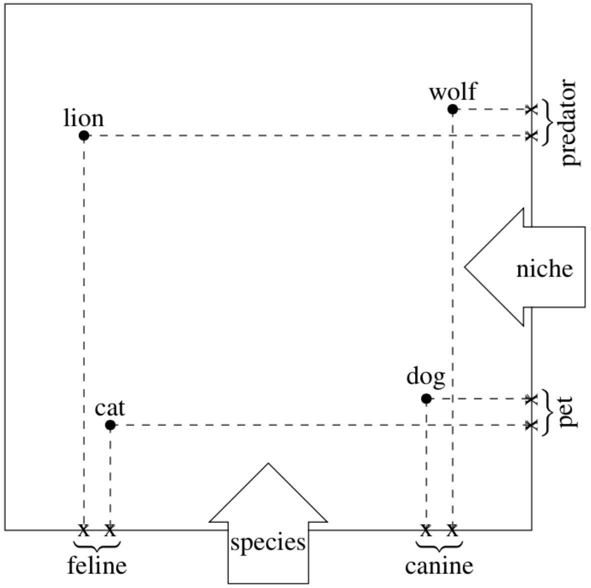
- Training on dialogue data can help:
  - (Kalchbrenner & Blunsom, 2013) RCNNs: 0.74 accuracy
  - But gives a domain/task-specific model

Meaning is Contextual

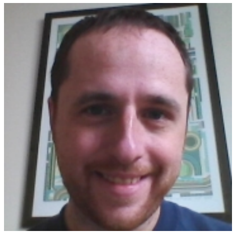


- Perhaps we need to account for **context**
- Distributional semantics & concept formation
  - (Agres, McGregor, Purver, Wiggins ICCC 2014)





Meaning is Contextual



- Perhaps we need to account for **context**
- Distributional semantics & concept formation
  - (Agres, McGregor, Purver, Wiggins ICCC 2014)
- Without context:

```
cat:
    dog mouse bone scratches mysterious mouse
    jazz bites
```

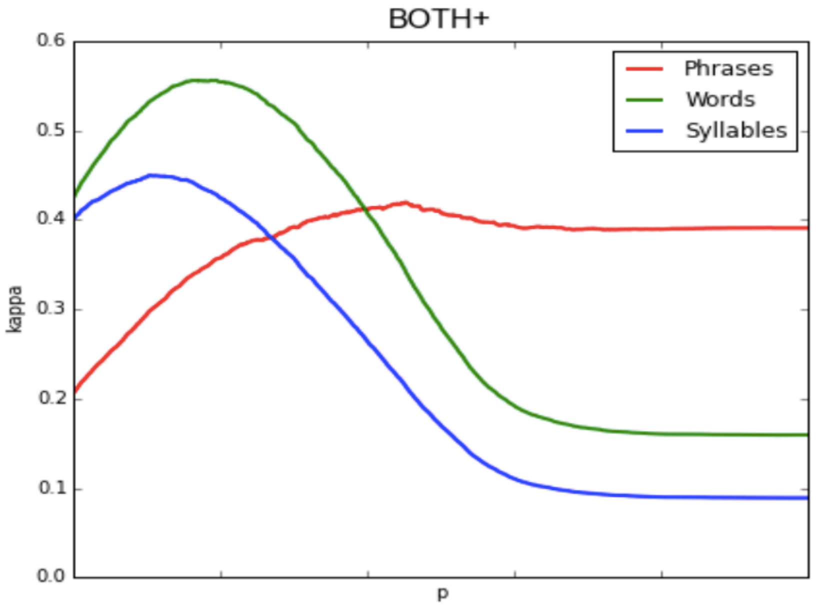
- With context:

```
cat dog lion wolf:
    bobcat raccoon boar hyena cougar opossum
    marten giraffe
```

What are the right units?

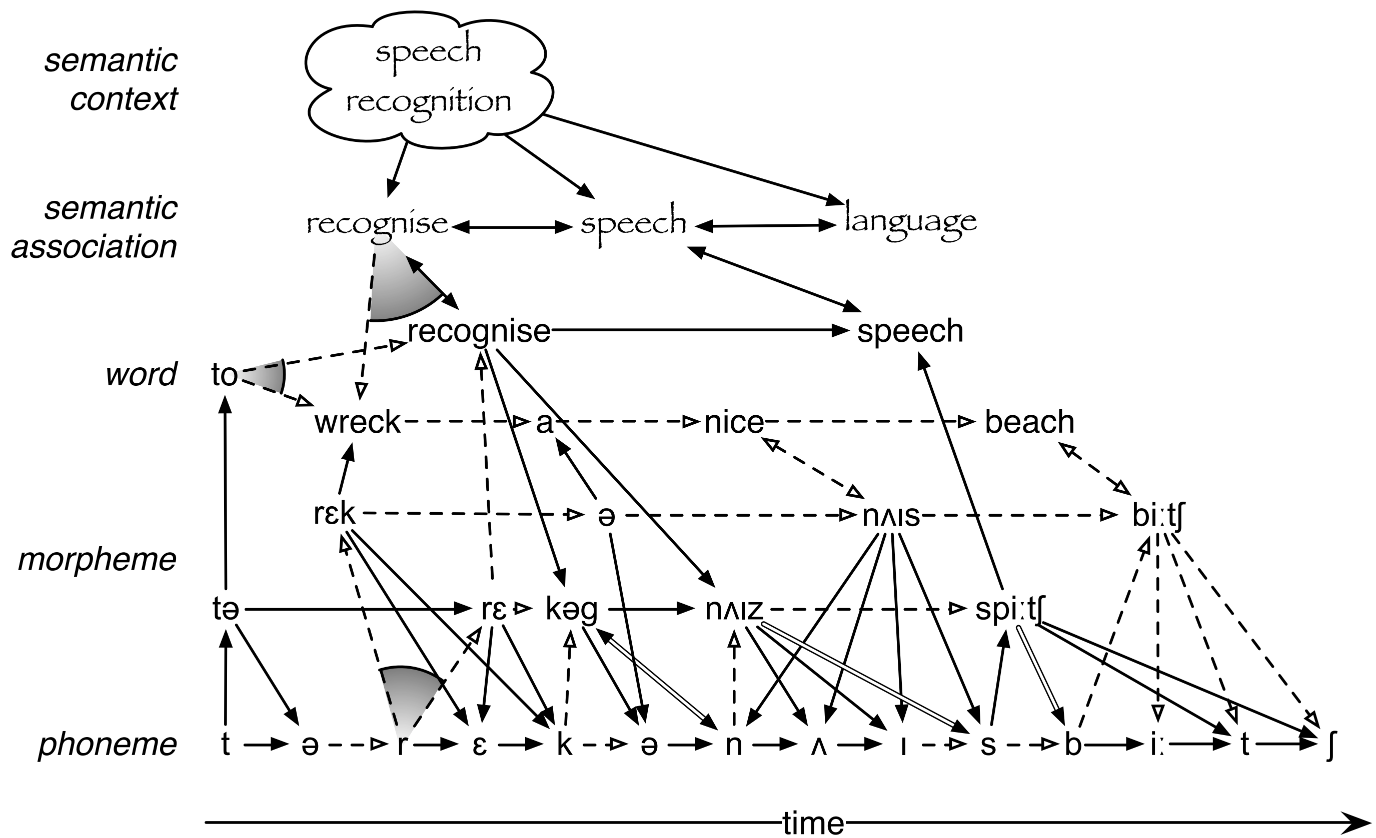
- Perhaps we need to learn from **contextual distributions**
- Which means we need to know the units of interest
  - (cf. Nishida "conversation quanta"?)

- Unsupervised, information-theoretic induction
  - (Griffiths et al, in prep)
  - Segment on changes:
    - information content
    - entropy
  - At different levels:
    - syllables 0.67 F1
    - words 0.71 F1



What are the right units?

- Scaling up to a hierarchical model



Thanks!

- To you and:
  - Shauna Concannon
  - Rose McCabe
  - Julian Hough
  - Arash Eshghi
  - Christine Howes
  - Dmitrijs Milajevs
  - Mehrnoosh Sadrzadeh
  - Dimitri Kartsaklis
  - Zheng Yuan
  - Pat Healey
  - Ruth Kempson
  - Kat Agres
  - Jamie Forth
  - Stephen McGregor
  - Geraint Wiggins
  - Sascha Griffiths