Predicting Adherence to Treatment for Schizophrenia from Dialogue Transcripts

Christine Howes, Matthew Purver, Rose McCabe, Patrick G. T. Healey, Mary Lavelle
Queen Mary University of London
Mile End Road, London E1 4NS
c.howes@qmul.ac.uk

Abstract

Recent work on consultations between outpatients with schizophrenia and psychiatrists has shown that adherence to treatment can be predicted by patterns of repair – specifically, the pro-activity of the patient in checking their understanding, i.e. patient clarification. Using machine learning techniques, we investigate whether this tendency can be predicted from high-level dialogue features, such as backchannels, overlap and each participant's proportion of talk. The results indicate that these features are not predictive of a patient's adherence to treatment or satisfaction with the communication, although they do have some association with symptoms. However, all these can be predicted if we allow features at the word level. These preliminary experiments indicate that patient adherence is predictable from dialogue transcripts, but further work is necessary to develop a meaningful, general and reliable feature set.

1 Introduction

How conversational partners achieve and maintain shared understanding is of crucial importance in the understanding of dialogue. One such mechanism, other initiated repair (Schegloff, 1992), where one conversational participant queries or corrects the talk of another, has been well documented in both general and task-based dialogues (Colman and Healey, 2011). However, how such shared understanding impacts beyond the level of the conversation has not typically been examined. Exceptions to this have highlighted the role of shared understanding in schizophrenia (McCabe et al., 2002; Themistocleous et al., 2009) and the association between psychiatrist-patient communication and adherence. McCabe et al. (in preparation) found that more patient clarification (i.e. other initiated repair) of the psychiatrist's talk was associated with better treatment adherence six months later. Clarification consists mainly of asking questions to clarify the meaning of the psychiatrist's utterance (checking understanding) and correcting something that the psychiatrist has said (getting the facts straight). Example 1, taken from a consultation, shows the patient requesting clarification of something the psychiatrist has just said about a possible side effect.

(1) **Dr:** Yep, well that is a possible side effect
Pat: Side effect?
Dr: Of the er haloperidol

The patient's request leads to additional explanation by the psychiatrist about the medication which can cause the possible side effect. More patient clarification reflects greater effort to reach a shared understanding. McCabe et al. (in preparation) found that for each unit increase in the patient clarification factor,[1] the odds of good (versus poor) adherence were increased by 5.8 (95% CI 1.3 to 25.8, p=0.02).

Explaining the link between communicative patterns of patients and adherence may create the possibility for new interventions to improve adherence, and has both clinical and theoretical implications.

[1] A regression factor weighted heavily towards patient clarifcations (as in e.g. 1).

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 79–83,
Seoul, South Korea, 5-6 July 2012.

However, there is no evidence regarding what factors influence patient clarification and may explain the link with adherence. If patient clarification is a measure of greater communicational effort, or engagement, then we might expect other dialogue measures, such as the amount of acknowledgements or other grounding cues (Traum and Allen, 1992), or the proportion of talk per person, to be correlated with other initiated repair and therefore similarly predictive of subsequent adherence behaviour. This is of particular importance if we wish to build a system to automatically predict possible (lack of) adherence from dialogue transcripts, especially given that the types of patient clarification which carry the highest weight in the patient clarification factor (next-turn repair initiators, Schegloff, 1992) are rare, occurring on average only 1.2 times per dialogue.

Further, although certain types of repair were shown to affect how patients reported they felt the conversation went, self-reports of symptoms and communicational factors are not predictive of adherence. Although micro-communicational behaviour (in the form of other initiated repair) does have a bearing on subsequent adherence behaviour, patients are unaware of this. Additional questions therefore concern whether we can predict patient's symptom levels and subjective analyses of the communication based only on overview dialogue factors.

2 Hypotheses

Factors which we would expect to index patient engagement, and thus be predictive of adherence to treatment are the amount of backchannel responses patients make, and the proportion of questions patients ask, both of which ought to be higher for the more engaged patients. We might also expect that such patients have a greater proportion of the talk overall, and/or longer turns on average, though note that this conversational pattern might also be one in which the patient is not engaged, as they might not be responding to the feedback from their consultant.

For the symptom scores (see below for details), we should expect that patients with high levels of negative symptoms (which includes loss of affect and poverty of speech) would produce less talk overall, and in general produce shorter turns. There should also be more noticeable gaps in the dialogues (defined as greater than approximately 200ms, (Heldner and Edlund, 2010)). Contrarily, for positive symptoms, (including hallucinations and delusions) patients ought to produce longer turns and have a greater proportion of the talk.

We also expect to see effects on how patients felt the conversation went from the amount of overlap, though as overlap can be both intended and interpreted as either interruptive or collaborative (as with e.g. overlapping backchannels) it is unclear which direction such a prediction should take.

3 Method

131 dialogues from outpatient consultations between patients and psychiatrists were analysed according to a number of factors. Each of these factors, detailed in table 1, below, is calculated for each dialogue participant (with the exception of pauses). Each patient featured in only one of the dialogues however, there were only 29 doctors in the study, so the same clinician may have featured in several of the dialogues with different patients. The consultations varied in length, with the shortest consisting of 61 turns (438 words) and the longest 881 turns (13178 words), with an average of 320.5 turns (2706.4 words). In addition, a third party was present in 47 of the consultations.

Following the consultation, each patient was asked questions from standard questionnaires to ascertain their level of symptoms, and their evaluation of aspects of the consultation. The positive and negative syndrome scale (PANSS) (Kay et al., 1987) assesses positive, negative and general symptoms on a 7-point scale of severity (1=absent – 7=extreme). Positive symptoms represent a change in the patients' behaviour or thoughts and include sensory hallucinations and delusional beliefs. Negative symptoms represent a withdrawal or reduction in functioning, including blunted affect, and emotional withdrawal and alogia (poverty of speech). Positive and negative subscale scores ranged from 7 (absent) – 49 (extreme), general symptoms (such as anxiety) scores ranged from 16 (absent) – 112 (extreme).

Patient satisfaction with the communication was assessed using the Patient Experience Questionnaire (PEQ) (Steine et al., 2001). Three of the five subscales (12 questions) were used as the others were

not relevant, having been developed for primary care. The three subscales were 'communication experiences', 'communication barriers' and 'emotions immediately after the visit'. For the communication subscales, items were measured on a 5-point Likert scale, with 1=disagree completely and 5=agree completely. The four items for the emotion scale were measured on a 7-point visual analogue scale, with opposing emotions were at either end. A higher score indicates a better experience.

Adherence to treatment was rated by the clinicians as good ($> 75\%$), average ($25 - 75\%$) or poor ($< 25\%$) six months after the consultation. Due to the low incidence of poor ratings (only 8 dialogues), this was converted to a binary score of 1 for good adherence (91 patients), and 0 otherwise (37). Ratings were not available for the remaining dialogues.

Measure	Description
Turns	Total number of turns
Words	Total number of words spoken
Proportion	Proportion of total talk in words (by each participant)
WordsPerTurn	Average length of turn in words
WhPerWord	Proportion of wh-words (e.g. what? who?) per word
OCRPerWord	Proportion of open class repair initiators (e.g. pardon? huh?) per word
BackchannelPerWord	Proportion of backchannels (e.g. uh-huh, yeah) per word
RepeatPerWord	Proportion of words repeated from preceding turn by other person
OverlapAny	Proportion of turns containing any overlapping talk
OverlapAll	Proportion of turns entirely overlapping another turn
QMark	Proportion of turns containing a question mark
TimedPause	Pause of more than approx 200ms, as marked on the transcripts

Table 1: Measures from outpatient consultations

3.1 Classification Experiments

We performed a series of classification experiments using the Weka machine learning toolkit (Hall et al., 2009) to predict each of the outcome measures outlined above (symptom measures, satisfaction measures, and adherence to treatment). In each case, outcome measures were converted to binary high/low scores on an equal frequency basis (i.e. providing approximately equal numbers of high and low instances). Features used were the high-level measures given in Table 1, and/or all unigrams extracted from the transcript; in both cases, features from doctor and patient were treated separately. Unigrams were produced by tokenising the lower-cased transcripts on white space; no stemming or stopword removal was performed, and feature values were binary i.e. indicating only presence or absence of the word spoken by the given speaker in the given dialogue.[2] Given the small size of our dataset (131 instances) and the large feature space resulting (> 6500 features), we selected features based on their predictive ability across the entire dataset (using Weka's CfsSubsetEval selector), reducing the number of features to 50-100. In order to avoid biasing towards doctor-specific features, we used only words spoken by patients in these experiments – each patient only features in one dialogue, so patient-specific vocabulary cannot help performance across dialogues. All unigram features thus selected were used in at least 3 dialogues.[3]

4 Results

Experiments including unigram features used LibSVM's support vector machine implementation (Chang and Lin, 2001) with a radial basis function kernel; experiments with only high-level features used J48 decision trees. In each case, experiments used 5-fold cross-validation.[4] In experiments predicting adherence, the distribution between positive and negative (i.e. good and bad adherence) made it impossible to balance the dataset - as this can be problematic for decision tree classifiers, we also present results for a downsampled dataset with only 71 instances but which provides balance. Performance is shown in Table 2 as overall percentage accuracy, and is compared to a majority-class baseline throughout; results which are significantly different at the 5% level according to a χ^2 test from a

[2]Experiments with frequency counts did not affect the results as reported.

[3]Bi- and tri-gram features were not extracted from this data because of the small amount of data available which we felt would result in models that suffered from overfitting (note that the same concern holds for the unigram features).

[4]Classifiers were trained on 80% and tested on 20% of the sample, with this was repeated 5 times over each possible 80/20 combination so as to test the whole dataset.

random distribution and the majority class distribution are shown marked with *.

	Baseline	Words	High-level
PANSS *positive*	51.1	87.0*	56.5*
PANSS *negative*	49.6	87.8*	56.5*
PANSS *general*	48.4	91.1*	54.0
PEQ *emotions*	51.9	89.1*	53.5
PEQ *communication*	50.8	79.8*	52.4
PEQ *comm. barriers*	51.6	90.6*	51.6
PEQ *overall*	50.8	90.6*	53.9
Adherence	73.2	91.1*	63.4
Adherence (balanced)	53.5	93.0*	52.1

Table 2: Percentage accuracies vs feature set

Results show good performance for all experiments when including lexical features, with all factors being predictable with around 90% accuracy with the exception of PEQ communication at just below 80%. However, using high-level features alone gives negligible performance, except for a small benefit on the PANSS negative and positive symptom measures, though contrary to our hypotheses the most important high-level features were OCR-PerWord by the doctor (negative) and WhWords by an other participant (positive).

Examination of the most predictive unigrams shows that sets selected for different outcome measures are different: for example, the 54 features selected for adherence and the 73 selected for PEQ overall have only 1 word in common (*"mates"*). Adherence-related words include words related to conditions, treatment and medication (*"schizophrenic"*, *"sickness"*, *"symptoms"*, *"worse"*, *"pains"*, *"flashbacks"*, *"sodium"*, *"chemical"*, *"monthly"*); PEQ-related words include those related to personal life (*"sundays"*, *"thursdays"*, *"television"*, *"sofa"*, *"wine"*, *"personally"*, *"played"*), and filled pauses (*"eerrmm"*, *"uhhm"*) – although more investigation is required to draw any firm conclusions from these. Table 3 shows the full lists for adherence and PEQ overall.

5 Discussion and Conclusions

The results show that although we can weakly predict symptoms at levels above chance using only high-level dialogue factors, we cannot do so for adherence, or satisfaction measures. Despite the link between patient other initiated repair and adherence, this is also not an effective predictor for our machine learning approach because of the scarcity of the phenomenon, and the fact that many of the consultations for which the patients subsequently exhibited good adherence behaviour do not feature a single patient clarification, which may be linked to psychiatrist clarity rather than lack of effort or engagement on the patient's part.

The high accuracies with lexical features show that some aspects of the consultations do enable accurate prediction of adherence, PEQ measures and symptoms. However, as the features which allow us to achieve such good results rely on specific words used, it is unclear how generalisable or interpretable such results are. The lexical features chosen do generalise over our dataset (in which individual patients appear only once), and exclude doctor talk, so cannot be simply picking out unique unigram signatures relating to individual patients or doctors; however, given the small size of the dataset used for this initial investigation with its constrained domain, genre and topics, and the use of the whole dataset to select predictive words, it is unclear whether these results will scale up to a larger dataset.

We therefore suspect that more general, higher-level dialogue features such as specific interaction phenomena (repair, question-answering) and/or more general models of topic may be required. While unigrams are too low-level to be explanatory and may not generalise, the dialogue features discussed are too high-level to be useful; we are therefore examining mid-level phenomena and models to capture the predictability while remaining general and providing more interpretable features and results. Although the word lists offer clues as to the relevance of specific words for the overall predictability, we would not like to leave it at that. Further experiments are therefore underway to investigate whether we can find a level of appropriate explanatory power and maximal predictivity using an interim level of analysis, for example with n-gram and part-of-speech-based models, topic models based on word distributions, and turn-taking phenomena. Additional experiments also look at the turn-level data to see if the patient led clarification factor can be directly extracted from the transcripts.

Adherence			PEQ overall			
air	grass	schizophrenic	20th	electric	onto	sometime
anyone	grave	sensation	ages	energy	overweight	son
balanced	guitar	sickness	angry	environment	oxygen	standing
bleach	h	simply	anxiety	experiencing	packed	stomach
build	hahaha	sodium	background	facilities	percent	suddenly
building	lager	stable	bladder	friendly	personally	sundays
busy	laying	stock	booked	helps	picture	suppose
challenge	lifting	symptoms	boy	ignore	played	table
chemical	lucky	talks	broken	immediately	programs	team
complaining	mates	teach	bus	increased	progress	television
cup	monthly	terminology	certificate	irritated	provide	thursdays
dates	mouse	throat	dead	kick	public	troubles
en	nowhere	virtually	deep	later	quid	uhhm
fill	pains	was	drunk	lee	radio	upsetting
finished	possibly	wave	earn	loose	realised	walks
fish	pr	weve	eeerrrr	low	reply	watchers
flashbacks	recent	worse	eerrmm	march	sat	wine
	removed	writing	eerrrmm	mates	shaky	
	ri			moments	sofa	

Table 3: Most predictive unigram features

References

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGDKDD Explorations*, 11(1):10–18.

M. Heldner and J. Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

S.R. Kay, A. Fiszbein, and L.A. Opfer. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin*, 13(2):261.

R. McCabe, C. Heath, T. Burns, S. Priebe, and J. Skelton. 2002. Engagement of patients with psychosis in the consultation: conversation analytic study. *British Medical Journal*, 325(7373):1148–1151.

R. McCabe, M. Lavelle, S. Bremner, D. Dodwell, P. G. T. Healey, R. Laugharne, S. Priebe, and A. Snell. in preparation. Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia.

E.A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, pages 1295–1345.

S. Steine, A. Finset, and E. Laerum. 2001. A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome. *Family practice*, 18(4):410–418.

M. Themistocleous, R. McCabe, N. Rees, I. Hassan, P. G. T. Healey, and S. Priebe. 2009. Establishing mutual understanding in interaction: An analysis of conversational repair in psychiatric consultations. *Communication & Medicine*, 6(2):165–176.

D.R. Traum and J.F. Allen. 1992. A speech acts approach to grounding in conversation. In *Second International Conference on Spoken Language Processing*.