Investigating Topic Modelling for Therapy Dialogue Analysis

Christine Howes, Matthew Purver and Rose McCabe
Queen Mary University of London
c.howes@qmul.ac.uk

Abstract

Previous research shows that aspects of doctor-patient communication in therapy can predict patient symptoms, satisfaction and future adherence to treatment (a significant problem with conditions such as schizophrenia). However, automatic prediction has so far shown success only when based on low-level lexical features, and it is unclear how well these can generalise to new data, or whether their effectiveness is due to their capturing aspects of style, structure or content. Here, we examine the use of *topic* as a higher-level measure of content, more likely to generalise and to have more explanatory power. Investigations show that while topics predict some important factors such as patient satisfaction and ratings of therapy quality, they lack the full predictive power of lower-level features. For some factors, unsupervised methods produce models comparable to manual annotation.

1 Introduction and Background

1.1 Therapy communication and outcomes

Aspects of doctor-patient communication have been shown to be associated with patient outcomes, in particular patient satisfaction, treatment adherence and health status (Ong et al., 1995). For patients with schizophrenia, non-adherence to treatment is a significant problem, with non-adherent patients having an average risk of relapse that is 3.7 times higher than adherent patients (Fenton et al., 1997). Some recent work suggests that a critical factor is conversation structure – *how* the communication proceeds. In consultations between out-patients with schizophrenia and their psychiatrists, McCabe et al. (tion) showed that patients who used more *other repair* — i.e. clarified what the doctor was saying — were more likely to adhere to their treatment six months later. However, outcomes are also affected by the content of the conversation – *what* is talked about. Using conversation analytic techniques, McCabe et al. (2002) show that doctors and patients have different agendas, made manifest in the topics that they talk about; on the same data, with topics annotated by hand, Hermann et al. (in prep) showed that patients attempt to talk about psychotic symptoms, but doctors focus more on medication issues. Importantly, more talk about medication from the patient increases the patient's chances of relapse in the six months following the consultation (Hermann et al., in prep).

1.2 Automatic prediction

Using machine learning techniques, Howes et al. (2012a,b) investigated whether outcomes such as adherence, evaluations of the consultation and symptoms can be predicted from therapy transcripts using features which can be extracted automatically. Their findings indicate that high-level features of the dialogue *structure* (backchannels, overlap etc) do not predict these outcomes to any degree of accuracy. However, by using all words spoken by patients as unigram lexical features, and selecting a subset based on correlation with outcomes over the training set, they were able to predict outcomes to reasonable degrees of accuracy (c. 70% for future adherence to treatment – see Howes et al., 2012a, for details).

These studies show that some aspects of therapy consultations which can be extracted automatically (thus removing the need for expert annotation) can enable accurate prediction of outcomes. However, as the successful features encode specific words spoken by the patient, it is unclear whether they relate to dialogue structure or content, or some combination of the two, and thus help little in explaining the results or providing feedback to help improve therapy effectiveness. It is also unclear

how generalisable such results are to larger datasets or different settings, given such specific features with a small dataset. More general models or features may therefore be required.

In this paper, we examine the role and extraction of *topic*. Topic provides a measure of content more general than lexical word features; by examining its predictive power, we hope to provide generalisable models while also shedding more light on the role of content vs structure. As content is known to be predictive of outcomes to some extent, identification and tracking of topics covered can provide useful information for clinicians, enabling them to better direct their discussions in time restricted consultations, and aid the identification of patients who may subsequently be at risk of relapse or non-adherence to treatment. However, annotating for topic by hand is a time-consuming and subjective process (topics must first be agreed on by researchers, and annotators subsequently trained on this annotation scheme); we therefore examine the use of automatic topic modelling.

1.3 Topic modelling

Probabilistic topic modelling using Latent Dirichlet Allocation (LDA; Blei et al., 2003) has been previously used to extract topics from large corpora of texts, e.g. web documents and scientific articles. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings (see e.g. Steyvers and Griffiths, 2007). LDA uses unsupervised learning methods, and learns the topic distributions from the data itself, by iteratively adjusting priors (see Blei, 2012, for an outline of the algorithms used in LDA). Such techniques have been applied to structured dialogue, such as meetings (Purver et al., 2006) and tutoring dialogues (Arguello and Rosé, 2006) with encouraging results.

In the clinical domain, probabilistic topic modelling has been applied to patients' notes to discover relevant clinical concepts and connections between patients (Arnold et al., 2010). In terms of clinical *dialogue*, there are few studies which apply unsupervised methods to learning topic models, though recently this has become an active field of exploration. Angus et al. (2012) apply unsupervised methods to primary care clinical dialogues, to visualise shared content in communication in this domain. However, their data relies on only six dialogues, with the three training dialogues being produced in a role play situation. It is unclear whether using constructed dialogues as the baseline measure maps reliably to genuine dialogues. Additonally, though they did find differences in the patterns of communication based on how the patient had rated the encounter, their task was a descriptive one, not a predictive one and it is unclear if or how their methodology would scale up, especially given that they selected their testing dialogues on the basis of the patient evaluations.

Cretchley et al. (2010) applied unsupervised techniques to dialogues between patients with schizophrenia and their carers (either professional carers or family members). Patients and carers were instructed to talk informally and given a list of general interest topics such as sport and entertainment. They split their sample into two pre-defined communication styles ("low- or high- activity communicators") and described differences in the most common words spoken by each type depending on both the type of carer and the type of communicator. Once again, however, this was a descriptive exercise, on a very small number of dyads, and in choosing to predefine the participants by the amount of communicative activity they undertook they may have missed ways to differentiate between groups of patients that can be extracted from the data, rather than being pre-theoretic.

1.4 Research questions

The preliminary studies outlined above demonstrate some of the issues arising from using unsupervised topic modelling techniques to look at clinical dialogues. One of the main issues is in the interpretation of results. Studies described above used visualisations of the data to find patterns; one question that therefore arises is whether we can usefully interpret "topics" without these – for example, just by examining the most common words in a topic. Another question concerns the limited evidence that different styles of communication can be demonstrated using unsupervised topic modelling, and that these differences have a bearing on, for example, the patient's evaluations of the communication or their symptoms. Our main questions here are therefore:

- Does identification of topic allow prediction of symptoms and/or therapy outcomes?
- If so, can automatic topic modelling be used instead of manual annotation?
- Does automatic modelling produce topics that are interpretable and/or comparable to human judgements?

2 Data

This study used data from a larger study investigating clinical encounters in psychosis (McCabe et al., tion), collected between March 2006 and January 2008. 31 psychiatrists agreed to participate. Patients meeting Diagnostic and Statistical Manual-IV (APA) criteria for a diagnosis of schizophrenia or schizoaffective disorder attending psychiatric outpatient and assertive outreach clinics in three centres (one urban, one semi-urban and one rural) were asked to participate in the study. After complete description of the study to the subjects, written informed consent was obtained from 138 (40%) of those approached. Psychiatrist-patient consultations were then audio-visually recorded using digital video. The dialogues were transcribed, and these transcriptions, consisting only of the words spoken, form our dataset here. The consultations ranged in length, with the shortest consisting of only 617 words (lasting approximately 5 minutes), and the longest 13816 (lasting nearly an hour). The mean length of consultation was 3751 words.

2.1 Outcomes

Patients were interviewed at baseline, immediately after the consultation, by researchers not involved in the patients' care, to assess their symptoms. Both patients and psychiatrists filled in questionnaires evaluating their experience of the consultation at baseline, and psychiatrists were asked to assess each patient's adherence to treatment in a follow-up interview six months after the consultation. The measures obtained are described in more detail below.

2.1.1 Symptoms

Independent researchers assessed patients' symptoms at baseline on the 30-item Positive and Negative Syndrome Scale (PANSS, Kay et al., 1987). The scale assesses positive, negative and general symptoms and is rated on a scale of 1-7 (with higher scores indicating more severe symptoms). 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). Inter-rater reliability using videotaped interviews for PANSS was good (Cohen's kappa=0.75).

2.1.2 Patient satisfaction

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 at either end. A higher score indicates a better experience.

2.1.3 Therapeutic relationship

The Helping Alliance Scale (HAS, Priebe and Gruyters, 1993) was used after the consultation to assess both patients' and doctors' experience of the therapeutic relationship. The HAS has 5 items in the clinician version and 6 items in the patient version, with questions rated on a scale of 1-10. Items cover aspects of interpersonal relationships between patients and clinician and aspects of their judgment as to the degree of common understanding and the capability to provide or receive the necessary help, respectively. The scores from the individual items were averaged to provide a single value, with lower scores indicating a worse therapeutic relationship.

2.1.4 Adherence to treatment

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 dia-

logues), this was converted to a binary score of 1 for good adherence (89 patients), and 0 otherwise (37). Ratings were not available for the remaining 12 dialogues.

2.2 Hand-coded topics

Hermann et al. (in prep) annotated all 138 consultations for topics. First, an initial list of categories was developed by watching a subset of the consultations. The dialogues were then manually segmented and topics assigned to each segment, with the list of topic categories amended iteratively to ensure best fit and coverage of all relevant topics. A subset of 12 consultations was coded independently by two annotators, such that every utterance (and hence every word) was assigned to a single topic; inter-rater reliability was found to be good using Cohen's kappa ($\kappa = 0.71$). The final list of topics used, with descriptions, is outlined in Table 1.

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, daily structure etc
04	Living situation	The life situation of the patient, including housing, finances, benefits, plans with life etc
05	Psychotic symptoms	Discussion on symptoms of psychosis such as hallucinations and delusional beliefs
06	Physical health	Any discussion on general physical health, physical illnesses, operations, etc
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions, compulsions, phobias etc
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 effects
10	Past illness	Discussion of past history of psychiatric illnesses, including previous admissions and relapses
11	Mental health services	Care coordinator, community psychiatric nurse, social worker or home treatment team etc
12	Other services	Primary care services, social services, DVLA, employment agencies, police, housing etc
13	General chat	Includes introductions; general topics; weather; holidays; end of appointment courtesies
14	Explanation about illness	Patients diagnosis, including doctor explanations and patients questions about their illness
15	Coping strategies	Discussions around coping strategies that the patient is using or the doctor is advising
16	Relapse indicators	Relapse indicators and relapse prevention, including early warning signs
17	Treatment	General and psychological treatments, advice on managing anxiety, building confidence etc
18	Healthy lifestyle	Any advice on healthy lifestyle such as dietary advice, exercise, sleep hygiene etc
19	Relationships	Family members, friends, girlfriends, neighbours, colleagues and relationships etc
20	Other	Anything else. Includes e.g. humour, positive comments and non-specific complaints

Table 1: Hand-coded topic names and descriptions

3 Topic Modelling

The transcripts from the same 138 consultations were analysed using an unsupervised probabilistic topic model. The model was generated using the MAchine Learning for LanguagE Toolkit (MALLET, McCallum, 2002), using standard Latent Dirichlet Allocation (Blei et al., 2003) with the notion of *document* corresponding to the transcribed sequence of words spoken (by any speaker) in one consultation. As is conventional (see e.g. Salton and McGill, 1986), stop words (common words which do not contribute to the content of the talk, such as 'the' and 'to') were removed. The number of topics was specified as 20 to match the number of topics used by the human annotators (see above),[1] and the default setting of 1000 Gibbs sampling iterations was used. As an uneven distribution of topics was observed in the hand-coded topic data (see below), automatic hyperparameter optimisation was enabled to allow the prominence of topics and the skewedness of their associated word distributions to vary to best fit the data.

3.1 Interpretation

The resulting topics (probability distributions over words) were then assessed by experts for their interpretability in the context of consultations between psychiatrists and out-patients with schizophrenia. The top 20 most probable words in each topic were presented to two groups independently — one group of experts in the area of psychiatric research (of whom some members were also involved

[1]This is, of course, an arbitrary decision, and future work should investigate different numbers of topics.

in developing the hand-coded topics), and one group of experts in the area of communication and dialogue (without specific expertise in the context of psychiatry) — and each group produced text descriptions of the topics they felt they corresponded to. The two groups' interpretations strongly agreed in 13 of the 20 topic assignments (65%) and partially agreed (i.e. there was some overlap in the interpretations) in a further 3 topic assignments (i.e. in total, 80%).

Having assigned a tentative interpretation to the top word lists for each topic, the two groups reconvened to examine the occurrences of the topics in the raw transcripts, in order to validate these interpretations within the context of the discussion. Excerpts from the dialogues were chosen on the basis of the proportion of words assigned to each topic in the final iteration of the LDA sampling algorithm. Four excerpts were examined for each of the 20 topics, and a final interpretation for each was agreed.

The ease of giving the topic lists of most common words a coherent "interpretation" varied greatly. Some topics were easily given compact descriptions, for example topics 6, 12 and 18, whilst other word lists appeared more disparate. The list of topics and interpretations can be seen in Table 2.

	Interpretation	Example words from top 20
0	Sectioning/crisis	hospital, police, locked
1	Physical health - side-effects of medication and other medical issues	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 beyond the illness setting (e.g. work, study)	progress, work, friends
5	Making sense of psychosis	god, talking, reason
6	Sleep patterns	sleep, bed, night
7	Social stressors - other people who are stressors or helpful under stress	home, thought, told
8	Physical symptoms - e.g. pain, hyperventilating	breathing, breathe, burning
9	Physical tests - Anxiety/stress arising from physical tests	blood, tests, stress
10	Psychotic symptoms - e.g. voices, etc.	voices, hearing, evil
11	Reasurrance/positive feedback - also possibly 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, negative feelings	feel, mood, depression
15	Medication issues	medication, drugs, reduce
16	External support - positive social support (e.g. work, family, people)	good, people, happy
17	Weight management - weight issues in the context of drug side-effects	weight, diet, exercise
18	Medication regimen - dose, timings etc	milligrams, tablets, dose
19	Leisure - social relationships/social life etc	mates, pub, birthday

Table 2: Interpretations of LDA topics

3.2 Distribution

Figure 1 shows the distribution of the different topics across the whole corpus; for the automatic LDA version, this is determined from the most likely assignment of observed words to topics. The distribution is highly skewed, with the largest topic (16) accounting for about a fifth of all the data, and the smallest topic (3) only 1.4%. Once stop words had been removed the corpus consisted of 78,723 tokens. Nearly 18,000 of these (17,957) were therefore most likely to be assigned to topic 16, with just over 1000 (1063) in the smallest topic.

As can be seen from Figure 1, the distribution of automatic topics is consistent with the distribution from the hand-coded topics (Kolmogorov-Smirnov $D = 0.300, p = 0.275$). However, it is not clear that the topics themselves correspond so well. For the hand-coded topics, the topic with the highest probability is *medication*, followed by *general chat* and then *psychotic symptoms*; for the LDA topics, the most likely is *external support*, followed by *medication regimen* and *social stressors* – with *psychotic symptoms* only appearing much further down the list.

3.3 Cross-correlations between hand-coded and automatic topics

We next examined the correspondence between automatic and hand-coded topics directly. Of course, because of the differences in methods, we do not expect these to be equivalent; but examining similarities and differences helps validate (or otherwise) the interpretations given to the LDA topics, and determine whether the topics in fact pick out different aspects of the dialogues in each case.

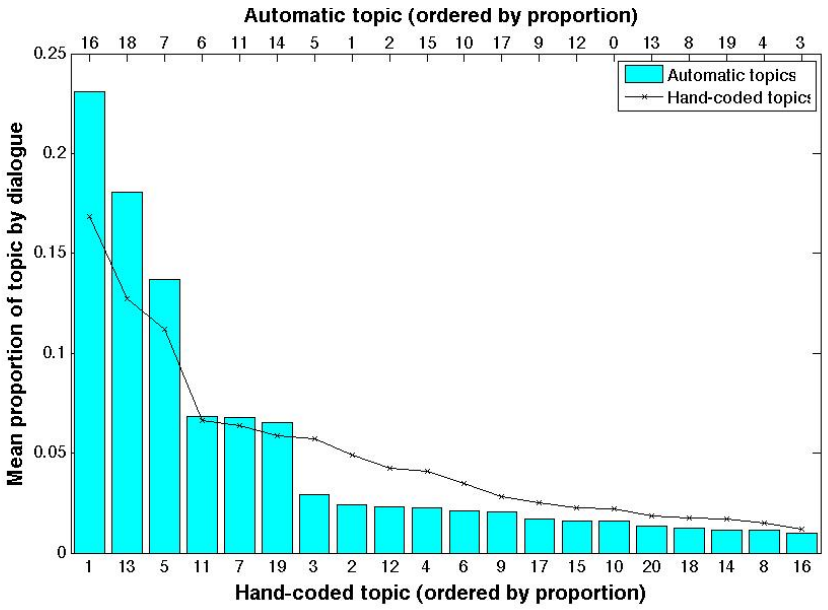

Figure 1: Distribution of topics

Table 3 shows correlations with coefficients greater than 0.3.[2] These correlations are calculated on the basis of the proportions of each topic in each dialogue; as such, these are overview figures across dialogues and do not tell us about topic assignment at a finer-grained level (for example, we know that highly correlated topics occur in the same dialogues, but not whether they occur in the same sequential sections of those 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

Table 3: Correlations between hand-coded and automatic topic distributions

From the data above we can see that some of the topics match up well, suggesting that in certain cases the LDA topic model is picking out similar aspects of the content. Examples are the high correlations between the hand and automatically coded *substance misuse* and *physical health* topics. Given the relative prominence of the two topics, the high correlation between *medication* and *medication regimen* suggests that the LDA topic model is picking out a subset of the talk on medication. This could be linked to the fact that though there may be many different ways of talking about medication (potentially depending on the type of drug, the patient's history etc) that are understandable to human annotators, there is a smaller set of talk about medication which refers to e.g. dosages which is being discovered by LDA topic modelling. Similar considerations may be at play with the link between *healthy lifestyle* and *weight management*, and *non-psychotic symptoms* and *sleep issues*.

More interestingly the hand-coded *psychotic symptoms* topic is highly correlated with two automatic topics about psychotic symptoms. Looking at the contexts of these topics, it appears that there

[2]Note that this is an arbitrary cut-off point; other smaller significant correlations also exist in the data.

may be differences in the ways people talk about their psychotic symptoms depending on whether they are describing the symptoms per se, or looking to make sense of their psychotic symptoms in a wider context.

Interesting differences in the two codings can also be seen in the correlations with *relationships*, which could illustrate different ways is which they are discussed, both negative (*ranting*), and positive (*leisure*). This suggests that the LDA topics are picking up additional factors of the communication in addition to the content.

4 Prediction of Target Variables

We now turn to examining the association between topics and the target variables we would like to predict: symptoms, doctor and patient evaluations of the therapy, and patient outcomes (specifically, adherence to treatment).

4.1 Correlations with symptoms

Patterns of symptoms are known to affect communication, and we therefore assessed whether there were correlations between what was talked about, as indexed by hand coded or automatically coded topic, and the three PANSS symptom scales (positive, negative, general).

	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

Table 4: Correlations between symptoms and topics

As can be seen from Table 4[3], for the hand coded topics, all three symptom scales were negatively correlated with *daily activities* (consultations with more ill patients contained less talk about *daily activities*) and positively correlated with talk about *psychotic symptoms*. Higher general symptoms were also associated with less talk about *healthy lifestyle*, and more about *suicide and self-harm*.

For the automatically extracted topics, consultations with patients with more positive symptoms had more talk in the categories of *ranting*, *making sense of psychosis*, *physical tests* and *psychotic symptoms*. Consultations with patients with worse negative symptoms had less talk about *weight management*. As with the hand-coded topics there was some overlap between positive and general symptoms, with general symptoms positively correlated with *ranting* and *making sense of psychosis*. These correlations also served as a validation measure of some of the topics, and their interpretations.

4.2 Classification experiments

We performed a series of classification experiments, to investigate whether the probability distributions of topics could enable automatic detection of patient and doctor evaluations of the consultation, symptoms and adherence. In each case, we used the Weka machine learning toolkit (Hall et al., 2009) to pre-process data, and a decision tree classifier (J48) and the support vector machine implementation Weka LibSVM (EL-Manzalawy and Honavar, 2005) as classifiers. Variables to be predicted were binarised into groups of equal size prior to analysis, and for the adherence measure a balanced

[3]Table 4 shows correlations above 0.2 only.

	Topics and Dr/P factors		Topics and P factors		Topics only		Dr/P factors only	
Measure	J48	SVM	J48	SVM	J48	SVM	J48	SVM
HAS Dr	**75.8**	**71.2**	47.0	56.8	50.8	56.8	**72.0**	**71.2**
HAS P	46.3	49.3	59.0	53.7	50.7	47.0	51.5	52.2
PANSS pos	58.0	59.5	58.8	49.6	**61.1**	58.0	45.8	59.5
PANSS neg	58.3	59.1	57.6	**62.1**	**61.4**	57.6	54.5	52.3
PANSS gen	51.9	55.0	55.0	57.3	55.7	59.5	51.9	53.4
PEQ comm	50.0	56.0	53.7	59.7	55.2	55.2	57.5	**61.2**
PEQ comm barr	50.7	**61.9**	56.0	50.7	52.2	52.2	49.3	**60.4**
PEQ emo	51.2	45.7	47.2	48.0	51.2	49.6	57.5	50.0
Adherence (balanced)	51.4	**66.2**	47.3	50.0	51.4	44.6	47.3	56.8

Table 5: Accuracy of hand-coded topics with different feature groups

	Topics and Dr/P factors		Topics and P factors		Topics only	
Measure	J48	SVM	J48	SVM	J48	SVM
HAS Dr	**75.0**	**75.0**	**62.9**	50.8	**65.2**	**62.9**
HAS P	49.3	48.5	50.7	50.7	53.7	47.0
PANSS pos	45.0	58.8	47.3	44.3	51.1	50.4
PANSS neg	50.8	52.3	56.1	56.1	48.5	50.8
PANSS gen	47.3	50.4	52.7	48.9	53.4	48.9
PEQ comm	51.5	56.0	54.5	50.7	56.7	53.7
PEQ comm barr	56.7	**60.4**	53.7	47.8	51.5	56.0
PEQ emo	57.5	49.6	48.8	51.2	52.8	53.5
Adherence (balanced)	47.3	54.1	47.3	44.6	47.3	51.4

Table 6: Accuracy of automatically extracted topics with different feature groups

subset of 74 cases was used. All experiments used 5-fold cross-validation, and the experiments using an SVM classifier used a radial bias function with the best values for cost and gamma determined by a grid search in each case.

Tables 5 and 6 show the accuracy figures for each predicted variable, using a variety of different feature subsets. Doctor factors are the gender and identity of the doctor. Patient factors are the gender and age of the patient, and also the total number of words spoken by both patient and doctor. Topic factors are the total number of words in that topic for the hand-coded topics; and an equivalent value for the automatic topics calculated by multiplying the topic's posterior probability for a dialogue by the total number of words.

From Tables 5 and 6[4] we can see that there are different patterns of results for the different measures. For the therapeutic relationship (HAS) measures, including doctor factors gives an accuracy of over 70% in all cases, with the identity of the psychiatrist the most important factor in the decision trees. However, although allowing us a reasonably good fit to the data, the inclusion of the doctor's identity as a feature means that this is not a generalisable result; we would not be able to utilise the information from this factor in predicting the HAS score of a consultation with a new doctor. In this respect, the 65% accuracy when using only the 20 coarse-grained automatic topics is encouraging. In the decision tree, the highest node is *social stressors*, with a high amount of talk in this category indicating a low rating of the therapeutic relationship from the doctor (66 low/21 high). If there was less talk about *social stressors*, the next highest node is *sleep patterns*, with more talk in this area indicating a greater likelihood of a good therapeutic relationship rating (29 high/3 low). Next, more talk about *non-psychotic symptoms* leads to low ratings (11 low/3 high), and more *reassurance*, leads to a better therapeutic relationship. Interestingly, automatic topics give better accuracy than manual topics when used alone.

For adherence, the best accuracy is achieved by a model which includes doctor features as well as hand-coded topics. Good physician communication is known to increase adherence (Zolnierek and DiMatteo, 2009) and in this sample, adherence was also related to the doctor's evaluation of the

[4]Accuracy values of over 60% are shown in bold.

therapeutic relationship, with 29 of the 37 non-adherent patients rated as having a poor therapeutic relationship by the doctor ($\chi 2 = 13.364, p < 0.001$).

Given this, it is surprising that we can predict the therapeutic relationship reasonably well using only automatic topics, but not adherence. Topics also do not appear to give useful performance when predicting patient ratings of the therapeutic relationship (HAS P), or patient evaluations of the consultation (PEQ), although doctor/patient factors seem to have some predictive power. Note that low-level lexical features have shown success in predicting both adherence and patient ratings (Howes et al., 2012a, achieved f-scores of around 70%).

The best predictors for the different types of symptoms are also low, but here the hand-coded topics do better than the automatic topics, with accuracies of 61% for both positive and negative symptoms. For positive symptoms, perhaps unsurprisingly, the decision tree only has one node; if there is more talk on the topic of *psychotic symptoms*, then the patient is likely to have higher positive symptoms (or vice versa). However, in this respect, especially given the cross-correlations discussed above, it is surprising that the automatic topics do not allow any prediction of symptoms at above chance levels. For negative symptoms, patients are likely to have more negative symptoms in consultations with little talk on either *healthy lifestyle* or *daily activities*.

5 Discussion

While both LDA and hand-coded topics seem to have some predictive power, they have different effects for different target variables. Automatic topics do not allow prediction of symptoms, where manual topics do – even though there is a correlation between their corresponding topics relating to psychotic symptoms. This may suggest that LDA used in this way is discovering topics which are a subset of the manual topics: discussion of symptoms may be wider and include more different conversational phenomena than suggested purely by symptom-related lexical items. On the other hand, LDA topics appear to be better at predicting evaluations of the therapeutic relationship; here, one possible explanation may be that LDA is producing "topics" which capture aspects of style or structure rather than purely content. Further investigation might reveal whether examination of the relevant LDA topics can reveal important aspects of communication style – particularly that of the doctor, given that doctor identity factors also improve prediction of this measure, and are related to patients subsequent adherence.

Although the results from this exploratory study are limited, they are encouraging. We have used only very coarse-grained notions of topics, and a simplistic document-style LDA model, so there is much potential for further research. Using a more dialogue-related model that takes account of topic sequential structure (e.g. Purver et al., 2006) or one that can incorporate stylistic material separately to content, as done for function vs content words by Griffiths and Steyvers (2004) should allow us to produce models that better describe the data and can be used to discover more directly what aspects of the communication between doctors and patients with schizophrenia are associated with their symptoms, therapeutic relationship and adherence behaviour.

References

Angus, D., B. Watson, A. Smith, C. Gallois, and J. Wiles (2012). Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLoS ONE 7*(6), 1–12.

Arguello, J. and C. Rosé (2006). Topic segmentation of dialogue. In *Proceedings of the HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, New York, NY.

Arnold, C., S. El-Saden, A. Bui, and R. Taira (2010). Clinical case-based retrieval using latent topic analysis. In *AMIA Annual Symposium Proceedings*, Volume 2010, pp. 26. American Medical Informatics Association.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM 55*(4), 77–84.

Blei, D., A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Cretchley, J., C. Gallois, H. Chenery, and A. Smith (2010). Conversations between carers and people with schizophrenia: a qualitative analysis using leximancer. *Qualitative Health Research 20*(12), 1611–1628.

EL-Manzalawy, Y. and V. Honavar (2005). *WLSVM: Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

Fenton, W., C. Blyler, and R. Heinssen (1997). Determinants of medication compliance in schizophrenia: Empirical and clinical findings. *Schizophrenia Bulletin 23*(4), 637.

Griffiths, T. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Science 101*, 5228–5235.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA data mining software: An update. *SIGDKDD Explorations 11*(1), 10–18.

Hermann, P., M. Lavelle, S. Mehnaz, and R. McCabe (in preparation). What do psychiatrists and patients with schizophrenia talk about in psychiatric encounters?

Howes, C., M. Purver, R. McCabe, P. G. T. Healey, and M. Lavelle (2012a). Helping the medicine go down: Repair and adherence in patient-clinician dialogues. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, Paris.

Howes, C., M. Purver, R. McCabe, P. G. T. Healey, and M. Lavelle (2012b). Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012 Conference)*, Seoul, South Korea, pp. 79–83. Association for Computational Linguistics.

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

McCabe, R., 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.

McCabe, R., 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.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ong, L., J. De Haes, A. Hoos, and F. Lammes (1995). Doctor-patient communication: a review of the literature. *Social science & medicine 40*(7), 903–918.

Priebe, S. and T. Gruyters (1993). The role of the helping alliance in psychiatric community care: A prospective study. *Journal of Nervous and Mental Disease 181*(9), 552–557.

Purver, M., K. Körding, T. Griffiths, and J. Tenenbaum (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, Sydney, Australia, pp. 17–24. Association for Computational Linguistics.

Salton, G. and M. McGill (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Steine, S., 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.

Steyvers, M. and T. Griffiths (2007). Probabilistic topic models. *Handbook of latent semantic analysis 427*(7), 424–440.

Zolnierek, K. and M. DiMatteo (2009). Physician communication and patient adherence to treatment: a meta-analysis. *Medical care 47*(8), 826.