A Generative Joint, Additive, Sequential Model of Topics and Speech Acts in Patient-Doctor Communication

Byron C. Wallace[†], **Thomas A. Trikalinos**[†], **M. Barton Laws**[†],
Ira B. Wilson[†] and **Eugene Charniak**[‡]
[†]Dept. of Health Services, Policy & Practice, Brown University, Providence, RI
[‡]Dept. of Computer Science, Brown University, Providence, RI
{byron_wallace, thomas_trikalinos, michael_barton_laws
ira_wilson, eugene_charniak}@brown.edu

Abstract

We develop a novel generative model of conversation that jointly captures both the topical content and the speech act type associated with each utterance. Our model expresses both token emission and state transition probabilities as log-linear functions of separate components corresponding to topics and speech acts (and their interactions). We apply this model to a dataset comprising annotated patient-physician visits and show that the proposed joint approach outperforms a baseline univariate model.

1 Introduction

Communication involves at least two aspects: the words one says and the *acts* one performs in saying them. Examples of the latter include asking questions, issuing commands, and so on. These are referred to as *speech acts* under the sociolinguistic theory of Austin (1955), which was further developed by Searle (1969; 1985). Recognizing speech acts is crucial to understanding communication because a speaker's meaning is only partially captured by the words they use; much of their intent is expressed implicitly via speech acts (Searle, 1969).

On this view, conversational utterances can be assigned both a topic and a speech act. The former describes the subject matter of what was said and the latter captures the "social act" (e.g., promising) performed by saying it. For example, the utterance "Obama won the election" is topically *political* and is an example of an *information giving* speech act. "Did Obama win the election?", meanwhile, belongs to the same topic but is a *question*. Both aspects are necessary to understand conversation.

Role	Utterance	Topic	Speech act
D	Let me just write down some of these issues here so I get them straight in my mind.	*Logistics*	*Commissive*
P	Doctor you ain't got to tell me nuttin'.	*Socializing*	*Directive*
P	I'm in very good hands when I'm around you.	*Socializing*	*Give Info.*
P	If push comes to a shove, you open the window and throw me out.	*Socializing*	*Humor/Levity*
D	I wanted to ask you, too -	*Biomedical*	*Conv. Mgmt.*
D	you know you had that colonic polyp -	*Biomedical*	*Ask Q.*
D	- is it two years from now that they're going to be doing the repeat?	*Biomedical*	*Ask Q.*
P	Yeah.	*Biomedical*	*Conv. Mgmt.*
D	We'll do the repeat coloscopy in about two years.	*Biomedical*	*Give Info.*

Table 1: An excerpt from a patient-doctor interaction, annotated with topic and speech act codes. The *D* and *P* roles denote doctor and patient, respectively. *Conv. Mgmt.* abbreviates *conversation management*; *Ask Q.* abbreviates *ask question*.

Previous computational work on speech acts – which we review in Section 6 – has modeled them in isolation (Perrault and Allen, 1980; Stolcke et al., 1998; Stolcke et al., 2000; Kim et al., 2010), i.e., independent of topical content. But a richer model would account for both speech acts *and* the contextualizing topic of each utterance. To this end, we develop a novel joint, generative model of topics and speech acts.

We focus on physician-patient communication as a motivating domain. This is of interest because

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1765–1775,
Seattle, Washington, USA, 18-21 October 2013.

it is widely appreciated that effective communication is an integral part of clinical practice (Irwin and Richardson, 2006; Makoul, 2001; Teutsch, 2003). We provide an excerpt of a conversation between a patient and their doctor annotated with topics and speech acts in Table 1. Such annotations can provide substantive insights into how doctors communicate with patients (Ong et al., 1995).

A concrete example of this is the use of topic and speech act codes to assess the efficacy of an intervention meant to influence physician-patient communication regarding adherence to antiretroviral (ARV) medication (Wilson et al., 2010). To measure the effect of the intervention, investigators performed a randomized control trial in which they quantified change in communication patterns by tallying the number of *information giving* speech acts that fell under the *ARV adherence* topic. Without assigning both topics and speech acts to utterances, this analysis would not have been possible.

In this work, we develop a novel component-based generative model for bivariate, sequentially structured problems. Our approach extends the recently proposed Sparse Additive Generative (SAGE) model (Eisenstein et al., 2011) and similar recently developed additive models (Paul and Dredze, 2012; Paul et al., 2013) to the case of supervised sequential tasks to capture the joint conditional influence of topics and speech acts, both with respect to token generation and state transitions. For brevity, we refer to this generative Joint, Additive, Sequential model as *JAS*. In contrast to previous work on speech acts, JAS provides a single, coherent generative model of conversations. And because it is component-based, this model provides a flexible framework for analyzing communication patterns. We demonstrate that JAS outperforms a generative univariate baseline in topic/speech act prediction. Further, we automatically reproduce an analysis of the aforementioned randomized control trial, and in doing so show that JAS reproduces the results more faithfully than a univariate approach.

2 The Markov-Multinomial Model

We begin by considering a baseline generative approach to modeling topics and speech acts independently. This simple approach was used by Stolcke et al. (2000) to model speech acts. It accounts for only a single output at each time point $y_t \in \mathcal{Y}$, and hence here we model topics and speech acts independently.

A straight-forward (albeit naïve) alternative would be to treat the Cartesian product of topics and speech acts as a single output space on which emissions and transitions are conditioned, but this space is too large and sparse for this approach to be practicable. We note that the *fully coupled HMM* (Brand et al., 1997) suffers from a similar exponential output state problem. The related *factorial HMM* (FHMM) (Ghahramani and Jordan, 1997; Van Gael et al., 2008), meanwhile, imposes unwarranted (in our case) independence assumptions with respect to state transitions along parallel chains, does not obviously lend itself to discrete observations (typically Gaussians are assumed), and does not scale well enough (in terms of training time) to be feasible for our application.

The Markov-Multinomial (MM) comprises two components; transitions and emissions. The former is modeled by making a first-order Markov assumption, specifically:

$$P(y_t|y_0, ..., y_{t-1}) = P(y_t|y_{t-1}) = \lambda_{y_{t-1},y_t} \quad (1)$$

Emissions can be modeled via a multinomial that captures the conditional probabilities of tokens given labels. Denoting an utterance (an utterance comprises the words corresponding to a single speech act; see Section 4) at time t by u_t and its label by y_t, and making the standard naïve assumption that words are generated independently conditioned on a label, we have:

$$P(u_t|y_t) = \prod_{w \in u_t} P(w|y_t) = \prod_{w \in u_t} \tau_{y_t,w} \quad (2)$$

Both sets of parameters (the $\boldsymbol{\lambda}$'s and the $\boldsymbol{\tau}$'s) can be estimated straight-forwardly using maximum likelihood (i.e., using observed counts). We can use Viterbi decoding (Rabiner and Juang, 1986) to make predictions for new sequences, as usual. To make both topic and speech act predictions, we simply induce models for each and make predictions independently.

3 *JAS*: A Joint, Additive, Sequential Model

An obvious shortcoming of the simple MM model outlined above is that it treats topics and speech acts

as statistically independent. They are not (as confirmed at statistical significance $p < .001$ using a χ^2 test). One would prefer a more expressive model that conditions topic and speech act transitions as well as the production of utterances jointly on both the current topic and the current speech act.

More specifically, we would like a model that reflects the assumption that some latent *intent* gives rise to both the topic and the speech act associated with an utterance. This is consistent with Searle's (1969) notion of *perlocutionary* effects; one performs speech acts with the aim of getting someone to do something. Intent gives rise to the current topic and speech act, and the current intent affects the next; this induces a correlation between adjacent topics and speech acts. This conceptual model is depicted graphically in the left-half of Figure 1.

The latent intent may be, e.g., to encourage a patient to take their medication more regularly. In our application the topical content may be *ARV adherence* and the type of speech act would be selected by the provider (presumably to maximize the likelihood of patient adherence). For example, she may opt to urge imperatively ("You really need to take your medicine") or to implore with a question ("Will you please remember to take your medicine?"). Because we have no way of explicitly modeling intent (it is never observed), we instead rely on variables for which we have annotations (i.e., the topics and speech acts; see Figure 1). We next describe the model in more detail.

We refer to the topic set by $\mathcal{Y}$, the speech act set by $\mathcal{S}$ and the vocabulary as $\mathcal{W}$. We denote the (log of the) background probability of word w by θ_w, and we will denote components corresponding to deviations from θ_w due to a specific topic (speech act) by η_w^y (η_w^s). Further, we include the component $\eta_w^{y,s}$ to capture interaction effects between topics and speech acts. We assume that the conditional probability of word w belonging to an utterance u_t with corresponding topic y_t and speech act s_t is log-linear with respect to these components, i.e.:

$$P(w|y_t, s_t) = \frac{1}{Z_w} exp\{\theta_w + \eta_w^{y_t} + \eta_w^{s_t} + \eta_w^{s_t,y_t}\} \quad (3)$$

Where Z_w is a normalizing term (implicitly conditioned on y_t and s_t) defined as:

$$Z_w = \sum_{w' \in \mathcal{W}} exp\{\theta_{w'} + \eta_{w'}^{y_t} + \eta_{w'}^{s_t} + \eta_{w'}^{s_t,y_t}\} \quad (4)$$

We make the standard naïve assumption that words are generated independently, given the topic and speech act of the utterance to which they belong:

$$P(u_t|y_t, s_t) = \prod_{w \in u_t} P(w|y_t, s_t) \quad (5)$$

The per-token emission probability just described falls under the additive generative family of models recently proposed by Eisenstein et al. (2011). However, in addition to conditional token emission probabilities, here we need also to model the *transition* probabilities such that the likelihood of transitioning to topic y_t (and to speech act s_t) reflects both the previous topic *and* the previous speech act, capturing the dependencies illustrated in Figure 1. To this end, we model topic and speech act transition probabilities as log-linear functions of the preceding topic and speech act.

We denote log of the background topic frequencies by $\boldsymbol{\pi}^{\mathcal{Y}}$, and components capturing the influence of transitioning to topic y_t due to the preceding topic and speech act by σ_{y_{t-1},y_t} and σ_{s_{t-1},y_t} respectively. We also include a component $\sigma_{(y_{t-1},s_{t-1}),y_t}$ that corresponds to the interaction effect on topic transition probability due to the preceding topic/speech act pair. We then model the topic transition probability (given the preceding states) as:

$$P(y_t|y_{t-1}, s_{t-1}) = \frac{1}{Z_y} exp\{\pi_{y_t}^{\mathcal{Y}} + \sigma_{y_{t-1},y_t} + \sigma_{s_{t-1},y_t} + \sigma_{(y_{t-1},s_{t-1}),y_t}\} \quad (6)$$

Where Z_y is a normalizing term for the topic transitions (implicitly conditioned on s_{t-1}, y_{t-1}):

$$Z_y = \sum_{y' \in \mathcal{Y}} exp\{\pi_{y'}^{\mathcal{Y}} + \sigma_{y_{t-1},y'} + \sigma_{s_{t-1},y'} + \sigma_{(y_{t-1},s_{t-1}),y'}\} \quad (7)$$

Similarly, denoting by $\boldsymbol{\pi}^{\mathcal{S}}$ log-transformed speech act background frequencies, and including analogous components as above that correspond to the influence of the preceding topic, speech act and their interaction on transitioning into speech act s_t, we

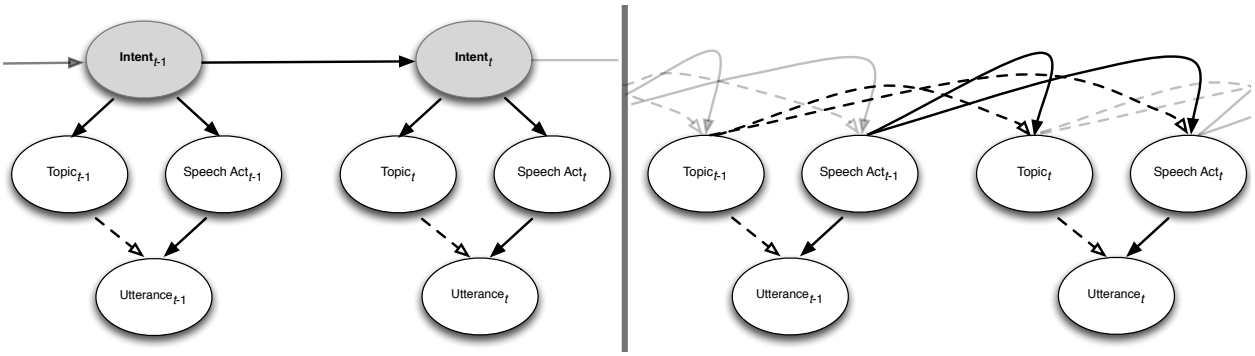

Figure 1: The generative story of utterances, depicted graphically. On the left we show our motivating conceptualization: a latent *intent* gives rise to both the topic and speech acts; these, in turn, jointly induce a distribution over words and transitions. On the right we show our operationalization of this concept. For clarity, we have denoted arrows capturing influence due to topics with dotted lines.

have:

$$P(s_t|s_{t-1}, y_{t-1}) = \frac{1}{Z_s} exp\{\pi^{\mathcal{S}}_{s_t} + \sigma_{s_{t-1},s_t} + \sigma_{y_{t-1},s_t} + \sigma_{(y_{t-1},s_{t-1}),s_t}\} \quad (8)$$

Where Z_s is a normalizing constant for speech acts analogous to Equation 7. Putting things together:

$$P(y_t, s_t|s_{t-1}, y_{t-1}, u_t) = P(u_t|y_t, s_t) \cdot P(y_t|y_{t-1}, s_{t-1}) \cdot P(s_t|s_{t-1}, y_{t-1}) \quad (9)$$

As implied by Figure 1, this model assumes that the topic and speech act at time t are conditionally independent given the preceding topic and speech act (y_{t-1} and s_{t-1}). This is intuitively agreeable because time intervenes as a blocking factor; conditioning the *current* topic on the *current* speech act (or vice versa) would contradict the fact that these occur simultaneously. Instead, the correlation is induced by the preceding topic/speech act pair. (That said, this is still a simplifying assumption, as one may instead choose to model speech act selection as conditional on topic (Traum and Larsson, 2003).)

Predictions can again be made via Viterbi decoding (Rabiner and Juang, 1986) over a matrix of *pairs* of joint topic/speech act states. The strategy of modeling (additive) components allows JAS to avoid problems due to sparsity in this large output space.

Model parameters can be estimated using standard optimization techniques. We fix the 'background' frequencies $\boldsymbol{\theta}$, $\boldsymbol{\pi}^{\mathcal{Y}}$, $\boldsymbol{\pi}^{\mathcal{S}}$ to the log of the corresponding observed proportions of words, topics and speech acts, respectively. For the remaining parameters, one can use descent-based optimization methods. The partial derivative for the topic-to-topic transition component $\sigma_{y,y'}$ with respect to the likelihood, for example, is:

$$\frac{\partial}{\partial \sigma_{y,y'}} = \sum_{s \in \mathcal{S}} C_{(y,s),y'} - P(y'|y, s) C_{(y,s),*} \quad (10)$$

Where $C_{(y,s),y'}$ denotes the observed count of transitions from topic/speech act pair (y, s) to y', and $C_{(y,s),*}$ denotes the total number of observed transitions out of this pair. The term $P(y'|y, s)$ is with respect to the current parameter estimates and is defined in Equation 6. The partial derivatives for the other component parameters (both transition and emission) are analogous. We use a Newton optimization method similar to the approach outlined by Eisenstein et al. (2011).[1] We assess convergence by calculating predictive performance on a held-out portion (5%) of the training dataset at each step, halting the descent when this declines.

4 Dataset

We use a corpus of patient-provider visits annotated with *Generalized Medical Interaction Anaylsis System* (GMIAS) codes. The GMIAS has been used to: characterize interaction processes in physician-patient communication about ARV adherence in the

[1] With the exception that we do not explicitly model the distribution over component variances.

Topic; *Speech act*	*Count (prevalence)*
ARV Adherence; Ask Q	2939 (0.013)
ARV Adherence; Commissive	245 (0.001)
ARV Adherence; Continuation	328 (0.001)
ARV Adherence; Conv. Management	4298 (0.018)
ARV Adherence; Directive	1650 (0.007)
ARV Adherence; Empathy	111 (0.000)
ARV Adherence; Give Information	12796 (0.055)
ARV Adherence; Humor/Levity	46 (0.000)
ARV Adherence; Missing/other	977 (0.004)
ARV Adherence; Social-Ritual	15 (0.000)
Biomedical; Ask Q	13753 (0.059)
Biomedical; Commissive	1049 (0.005)
Biomedical; Continuation	1005 (0.004)
Biomedical; Conv. Management	17611 (0.076)
Biomedical; Directive	4617 (0.020)
Biomedical; Empathy	423 (0.002)
Biomedical; Give Information	54231 (0.233)
Biomedical; Humor/Levity	255 (0.001)
Biomedical; Missing/other	4426 (0.019)
Biomedical; Social-Ritual	119 (0.001)
Logistics; Ask Q	5517 (0.024)
Logistics; Commissive	2308 (0.010)
Logistics; Continuation	435 (0.002)
Logistics; Conv. Management	9672 (0.042)
Logistics; Directive	5148 (0.022)
Logistics; Empathy	100 (0.000)
Logistics; Give Information	23351 (0.101)
Logistics; Humor/Levity	135 (0.001)
Logistics; Missing/other	2732 (0.012)
Logistics; Social-Ritual	285 (0.001)
Missing/other; Ask Q	820 (0.004)
Missing/other; Commissive	70 (0.000)
Missing/other; Continuation	1173 (0.005)
Missing/other; Conv. Management	1605 (0.007)
Missing/other; Directive	523 (0.002)
Missing/other; Empathy	48 (0.000)
Missing/other; Give Information	3994 (0.017)
Missing/other; Humor/Levity	27 (0.000)
Missing/other; Missing/other	12103 (0.052)
Missing/other; Social-Ritual	69 (0.000)
Psycho-Social; Ask Q	2933 (0.013)
Psycho-Social; Commissive	164 (0.001)
Psycho-Social; Continuation	208 (0.001)
Psycho-Social; Conv. Management	4433 (0.019)
Psycho-Social; Directive	787 (0.003)
Psycho-Social; Empathy	262 (0.001)
Psycho-Social; Give Information	15521 (0.067)
Psycho-Social; Humor/Levity	63 (0.000)
Psycho-Social; Missing/other	1199 (0.005)
Psycho-Social; Social-Ritual	36 (0.000)
Socializing; Ask Q	1283 (0.006)
Socializing; Commissive	79 (0.000)
Socializing; Continuation	85 (0.000)
Socializing; Conv. Management	2166 (0.009)
Socializing; Directive	222 (0.001)
Socializing; Empathy	73 (0.000)
Socializing; Give Information	8981 (0.039)
Socializing; Humor/Levity	306 (0.001)
Socializing; Missing/other	849 (0.004)
Socializing; Social-Ritual	1685 (0.007)

Table 2: Topic/speech act pairs and their counts.

context of an intervention trial (Wilson et al., 2010); analyze communication about sexual risk behavior (Laws et al., 2011a); elucidate the association of visit length with constructs of patient-centeredness (Laws et al., 2011b); and to describe provider-patient communication regarding ARV adherence compared with communication about other issues (Laws et al., 2012). GMIAS annotation is described at length elsewhere,[2] but we summarize it here for completeness.

GMIAS segments conversation into *utterances*. An utterance is here defined as a single completed speech act. Previous coding systems have simply defined an utterance as conveying a single thought (Roter and Larson, 2002) or any independent or unrestrictive dependent clause of a sentence (Ford and Ford, 1995). Stolcke et al. (2000) followed Meteer et al. (1995) in using "sentence-level units". These definitions provide helpful guidance to coders, but many speech acts are poorly formed grammatically, and cannot be described as a "clause". Further, some speech acts cannot be said to convey a "thought" (or sentence) at all, but rather are pre-syntactical (e.g., interjections and non-lexical utterances like laughter). In any case, most natural segmentations of conversations probably largely agree with intuition, and are not likely to differ substantially.

The model we develop in this work assumes that transcripts have been manually segmented. While this comes at some cost, segmenting is still much cheaper than *annotating* transcripts. Manually annotating a single visit with GMIAS codes takes 2-4 hours and must be performed by someone with substantive domain expertise. By contrast, segmenting transcripts into utterances takes at most 1/4th of the time as annotation and can be done by a less highly-skilled individual. That said, in future work we hope to explore incorporating automatic segmentation methods (Galley et al., 2003; Eisenstein and Barzilay, 2008) into our approach.

Each utterance is assigned a single topic code and a single speech act code. Inter-rater agreement has been observed to be relatively high for this task: Kappa between three trained annotators and a reference annotation ranged from 0.89 to 1.0 for topics and 0.81 to 0.95 for speech acts. We next de-

[2] https://sites.google.com/a/brown.edu/m-barton-laws/home/gmias

scribe the topics and speech acts we consider in more detail; Table 2 enumerates all pairs of these and their respective counts in the corpus. We note that GMIAS defines a hierarchy of both topic and speech act codes, but here we only attempt to capture the highest level codes in these hierarchies.

Topics comprise six major categories: *ARV adherence*, *biomedical*, *logistics*, *missing/other*, *psycho-social* and *socializing*. *Antiretroviral (ARV) adherence* applies to utterances that address ARV medication usage. *Biomedical* utterances subsume clinical observations and diagnostic conclusions. Utterances that concern the business of conducting a physical examination fall under *logistics*. The *missing/other* topic covers a few cases, including utterances that are effectively outside of the GMIAS universe and inaudible utterances; however we note that *missing/other* is a topic explicitly assigned by human annotators. The *psycho-social* topic includes such issues as substance abuse, recovery, employment and relationships. Finally, *socializing* refers to casual conversation unrelated to the business of the medical visit, and to social rituals such as greetings.

There are 10 speech acts:[3] *ask question*, *commissive*, *continuation*, *conversation management*, *directive*, *empathy*, *give information*, *humor/levity*, *missing/other*, and *social-ritual*. *Ask question* is self-explanatory. Utterances in which the speaker makes a promise or resolves to take action are *commissives*. A *continuation* refers to the completion of a previously interrupted speech act (these are rare). *Conversation management* describes utterances that facilitate turn-taking or guide discussion ('talk about talk'). *Directives* refer to statements that look to control or influence the behavior of the interlocutor. Utterances that express responses to emotions, concerns or feelings are coded under *empathy*. Communication of (purported) facts falls under *give information*. *Humor/levity* captures jokes and jovial conversation. *Missing/other* is the same as for topics. Finally, *social-ritual* utterances represent formalities (e.g., "thank you").

The corpus we use includes 360 GMIAS annotated patient-provider interactions (median length: 605 utterances). This data originated as part of a study designed to assess the role of the patient-provider relationship in explaining racial/ethnic disparities in HIV care. Study subjects were HIV care providers and their patients at four US care sites. The group responsible for the data are awaiting a decision from the institutional review board (IRB) regarding whether we can make this data publicly available in some form.

5 Experimental Results

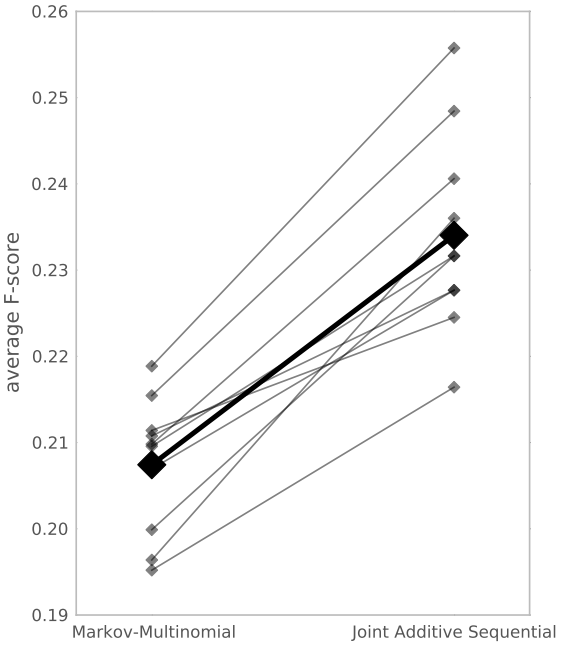

Figure 2: Mean F-scores across all topic/speech act pairs for the Markov-Multinomial (MM; left) and the proposed Joint Additive Sequential (JAS; right) models. The thick black line shows the mean difference over ten different folds; the thin grey lines describe per-fold differences. The proposed JAS model outperforms the baseline MM model for all folds

Our evaluation includes two parts.[4] First, we perform standard cross-validation over the aforementioned 360 annotated interactions, evaluating F-measure for each topic/speech act pair. Second, we look to automatically reproduce an analysis of

[3]These are high-level speech acts; technically each constitutes a *category* of speech act types.

[4]Source code at: https://github.com/bwallace/JAS; unfortunately we do not yet have permission to post the data.

a randomized control trial that assessed the efficacy of an intervention meant to alter physician-patient communication. We show that JAS outperforms the baseline approach with respect to both tasks.

We emphasize that while we are here comparing predictive performance, we are specifically interested in fully generative models of conversations due to the longer-term applications we have in mind. We would like, e.g., to use this model to assess the variation in communicative approaches across different doctors, and generative models are more naturally amenable to answering such exploratory questions. Indeed, perhaps the main strength of the additive component based sequential model we have proposed here is that it will allow us to easily incorporate physician-specific parameters that capture deviations in provider speech act and/or topic transition patterns. Further, we may soon have access to many unannotated transcripts, and we would like to learn from these; generative approaches allow straight-forward exploitation of unlabeled data. For these reasons, we did not experiment with discriminative models, e.g., Dynamic Conditional Random Fields (DCRFs) (Sutton et al., 2007) for this work.

5.1 Cross-fold Validation

Our aim is to measure model performance in terms of correctly identifying both the topic and speech act corresponding to each utterance. We quantify this via the F-score calculated for each topic/speech act pair that is observed at least once. One can see in Table 2 that many such pairs have low prevalence; this can result in undefined F-scores (e.g., when no utterances are assigned to a given pair). In this case, it is reasonable to treat these as zero values, as is commonly done (Forman and Scholz, 2010). This penalizes models when they completely fail to identify an entire class of utterances.

We first report macro-averages, that is, averages of the individual topic/speech act pair F-scores. Figure 2 displays the macro-averaged F-score for each of the 10 folds (grey lines connect folds) and the average of these (thick black line). The JAS model achieves an average macro-averaged F-score of .234 versus the .207 achieved by baseline Markov-Multinomial (MM) model; JAS outperforms MM on every fold.

For a more granular picture, Figure 3 displays average F-score differences with respect to every individual topic and speech act pair for which this difference was non-zero. This is the (signed) difference of the F-score achieved using JAS minus that achieved using the MM model; black lines thus correspond to pairs for which JAS outperformed MM, and red lines to pairs for which MM outperformed JAS. The latter achieves an improvement of $>= .05$ for 10 pairs, and results in an F-score of $> .02$ below that attained by MM only once.

The relatively low F-scores for the metrics quantifying performance with respect to the cross of topic and speech act codes belie relatively good *overall* (marginal) predictive performance. That is, we achieve much better performance with respect to metrics that measure topic and speech act predictions independently of one another. This is due to the very large output space under consideration (see Table 2). Specifically, averaged over ten runs, the MM model achieves a marginal mean topic F-score of .667 and marginal mean speech act F-score of .516. JAS begets a marginal mean topic F-score of .661 and a marginal mean speech act F-score of .544; hence the JAS model incurs an F-score loss of .006 (a 0.9% decrease) with respect to marginal topic code prediction, but improves the marginal speech act F-score by .028 (a 5.4% increase).

5.2 (Re-)Analysis of Randomized Control Trial

We also evaluated performance by tallying model predictions over 116 held-out cases collected from a randomized, cross-over study of an intervention aimed at improving physicians knowledge of patients anti-retroviral (ARV) adherence (Wilson et al., 2010). The intervention was a report given to the physician before a routine office visit that contained information regarding the patients ARV usage and their beliefs about ARV therapy. To explore the efficacy of this intervention, 58 paired (116 total) audio recorded visits were annotated with GMIAS; 58 correspond to visits before which the provider was not provided with the report (control cases), while the other 58 correspond to visits before which they were (intervention cases).

Wilson et al. (2010) demonstrated that the intervention indeed increased adherence-related dialogue, and specifically the number of *information giving* speech acts performed by the physician un-

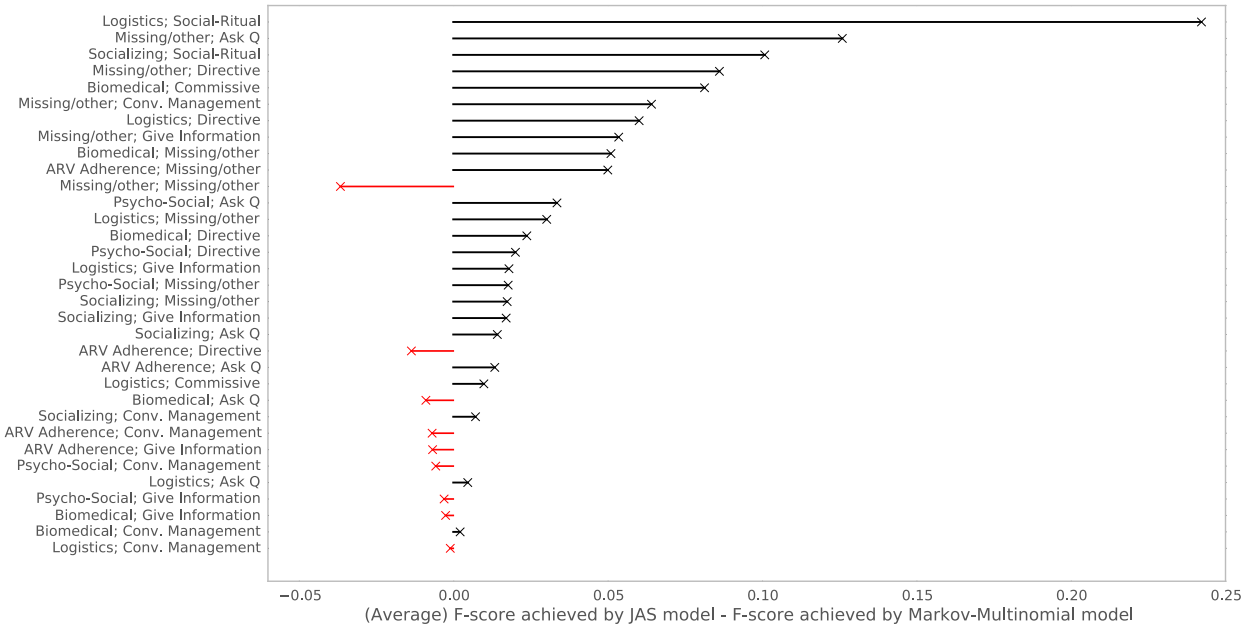

Figure 3: Average difference in F-scores corresponding to specific topic/speech act pairs, sorted by magnitude. Black lines (extending rightward) represent pairs for which JAS outperforms the baseline model; red lines (leftward) are pairs for which baseline performs better.

True		MM		JAS	
control	intervention	control	intervention	control	intervention
10 (4, 28)	23 (11, 39)	13 (5, 33)	27 (16, 44)	12 (5, 28)	23 (14, 40)

Table 3: Utterance counts {Median (25th, 75th percentile)} for the *ARV/information giving* topic and speech act pair. We show the 'gold standard' (True) tallies, which were assigned by humans, and the counts taken using the two models, MM and JAS. The JAS model predictions are closer to the true numbers.

derneath this topic. We attempted to reproduce this finding using automated rather manual annotations. To this end, we trained MM and JAS models over the aforementioned 360 annotated visits and then used this model to generate topic and speech act code predictions for the utterances comprising the 116 held-out visits used for the analysis (these were *not* part of the training set). We then assessed the direction and magnitude of the change in the number of *ARV adherence/information giving* utterances in the paired control versus intervention cases. We compared the results for this analysis calculated using the true (manually assigned) codes to the results calculated using the predicted codes.

Following the original analysis (Wilson et al., 2010), we report the median number of *ARV/information giving* utterances and corresponding 25th and 75th percentiles over the 58 control and intervention visits, as counted using the true (human) annotations and using the codes predicted by the MM and JAS models. These are reported in Table 3. The JAS model predictions better match the true labels in all except one case (the lower 25th for the controls, for which it predicts the same number as the MM model).

6 Related work

There is a relatively long history of research into modeling conversational speech acts in computational linguistics. Perrault and Allen (1980) conducted pioneering work on computationally formalizing speech acts, though their work pre-dates statistical NLP and is therefore not directly relevant to the present work.

Stolcke et al. (2000; 1998) proposed a probabilistic approach to modeling conversational speech acts based on the Hidden Markov Model (HMM) (Rabiner and Juang, 1986). They were interested in modeling an unrestricted set of conversations, and did not impose a hierarchy on the speech acts; they

therefore enumerated many more speech acts (42) than we do in the present work (recall that we use 10 'high-level' speech acts).[5] Their model has served as the baseline approach in the present work. Stolcke et al. also considered jointly performing speech *recognition* and speech act classification.

Others have investigated visual structures of patient-provider interactions to qualitatively assess communication in care. Specifically, (Cretchley et al., 2010) leveraged concept maps to explore conversations between people with schizophrenia and their carers. Briefly, this approach allowed them to (qualitatively) identify two distinct conversational strategies used by care-takers and their patients. Angus et al. (Angus et al., 2012) presented a similar approach in which they used text visualization software to explore patterns of (inferred) topics in consultations.

Another thread of research has investigated classifying speech acts in emails into one of a small set of "email speech acts", e.g., *request*, *propose*, *commit* (Cohen et al., 2004; Goldstein et al., 2006). Cohen et al. (2004) demonstrated that good performance can be achieved for this task via existing text classification technologies. Elsewhere, researchers have explored automatically inferring "speech acts" in various other online social mediums, including message board posts (Qadir and Riloff, 2011), Wikipedia talk pages (Ferschke et al., 2012) and Twitter (Zhang et al., 2012).

A separate line of inquiry concerns classifying dialogue acts in chat. Researchers have attempted dialogue act classification both for 1-on-1 (Kim et al., 2010) and multi-party (Kim et al., 2012; Clark and Popescu-Belis, 2004) online chats. Ang et al. (2005) considered the task of jointly segmenting and classifying utterances comprising multiparty meetings, while Hsueh and Moore (2006) proposed analogous methods for *topic* segmentation and labeling (other works on topic segmentation include (Galley et al., 2003) and (Eisenstein and Barzilay, 2008)). Incorporating such segmentation methods into the proposed model (rather than relying on inputs to be manually segmented beforehand) would be a natural extension of this work.

Additive component models of text have recently gained traction (Eisenstein et al., 2011; Paul, 2012; Paul and Dredze, 2012; Paul et al., 2013). To our knowledge, this is the first extension of supervised additive component models to a sequential task.[6]

7 Conclusions and Future Directions

We have proposed a novel Joint, Additive, Sequential (JAS) model of conversational topics and speech acts. In contrast to previous approaches to modeling conversational exchanges, this model factors both the current topic and the current speech act into token emission *and* state transition probabilities. We demonstrated that this model consistently outperforms a univariate generative baseline that treats speech acts and topics independently. Furthermore, we showed JAS can automatically re-produce the analysis of a randomized control trial designed to assess the efficacy of an intervention to alter physician communication habits with high-fidelity.

The generative component-based framework we have introduced in this work provides a means of exploring factors in patient-physician communication. One limitation of the model we have presented is that it makes several simplifying assumptions around dialogue. For example, we have ignored non-linearities and 'back-channels' in conversation, and we have ignored differences across physicians with respect to communication styles.

Going forward, we hope to address these limitations. We also plan on extending this model to investigate qualitative questions surrounding patient-physician communication quantitatively. For example, we are interested in investigating how communication varies across hospitals and physicians. To explore this, we can add additional components to the transition probability terms corresponding to different hospitals and doctors. Ultimately, we would like to correlate patterns in physician communication (as gleaned from the model) with objective, measured health outcomes (e.g., patient satisfaction and adherence to ARVs).

[5] We note that only 8 of the 42 speech acts appeared with greater than 1% frequency in Stolcke et al.'s corpus.

[6] Though Paul (2012) recently proposed 'mixed-membership' Markov models for *unsupervised* conversation modeling.

8 Acknowledgements

The authors thank members of the Brown Laboratory for Linguistic Information Processing (BLLIP) and Kevin Small for providing helpful feedback on earlier versions of this work. We also thank the three anonymous EMNLP reviewers for insightful comments. This work was partially supported by the National Institute of Mental Health (2 K24MH092242, R34MH089279 and R01MH083595) and by NIDA (R01DA015679).

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP*, volume 1, pages 1061–1064.

Daniel Angus, Bernadette Watson, Andrew Smith, Cindy Gallois, and Janet Wiles. 2012. Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PloS one*, 7(6).

John Langshaw Austin. 1955. *How to do things with words*, volume 88. Harvard University Press.

Matthew Brand, Nuria Oliver, and Alex Pentland. 1997. Coupled hidden Markov models for complex action recognition. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 994–999. IEEE.

Alexander Clark and Andrei Popescu-Belis. 2004. Multi-level dialogue act tags. In *Proc. SIGdial*, pages 163–170.

William W Cohen, Vitor R Carvalho, and Tom M Mitchell. 2004. Learning to classify email into speech acts. In *Proceedings of EMNLP*, volume 4. sn.

Julia Cretchley, Cindy Gallois, Helen Chenery, and Andrew Smith. 2010. Conversations between carers and people with schizophrenia: a qualitative analysis using leximancer. *Qualitative Health Research*, 20(12):1611–1628.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343. Association for Computational Linguistics.

J. Eisenstein, A. Ahmed, and E.P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 777–786. Citeseer.

Jeffrey D Ford and Laurie W Ford. 1995. The role of conversations in producing intentional change in organizations. *Academy of Management Review*, pages 541–570.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. volume 12, pages 49–57. ACM.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 562–569. Association for Computational Linguistics.

Zoubin Ghahramani and Michael I Jordan. 1997. Factorial hidden Markov models. *Machine learning*, 29(2-3):245–273.

Jade Goldstein, Andrew Kwasinski, Paul Kingsbury, R Sabin, and Albert McDowell. 2006. Annotating subsets of the enron email corpus. In *Proceedings of the Third Conference on Email and Anti-Spam*. Citeseer.

P-Y Hsueh and Johanna D Moore. 2006. Automatic topic segmentation and labeling in multiparty dialogue. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 98–101. IEEE.

Richard S Irwin and Naomi D Richardson. 2006. Patient-focused careusing the right tools. *CHEST Journal*, 130(1_suppl):73S–82S.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871. Association for Computational Linguistics.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2012. Classifying dialogue acts in multi-party live chats.

Michael Barton Laws, Ylisabyth S Bradshaw, Steven A Safren, Mary Catherine Beach, Yoojin Lee, William Rogers, and Ira B Wilson. 2011a. Discussion of sexual risk behavior in HIV care is infrequent and appears ineffectual: a mixed methods study. *AIDS and Behavior*, 15(4):812–822.

Michael Barton Laws, Lauren Epstein, Yoojin Lee, William Rogers, Mary Catherine Beach, and Ira B Wilson. 2011b. The association of visit length and measures of patient-centered communication in HIV care: A mixed methods study. *Patient Education and Counseling*, 85(3):e183–e188.

Michael Barton Laws, Mary Catherine Beach, Yoojin Lee, William H Rogers, Somnath Saha, P Todd Korthuis, Victoria Sharp, and Ira B Wilson. 2012. Provider-patient adherence dialogue in HIV care: results of a multisite study. *AIDS and Behavior*, pages 1–12.

Gregory Makoul. 2001. Essential elements of communication in medical encounters: the kalamazoo consensus statement. *Academic Medicine*, 76(4):390–393.

Marie W Meteer, Ann A Taylor, Robert MacIntyre, and Rukmini Iyer. 1995. *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania.

Lucille ML Ong, Johanna CJM De Haes, Alaysia M Hoos, and Frits B Lammes. 1995. Doctor-patient communication: a review of the literature. *Social science & medicine*, 40(7):903–918.

Michael Paul and Mark Dredze. 2012. Factorial lda: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems 25*, pages 2591–2599.

Michael J. Paul, Byron C. Wallace, and Mark Dredze. 2013. What affects patient (dis)satisfaction? analyzing online doctor ratings with a joint topic-sentiment model. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Michael J Paul. 2012. Mixed membership Markov models for unsupervised conversation modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 94–104. Association for Computational Linguistics.

C Raymond Perrault and James F Allen. 1980. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6(3-4):167–182.

Ashequl Qadir and Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 748–758. Association for Computational Linguistics.

Lawrence Rabiner and B Juang. 1986. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Debra Roter and Susan Larson. 2002. The roter interaction analysis system (rias): utility and flexibility for analysis of medical interactions. *Patient education and counseling*, 46(4):243–251.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

John R Searle. 1985. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *The Journal of Machine Learning Research*, 8:693–723.

Carol Teutsch. 2003. Patient-doctor communication. *The medical clinics of North America*, 87(5):1115.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

Jurgen Van Gael, Yee Whye Teh, and Zoubin Ghahramani. 2008. The infinite factorial hidden Markov model. In *Neural Information Processing Systems*, volume 21.

Ira B Wilson, M Barton Laws, Steven A Safren, Yoojin Lee, Minyi Lu, William Coady, Paul R Skolnik, and William H Rogers. 2010. Provider-focused intervention increases adherence-related dialogue, but does not improve antiretroviral therapy adherence in persons with HIV. *Journal of acquired immune deficiency syndromes*, 53(3):338.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2012. Towards scalable speech act recognition in twitter: Tackling insufficient training data. *EACL 2012*, page 18.