Unsupervised Modeling of Twitter Conversations

Ritter, Alan; Cherry, Colin; Dolan, Bill

Unsupervised Modeling of Twitter Conversations

Alan Ritter*
Computer Sci. & Eng.
University of Washington
Seattle, WA 98195
aritter@cs.washington.edu

Colin Cherry*
National Research Council Canada
Ottawa, Ontario, K1A 0R6
Colin.Cherry@nrc-cnrc.gc.ca

Bill Dolan
Microsoft Research
Redmond, WA 98052
billdol@microsoft.com

Abstract

We propose the first unsupervised approach to the problem of modeling dialogue acts in an open domain. Trained on a corpus of noisy Twitter conversations, our method discovers dialogue acts by clustering raw utterances. Because it accounts for the sequential behaviour of these acts, the learned model can provide insight into the shape of communication in a new medium. We address the challenge of evaluating the emergent model with a qualitative visualization and an intrinsic conversation ordering task. This work is inspired by a corpus of 1.3 million Twitter conversations, which will be made publicly available. This huge amount of data, available only because Twitter blurs the line between chatting and publishing, highlights the need to be able to adapt quickly to a new medium.

1 Introduction

Automatic detection of dialogue structure is an important first step toward deep understanding of human conversations. **Dialogue acts**[1] provide an initial level of structure by annotating utterances with shallow discourse roles such as "statement", "question" and "answer". These acts are useful in many applications, including conversational agents (Wilks, 2006), dialogue systems (Allen et al., 2007), dialogue summarization (Murray et al., 2006), and flirtation detection (Ranganath et al., 2009).

Dialogue act tagging has traditionally followed an annotate-train-test paradigm, which begins with the design of annotation guidelines, followed by the collection and labeling of corpora (Jurafsky et al., 1997; Dhillon et al., 2004). Only then can one train a tagger to automatically recognize dialogue acts (Stolcke et al., 2000). This paradigm has been quite successful, but the labeling process is both slow and expensive, limiting the amount of data available for training. The expense is compounded as we consider new methods of communication, which may require not only new annotations, but new annotation guidelines and new dialogue acts. This issue becomes more pressing as the Internet continues to expand the number of ways in which we communicate, bringing us e-mail, newsgroups, IRC, forums, blogs, Facebook, Twitter, and whatever is on the horizon.

Previous work has taken a variety of approaches to dialogue act tagging in new media. Cohen et al. (2004) develop an inventory of dialogue acts specific to e-mail in an office domain. They design their inventory by inspecting a large corpus of e-mail, and refine it during the manual tagging process. Jeong et al. (2009) use semi-supervised learning to transfer dialogue acts from labeled speech corpora to the Internet media of forums and e-mail. They manually restructure the source act inventories in an attempt to create coarse, domain-independent acts. Each approach relies on a human designer to inject knowledge into the system through the inventory of available acts.

As an alternative solution for new media, we propose a series of **unsupervised** conversation models, where the discovery of acts amounts to clustering utterances with similar conversational roles. This avoids manual construction of an act inventory, and allows the learning algorithm to tell us something about how people converse in a new medium.

*This work was conducted at Microsoft Research.

[1] Also called "speech acts"

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 172–180,
Los Angeles, California, June 2010.

There is surprisingly little work in unsupervised dialogue act tagging. Woszczyna and Waibel (1994) propose an unsupervised Hidden Markov Model (HMM) for dialogue structure in a meeting scheduling domain, but model dialogue state at the word level. Crook et al. (2009) use Dirichlet process mixture models to cluster utterances into a flexible number of acts in a travel-planning domain, but do not examine the sequential structure of dialogue.[2]

In contrast to previous work, we address the problem of discovering dialogue acts in an informal, open-topic domain, where an unsupervised learner may be distracted by strong topic clusters. We also train and test our models in a new medium: Twitter. Rather than test against existing dialogue inventories, we evaluate using qualitative visualizations and a novel conversation ordering task, to ensure our models have the opportunity to discover dialogue phenomena unique to this medium.

2 Data

To enable the study of large-data solutions to dialogue modeling, we have collected a corpus of 1.3 million conversations drawn from the microblogging service, Twitter. [3] To our knowledge, this is the largest corpus of naturally occurring chat data that has been available for study thus far. Similar datasets include the NUS SMS corpus (How and Kan, 2005), several IRC chat corpora (Elsner and Charniak, 2008; Forsyth and Martell, 2007), and blog datasets (Yano et al., 2009; Gamon et al., 2008), which can display conversational structure in the blog comments.

As it characterizes itself as a micro-blog, it should not be surprising that structurally, Twitter conversations lie somewhere between chat and blogs. Like blogs, conversations on Twitter occur in a public environment, where they can be collected for research purposes. However, Twitter posts are restricted to be no longer than 140 characters, which keeps interactions chat-like. Like e-mail and unlike IRC, Twitter conversations are carried out by replying to specific posts. The Twitter API provides a link from each reply to the post it is responding to, allowing accurate thread reconstruction without requiring a conversation disentanglement step (Elsner and Charniak, 2008). The proportion of posts on Twitter that are conversational in nature are somewhere around 37% (Kelly, 2009).

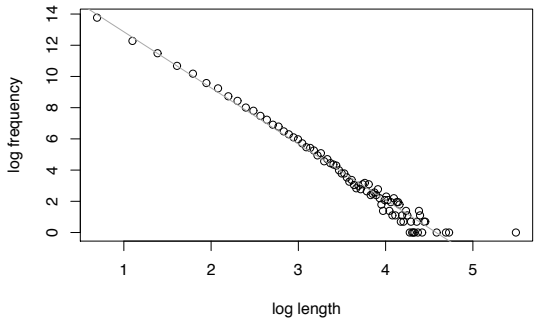

Figure 1: Conversation length versus frequency

To collect this corpus, we crawled Twitter using its publicly available API. We monitored the *public timeline*[4] to obtain a sample of active Twitter users. To expand our user list, we also crawled up to 10 users who had engaged in dialogue with each seed user. For each user, we retrieved all posts, retaining only those that were in reply to some other post. We recursively followed the chain of replies to recover the entire conversation. A simple function-word-driven filter was used to remove non-English conversations.

We crawled Twitter for a 2 month period during the summer of 2009. The resulting corpus consists of about 1.3 million conversations, with each conversation containing between 2 and 243 posts. The majority of conversations on Twitter are very short; those of length 2 (one status post and a reply) account for 69% of the data. As shown in Figure 1, the frequencies of conversation lengths follow a power-law relationship.

While the style of writing used on Twitter is widely varied, much of the text is very similar to SMS text messages. This is likely because many users access Twitter through mobile devices. Posts are often highly ungrammatical, and filled with spelling errors. In order to illustrate the spelling variation found on Twitter, we ran the Jcluster word clustering algorithm (Goodman, 2001) on our cor-

[2] The Crook et al. model should be able to be combined with the models we present here.

[3] Will be available at `http://www.cs.washington.edu/homes/aritter/twitter_chat/`

[4] `http://twitter.com/public_timeline` provides the 20 most recent posts on Twitter

coming comming
enough enought enuff enuf
be4 b4 befor before
yuhr yur your yor ur youur yhur
msgs messages
couldnt culdnt cldnt cannae cudnt couldent
about bou abt abour abut bowt

Table 1: A sample of Twitter spelling variation.

pus, and manually picked out clusters of spelling variants; a sample is displayed in Table 1.

Twitter's noisy style makes processing Twitter text more difficult than other domains. While moving to a new domain (e.g. biomedical text) is a challenging task, at least the new words found in the vocabulary are limited mostly to verbs and nouns, while function words remain constant. On Twitter, even closed-class words such as prepositions and pronouns are spelled in many different ways.

3 Dialogue Analysis

We propose two models to discover dialogue acts in an unsupervised manner. An ideal model will give insight into the sorts of conversations that happen on Twitter, while providing a useful tool for later processing. We first introduce the summarization technology we apply to this task, followed by two Bayesian extensions.

3.1 Conversation model

Our base model structure is inspired by the content model proposed by Barzilay and Lee (2004) for multi-document summarization. Their sentence-level HMM discovers the sequence of topics used to describe a particular type of news event, such as earthquakes. A news story is modeled by first generating a sequence of hidden topics according to a Markov model, with each topic generating an observed sentence according to a topic-specific language model. These models capture the sequential structure of news stories, and can be used for summarization tasks such as sentence extraction and ordering.

Our goals are not so different: we wish to discover the sequential dialogue structure of conversation. Rather than learning a disaster's location is followed by its death toll, we instead wish to learn that a question is followed by an answer. An initial

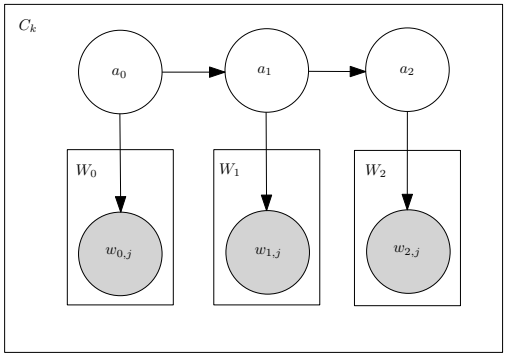

Figure 2: Conversation Model

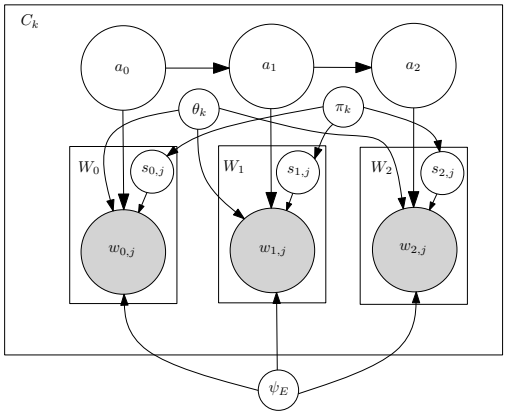

Figure 3: Conversation + Topic Model

conversation model can be created by simply applying the content modeling framework to conversation data. We rename the hidden states *acts*, and assume each post in a Twitter conversation is generated by a single act.[5] During development, we found that a unigram language model performed best as the act emission distribution.

The resulting conversation model is shown as a plate diagram in Figure 2. Each conversation C is a sequence of acts a, and each act produces a post, represented by a bag of words shown using the W plates. The number of acts available to the model is fixed; we experimented with between 5 and 40. Starting with a random assignment of acts, we train our conversation model using EM, with forward-backward providing act distributions during the expectation step. The model structure in Figure 2 is

[5]The short length of Twitter posts makes this assumption reasonable.

sadly no. some **pasta bake**, but **coffee** and **pasta bake** is not a contender for **tea** and **toast**... .
yum! **Ground beef tacos**? We 're **grilling** out. **Turkey dogs** for me, a **Bubba Burger** for my dh, and **combo** for the kids.
ha! They gotcha! You had to think about **Arby's** to write that tweet. **Arby's** is conducting a psychlogical study. Of **roast beef**.
Rumbly tummy soon to be tamed by **Dominos** for **lunch**! **Nom nom nom**!

Table 2: Example of a topical cluster discovered by the EM Conversation Model.

similar to previous HMMs for supervised dialogue act recognition (Stolcke et al., 2000), but our model is trained unsupervised.

3.2 Conversation + Topic model

Our conversations are not restricted to any particular topic: Twitter users can and will talk about anything. Therefore, there is no guarantee that our model, charged with discovering clusters of posts that aid in the prediction of the next cluster, will necessarily discover dialogue acts. The sequence model could instead partition entire conversations into topics, such as *food*, *computers* and *music*, and then predict that each topic self-transitions with high probability: if we begin talking about food, we are likely to continue to do so. Since we began with a content model, it is perhaps not surprising that our Conversation Model tends to discover a mixture of dialogue and topic structure. Several high probability posts from a topic-focused cluster discovered by EM are shown in Table 2. These clusters are undesirable, as they have little to do with dialogue structure.

In general, unsupervised sentence clustering techniques need some degree of direction when a particular level of granularity is desired. Barzilay and Lee (2004) mask named entities in their content models, forcing their model to cluster topics about earthquakes in general, and not instances of specific earthquakes. This solution is not a good fit for Twitter. As explained in Section 2, Twitter's noisiness resists off-the-shelf tools, such as named-entity recognizers and noun-phrase chunkers. Furthermore, we would require a more drastic form of preprocessing in order to mask all topic words, and not just alter the topic granularity. During development, we explored coarse methods to abstract away content while maintaining syntax, such as replacing tokens with either parts-of-speech or automatically-generated word clusters, but we found that these approaches degrade model performance.

Another approach to filtering out topic information leaves the data intact, but modifies the model to account for topic. To that end, we adopt a Latent Dirichlet Allocation, or LDA, framework (Blei et al., 2003) similar to approaches used recently in summarization (Daumé III and Marcu, 2006; Haghighi and Vanderwende, 2009). The goal of this extended model is to separate content words from dialogue indicators. Each word in a conversation is generated from one of three **sources**:

- The current post's dialogue act
- The conversation's topic
- General English

The extended model is shown in Figure 3.[6] In addition to act emission and transition parameters, the model now includes a conversation-specific word multinomial θ_k that represents the topic, as well as a universal general English multinomial ψ_E. A new hidden variable, s determines the source of each word, and is drawn from a conversation-specific distribution over sources π_k. Following LDA conventions, we place a symmetric Dirichlet prior over each of the multinomials. Dirichlet concentration parameters for act emission, act transition, conversation topic, general English, and source become the hyper-parameters of our model.

The multinomials θ_k, π_k and ψ_E create non-local dependencies in our model, breaking our HMM dynamic programing. Therefore we adopt Gibbs sampling as our inference engine. Each hidden variable is sampled in turn, conditioned on a complete assignment of all other hidden variables throughout the data set. Again following LDA convention, we carry out collapsed sampling, where the various multinomials are integrated out, and are never explicitly estimated. This results in a sampling sequence where for each post we first sample its act, and then sample a source for each word in the post. The hidden act and source variables are sampled according to the following transition distributions:

[6] This figure omits hyperparameters as well as act transition and emission multinomials to reduce clutter. Dirichlet priors are placed over all multinomials.

$$P_{trans}(a_i|\mathbf{a}_{-i}, \mathbf{s}, \mathbf{w}) \propto$$
$$P(a_i|\mathbf{a}_{-i}) \prod_{j=1}^{W_i} P(w_{i,j}|\mathbf{a}, \mathbf{s}, \mathbf{w}_{-(i,j)})$$

$$P_{trans}(s_{i,j}|\mathbf{a}, \mathbf{s}_{-(i,j)}, \mathbf{w}) \propto$$
$$P(s_{i,j}|\mathbf{s}_{-(i,j)}) P(w_{i,j}|\mathbf{a}, \mathbf{s}, \mathbf{w}_{-(i,j)})$$

These probabilities can be computed analogously to the calculations used in the collapsed sampler for a bigram HMM (Goldwater and Griffiths, 2007), and those used for LDA (Griffiths and Steyvers, 2004).

Note that our model contains five hyperparameters. Rather than attempt to set them using an expensive grid search, we treat the concentration parameters as additional hidden variables and sample each in turn, conditioned on the current assignment to all other variables. Because these variables are continuous, we apply slice sampling (Neal, 2003). Slice sampling is a general technique for drawing samples from a distribution by sampling uniformly from the area under its density function.

3.3 Estimating Likelihood on Held-Out Data

In Section 4.2 we evaluate our models by comparing their probability on held-out test conversations. As computing this probability exactly is intractable in our model, we employ a recently proposed Chibb-style estimator (Murray and Salakhutdinov, 2008; Wallach et al., 2009). Chibb estimators estimate the probability of unseen data, $P(\mathbf{w})$ by selecting a high probability assignment to hidden variables $\mathbf{h}^*$, and taking advantage of the following equality which can be easily derived from the definition of conditional probability:

$$P(\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{h}^*)}{P(\mathbf{h}^*|\mathbf{w})}$$

As the numerator can be computed exactly, this reduces the problem of estimating $P(\mathbf{w})$ to the easier problem of estimating $P(\mathbf{h}^*|\mathbf{w})$. Murray and Salakhutdinov (2008) provide an unbiased estimator for $P(\mathbf{h}^*|\mathbf{w})$, which is calculated using the stationary distribution of the Gibbs sampler.

3.4 Bayesian Conversation model

Given the infrastructure necessary for the Conversation+Topic model described above, it is straightforward to also implement a Bayesian version of of the conversation model described in Section 3.1. This amounts to replacing the add-x smoothing of dialogue act emission and transition probabilities with (potentially sparse) Dirichlet priors, and replacing EM with Gibbs sampling. There is reason to believe that integrating out multinomials and using sparse priors will improve the performance of the conversation model, as improvements have been observed when using a Bayesian HMM for unsupervised part-of-speech tagging (Goldwater and Griffiths, 2007).

4 Experiments

Evaluating automatically discovered dialogue acts is a difficult problem. Unlike previous work, our model automatically discovers an appropriate set of dialogue acts for a new medium; these acts will not necessarily have a close correspondence to dialogue act inventories manually designed for other corpora. Instead of comparing against human annotations, we present a visualization of the automatically discovered dialogue acts, in addition to measuring the ability of our models to predict post order in unseen conversations. Ideally we would evaluate performance using an end-use application such as a conversational agent; however as this is outside the scope of this paper, we leave such an evaluation to future work.

For all experiments we train our models on a set of 10,000 randomly sampled conversations with conversation length in posts ranging from 3 to 6. Note that our implementations can likely scale to larger data by using techniques such as SparseLDA (Yao et al., 2009). We limit our vocabulary to the 5,000 most frequent words in the corpus.

When using EM, we train for 100 iterations, evaluating performance on the test set at each iteration, and reporting the maximum. Smoothing parameters are set using grid search on a development set.

When performing inference with Gibbs Sampling, we use 1,000 samples for burn-in and take 10 samples at a lag of 100. Although using multiple samples introduces the possibility of poor results due to "act drift", we found this not to be a problem in practice; in fact, taking multiple samples substantially improved performance during development.

Recall that we infer hyperparameters using slice

sampling. The concentration parameters chosen in this manner were always sparse (< 1), which produced a moderate improvement over an uninformed prior.

4.1 Qualitative Evaluation

We are quite interested in what our models can tell us about how people converse on Twitter. To visualize and interpret our competing models, we examined act-emission distributions, posts with high-confidence acts, and act-transition diagrams. Of the three competing systems, we found the Conversation+Topic model by far the easiest to interpret: the 10-act model has 8 acts that we found intuitive, while the other 2 are used only with low probability. Conversely, the Conversation model, whether trained by EM or Gibbs sampling, suffered from the inclusion of general terms and from the conflation of topic and dialogue. For example, the EM-trained conversation model discovered an "act" that was clearly a collection of posts about food, with no underlying dialogue theme (see Table 2).

In the remainder of this section, we reproduce our visualization for the 10-act Conversation+Topic model. Word lists summarizing the discovered dialogue acts are shown in Table 3. For each act, the top 40 words are listed in order of decreasing emission probability. An example post, drawn from the set of highest-confidence posts for that act, is also included. Figure 4 provides a visualization of the matrix of transition probabilities between dialogue acts. An arrow is drawn from one act to the next if the probability of transition is above 0.15.[7] Note that a uniform model would transition to each act with probability 0.10. In both Table 3 and Figure 4, we use intuitive names in place of cluster numbers. These are based on our interpretations of the clusters, and are provided only to benefit the reader when interpreting the transition diagram.[8]

From inspecting the transition diagram (Figure 4), one can see that the model employs three distinct acts to initiate Twitter conversations. These initial acts are quite different from one another, and lead to

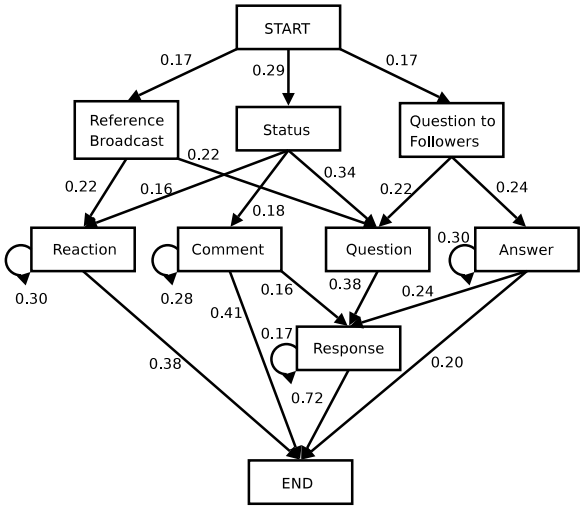

Figure 4: Transitions between dialogue acts. See table 3 for word lists and example posts for each act

different sets of possible responses. We discuss each of these in turn.

The *Status* act appears to represent a post in which the user is broadcasting information about what they are currently doing. This can be seen by the high amount of probability mass given to words like *I* and *my*, in addition to verbs such as *go* and *get*, as well as temporal nouns such as *today*, *tomorrow* and *tonight*.

The *Reference Broadcast* act consists mostly of usernames and urls.[9] Also prominent is the word *rt*, which has special significance on Twitter, indicating that the user is re-posting another user's post. This act represents a user broadcasting an interesting link or quote to their followers. Also note that this node transitions to the *Reaction* act with high probability. *Reaction* appears to cover excited or appreciative responses to new information, assigning high probability to *!*, *!!*, *!!!*, *lol*, *thanks*, and *haha*.

Finally *Question to Followers* represents a user asking a question to their followers. The presence of the question mark and WH question words indicate a question, while words like *anyone* and *know* indicate that the user is asking for information or an opinion. Note that this is distinct from the *Question* act, which is in response to an initial post.

Another interesting point is the alternation be-

[7] After setting this threshold, two Acts were cut off from the rest of the graph (had no incoming edges), and were therefore removed

[8] In some cases, the choice in name is somewhat arbitrary, ie: answer versus response, reaction versus comment.

[9] As part of the preprocessing of our corpus we replaced all usernames and urls with the special tokens *-usr-* and *-url-*.

Status	I . to ! my , is for up in ... and going was today so at go get back day got this am but Im now tomorrow night work tonight off morning home had gon need !! be just getting
	I just changed my twitter page bkgornd and now I can't stop looking at it, lol!!
Question to Followers	? you is do I to -url- what -usr- me , know if anyone why who can " this or of that how does - : on your are need any rt u should people want get did have would tell
	anyone using google voice? just got my invite, should i?? don't know what it is? -url- for the video and break down
Reference Broadcast	-usr- ! -url- rt : -usr-: - " my the , is (you new – ? !!) this for at in follow of on ¡ lol u are twitter your thanks via !!! by :) here 2 please check
	rt -usr-: -usr- word that mac lip gloss give u lock jaw! lol
Question	? you what ! are is how u do the did your that , lol where why or ?? hey about was have who it in so haha on doing going know good up get like were for there :) can
	DWL!! what song is that??
Reaction	! you I :) !! , thanks lol it haha that love so good too your thank is are u !!! was for :d me -usr- ¡ hope ? my 3 omg ... oh great hey awesome - happy now aww
	sweet! im so stoked now!
Comment	you I . to , ! do ? it be if me your know have we can get will :) but u that see lol would are so want go let up well need - come ca make or think them
	why are you in tx and why am I just now finding out about it?! i'm in dfw, till I get a job. i'll have to come to Htown soon!
Answer	. I , you it " that ? is but do was he the of a they if not would know be did or does think) like (as have what in are - no them said who say '
	my fave was "keeping on top of other week"
Response	. I , it was that lol but is yeah ! haha he my know yes you :) like too did well she so its ... though do had no - one as im thanks they think would not good oh
	nah im out in maryland, leaving for tour in a few days.

Table 3: Word lists and example posts for each Dialogue Act. Words are listed in decreasing order of probability given the act. Example posts are in *italics*.

tween the personal pronouns *you* and *I* in the acts due to the focus of conversation and speaker. The *Status* act generates the word *I* with high probability, whereas the likely response state *Question* generates *you*, followed by *Response* which again generates *I*.

4.2 Quantitative Evaluation

Qualitative evaluations are both time-consuming and subjective. The above visualization is useful for understanding the Twitter domain, but it is of little use when comparing model variants or selecting parameters. Therefore, we also propose a novel quantitative evaluation that measures the intrinsic quality of a conversation model by its ability to predict the ordering of posts in a conversation. This measures the model's predictive power, while requiring no tagged data, and no commitment to an existing tag inventory.

Our test set consists of 1,000 randomly selected conversations not found in the training data. For each conversation in the test set, we generate all $n!$ permutations of the posts. The probability of each permutation is then evaluated as if it were an unseen conversation, using either the forward algorithm (EM) or the Chibb-style estimator (Gibbs). Following work from the summarization community (Barzilay and Lee, 2004), we employ Kendall's τ to measure the similarity of the max-probability permutation to the original order.

The Kendall τ rank correlation coefficient measures the similarity between two permutations based on their agreement in pairwise orderings:

$$\tau = \frac{n_+ - n_-}{\binom{n}{2}}$$

where n_+ is the number of pairs that share the same order in both permutations, and n_- is the number that do not. This statistic ranges between -1 and +1, where -1 indicates inverse order, and +1 indicates identical order. A value greater than 0 indicates a positive correlation.

Predicting post order on open-domain Twitter conversations is a much more difficult task than on topic-focused news data (Barzilay and Lee, 2004). We found that a simple bigram model baseline does very poorly at predicting order on Twitter, achieving only a weak positive correlation of $\tau = 0.0358$ on our test data as compared with 0.19-0.74 reported by Barzilay and Lee on news data.

Note that τ is not a perfect measure of model quality for conversations; in some cases, multiple order-

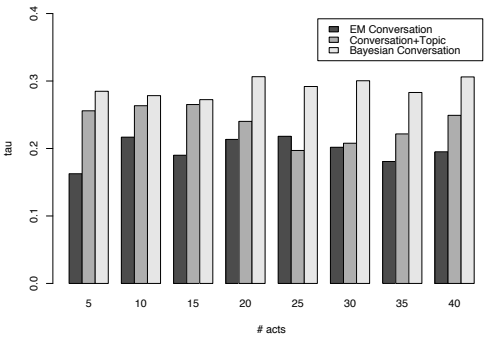

Figure 5: Performance at conversation ordering task.

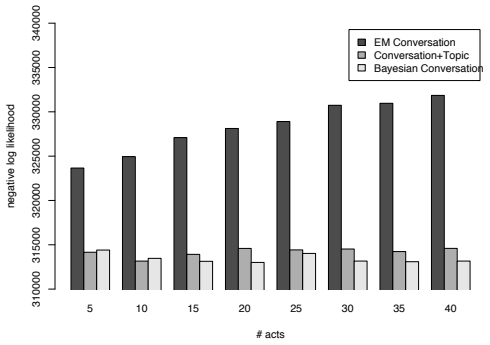

Figure 6: Negative log likelihood on held out test data (smaller values indicate higher likelihood).

ings of the same set of posts may form a perfectly acceptable conversation. On the other hand, there are often strong constraints on the type of response we might expect to follow a particular dialogue act; for example, answers follow questions. We would expect an effective model to use these constraints to predict order.

Performance at the conversation ordering task while varying the number of acts for each model is displayed in Figure 5. In general, we found that using Bayesian inference outperforms EM. Also note that the Bayesian Conversation model outperforms the Conversation+Topic model at predicting conversation order. This is likely because modeling conversation content as a sequence can in some cases help to predict post ordering; for example, adjacent posts are more likely to contain similar content words. Recall though that we found the Conversation+Topic model to be far more interpretable.

Additionally we compare the likelihood of these models on held out test data in Figure 6. Note that the Bayesian methods produce models with much higher likelihood.[10] For the EM models, likelihood tends to decrease on held out test data as we increase the number of hidden states, due to overfitting.

5 Conclusion

We have presented an approach that allows the unsupervised induction of dialogue structure from naturally-occurring open-topic conversational data. By visualizing the learned models, coherent patterns emerge from a stew of data that human readers find difficult to follow. We have extended a conversation sequence model to separate topic and dialogue words, resulting in an interpretable set of automatically generated dialogue acts. These discovered acts have interesting differences from those found in other domains, and reflect Twitter's nature as a micro-blog.

We have introduced the task of conversation ordering as an intrinsic measure of conversation model quality. We found this measure quite useful in the development of our models and algorithms, although our experiments show that it does not necessarily correlate with interpretability. We have directly compared Bayesian inference to EM on our conversation ordering task, showing a clear advantage for Bayesian methods.

Finally, we have collected a corpus of 1.3 million Twitter conversations, which we will make available to the research community, and which we hope will be useful beyond the study of dialogue. In the future, we wish to scale our models to the full corpus, and extend them with more complex notions of discourse, topic and community. Ultimately, we hope to put the learned conversation structure to use in the construction of a data-driven, conversational agent.

Acknowledgements

We are grateful to everyone in the NLP and TMSN groups at Microsoft Research for helpful discussions and feedback. We thank Oren Etzioni, Michael Gamon, Mausam and Fei Wu, and the anonymous reviewers for helpful comments on a previous draft.

[10] Likelihood of the test data is estimated using the Chibb Style estimator described in (Murray and Salakhutdinov, 2008; Wallach et al., 2009). This method under-estimates likelihood in expectation. The maximum likelihood (EM) likelihoods are exact.

References

James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Taysom. 2007. Plow: a collaborative task learning agent. In *Proceedings of AAAI*.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

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

Nigel Crook, Ramon Granell, and Stephen Pulman. 2009. Unsupervised classification of dialogue acts using a Dirichlet process mixture model. In *Proceedings of SIGDIAL*, pages 341–348.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of ACL*.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, International Computer Science Institute.

Micha Elsner and Eugene Charniak. 2008. You talking to me? A corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-HLT*.

Eric N. Forsyth and Craig H. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *Proceedings of ICSC*.

Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian Knig. 2008. Blews: Using blogs to provide context for news articles. In *Proceedings of ICWSM*.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, pages 744–751.

Joshua T. Goodman. 2001. A bit of progress in language modeling. Technical report.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proc Natl Acad Sci*, 101 Suppl 1:5228–5235.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of HLT-NAACL*, pages 362–370.

Yijue How and Min-Yen Kan. 2005. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of HCII*.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of EMNLP*, pages 1250–1259.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado Institute of Cognitive Science.

Ryan Kelly. 2009. Pear analytics twitter study. Whitepaper, August.

Iain Murray and Ruslan Salakhutdinov. 2008. Evaluating probabilities under high-dimensional latent variable models. In *Proceedings of NIPS*, pages 1137–1144.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of HLT-NAACL*, pages 367–374.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31:705–767.

Rajesh Ranganath, Dan Jurafsky, and Dan Mcfarland. 2009. It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of EMNLP*, pages 334–342.

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

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. 2009. Evaluation methods for topic models. In *Proceedings of ICML*, page 139.

Yorick Wilks. 2006. Artificial companions as a new kind of interface to the future internet. In *OII Research Report No. 13.*

M. Woszczyna and A. Waibel. 1994. Inferring linguistic structure in spoken language. In *Proceedings of ICSLP*.

Tae Yano, William W. Cohen, and Noah A. Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of NAACL*, pages 477–485.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of KDD*, pages 937–946.