Unsupervised Modeling of Dialog Acts in Asynchronous Conversations

Shafiq Joty[*] and **Giuseppe Carenini**
Department of Computer Science
University of British Columbia, Canada
{rjoty, carenini}@cs.ubc.ca

Chin-Yew Lin
Microsoft Research Asia
Beijing, China
cyl@microsoft.com

Abstract

We present unsupervised approaches to the problem of modeling dialog acts in asynchronous conversations; i.e., conversations where participants collaborate with each other at different times. In particular, we investigate a graph-theoretic deterministic framework and two probabilistic conversation models (i.e., HMM and HMM+Mix) for modeling dialog acts in emails and forums. We train and test our conversation models on (a) temporal order and (b) graph-structural order of the datasets. Empirical evaluation suggests (i) the graph-theoretic framework that relies on lexical and structural similarity metrics is not the right model for this task, (ii) conversation models perform better on the graph-structural order than the temporal order of the datasets and (iii) HMM+Mix is a better conversation model than the simple HMM model.

1 Introduction

Email and discussion fora are examples of social media that have become extremely popular for allowing people to discuss events, issues, queries and personal experiences [Baron, 2008]. Although people communicate in writing (as opposed to speech) in these media, the nature of the interaction is conversational, in the sense that once a participant initiated the communication all the other contributions are replies to previous ones. That is, the participants take turns and each turn consists of a joint action of writing and reading (though at different times). In such conversations participants interact with each other in complex ways, performing different **dialog acts (DA)**[1] (e.g., 'question', 'answer', 'request') to achieve their communicative goals. The two-part structures across messages (or posts) like 'question-answer', 'request-grant', 'complement-downplayer' (e.g., congrats-thanks), 'greeting-greeting' (e.g., hi-hello) are called **adjacency pairs**.

Uncovering the rich conversational structure is an important step towards deep conversation analysis in these media. Annotating utterances with DAs (Figure 1) provides an initial level of structure and has been shown to be useful for spoken dialog, in many applications including meeting summarization [Murray *et al.*, 2006; 2010], collaborative task learning agents [Allen *et al.*, 2007], artificial companions for people to use the Internet [Wilks, 2006] and flirtation detection in speed-dates [Ranganath *et al.*, 2009]. We believe that similar benefits will also hold for written asynchronous conversation.

Traditional approaches to DA tagging have been mostly supervised [Stolcke *et al.*, 2000; Bangalore *et al.*, 2006]. This learning strategy has been quite successful, but it is very domain specific and labor intensive. Arguably, as the number of social media grows (e.g., email, blogs, Facebook) and the number of communicative settings in which people interact through these media also grows, the supervised paradigm of 'label-train-test' becomes too expensive and unrealistic. Every novel use of a new media may require not only new annotations, but possibly also new annotation guidelines and new DA tagsets. In contrast, the approach we present in this paper is to adopt an unsupervised paradigm, which is more robust across new forms of media and new domains. In particular, we investigate a graph-theoretic deterministic framework and a set of probabilistic conversation models for DA tagging. Note that our unsupervised models do not label each sentence with a specific DA, but they cluster sentences so that each cluster should contain sentences expressing the same DA. The DA label for each cluster needs to be then determined through other means, which we do not explore in this paper, but may include minimal supervision.

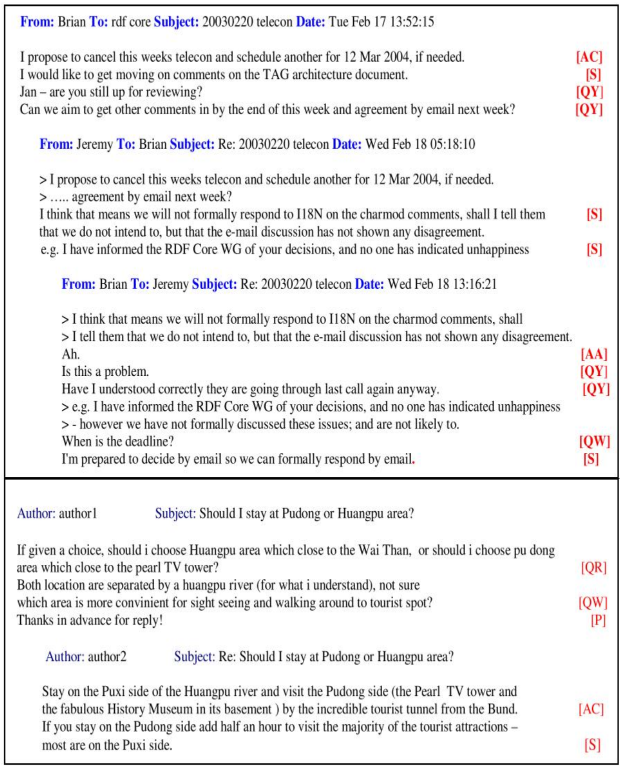
The graph-theoretic framework, which has already been successfully applied to several NLP tagging tasks (e.g., [Malioutov and Barzilay, 2006; Joty *et al.*, 2010; Elsner and Charniak, 2010]), clusters sentences that are similar to each other by only relying on lexical and structural similarity between the sentences, without considering the sequential nature of the conversation. At the end of this process, sentences in the same cluster receive the same DA tag. The performance of this model crucially depends on how one measures the similarity between two sentences.

Quite differently, the conversation models frame DA tagging as a sequence-labeling problem that can be solved with variations of the Hidden Markov Model (HMM) paradigm, by assuming that a conversation is a sequence of hidden DAs, with each DA emitting an observed sentence. In this case, the performance of a model depends on the accuracy of both the sequence dependencies between hidden DAs (e.g., 'question'

[*] This work was conducted at Microsoft Research Asia.

[1] Also known as 'speech act' in the literature.

followed by ‘answer’, ‘request’ followed by ‘accept’) and the act emission distribution. While the idea of using probabilistic techniques for sequence labeling to perform DA tagging is not new, we make several key contributions in showing how it can be effectively applied to asynchronous conversations by dealing with critical limitations in previous work.



From: Brian **To:** rdf core **Subject:** 20030220 telecon **Date:** Tue Feb 17 13:52:15

I propose to cancel this weeks telecon and schedule another for 12 Mar 2004, if needed. [AC]

I would like to get moving on comments on the TAG architecture document. [S]

Jan – are you still up for reviewing? [QY]

Can we aim to get other comments in by the end of this week and agreement by email next week? [QY]

From: Jeremy **To:** Brian **Subject:** Re: 20030220 telecon **Date:** Wed Feb 18 05:18:10

> I propose to cancel this weeks telecon and schedule another for 12 Mar 2004, if needed.
> ….. agreement by email next week?

I think that means we will not formally respond to I18N on the charmod comments, shall I tell them that we do not intend to, but that the e-mail discussion has not shown any disagreement. [S]

e.g. I have informed the RDF Core WG of your decisions, and no one has indicated unhappiness [S]

From: Brian **To:** Jeremy **Subject:** Re: 20030220 telecon **Date:** Wed Feb 18 13:16:21

> I think that means we will not formally respond to I18N on the charmod comments, shall
> I tell them that we do not intend to, but that the e-mail discussion has not shown any disagreement.

Ah. [AA]

Is this a problem. [QY]

Have I understood correctly they are going through last call again anyway. [QY]

> e.g. I have informed the RDF Core WG of your decisions, and no one has indicated unhappiness
> - however we have not formally discussed these issues; and are not likely to.

When is the deadline? [QW]

I'm prepared to decide by email so we can formally respond by email. [S]

Author: author1 Subject: Should I stay at Pudong or Huangpu area?

If given a choice, should i choose Huangpu area which close to the Wai Than, or should i choose pu dong area which close to the pearl TV tower? [QR]

Both location are separated by a huangpu river (for what i understand), not sure which area is more convinient for sight seeing and walking around to tourist spot? [QW]

Thanks in advance for reply! [P]

Author: author2 Subject: Re: Should I stay at Pudong or Huangpu area?

Stay on the Puxi side of the Huangpu river and visit the Pudong side (the Pearl TV tower and the fabulous History Museum in its basement) by the incredible tourist tunnel from the Bund. [AC]

If you stay on the Pudong side add half an hour to visit the majority of the tourist attractions – most are on the Puxi side. [S]

Figure 1: Sample truncated email (top) and forum (down) conversations from our corpora annotated with dialog acts (right most column). The tags are defined in Table 1.

In synchronous conversations (e.g., meeting, telephone), the conversational flow has a sequential structure (e.g., ‘question’ followed by ‘answer’, ‘request’ followed by ‘grant’). Therefore, one can accurately learn the sequence dependencies in a sequence labeler from the temporal ordering (i.e., arrival time) of the utterances. But, in asynchronous conversations (e.g., email, forum), the conversational flow is not sequential. Hence, the temporal ordering of the utterances is not the only option and could even be misleading. For example, consider the sample email and forum conversations shown in Figure 1. One can notice that if we arrange the sentences (excluding the quoted sentences) as they arrive, it is hard to capture the referential relation between sentences of different posts. This could lead to inaccurate sequential dependencies when we apply a sequence-labeler on this (temporal) order of the sentences. It is our key hypothesis that the graph structure of the conversations is rather more effective to accurately learn the sequence dependencies. If two sentences are close in the thread structure of the conversations, they are likely to be related (and to express related DAs), independently from their arrival time-stamp. We will show that this is the case for both fora and email, in particular when for email we use a finer level graph structure of the conversations (i.e., the Fragment Quotation Graph (FQG) [Carenini *et al.*, 2008]).

Previous work in unsupervised DA tagging mostly uses only unigram language model as the act emission distribution [Ritter *et al.*, 2010]. However, there are other features that proved to be beneficial in the supervised settings, such as ‘speaker’, ‘relative position’ and ‘sentence length’ [Jeong *et al.*, 2009; Kim *et al.*, 2010]. In our work, we expand the set of sentence features considered in the act emission distribution, including not only unigrams but also the speaker, its relative position and length. Another limitation of the approach presented in [Ritter *et al.*, 2010] is that their conversation (HMM) model, with the unigram language model as the emission distribution, tends to find topical clusters, in addition to DA clusters. They address this problem with an HMM+Topic model, which tries to separate the topic words from the DA indicators. In this paper, we propose a more adequate HMM+Mix model, which not only explains away the topics, but also improves the act emission distribution by defining it as a mixture model.

In what is to the best of our knowledge the first quantitative evaluation of unsupervised DA tagging for asynchronous conversations, we show that (i) graph-theoretic framework is not the right model for this task, (ii) conversation models perform better on the graph-structural order than the temporal order of the conversations and (iii) HMM+Mix is a better conversation model than the simple HMM model.

2 Related Work

There has been little work so far on DA recognition in asynchronous conversation. [Cohen *et al.*, 2004] was the first to use the notion of ‘email speech act’ to classify the *emails* (not sentences) in the office domain. However, their inventory of DAs is specific to that domain (e.g., deliver, meeting) and they use a supervised approach to identify them. [Ravi and Kim, 2007] also use supervised learning methods to detect the question and answer passages in email discussions.

[Jeong *et al.*, 2009] use semi-supervised boosting to tag the *sentences* in emails and forums with DAs by adapting knowledge from annotated spoken conversations (i.e., MRDA meeting and DAMSL telephone conversation). They derive a coarse, domain-independent inventory of twelve DAs from the MRDA tagset. Then, given an utterance represented as a set of trees (i.e., dependency tree, n-gram tree and part of speech (POS) tree), the boosting algorithm iteratively learns the best feature set (i.e., sub-trees) that minimizes the errors in the training data. Although, this work tries to adapt from synchronous conversations to asynchronous conversations, it still relies on annotated corpora. Furthermore, it does not consider the sequential structure of conversations, something we successfully exploit in this paper.

[Ritter *et al.*, 2010] propose three HMM-based unsupervised conversation models for DA recognition in Twitter. In particular, they use HMM, Bayesian HMM and HMM+Topic models to cluster the Twitter posts (not the sentences) into DAs. Since, they use a unigram model to define the emission distribution, their simple HMM model finds topical clusters in addition to DA clusters (i.e., sentences are similar because they talk about the same topic not because they play the same discourse role). The HMM+Topic model tries to separate the DA indicator from the topic words. By visualizing the type of conversations found by the two models they show that the output of the HMM+Topic model is more interpretable than that of the HMM one, however, their classification accuracy is not empirically evaluated. Therefore, it is not clear whether these models are actually useful (i.e., beat the baseline), and which of the two models is a better DA tagger.

Our conversation models were inspired by the models of [Ritter *et al.*, 2010], but, we improve on those by making the following four key contributions: (i) We model at the finer level of granularity of the sentence (as opposed to a whole post) with a richer feature set including not only unigram but also sentence relative position, sentence length and speaker. (ii) Our models exploit the graph structure of the conversations. (iii) Our HMM+Mix model not only explains away the topics (like HMM+Topic does), but also improves the emission distribution by defining it as a mixture model [Bishop, 2006]. (iv) We provide classification accuracy of the models on two corpora (email and forum) by applying the 1-to-1 metric from [Elsner and Charniak, 2010].

3 Data Preparation

3.1 Dataset Selection and Clean up

We have used the same DA tagset and test datasets used in [Jeong *et al.*, 2009]. The tagset containing 12 act categories with their relative frequency in the email and forum test corpora is shown in Table 1. This inventory of DAs is originally adopted from the MRDA tagset [Dhillon *et al.*, 2004]. This tagset is different from the prior work on DA recognition in asynchronous conversations (e.g., [Cohen *et al.*, 2004; Ravi and Kim, 2007]), since it is domain independent and suitable for sentence level annotation. Our test datasets include: (i) 40 email threads of the BC3 corpus [Ulrich *et al.*, 2008] which were originally taken from the W3C corpus, and (ii) 200 forum threads from the TripAdvisor travel forum site[2]. The act categories have similar distribution in the two corpora. The κ agreements between two human annotators were 0.79 for email dataset and 0.73 for forum dataset.

Due to privacy issues, there are only a few email corpora available for training an unsupervised system (e.g., Enron, W3C). Since it is preferable to train and test such a system on similar data, we choose the W3C email corpus to train our models[3]. W3C contains $23,957$ email threads, however, the raw data is too noisy to directly inform our models, as it contains system messages and signatures. We cleaned up the data with the intention to keep only the headers, bodies and quotations. By processing the headers, we then reconstruct the thread structure of the email conversations. For the forum data, we crawled $25,000$ forum threads from the same travel forum site i.e., TripAdvisor. Our forum data is much less noisy, but does not contain any thread structure.

Tag	Description	Email	Forum
S	Statement	69.56%	65.62%
P	Polite mechanism	6.97%	9.11%
QY	Yes-no question	6.75%	8.33%
AC	Action motivator	6.09%	7.71%
QW	Wh-question	2.29%	4.23%
A	Accept response	2.07%	1.10%
QO	Open-ended question	1.32%	0.92%
AA	Acknowledge and appreciate	1.24%	0.46%
QR	Or/or-clause question	1.10%	1.16%
R	Reject response	1.06%	0.64%
U	Uncertain response	0.79%	0.65%
QH	Rhetorical question	0.75%	0.08%

Table 1: Dialog act tag categories and their relative frequency.

3.2 Dealing with the Conversational Structure

In the probabilistic models, the sequence dependencies between DAs can be learned either from the simple temporal order of the contributions to the conversation, or from the more refined graph-structure of the conversation threads. We create a temporal ordered conversation by simply arranging its posts based on their temporal relation (i.e., arrival time). We create the graph-structural order of the conversations in two steps. First, we discover the graph structure of the conversations. Second, we derive our data from this structure. Below, we describe these two steps for emails and fora.

We extract the finer level graph structure of email conversations in the form of Fragment Quotation Graph (FQG). We demonstrate how to build a FQG (Figure 2 (b)) through the example email thread involving 7 emails (Figure 2 (a)) taken from our corpus. For convenience we do not show the real content but abbreviate them as a sequence of fragments.

In the first pass, by processing the whole thread, we identify the new (i.e., quotation depth 0) and quoted (i.e., quotation depth $>$ 0) fragments based on the usage of quotation (e.g., '$>$', '$\&gt$') marks. For instance, email E_3 contains two new fragments (f, g), and two quoted fragments (d, e) of depth 1. E_2 contains abc (quoted) and de (new) fragments. Then in the second step, we compare the fragments with each other and based on the overlap we find the distinct fragments. If necessary we split the fragments in this step. For example, de in E_2 is divided into d and e distinct fragments when compared with the fragments of E_3. This process gives 15 distinct fragments which constitute the vertices of the FQG. In the third step, we compute the edges, which represent referential relations between fragments. For simplicity we assume that any new fragment is a potential reply to its neighboring quoted fragments. For example, for the fragments of E_4 we create two edges from h ((h,a),(h,b)) and one edge from i ((i,b)). We then remove the redundant edges. In E_6 we found the edges (n,h), (n,a) and (n,m). As (h,a) is already there we exclude (n,a). If an email does not quote anything, then its fragments are connected to the fragments of the email to which it replies, revealing the original 'reply-to' relation.

TripAdvisor's forum conversations do not contain any thread structure and people hardly quote others' utterances in this forum. However, we have noticed that participants almost always respond to the initial post of the thread, and mention

[2] http://tripadvisor.com

[3] In contrast, [Jeong *et al.*, 2009] train on Enron and test on BC3.

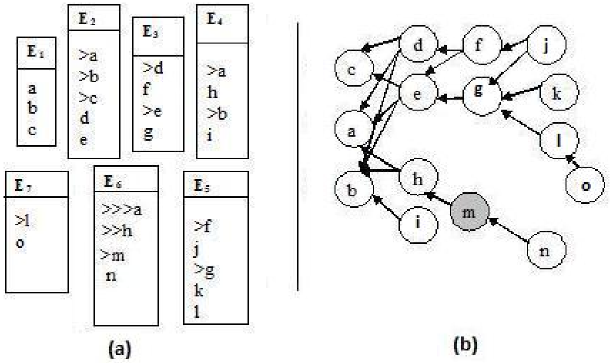


Figure 2: Fragment Quotation Graph for emails

other participants' name to respond to their post. Therefore, we create the graph structure of a forum conversation with the simple assumption that a post usually responds to the initial post unless it mentions other participants' names.

In the graph structure of the conversations, one can notice that the paths (e.g., c-d-f-j in Figure 2 (b)) capture the adjacency relations between email fragments or forum posts. And as we noted in the introduction, our assumption is that if two sentences are close in the thread structure of the conversations, they are likely to be related (and to express related DAs). Based on this assumption, we prepare our graph-structural conversations by arranging the sentences according to the paths of the graph. Note that, in this order, the sentences in the common nodes shared by multiple paths (e.g., c, e, g) are duplicated. In Section 5, we describe how our conversational models deal with the duplicated sentences.

4 Graph-theoretic Framework

Our first model for dialog act tagging is built on a graph-theoretic framework, which has been successfully used in many NLP tasks, including topic segmentation of spoken lectures [Malioutov and Barzilay, 2006] and emails [Joty *et al.*, 2010], and disentanglement of multi-party chat [Elsner and Charniak, 2010]. We investigate whether the same framework can be adapted to clustering sentences of a forum or email conversation into dialog acts.

4.1 Algorithm Description

In this framework, at first we form a complete similarity graph $G = (V, E)$, where the nodes V represent the sentences of a conversation and the edge-weights represent the similarity between two nodes (i.e., for an edge $(u, v) \in E$, edge-weight $w(u, v)$ represents how similar the sentences u and v are). We then formulate the clustering problem as a N-mincut graph-partitioning problem with the intuition that sentences in a cluster should be similar to each other, while sentences in different clusters should be dissimilar. To do this, we try to optimize the 'normalized cut' criterion:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(B, A)}{assoc(B, V)} \qquad (1)$$

where $cut(A, B) = \Sigma_{u \in A, v \in B} w(u, v)$ is the total connection from nodes in partition A to nodes in partition B, $assoc(A, V) = \Sigma_{u \in A, t \in V} w(u, t)$ is the total connection from nodes in partition A to all nodes in the graph and $assoc(B, V)$ is similarly defined. Previous research on graph-based clustering (e.g., [Malioutov and Barzilay, 2006], [Shi and Malik, 2000]) has shown that the 'normalized cut' criterion is more appropriate than just the 'cut' criterion which accounts only for total edge weight connecting A and B and therefore, favors cutting small sets of isolated nodes in the graph. However, solving 'normalized cut' is NP-complete. Hence, we approximate the solution following [Shi and Malik, 2000], which is time efficient and has been successfully applied to image segmentation. Note that, depending on the task, the performance of this framework depends on how one measures the similarity between two sentences.

Unigrams or bag-of-words (BOW) features have been quite extensively used in the previous work on DA recognition (e.g., [Ritter *et al.*, 2010; Kim *et al.*, 2010]). To measure the BOW-based similarity between two sentences we represent each as a vector of TF.IDF values of the words and compute the cosine of the angle in between the vectors. Since this framework with BOW features has been quite successful for finding topical clusters [Malioutov and Barzilay, 2006], even though we retain the stop-words and punctuations which are arguably useful for finding DA clusters, it may still find topical clusters in addition to DA clusters. In an attempt to abstract away the topic words, we mask the nouns[4] in the sentences and measure the BOW-based similarity as before.

The BOW similarity metric described above does not consider the order of the words. One can use n-gram co-occurrences to account for the order of the words. The Word Subsequence Kernel (WSK) [Cancedda *et al.*, 2003], which is an improvement over n-gram co-occurrences, considers the order by transforming the sentences into higher dimensional spaces and then measuring the similarity in that space. Extended String Subsequence Kernel (ESK) [Hirao *et al.*, 2004] which is a simple extension of WSK allows one to incorporate word-specific syntactic and/or semantic (e.g., word sense, part of speech (POS)) information into WSK. Since [Jeong *et al.*, 2009] found n-grams and POS-tags useful, we implement the WSK and the ESK with POS-tags of the words.

The work of [Jeong *et al.*, 2009] also suggests that subtrees of the dependency tree are important features for DA tagging. We measure the dependency similarity between two sentences by extracting the Basic Elements (BE) and counting the number of co-occurrences. The "head-modifier-relation" triples, extracted from the dependency trees of the sentences are considered as BEs in our experiment. The triples encode some syntactic and semantic information and one can quite easily decide whether any two units match considerably more easily than with longer units [Hovy *et al.*, 2005].

Like dependency tree, the sub-trees of the syntactic tree may also be important indicators for DA tagging. To measure the syntactic similarity between two sentences we first parse the sentences into the syntactic trees using Charniak parser and then compute the similarity between the two trees using

[4] Since nouns are arguably the most indicative for topics.

the Tree Kernel (TK) function [Collins and Duffy, 2001].

We also measure a combined similarity (**All** in Table 2) between two sentences by taking a linear combination of the above mentioned similarity metrics.

4.2 Evaluation of the Graph-theoretic Clustering

We wish to compare the DAs automatically discovered by our models with the human annotations. However, unsupervised clustering techniques do not assign any label to the clusters. Therefore, metrics widely used in supervised classification, such as κ statistic, F_1 score, are not applicable. In this paper, we propose to use the 1-to-1 metric introduced recently by [Elsner and Charniak, 2010]. Given two annotations (model's output and human annotation), it pairs up the clusters from the two annotations in a way that maximizes (globally) the total overlap and then reports the percentage of overlap.

The number of DAs (clusters) available to the systems was fixed to 12. Table 2 shows the 1-to-1 accuracy of the graph-theoretic framework with various similarity metrics. At the right most column we show the baseline (BL) system that considers all the utterances of a corpus as 'statement', since 'statement' is the majority class in both corpora.

	BOW	BOW-M	WSK	ESK-P	BE	TK	All	BL
Email	62.6	34.3	64.7	24.8	39.1	22.5	26.0	69.6
Forum	65.0	38.2	65.8	36.3	46.0	30.1	32.2	65.6

Table 2: 1-to-1 accuracy for different similarity metrics in the graph-theoretic framework. BOW-M refers to BOW with masking the nouns and ESK-P refers to ESK with POS.

One can notice that all the systems fail to beat the baseline indicating that this framework is not the right model for recognizing DAs in these corpora. When we compare the BOW and the BOW-M (i.e., BOW with masking the nouns) systems, we can observe that BOW performs way better than the BOW-M. This indicates that masking the nouns in an attempt to abstract away the topic words degrades the performance substantially. The WSK system performs slightly better than the BOW system meaning that considering the order of the words in the similarity metric is useful. However, when we add the POS of the words in the ESK (ESK-P in table 2), we get large decrease in the accuracy. This means that the POS similarity between sentences has adverse effect on clustering sentences into DAs. The results of the BE and TK systems indicate that the shallow syntactic (i.e., dependency) and the deep syntactic similarity between sentences also are not useful for recognizing DAs in this framework.

5 Probabilistic Conversation Models

The graph-theoretic clustering framework discussed above has three main limitations when applied to find DAs in conversations. First, this framework does not model the potentially informative sequential structure of the conversation (e.g., 'question' followed by 'answer', 'request' followed by 'accept'). Second, this framework seems to be still confused by topical clusters even when nouns are masked, word order, POS and syntactic features are considered. Third, unlike our conversation models (discussed below), this framework does not allow us to incorporate other important features (speaker, relative position, length) in a principled way. To address these limitations we propose two probabilistic conversation models which assume that a conversation is a Markov sequence of hidden DAs, with each DA emitting an observed sentence.

5.1 HMM Conversation Model

Figure 3 shows our first conversation model in plate notation. A conversation C_k is a sequence of hidden DAs D_i, and each DA produces an observed sentence X_i, represented by its (i) bag-of-words (i.e., unigrams) shown in the W_i plate, (ii) author or speaker S_i, (iii) relative position P_i, and (iv) length L_i. These features are discrete valued, therefore, we model them as multinomial distributions. Following [Ritter *et al.*, 2010], for unigrams, we limit our vocabulary to the $5,000$ most frequent words in the corpus. The relative position of a sentence is computed by dividing its position in the post with the number of sentences in the post. We then convert the relative positions and lengths to a sequence of natural numbers.

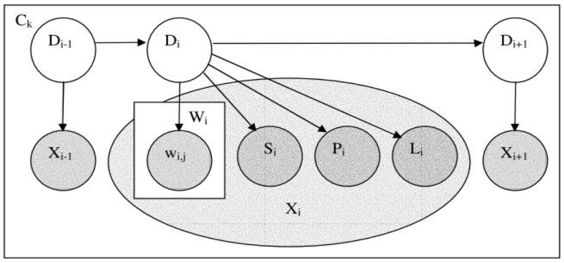


Figure 3: HMM conversation model

We place a symmetric Dirichlet prior with hyperparameter $\alpha = 2$ over each of the six multinomials (i.e., distributions over initial states, transitions, unigrams, speakers, positions and lengths)[5]. We compute maximum a posterior (MAP) estimate of the parameters using the Baum-Welch (EM) algorithm with forward-backward providing the smoothed node and edge marginals for each sequence in E-step. Specifically, given n-th sequence $X_{n,1:T_n}$, forward-backward computes:

$$\gamma_{n,i}(j) := p(D_i = j | X_{n,1:T_n}, \theta) \quad (2)$$

$$\xi_{n,i}(j,k) := p(D_{i-1} = j, D_i = k | X_{n,1:T_n}, \theta) \quad (3)$$

Where the local evidence is given by:

$$p(X_i|D_i) = \left[\prod_j p(W_{i,j}|D_i) \right] p(S_i|D_i)p(P_i|D_i)p(L_i|D_i) \quad (4)$$

5.2 HMM+Mix Conversation Model

Our HMM conversation model described above is similar to the conversation model of [Ritter *et al.*, 2010] except that, in addition to the unigrams of a sentence, we also use its relative

[5] We do not show the parameters and the hyperparameters in figure 3 and 4 to reduce visual clutter.

position, speaker and length to define the emission distributions. However, as they describe, without additional guidance an unsupervised clustering model may find clusters which are not desired. For example, their model with unigrams finds some topical clusters. The features (i.e., unigrams, speaker, relative position, length) used in our model are also indicators for finding topical clusters [Joty *et al.*, 2011]. Therefore, our model with the above features may also find some unwanted clusters. Like them we have also noticed in our graph-theoretic framework that masking nouns in an attempt to abstract away the topics even degrades the performance. As a solution [Ritter *et al.*, 2010] propose the HMM+Topic model to separate the topic words from the DA indicators. In this model, a word is generated from one of these three hidden sources: (i) DA, (ii)Topic and (iii) General English. In contrast, we propose the HMM+Mix model where the emission distribution is defined as a mixture model. This model has two main advantages over the HMM+Topic model: (a) By defining the emission distribution as a mixture model we not only indirectly explain away the topic but also enrich the emission distribution, since the mixture models (i.e., observed W, S, P, L conditioned on hidden M) can define finer distributions [Bishop, 2006] and (b) Learning and inference in this model is much easier (using EM) without requiring approximate inference techniques such as Gibbs sampling.

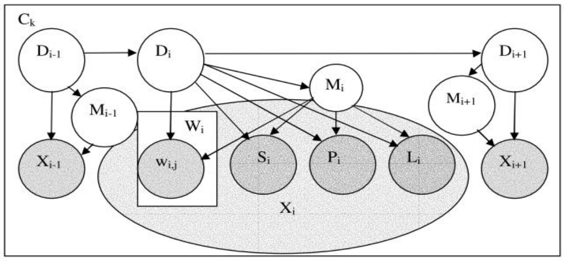


Figure 4: HMM+Mix conversation model

Figure 4 shows the extended HMM+Mix model. In this model, emission distributions are mixtures of multinomials with $M_i \in \{1, \ldots, M\}$ representing the mixture component.

One can use EM to compute the MAP estimate of the parameters in this model, where the local evidence is given by:

$$p(X_i|D_i) = \sum_{M_i} p(M_i|D_i)p(X_i|D_i, M_i) \qquad (5)$$

Where $p(X_i|D_i, M_i) =$
$\left[\prod_j p(W_{i,j}|D_i, M_i)\right] p(S_i|D_i, M_i)p(P_i|D_i, M_i)p(L_i|D_i, M_i)$
In this model, in E-step, in addition to $\gamma_{n,i}(j)$ and $\xi_{n,i}(j, k)$ (equation 2, 3), for each sequence we also need to compute:

$$\tau_{n,i}(j, k) := p(D_i = j, M_i = k|X_{n,1:T_n}, \theta) \qquad (6)$$

One can show that this is given by the following expression:

$$\tau_{n,i}(j, k) := \gamma_{n,i}(j) \frac{p(M_i = k|D_i = j)p(X_i|D_i = j, M_i = k)}{\sum_m p(M_i = m|D_i = j)p(X_i|D_i = j, M_i = m)}$$

In EM, we must ensure that we initialize the parameters carefully, to minimize the chance of getting stuck in poor local optima. We use multiple (10) restarts and pick the best solution based on the likelihood of the data. For the first restart we ignore the Markov dependencies, and estimate the observation parameters using the standard mixture model estimation method (EM) and use it to initialize other parameters. For the other 9 restarts we randomly initialize the parameters. We use this process of initialization in an attempt to ensure that our models are at least as good as the mixture model which ignores the sequential structure of a conversation.

With the learned parameters, given an observed test sequence (i.e., conversation), we use Viterbi decoding to infer the most probable DA sequence. As described in section 3, in the graph-structural order of the conversations, the sentences in the common nodes are duplicated in the sequence. As a result, when we run Viterbi decoding on it, for the same sentence we get multiple DA assignments (one for each position). We take the maximum vote to finally assign its DA.

6 Evaluation of the Conversation Models

For all experiments we train our models on a set of $12,000$ conversations having at least two posts. These are randomly sampled from each of the corpora. We do this 50 times, evaluating performance on the test sets at each time. The number of DAs available to the models was set to 12. In the HMM+Mix model, the number of mixture component (M)[6] was set to 3.

Table 3 presents the mean 1-to-1 accuracy of the conversation models and the baseline on the temporal order and the graph-structural order of the datasets. Notice that the conversation models beat the baseline in both corpora, proving their effectiveness in DA recognition. We get similar results for the models when applied to email and forum corpora. Both models benefit from the graph-structural order of the conversations (p<0.05). We can conclude that our models learn better sequential dependencies with the graph-structural order of the conversations. Therefore, the finer referential structure of email conversations in the form of FQG and the assumed referential structure of forum conversations have been proved to be beneficial for recognizing DAs in these corpora. By comparing the performance of the models on the datasets we can see that the HMM+Mix model outperforms the HMM model (p<0.05). This indicates that the mixture model as the act emission distribution not only explains the topics away but also defines a finer observation model.

	Email		Forum	
	Temporal	Graph	Temporal	Graph
Baseline	70.00	70.00	66.00	66.00
HMM	73.45	76.81	69.67	74.41
HMM+Mix	76.73	**79.66**	75.61	**78.35**

Table 3: Mean 1-to-1 accuracy of the conversation models

[6] We experimented with M={1, 2, 3, 4, 5}, and got the highest accuracy with M=3.

7 Conclusion and Future Work

In our investigation of approaches for modeling DAs in asynchronous conversations we have made several key contributions. We apply a graph-theoretic framework to the DA tagging task and compare it with probabilistic sequence-labeling models. Then, we show how in the probabilistic models the sequence dependencies can be more effectively learned by taking the conversational structure into account. After that, we successfully expand the set of sentence features considered in the act emission distribution. Finally, we improve the act emission distribution by applying a mixture model. Quantitative evaluation with human annotations shows that while the graph-theoretic framework is not the right model for this task, the probabilistic conversation models (i.e., HMM and HMM+Mix) are quite effective and their benefits are more pronounced with graph-structural conversational order as opposed to the temporal one. Comparison of the outputs of these models reveals that HMM+Mix model can predict the DAs better than the HMM model. In the future, we wish to experiment with the Bayesian versions of these conversation models and also apply our models to other conversational modalities.

Acknowledgments

This work was supported in part by NSERC CGS-D award, NSERC BIN project and Institute for Computing, Information and Cognitive Systems (ICICS) at UBC.

References

[Allen *et al.*, 2007] J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, and W. Taysom. Plow: A collaborative task learning agent. In *AAAI-07*, pages 22–26, 2007.

[Bangalore *et al.*, 2006] S. Bangalore, G. Di Fabbrizio, and A. Stent. Learning the structure of task-driven human-human dialogs. In *ACL-44*, pages 201–208. ACL, 2006.

[Baron, 2008] Naomi S. Baron. Always on: Language in an online and mobile world. *Oxford ; New York : Oxford University Press*, ISBN 978-0-19-531305-5, 2008.

[Bishop, 2006] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Cancedda *et al.*, 2003] N. Cancedda, E. Gaussier, C. Goutte, and J. M. Renders. Word sequence kernels. *JMLR*, 3:1059–1082, 2003.

[Carenini *et al.*, 2008] G. Carenini, R. T. Ng, and X. Zhou. Summarizing emails with conversational cohesion and subjectivity. In *ACL-08*, pages 353–361, OH, 2008. ACL.

[Cohen *et al.*, 2004] William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. Learning to classify email into "speech acts". In *EMNLP*, pages 309–316, 2004.

[Collins and Duffy, 2001] Michael Collins and Nigel Duffy. Convolution Kernels for Natural Language. In *NIPS-2001*, pages 625–632, Vancouver, Canada, 2001.

[Dhillon *et al.*, 2004] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. Meeting Recorder Project: Dialog Act Labeling Guide. Technical report, ICSI Tech. Report, 2004.

[Elsner and Charniak, 2010] Micha Elsner and Eugene Charniak. Disentangling chat. *Computational Linguistics*, 36:389–409, 2010.

[Hirao *et al.*, 2004] T. Hirao, , J. Suzuki, H. Isozaki, and E. Maeda. Dependency-based sentence alignment for multiple document summarization. In *Proceedings of Coling 2004*, pages 446–452, Geneva, Switzerland, 2004.

[Hovy *et al.*, 2005] E. Hovy, C. Y. Lin, and L. Zhou. A BE-based Multi-document Summarizer with Query Interpretation. In *DUC-05*, Vancouver, Canada, 2005.

[Jeong *et al.*, 2009] Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. Semi-supervised speech act recognition in emails and forums. In *EMNLP'09*, 2009.

[Joty *et al.*, 2010] S. Joty, G. Carenini, G. Murray, and R. T. Ng. Exploiting conversation structure in unsupervised topic segmentation for emails. In *EMNLP'10*, USA., 2010.

[Joty *et al.*, 2011] S. Joty, G. Carenini, G. Murray, and R. T. Ng. Supervised topic segmentation of email conversations. In *ICWSM'11*, Barcelona, 2011. AAAI.

[Kim *et al.*, 2010] Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in one-on-one live chats. In *EMNLP'10*. ACL, 2010.

[Malioutov and Barzilay, 2006] Igor Malioutov and Regina Barzilay. Minimum cut model for spoken lecture segmentation. In *ACL'06*, pages 25–32, Sydney, Australia, 2006.

[Murray *et al.*, 2006] G. Murray, S. Renals, J. Carletta, and J. Moore. Incorporating speaker and discourse features into speech summarization. In *HLT-NAACL'06*, 2006.

[Murray *et al.*, 2010] Gabriel Murray, Giuseppe Carenini, and Raymond T. Ng. Generating and validating abstracts of meeting conversations: a user study. In *INLG'10*, 2010.

[Ranganath *et al.*, 2009] R. Ranganath, D. Jurafsky, and D. Mcfarland. Its not you, its me: Detecting flirting and its misperception in speed-dates. In *EMNLP-09*, 2009.

[Ravi and Kim, 2007] Sujith Ravi and Jihie Kim. Profiling student interactions in threaded discussions with speech act classifiers. In *AIED'07*, LA, USA, 2007.

[Ritter *et al.*, 2010] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *HLT: NAACL'10*, LA, California, 2010. ACL.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Stolcke *et al.*, 2000] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26:339–373, 2000.

[Ulrich *et al.*, 2008] J. Ulrich, G. Murray, and G. Carenini. A publicly available annotated corpus for supervised email summarization. In *EMAIL'08 Workshop*. AAAI, 2008.

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