Modeling Dialogue Acts with Content Word Filtering and Speaker Preferences

Yohan Jo and **Michael Miller Yoder** and **Hyeju Jang** and **Carolyn P. Rosé**
Language Technologies Institute
Carnegie Mellon University
{yohanj,yoder,hyejuj,cprose}@cs.cmu.edu

Abstract

We present an unsupervised model of dialogue act sequences in conversation. By modeling topical themes as transitioning more slowly than dialogue acts in conversation, our model de-emphasizes content-related words in order to focus on conversational function words that signal dialogue acts. We also incorporate speaker tendencies to use some acts more than others as an additional predictor of dialogue act prevalence beyond temporal dependencies. According to the evaluation presented on two dissimilar corpora, the CNET forum and NPS Chat corpus, the effectiveness of each modeling assumption is found to vary depending on characteristics of the data. De-emphasizing content-related words yields improvement on the CNET corpus, while utilizing speaker tendencies is advantageous on the NPS corpus. The components of our model complement one another to achieve robust performance on both corpora and outperform state-of-the-art baseline models.

1 Introduction

Dialogue acts (DAs), or speech acts, represent the intention behind an utterance in conversation to achieve a conversational goal (Austin, 1975). Modeling conversations as structured DA sequences is a step toward the automated understanding of dialogue, useful for dialogue agents (Traum, 1999; Louwerse et al., 2002) and the processing of informal online conversational data (Misra and Walker, 2013; Vosoughi and Roy, 2016). Distributions of DAs can also be used as predictors of conversational outcome measures such as student learning in tutoring systems (Litman and Forbes-Riley, 2006) and engagement in meetings (Wrede and Shriberg, 2003). Unsupervised models for DA recognition may substitute or aid costly human annotation. We present an unsupervised model of DA sequences in conversation that overcomes limitations of prior models.

The first improvement our model offers is separating out content-related words to emphasize words more relevant to DAs. DAs are associated more closely with style and function words such as discourse markers and light verbs than with content words, which are more related to the propositional content (Erkens and Janssen, 2008; O'Shea et al., 2012). However, separating out content words is not standard in our field. For example, in some rule-based semantic and pragmatic parsing, the content and function of dialogue acts are not formally distinguished in the formalization (Becker et al., 2011), especially in domain-specific applications in dialogue systems (Gavaldà, 2004). A separation between content and function is useful for making cross-domain or cross-task generalizations about conversational processes.

Our model filters out content words by implementing the assumption that conversations proceed against a backdrop of underlying topics that transition more slowly than DAs or that are constant throughout. Based on a difference in transition speed, two types of language models are learned: foreground language models that capture DA-related words and background language models for content words. Although some existing models assume a background or domain-specific language model to filter out words unrelated to DAs (Lee et al., 2013; Paul, 2012; Ritter et al., 2010), they either require domain labels or do not learn topics underlying conversations.

The second improvement offered by our model is inclusion of speaker preferences, or tendencies to use some DAs more than others. Prior mod-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2179–2189
Copenhagen, Denmark, September 7–11, 2017.

els of DAs in conversation often rely on the discourse property of conditional relevance (Levinson, 1983; Martin and Rose, 2003), i.e., tendencies for sequences of conversational DAs such as questions followed by answers, greetings followed by greetings, and invitations followed by acceptances (Sidnell, 2011). Though conditional relevance, which motivates the use of Markov models for inducing DA representations, is one stable signal to discover DAs in discourse data (Brychcín and Král, 2017; Lee et al., 2013), there are reasons that it is a less strong signal than ultimately desired. One of the reasons is that the DA of an utterance depends not only on the preceding DA, but also on the speaker's personal style (Appling et al., 2013) or preferences for certain DAs. Our model explicitly accounts for speaker preferences as a factor in determining the DA of an utterance.

Our model also includes additional structure to account for assumptions about distribution and packaging of observed DAs in running discourse. First, one utterance can involve more than one DA (Levinson, 1983); for example, asking a question in a forum may involve introducing the speaker, explaining the problem, etc. Hence, we assume that DAs operate on more than one level simultaneously, and an utterance-level DA is a mixture of finer-grained sentence-level DAs. Second, online conversations often have multi-level structure, branching into multiple conversational threads using replies. Our model supports conversations that have such multi-level structure.

To illustrate the generalizability of our model, we evaluate it on two corpora with very different characteristics in terms of utterance length, the number of speakers per conversation, and the domain: CNET and NPS Chat Corpus. We evaluate the DA recognition accuracy of our model and compare the result with other latest models. As we tune the model parameters for each corpus, we use our model as a lens to understand the relationship between the nature of conversations and effective model components for identifying DAs, which may inform future model design.

For the remainder of the paper, we will discuss prior work on dialogue acts and existing models (Section 2) and explain our model design (Section 3). Then we will describe our evaluation method and corpora (Section 4) and discuss the lessons learned from our empirical investigation (Section 5). We conclude the paper in Section 6.

2 Related Work

Austin (1975) makes a distinction between the illocutionary, social intention of an utterance (as seen in the indirect sentence "Can you pass the salt?") and the locutionary act of an utterance, which includes the ostensible surface-level meaning of the words. DAs are commonly thought of as describing illocutionary actions in talk. Example DAs used in computational systems include *yes-no question*, *statement*, *backchannel*, and *opinion* (Jurafsky et al., 1998).

Winograd and Flores (1986) were some of the first to conceptualize DAs with state transitions as a model for conversation. Similarly, contemporary unsupervised DA models often use a hidden Markov model (HMM) to structure a generative process of utterance sequences (Ritter et al., 2010). It is commonly assumed that each hidden state corresponds to a DA, but different approaches use different representations for states.

One common representation of a state is a multinomial distribution over words, from which words related to DAs are generated. Often, this generative process includes domain- or content-related language models that are independent of states and used to filter out words unrelated to DAs (Lee et al., 2013; Ritter et al., 2010). However, these language models have some limitations. For instance, Lee et al. (2013) rely on domain labels for learning domain-specific language models, which may require human annotation, whereas our model learns them without labels. Ritter et al. (2010) learn conversation-specific language models to filter out content words. We take a different approach, simultaneously learning content-related topics underlying the entire corpus and filtering out these content words. Although most models incorporate a general language model to separate out common words (Lee et al., 2013; Paul, 2012; Ritter et al., 2010), we do not learn it because we assume that common words are relevant to DAs.

Word embedding vector representations have also been researched as the outputs of latent states. For example, Brychcín and Král (2017) represent an utterance as a weighted sum of word vectors from GloVe[1]. Each utterance vector is generated from a Gaussian distribution that parameterizes a latent state. This model has been shown to capture

[1] https://nlp.stanford.edu/projects/glove/

DAs effectively for short utterances.

DAs are not completely determined by preceding DAs (Levinson, 1983), and this difficulty can be overcome partly by modeling speaker style, as there is evidence that each speaker has preferences for certain DAs (Appling et al., 2013). Joty et al. (2011) model speakers as outputs generated by an HMM, but this structure makes it hard to adjust the contribution of speaker preferences and may overestimate the influence of speakers. We model speaker preferences more directly such that the preceding DA and the speaker together determine an utterance's probability distribution over DAs.

One reason for the nondeterministic nature of DAs is that one utterance can involve more than one DA (Levinson, 1983); this suggests that one language model per DA may not be enough. Paul (2012) represents latent states as mixtures of topics, but there is no one-to-one relationship between states and DAs. Joty et al. (2011) assume that words are drawn individually from a fixed number of language models specific to each DA. However, we observe that one sentence usually performs a consistent finer-grained act, so we constrain each sentence in an utterance to one language model. Thus, utterances, which may consist of multiple sentences, are represented as a mixture of finer-grained sentence-level DAs.

Word order in an utterance may play an important role in determining a DA, as in the difference between "I am correct" and "am I correct". Ezen-Can and Boyer (2015) compute the similarity between utterances based on word order using a Markov random field and cluster similar utterances to identify DAs. This model, however, does not consider transitions between clusters.

Online conversations often have asynchronous, multi-level structure (e.g., nested replies). In Joty et al. (2011)'s model, individual reply structure paths from the first utterance to terminal utterances are teased apart into separate sequential conversations by duplicating utterances. However, this method counts the same utterance multiple times and requires an aggregation method for making a final decision of the DA for each utterance. We address multi-level structure without duplicating utterances.

The properties of the models explained so far are summarized in Table 1.

The relative importance of each structural component in a model may not be identical across all corpora. Differences, especially as they are attributed to meaningful contextual variables, can be interesting both practically and theoretically. One contribution of our work is considering how differences in these kinds of contextual variables lead to meaningful differences in the utility of our different modeling assumtions. More typical work in the field has emphasized methodological concerns such as minimization of parameter tuning, for example, by using a hierarchical Dirichlet process to determine the number of latent DAs automatically (Lee et al., 2013; Ritter et al., 2010) or by simply assuming that a word is equally likely to be DA-related or general (Paul, 2012). While these efforts are useful, especially when maximizing the likelihood of the data, searching for the optimal values of parameters for DA recognition may allow us to better understand the contribution of each model component depending on the characteristics of the dialogue, which in turn can inform future model design.

	Sp	Tr	LM	ML	M
Brychcín and Král (2017)	N	Y	-	N	N
Ezen-Can and Boyer (2015)	N	N	-	N	N
Lee et al. (2013)	N	Y	GD	N	N
Paul (2012)	N	Y	G	N	Y
Joty et al. (2011)	Y	Y	U	Y	Y
Ritter et al. (2010)	N	Y	GD	N	N
Our model	Y	Y	D	Y	Y

Table 1: Properties of baseline models. **(Columns)** Sp: speaker preferences, Tr: DA transitions, LM: language models unrelated to DAs (G: general background, D: domain-specific, U: unspecified), ML: multi-level structure support, M: mixture of language models for DAs.

3 Model

Our model, CSM (content word filtering and speaker preferences model), is based on an HMM combined with components for content word filtering and speaker preferences. In the model, each latent state represents an utterance-level DA as a mixture of *foreground topics*, each of which represents a sentence-level DA. Each sentence in an utterance is assigned one foreground topic. To filter content words, there is a set of *background topics* shared across conversations, and each conversation is assigned a background topic that underlies the whole conversation.

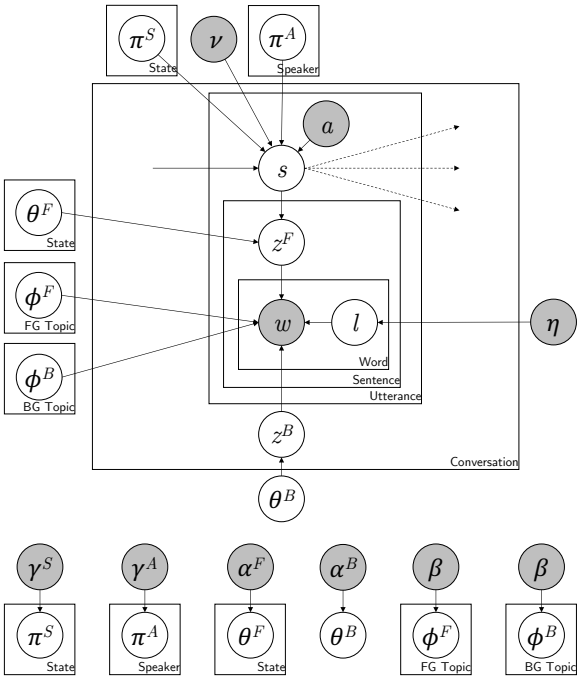

Figure 1: Graphical representation. Shaded nodes represent observable variables.

A transition between states is defined on every parent-child utterance pair, supporting multi-level structure. The state of an utterance is dependent on both its parent's state and its speaker. Speakers are specific to each conversation, i.e., a speaker participating in multiple conversations is treated as different speakers for different conversations. The graphical representation of CSM is in Figure 1.

The formal generative process of conversations is as follows:

- For each speaker a, draw a preference distribution over states $\pi_a^A \sim \text{Dir}(\gamma^A)$.
- For each state s
  - ▷ Draw a transition probability distribution over states $\pi_s^S \sim \text{Dir}(\gamma^S)$.
  - ▷ Draw a probability distribution over foreground topics $\theta_s^F \sim \text{Dir}(\alpha^F)$.
- For each foreground topic t, draw a probability distribution over words $\phi_t^F \sim \text{Dir}(\beta)$.
- For each background topic t, draw a probability distribution over words $\phi_t^B \sim \text{Dir}(\beta)$.
- For the corpus, draw a distribution over background topics $\theta^B \sim \text{Dir}(\alpha^B)$.
- For each conversation
  - ▷ Draw a background topic $z^B \sim \text{Cat}(\theta^B)$.
  - ▷ For each utterance u, with its speaker a_u, its parent p, and the parent's state s_p,
    - ♦ Draw a state $s_u \sim \text{Cat}(\nu\pi_{s_p}^S + (1-\nu)\pi_{a_u}^A)$.
    - ♦ For each sentence
      - □ Draw a foreground topic $z^F \sim \text{Cat}(\theta_{s_u}^F)$.
      - □ For each word
        - · Draw an indicator of "foreground" or "background" $l \sim \text{Cat}((\eta, 1-\eta))$.
        - · If l is "foreground", draw a word $w \sim \text{Cat}(\phi_{z^F}^F)$.
        - · If l is "background", draw a word $w \sim \text{Cat}(\phi_{z^B}^B)$.

According to this model, content words are separated out into background topics in several ways. A background topic does not transition as frequently as foreground topics within a conversation. Accordingly, words that are consistently used across utterances in a conversation are likely to be clustered into the background topic z^B, whereas words whose use is sensitive to the previous state and the speaker are likely to be clustered into foreground topics z^F. However, common function words, such as pronouns, prepositions, and punctuations, may also be separated out. Hence, η, the probability of a word being foreground, adjusts the degree of filtering. The higher the η value, the more words are likely to be generated from a foreground topic, and thus the more function words are included in foreground topics, leaving background topics with content words. Hence, we may set η high if we believe function words play an important role in DAs in a corpus and low otherwise. Note that $\eta = 0.5$ is equivalent to the assumption of existing models that a word is equally likely to be foreground or background (Lee et al., 2013; Paul, 2012). Background topics capture content words underlying the corpus, as they are shared across conversations.

Speaker preferences are captured as a probability distribution over DAs (π^A), which, along with the preceding state, affects the probability of the current state. ν adjusts the contribution of the speaker's preferences; the higher ν, the weaker the contribution. So, we may set ν low if the role or conversational style of each speaker is believed to be invariant and each speaker is expected to conduct specific DAs. If there is not enough such evidence and the conversation is driven without specific roles of the speakers, then we may set ν high. We find that corpora have different optimal values

N_{ij}^{SS}	Transition from state i to state j
N_{ij}^{AS}	Assignment of speaker i to state j
N_{ij}^{SF}	Assignment of state i to foreground topic j
N_{j}^{B}	Assignment to background topic j
N_{ij}^{FW}	Assignment of foreground topic i to word j
N_{ij}^{BW}	Assignment of background topic i to word j

Table 2: Descriptions of counter matrices.

of ν depending on the conversational characteristics.

We use collapsed Gibbs sampling for inference to integrate out π^S, π^A, θ^F, θ^B, ϕ^F, and ϕ^B. Given conversation text with speakers for each utterance, along with the hyperparameters, ν, and η, the Gibbs sampler estimates the following variables using counter matrices explained in Table 2:

$$\pi_{ij}^S = \frac{N_{ij}^{SS} + \gamma^S}{\sum_{j'}(N_{ij'}^{SS} + \gamma^S)}, \pi_{ij}^A = \frac{N_{ij}^{AS} + \gamma^A}{\sum_{j'}(N_{ij'}^{AS} + \gamma^A)}$$

$$\theta_{ij}^F = \frac{N_{ij}^{SF} + \alpha^F}{\sum_{j'}(N_{ij'}^{SF} + \alpha^F)}, \theta_{j}^B = \frac{N_{j}^{B} + \alpha^B}{\sum_{j'}(N_{j'}^{B} + \alpha^B)}$$

$$\phi_{ij}^F = \frac{N_{ij}^{FW} + \beta}{\sum_{j'}(N_{ij'}^{FW} + \beta)}, \phi_{ij}^B = \frac{N_{ij}^{BW} + \beta}{\sum_{j'}(N_{ij'}^{BW} + \beta)}.$$

We may use slice sampling (Neal, 2003) to estimate ν and η too, but the estimated values of ν and η may not be optimal for DA recognition. We can also obtain state assignments for utterances by taking a sample from the Gibbs sampler. Detailed derivation for Gibbs sampling and the code are available online[2].

4 Evaluation

This section describes our evaluation method and settings.

4.1 Task and Metrics

We evaluate our model in terms of accuracy in utterance-level DA recognition. Since the output of the model is assignments to discovered states for utterances, not pre-determined DA labels, we use a clustering evaluation method, as adopted by previous work on unsupervised DA modeling. Specifically, we use homogeneity, completeness, and v-measure as metrics (Rosenberg and Hirschberg, 2007). Homogeneity represents the degree to which utterances assigned to the same cluster by the model share the same DA in the labeled corpus. Completeness represents the degree to which utterances that have the same DA according to the gold standard are assigned to the same cluster. V-measure is the harmonic mean of homogeneity and completeness. These metrics are easy to interpret and have been demonstrated to be invariant to dataset size and number of clusters. This enables a meaningful comparison of accuracy across different corpora.

[2] https://github.com/yohanjo/Dialogue-Acts

	CNET	NPS
# conversations	310	15
# utterances	1,332	10,567
# DAs	12	15
# domains	24	-
Median # utterances/conversation	3	706
Median # words/utterance	51	2
Median # speakers/conversation	2	94

Table 3: Corpora statistics.

CNET	NPS
Question-Question	Accept
Question-Add	Bye
Question-Confirmation	Clarify
Question-Correction	Continuer
Answer-Answer	Emotion
Answer-Add	Emphasis
Answer-Confirmation	Greet
Answer-Correction	Reject
Answer-Objection	Statement
Resolution	System
Reproduction	yAnswer
Other	nAnswer
	whQuestion
	ynQuestion
	Other

Table 4: Dialogue act tags in the corpora.

4.2 Corpora and Preprocessing

We evaluate on two corpora: CNET and NPS Chat (see Table 3 for statistics).

CNET (Kim et al., 2010) is a set of post threads from the Operating System, Software, Hardware, and Web Development sub-forums of CNET. This corpus is tagged with 12 DAs, including *Question-Question*, *Question-Confirmation*, *Answer-Add*, *Resolution*, and *Other* (Table 4). Note that question- and answer-related DAs are two-level. Most posts are tagged with one DA; in case a post is tagged with multiple DAs, we choose the first DA in the meta-data[3]. Each post is considered an

[3] Some tagging systems, such as the DAMSL-style, break down an utterance that has multiple DAs.

utterance and each thread as a conversation. Each thread has only a few posts (median 3) and involves a few speakers (median 2). Since there are many URLs, email addresses, and numbers in text, we replace them with special tokens using regular expressions, and tokenize with the Stanford PTBTokenizer included in Stanford Parser 3.7.0[4].

NPS Chat (Forsythand and Martell, 2007) is a set of conversations from various online chat services. This corpus is tagged with 15 DAs, including *Emotion*, *System*, and *whQuestion* (Table 4). Every turn is tagged with a DA and considered an utterance. Each conversation is long (median 706 utterances) and involves many speakers (median 94). This corpus has already been tokenized, so we only replace usernames with a special token. Conversations in NPS have no reply structure, but we build in multi-level structure, simply treating an utterance that mentions another user as a child of the nearest utterance of the mentioned user. We compare the DA accuracy of the multi-level structure and the original linear structure in Section 5.

4.3 Models and Parameters

We set the numbers of states and background topics to the numbers of DAs and domains, respectively, if these numbers are available. For NPS, we search for the optimal number of background topics between 1 and 2, because there are only a few conversations. The optimal number of foreground topics is chosen among multiples of five between the number of states and four times the number of states, and the weights for state transition (ν) and foreground topics (η) are chosen among multiples of 0.1. For Dirichlet hyperparameters, we use $\alpha^F = 0.1, \gamma^A = 0.1, \beta = 0.001$ to induce sparsity, and $\gamma^S = 1, \alpha^B = 1$ for the uniform distribution over all configurations.

We randomly split each corpus into five groups and use three groups for training, one for parameter tuning, and one for testing. We run 5-fold cross-validation and report the average optimal parameter values and accuracy across the folds. The number of sampling iterations was chosen such that the log-likelihood of the data has converged. For each fold, we take 10 samples during inference on the test data with interval of 10 iterations and compute the mean and standard deviation of the 50 samples from all folds.

[4] https://nlp.stanford.edu/software/lex-parser.html

We compare our model with the three most recent unsupervised models we surveyed. The baseline models and settings are as follows.

Gaussian mixture HMM (Brychcín and Král, 2017), based on an HMM, has a characteristic output representation: utterance vectors. These vectors are generated from Gaussian distributions instead of using language models as in most existing models. After following their same preprocessing steps, we trained a model on the training data, chose the optimal word vector dimensionality on the validation data (among 50, 100, 200, and 300, as used in the original model), and performed inference on the test data. We used the original source code from the authors for training and modified the code for inference.

MRF-based clustering (Ezen-Can and Boyer, 2015) considers word order within an utterance to calculate similarity between utterances using an MRF. Then k-medoids clustering is conducted based on the similarity scores, resulting in clusters that represent DAs. The similarity score between two utterances is asymmetric, so we took the average value of each direction and inversed it to obtain the distance between two utterances. We trained a model on the training data, chose the optimal parameter values ($\lambda_i, \lambda_t, \alpha_d$ in the original paper) on the validation data, and assigned clusters to the test data. We implemented the algorithm since the original code was not available.

HDP-HMM (Lee et al., 2013) is based on an HMM, and each word comes from either the state-specific, general background, or domain-specific language model. HDP-HMM automatically decides the number of states using a hierarchical Dirichlet process, but we manually set the number of DAs in our experiment, assuming that we know the number of DAs of interest. We trained a model on the training data and performed inference on the test data; the validation data was not used since there are no parameters to tune. We used the original source code from the authors for training and modified the code for inference.

5 Results

Accuracy of DA recognition in terms of homogeneity, completeness, and v-measure on both corpora is summarized in Table 5. We also tested the following configurations:

- **CSM + Domain** uses true domain labels when learning background topics by force-

Model	CNET			NPS		
	H	C	V	H	C	V
Brychcín and Král (2017)	.13± .00	.09± .00	.10± .00	.24± .10	**.33± .06**	.28± .08
Ezen-Can and Boyer (2015)	.03± .00	.37± .00	.05± .00	.26± .00	.33± .00	.28± .00
Lee et al. (2013)	.09± .03	.16± .03	.11± .03	**.36± .02**	.28± .02	.31± .02
CSM	.24± .03	**.38± .04**	**.29± .03**	.35± .04	.31± .04	**.33± .04**
CSM + Domain	**.27± .02**	.33± .11	.29± .05		N/A	
CSM - Speaker	.24± .03	**.38± .04**	**.29± .03**	.21± .03	.19± .05	.20± .04
CSM - Multi-level	.23± .04	.33± .06	.27± .04	.35± .02	.30± .04	.32± .03
CSM - Background Topics	.15± .03	.11± .02	.12± .02	.35± .04	.31± .04	**.33± .04**

Table 5: Accuracy of DA recognition (the higher the better). Smaller numbers are population standard deviations. **(Columns)** H: homogeneity, C: completeness, V: v-measure. Optimal parameter values for CSM: # foreground topics=34, $\eta = .86$, $\nu = 1.00$ for CNET and # foreground topics=35, $\eta = 1.00$, $\nu = 0.58$ for NPS.

fully assigning a conversation the background topic corresponding to the true label.

- **CSM - Speaker** does not use speaker preferences by setting $\nu = 1$.
- **CSM - Multi-level** ignores multi-level structure; that is, utterances in each conversation are ordered by time.
- **CSM - Background Topics** uses only one background topic.

Overall, our model performs significantly better than the baselines for CNET and marginally better for NPS. The baseline models show a large variance in performance depending on the characteristics of the corpus. In contrast, our model has a low variance between the corpora, because the content word filtering, distinction between utterance-level and sentence-level DAs, and speaker preferences complement one another to adapt to different corpora. For example, content word filtering and DA level distinction play more significant roles than speaker preferences on CNET, whereas their effects are reversed on NPS. The details will be described later with qualitative analyses.

There may be several reasons for the poor performance of the baseline models on CNET. First, in our model, each utterance-level DA (latent state) is a probability distribution over sentence-level DAs (foreground topics), which better captures multiple sentence-level DAs in long utterances as in CNET. The utterances in CNET are long and may be too complex for the baseline models, which use a simpler representation for utterance-level DAs. Another reason for the low performance could be that the baseline models do not filter out content words as our model does.

BT0	drive partition drives partitions c
BT1	router wireless network connected connection
BT2	vista camera canon windows scanner
BT3	drive ipod touch data recovery
BT4	speakers firewall sound still no
BT5	/ \blaster dos drive
BT6	windows cd i xp boot
BT7	page xp sp3 ! content
BT8	ram mhz 1gb 512mb screen
BT9	his rupesh to company he
BT10	xp drive drivers new hard
BT11	tv port cpu motherboard grounded
BT12	file files copy external mac
BT13	" password flash ##NUMBER## ?
BT14	fan fans cpu case air
BT15	ram card 2.4 graphics nvidia
BT16	registry file shutdown machines screen
BT17	div site % ie6 firefox
BT18	printer sound would card contact
BT19	hosting web hostgator they host
BT20	ubuntu linux memory boot reader
BT21	mac compression archive format trash
BT22	bluetooth router wireless laptop 802.11
BT23	email address account mail bounce

Table 6: Background topics learned from CNET. **(Columns)** Left: topic index, right: top 5 words.

In the remainder of this section, we describe our qualitative analysis on the results. All examples shown in the analysis are from the result with the optimal parameter values for the first fold.

Filtering content words Our model effectively separates content words from DA-related words without using the domain label of each conversation. As an example, the background topics

learned by our model from CNET are shown in Table 6. These topics are clearly related to the subjects of the forum, rather than reflecting DAs, and the topics are distinctive from one another and cohesive in themselves.

The main purpose of learning background topics is to filter out content words and retain DA-related words as foreground. The learned background topics serve this purpose well, as these topics increase v-measure by 0.17 (CSM vs. CSM - Background Topics). It is also promising that the background topics learned without domain labels perform as well as when they are learned with domain labels (CSM vs. CSM + Domain), because domain labels may not always be available.

Function words play an important role in DAs in CNET as indicated by the high optimal value of $\eta = 0.86$ (the probability of a word being foreground). The higher η means more function words are included in foreground topics, leaving background topics with content words (Section 3). The high η is evidence contrary to the common practice of designating a general background topic to filter out common words and assuming that a word is equally likely to be foreground or background (Lee et al., 2013; Paul, 2012).

The effectiveness of our method of separating background topics turns out to diminish when there are no consistent conversational topics within and across conversations as in NPS. Our model learns not to use background topics ($\eta = 1$) for NPS, because background topics may filter out function words and DA-related words that occur more consistently throughout a conversation than content words do.

Mixture of foreground topics As a consequence of filtering out content words, the foreground topics reflect various acts in conversation. Some of the learned foreground topics from CNET are shown in Table 7a. These topics capture important sentence-level DAs that constitute utterance-level DAs that are assigned to each post in CNET. For example, *Question-Question* is an utterance-level DA that often starts a conversation, and conducting this DA typically includes multiple finer-grained acts, such as explaining the environment and situation, asking a question, and thanking, as shown in the post:

> I am currently running Windows XP Media Edition on a 500G hard drive. (FT20) / I want to move my XP to it's own partition, move all of my files(music, games, work) to another, and then install the Windows 7 beta on another partition. (FT10) / I don't know if this is possible or not, but I have access to Partition Magic 8, and am wondering if I can do it with that or not. (FT10) / I am not worried about installing 7 on another partition, but am not sure if I can move my files onto a separate one while keeping XP intact. (FT10) / Any help is great, thank you. (FT17)

Environments (FT20)	. i a ##NUMBER## and have - rrb- xp -lrb- : windows my is the dell vista
Error msgs (FT12)	. the # * messages / : it log
Asking (FT19)	any help you ? ! . appreciated i suggestions
Thanking (FT17)	thanks . for the ! in advance help your all response
Problem (FT8)	: \file is the c corrupted following missing or error
Wishes (FT14)	. bob good luck
Reference (FT5)	##URL##
Praise (FT1)	. thank you ~ sovereign , and are excellent recommendations
Explanation (FT10)	the . to , i and a it you is that of

(a) Foreground topics learned from CNET.

Wh question (FT7)	##USERNAME## ? how you are u good is round where who . ??
Wh question (FT27)	##USERNAME## ? you i u what how , ok 'm for up do have
YN question (FT1)	chat any wanna / me pm to ? anyone f guys m want here
Greeting (FT5)	##USERNAME## hi hey :) hello wb ! ... hiya ty
Laughing (FT0)	##USERNAME## lol lmao yes ! hey up !!!! ?
Laughing (FT12)	lol ##USERNAME## haha ! brb omg nite hiyas hb :p !!! . ha lmfao
Emotion (FT30)	ok ! im lol my its in " ... oh always
System logs (FT25)	part join

(b) Foreground topics learned from NPS.

Table 7: Foreground topics learned from the corpora. **(Columns)** Left: interpretation (topic index), right: top words truncated for clarity.

Likewise, the *Answer-Answer* DA includes finer acts such as wishes or URLs, as in the posts:

> Simple - Download and install the Vista Rebel XT drivers from canon usa.com. (FT10) / Once installed.....go to camera menu and switch the communication to Print/PTP. (FT10) / Don't forget to switch it back if you're connecting to an XP machine. (FT10) / Good Luck (FT14)

> http://forums.microsoft.com/MSDN/ShowPost.aspx?PostID=1996406&SiteID=1 (FT5)

When a problem is resolved, the *Resolution* DA may be performed with thanking and praising:

> Excellent summary Thank you. (FT1) / Sounds like at some point it's worth us making the transition to a CMS... (FT10)

FT10 covers explanations and statements, as well as long sentences. The distinction between two levels DAs is effective for CNET, as our model beats the baselines significantly.

The foreground topics learned from NPS also reflect DAs in the corpus (Table 7b). The distinction between utterance-level and sentence-level DAs is not beneficial for NPS because each utterance is short and usually conducts only one DA. As a consequence, the model has difficulty grouping foreground topics (i.e., sentence-level DAs) that are related to one another into the same utterance-level DAs (i.e., states); for CNET, on the other hand, foreground topics that co-occur in the same utterance tend to cluster to the same state.

The DAs of some foreground topics not shown in Table 7 are difficult to interpret, and those topics possibly capture aspects of sentences other than DAs. However, they do not have undue influence in our model.

Speaker preferences Speaker preferences substantially increase the v-measure by 0.13 for NPS (CSM vs. CSM - Speaker). Notably, speaker preferences complement the mixture of sentence-level DAs, which is not good at clustering related sentence-level DAs into the same utterance-level DA for short utterances. More specifically, each speaker is modeled to have sparse preferences for utterance-level DAs (i.e., states), so foreground topics used by the same speaker, often representing the same utterance-level DA, tend to cluster to the same state.

Speaker preferences also capture the characteristic styles of some speakers. Among speakers who are found to have sparse preferences by our model, some actively express reactions and often mark laughter (FT12). Others frequently agree (FT0), greet everyone (FT5), or have many questions (FT7, FT27). Accordingly, the model finds a relatively high optimal weight for speaker preferences in NPS ($\nu = 0.58$).

In contrast, CNET benefits little from speaker preferences ($\nu = 1$), partly because there is not enough information about each speaker in such short conversations. Speakers also show little preference for DAs as defined in the corpus. For instance, while a conversation initiator tends to ask questions in successive posts, these questions are annotated as different DAs (e.g., *Question-Question*, *Question-Add*, *Question-Confirmation*, etc.) depending on the position of the post within the conversation.

Multi-level structure Our model's ability to account for multi-level structure improves the accuracy of DA recognition for both corpora (CSM vs. CSM - Multi-level). For NPS, where multi-level structure is not explicit, this improvement comes from simple heuristics for inferring multi-level structure based on user mentions.

Sentence length and foreground topics In our model, all words in the same sentence are assigned to the same foreground topic, just as many existing models assign one utterance one topic. Topic assignment is based on similarity of words in a sentence to other sentences in that topic, and short sentences often find similar sentences more easily than long sentences do. Therefore, learned topics tend to be characteristic of short sentences that are similar enough to form the separate topics, and as a result, long sentences may be assigned the same topic regardless of the DA actually performed.

6 Conclusion

We have presented an unsupervised model of DAs in conversation that separates out content words to better capture DA-related words and that incorporates speaker preferences. Our model also uses a mixture of sentence-level DAs for utterance-level DAs and supports multi-level thread structure. We find that different characteristics of conversation require different modeling assumptions for DA recognition. Unlike the baseline models, which show a large variance in performance across corpora, our model is robust for both corpora used in the evaluation due to the model components complementing one another. Specifically, content word filtering is found to be effective when each conversation has a consistent conversational topic, and the separation between sentence-level and utterance-level DAs is beneficial for long utterances. Speaker preferences are found to be helpful when speakers have characteristic styles of conversation. These findings in addition to the fact that many function words are not filtered out as background may help inform future model design.

Acknowledgments

This research was supported by the Kwanjeong Educational Foundation, NIH grant R01HL122639, and NSF grant IIS-1546393.

References

D Scott Appling, Erica J Briscoe, Heather Hayes, and Rudolph L Mappus. 2013. Towards automated personality identification using speech acts. In *AAAI Workshop - Technical Report*.

John L Austin. 1975. *How to Do Things with Words*. Harvard University Press.

Lee Becker, Wayne H Ward, Sarel van Vuuren, and Martha Palmer. 2011. DISCUSS: a dialogue move taxonomy layered over semantic representations. *Proceedings of the Ninth International Conference on Computational Semantics*, pages 310–314.

Tomáš Brychcín and Pavel Král. 2017. Unsupervised dialogue act induction using gaussian mixtures. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 485–490, Valencia, Spain. Association for Computational Linguistics.

Gijsbert Erkens and Jeroen Janssen. 2008. Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer-Supported Collaborative Learning*, 3(4):447–470.

Aysu Ezen-Can and Kristy Elizabeth Boyer. 2015. Understanding Student Language: An Unsupervised Dialogue Act Classification Approach. *JEDM - Journal of Educational Data Mining*, 7(1):51–78.

Eric N Forsythand and Craig H Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26.

Marsal Gavaldà. 2004. *Soup: A Parser for Real-World Spontaneous Speech*. Springer Netherlands, Dordrecht.

Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, pages 1807–1813. AAAI Press.

Daniel Jurafsky, Rebecca Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, Paul Taylor, and C. Van Ess-Dykema. 1998. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202, Uppsala, Sweden. Association for Computational Linguistics.

Donghyeon Lee, Minwoo Jeong, Kyungduk Kim, Seonghan Ryu, and Gary Geunbae Lee. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2451–2464.

Stephen C Levinson. 1983. Conversational structure. In *Pragmatics*, chapter 6, pages 284–333. Cambridge University Press.

Diane J Litman and Katherine Forbes-Riley. 2006. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12(02):161–176.

Max Louwerse, Art Graesser, Andrew Olney, and the Tutoring Research Group. 2002. Good Computational Manners: Mixed-Initiative Dialog in Conversational Agents. *Etiquette for Human-Computer Work: Papers from the AAAI Fall Symposium*, pages 71–76.

J R Martin and David Rose. 2003. *Negotiation: interacting in dialogue*. New Century Series. Bloomsbury Academic.

Amita Misra and Marilyn Walker. 2013. Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue. *Proceedings of the SIGDIAL 2013 Conference*, (August):41–50.

Radford M Neal. 2003. Slice sampling. *The Annals of Statistics*, 31(3):705–767.

James O'Shea, Zuhair Bandar, and Keeley Crockett. 2012. *A Multi-classifier Approach to Dialogue Act Classification Using Function Words*. Springer Berlin Heidelberg, Berlin, Heidelberg.

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, Jeju Island, Korea. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Jack Sidnell. 2011. *Conversation Analysis: An Introduction*. Language in Society. Wiley.

David R. Traum. 1999. *Speech Acts for Dialogue Agents*. Springer Netherlands, Dordrecht.

Soroush Vosoughi and Deb Roy. 2016. Tweet Acts : A Speech Act Classifier for Twitter. *Proceedings of the 10th AAAI Conference on Weblogs and Social Media*, (ICWSM):1–4.

Terry Winograd and Fernando Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Language and being. Ablex Publishing Corporation.

Britta Wrede and Elizabeth Shriberg. 2003. Relationship between dialogue acts and hot spots in meetings. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 180–185. IEEE.