Dramatic Conversation Disentanglement

Kent K. Chang, Danica Chen and **David Bamman**
University of California, Berkeley
{kentkchang,danicachen2019,dbamman}@berkeley.edu

Abstract

We present a new dataset for studying conversation disentanglement in movies and TV series. While previous work has focused on conversation disentanglement in IRC chatroom dialogues, movies and TV shows provide a space for studying complex pragmatic patterns of floor and topic change in face-to-face multiparty interactions. In this work, we draw on theoretical research in sociolinguistics, sociology, and film studies to operationalize a conversational thread (including the notion of a floor change) in dramatic texts, and use that definition to annotate a dataset of 10,033 dialogue turns (comprising 2,209 threads) from 831 movies. We compare the performance of several disentanglement models on this dramatic dataset, and apply the best-performing model to disentangle 808 movies. We see that, contrary to expectation, average thread lengths do not decrease significantly over the past 40 years, and characters portrayed by actors who are women, while underrepresented, initiate more new conversational threads relative to their speaking time.

1 Introduction

Movie and TV dialogues, or dramatic dialogues more generally, have offered linguists a wealth of resources to study conversational behaviors (Lakoff and Tannen, 1984; He and Herman, 1998; Richardson, 2010), including within NLP (Danescu-Niculescu-Mizil and Lee, 2011; Ramakrishna et al., 2017; Sap et al., 2017; Azab et al., 2019). While dramatic dialogues do not necessarily mimic conversations in real life, they present complex pragmatic and sociolinguistic phenomena that warrant study; given the widespread viewership of movies and TV, what appears on screen—both visually and in dialogue—can have a real social impact in the world (Rosen, 1973; hooks, 1992; Heldman, 2016).

An important feature of such dialogues is that they are *entangled*. In his work on dramatic dialogues, McKee (2016, p. 3) articulates a speech act view: in screenplays, "all talk responds to a need, engages a purpose, and performs an action." In any given scene in a movie or TV show, then, we can often see multiple needs expressed by different characters in one sequence of conversation. Consider this scene from *Young Sheldon* (Fig. 1): Missy relays a message, Sheldon wants to know where his bow tie is, and Georgie seeks to avoid showing up with Sheldon at school—each of them starts a new conversational thread (or, subconversation) with their speech act that reflects those different intents. In the screenplay, there is no explicit structure that indicates where each subconversation starts and ends. If, however, we could disentangle dramatic conversations, we could ask such questions as: What kind of characters get to *start* a thread? How long do conversations tend to last? Answers to those questions can enhance our understanding of cultural representations on screen.

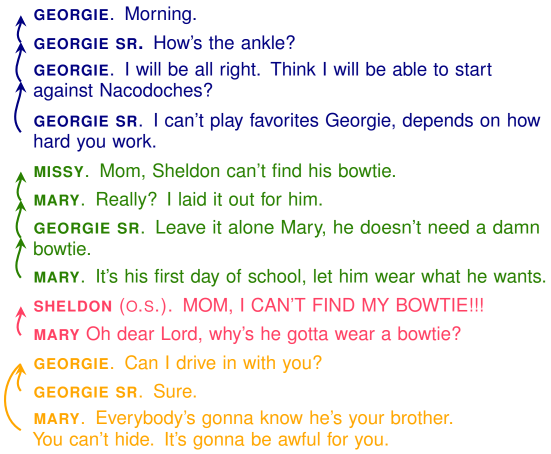

Figure 1: Example of dramatic conversations, taken from a scene in *Young Sheldon*. Speaker labels are in boldface and SMALL CAPS. Curved arrow lines indicate the reply-to relations between dialogue lines. Each thread is distinguished by colors.

Much of the work on conversation disentanglement in NLP has studied Internet Relay Chat (IRC)

Findings of the Association for Computational Linguistics: ACL 2023, pages 4020–4046
July 9-14, 2023 ©2023 Association for Computational Linguistics

logs, most notably #Linux (Elsner and Charniak, 2008) and #Ubuntu (Kummerfeld et al., 2019). IRC logs are in a different domain from dramatic conversations, and some salient features in IRC, such as invoking a username to indicate replies to that user, are not found in screenplays. Conversely, there is no equivalent for "off-screen" speakers in IRC. Given the face-to-face nature of conversations in drama, movie and TV characters can start new conversational threads by entering the scene. Those major differences mean chat logs may be insufficient to train models that disentangle drama.

To bridge the gap, we present in this work a new annotated dataset to support the study of conversation disentanglement in the domain of movies and TV shows. We draw heavily on the theoretical resources found in film studies, sociology, and linguistics as we design our annotation framework, with particular attention paid to the semantic and pragmatic signals of the start of a new thread.

In this work we make the following contributions:

- We draw on theoretical research in sociolinguistics, sociology, and film studies to operationalize a conversational thread (including the notion of a floor change) in dramatic texts, and use that definition to annotate a dataset of 10,033 dialogue turns (comprising 2,209 threads) from 831 movies. All annotations are freely available for public use under a CC BY-NC-SA license on GitHub.[1]

- We compare the performance of several disentanglement models on this dramatic dataset to see if model architectures designed for or models trained on Kummerfeld et al. (2019) perform well in the domain of drama.

- We apply the best-performing model to analyze and disentangle 808 films in SCRIPTBASE-J (Gorinski and Lapata, 2015, 2018), investigating both the relationship between historical thread length and intensified continuity style (Bordwell, 2002) and the relationship between gender and power in floor claiming. In this data, we see that, unlike shot lengths, average thread lengths do not decrease significantly over the past 40 years (contrary to expectation), and characters portrayed by actors who are women, while underrepresented, initiate more new conversational threads relative to their speaking time.

2 Related work

Conversation disentanglement. Conversation disentanglement seeks to identify threads (or, clusters, subconversations) in a sequence of utterances. Conceptually, this task requires a robust operationalization of *thread*, which is usually understood as related to topic or floor change (O'Neill and Martin, 2003; Shen et al., 2006; Jiang et al., 2018). Elsner and Charniak (2008, 2010) considered this problem in the context of chat history, which has been extensively studied since (for a recent survey, see Gu et al., 2022).

In terms of modeling, there are two popular approaches for this task (Zhu et al., 2021): two steps (models link individual utterances first, and then we recover thread membership) or end-to-end (models predict thread membership directly). Our work adopts the two-step method: we first calculate the similarity score to identify the reply-to relations between two utterances (the link prediction task), and apply a greedy clustering algorithm to put utterances that reply to one another into threads (the clustering task). For the two-step method, there have been attempts to adopt a multi-task learning setup: At training time, when gold cluster information is available, this auxiliary task calculates another loss function dedicated to thread prediction, which can be used to improve the performance of link prediction, the main task, with which we also experimented.

Datasets for conversation disentanglement are currently limited. In Mahajan and Shaikh's (2021) comprehensive survey on multi-party dialogue understanding, most datasets are not curated with this purpose in mind. Kummerfeld et al. (2019)'s corpus, built on annotations of IRC chat logs, has been the standard benchmark dataset for this task. Liu et al. (2020, p. 3871) released a dataset of movie dialogues, where they "collect 869 movie scripts that explicitly indicate the plot changing" and "extract 56,562 sessions from the scripts and manually intermingle these sessions to construct a synthetic dataset." In this work, we present a new annotated dataset built on movies and TV shows, adding a spoken, scripted corpus to facilitate this line of work. Instead of inferring from the narrative structure and threading conversations through

[1] https://github.com/kentchang/dramatic-conversation-disentanglement

a synthetic process, we developed an annotation framework (described in §4).

Theoretical approaches to conversation. Conversation has been extensively theorized and studied in sociology, linguistics, and film studies, which this work draws on. For our annotation scheme, Goffman's (1963) idea that conversation is a form of focused interaction and McKee's (2016) speech act view on dramatic dialogues provide us with the theoretical foundation.

The ideas related to *focus* and *topic* are further explored in the following: Ervin-Tripp (1964), who considers the surface and semantic features of sustained attention in conversational organization; Roberts (1996), who expounds on "topic under discussion" in pragmatics; Ng and Bradac (1993), who see topic change and floor claiming from the perspective of power dynamics between speakers involved. More broadly, Sacks et al. (1974); Goodwin (1981) detail the organizational and pragmatic principles for conversation, and He and Herman (1998) consider them specifically in the context of drama.

This work is further motivated by the social implications of conversations taking place on screen, and we find the following particularly relevant: Richardson (2010) carried out a sociolinguistic study on TV dialogues; Silverman (1988); Boon (2008); O'Meara (2019) highlights related issues as they pertain to gender and race.

Modeling multi-party conversation structure. The interaction structure between speakers in a multi-party conversation is shown to be useful for conversation disentanglement (Mayfield et al., 2012). More recently, various neural architectures have been proposed to encode utterances and the hierarchical structure of conversations (Jiang et al., 2018; Henderson et al., 2020; Wu et al., 2020; Yu and Joty, 2020). Pre-training with self-supervised tasks (Zhu et al., 2020; Wang et al., 2020; Gu et al., 2021) is also used to derive contextual embeddings while factoring in the conversational structure (replacing, for example, next sentence prediction with next *utterance* prediction). With or without self-supervision, additional embedding layers or attention mechanisms have been proposed to encode the information of conversation structure. Gu et al. (2020); Liu et al. (2021) incorporated speaker embeddings and Sang et al. (2022) and Ma et al. (2022) emphasize the interaction between speakers

```
EXT. ONE OF THE EXITS - MADISON SQUARE GARDEN - NIGHT -

Emily and Junior are standing, waiting for Kane.

                      JUNIOR
           Is Pop Governor yet, Mom?

Just then, Kane appears, with Reilly and several other men.
Kane rushes toward Emily and Junior, as the men politely greet
Emily.

                      KANE
           Hello, Butch!  Did you like your
           old man's speech?

                      JUNIOR
           Hello, Pop!  I was in a box.  I
           could hear every word.
```

Figure 2: Example of the standard screenplay format (*Citizen Kane*). `EXT. ONE OF THE EXITS` is the scene header. The dialogue portions are usually indented, while action statements ("*Emily and Junior are standing*") are not indented. Speaker labels are found in cue lines in all caps, followed by their dialogue lines.

through attention. For our experiments, as a baseline, we start with a simple and intuitive approach, where we train embedding layers based on structural features like the distance between utterances, and concatenate relevant embedding vectors with standard contextual embeddings from pre-trained language models that we can fine tune for this task.

In this work, conversation structure is situated in the specific domain of films and TV shows, spoken by characters in a scene. The structural richness of screenplays makes it an interesting textual representation of movies in NLP (Bhat et al., 2021; Chen et al., 2022), and here, we use screenplays to create an annotated dataset to facilitate future work on conversation disentanglement.

3 Data

To study conversation disentanglement in drama, we consider 831 titles: 340 movies and 491 TV series,[2] randomly sampling one scene from each title for annotation. Movies are taken from SCRIPTBASE-J (Gorinski and Lapata, 2018), based on IMSDB,[3] because of its extensive coverage of genres and temporal span, along with rich metadata that can adequately support NLP research related to movies. Since TV dialogues have different linguistic styles from movie dialogues (Nelmes, 2010), we curated a new dataset, TVPILOTS, using teleplays of pilot episodes made available by TV Writing.[4]

All screenplays and teleplays from these sources come in the standard format (Fig. 2): They have

[2]The complete list of titles we used can be found in our GitHub repository.
[3]https://imsdb.com/
[4]https://sites.google.com/site/tvwriting/

distinct scene headers, speaker labels, and other typographical features for annotators to distinguish between an action statement and a dialogue line. Since those features are consistent, we have reliable scene markers and speaker labels, and annotators can refer back to the original screenplays when necessary. Action statements also helped annotators better understand the scene.

4 Disentangling drama

4.1 Task definition

We consider conversation disentanglement applied to the domain of scripted conversations in TV series and movies. On the highest level, given a segmented scene in a screenplay, we want to identify *threads* in a conversation between multiple characters. Intuitively, in a scene, a character can change the subject or redirect other characters' attention to themselves, which, in our formulation, means they start a new conversational thread. Threading drama can help us understand conversational patterns, such as who gets to start a new conversation, who dominates the conversation, and how long an average conversation lasts.

The interaction structure involves an *utterance of interest* (UOI) and its *parent utterance*. An utterance is a dialogue line spoken by a character. In this work, an utterance of interest holds a directed edge to its parent utterance:

$$u_{\text{interest}} \rightarrow u_{\text{parent}} \tag{1}$$

Each utterance of interest has only one parent, but one parent can have multiple children. A thread is the transitive closure of all such pairwise links.

4.2 Threading dramatic conversations

In defining conversational threads in drama, we first adapt Goffman's (1963, p. 24) definition of *conversation*:[5] a thread is a kind of *focused interaction*: one where "persons gather close together and openly cooperate to sustain a single focus of attention, typically by taking turns at talking." Second, we assume conversations in a given scene in drama are *entangled* and there is often more than one *thread* in any given conversation. Taken together, we define a thread to be a cluster of semantically and pragmatically coherent utterances that are part of a conversation. Those utterances share a single, sustainable focus of attention (cf. Ervin-Tripp, 1964; Sacks et al., 1974), either on a character (who has other characters' attention) or a topic (often related to the wants and needs of a character), as well as other observable contextual relations (O'Neill and Martin, 2003).[6]

In a conversation, attention can be paid to a character (who has the floor) or a topic (why they are having this conversation):

Floor. Elsner and Charniak (2010, p. 392) describe the start of a new conversational thread as the process of participants (or in our case, characters) "hav[ing] refocused their attention ... away from whoever held the floor in the parent conversation." Like in Goffman (1963), *attention* is key: the character holding the floor can safely assume that they have attention from others. Such attention is singular and must be sustained throughout the thread; or someone else has the floor. Sheldon's line off screen in Fig. 1 further demonstrates how characters can start a thread by gaining the floor from other characters present in the scene, so they can express their need in their own voice. His line, coming out of nowhere, also does not respond to the speech acts of others, making it a typical example of thread starters in drama.

Topic. Since we follow McKee (2016) and see dialogues as speech acts, we can often relate the topic of a thread to the desire or intent of the character who started the conversation: characters can express their own needs, or respond to someone else's needs, and the need acts as the driving passion of the dramatic conversation. Operationally, a topic change within a scene usually occurs when the conversation is no longer about the original speech act that starts the scene. In Fig. 1, Georgie's line, "Can I drive in with you?" is an example of topic change and a new expressed intent from a different character in the scene.

For further details and examples, see Appendix B.

4.3 Annotation process

Prior to annotation, all the plays in TVPilots are OCR'd, and all plays were pre-processed follow-

[5] We note that thread in other work might be called *sub-conversation*, or simply *conversation*, as opposed to a stream of messages described in e.g., Elsner and Charniak, 2008. Given the complexity of drama, we chose not to overload the term of *conversation* and preferred *thread*, which also captures their interleaving nature.

[6] In our annotation framework, threads can cross each other or be resumed later, but we note only 0.007% of utterances we annotated exhibit this phenomenon.

	Metric	Agreement
Link	Pairwise exact match acc.	92.81
Cluster	Adjusted Random Index	79.71
	1−VI	94.45
	Shen F_1	84.52
	One-to-One	81.46
	Exact match F_1	53.33

Table 1: Inter-annotator agreement.

ing Bhat et al. (2021): we extracted the content of structural components (scene headers, character cue lines, dialogue and action lines) in each screenplay and store this information as tabular data. We further segmented each dialogue line into sentences with spaCy 3.3.0 (Honnibal and Johnson, 2015), which allows us to annotate at a greater granularity, since sentences in the same dialogue turn can reply to different previous sentences. We also assigned each scene, action line, dialogue turn, and each sentence in a turn an ID.

All annotations were carried out by the authors of the paper in the span of six months. We spent two months on pilot runs as we revised annotation guidelines and discussed edge cases, and the rest on independent annotating. On average, it took an hour to work through 500 dialogue and action lines. Our agreement rate is reported in Table 1; we considered standard metrics (described in Appendix A) based on 3,271 jointly annotated lines. Our agreement rate is comparable to previous work (Kummerfeld et al., 2019).

5 Experiments

We compare seven models to see how well existing model architectures that have been proposed for conversation disentanglement perform in the domain of dramatic texts. We would also like to know if models trained on Kummerfeld et al.'s (2019) data, but evaluated on our dramatic data, can leverage its size (seven times larger than ours) to compensate for the striking difference in domain.

5.1 Notation

We define a dataset $\mathcal{D} = \{(c_j^{\mathcal{S}_{i,j}}, u_j^-, u_j^+, u_i)\}_{i=1}^{|\mathcal{D}|}$ where:

- i is the index of an UOI, j that of a candidate parent utterance; an utterance u is a sentence of a dialogue line
- $\mathcal{S}_{i,j}$ denotes the scene both u_i and u_j are in
- c_j is the context of u_j, defined as all the dialogue and action lines preceding u_j in $\mathcal{S}_{i,j}$
- u_j^+ is a true parent, and u_j^- is a negative example ($u_j \in \mathcal{S}_{i,j}$)
- $u_i = \{t_i, k_i, w_i^1, w_i^2, \ldots, w_i^m\}$ is an utterance of m tokens spoken by character k in turn t; turn information is often given in the play parenthetically as `(CONT'D)`

5.2 Models

We consider the following models:

Previous. Adapting Kummerfeld et al. (2019), we connect all UOIs to their immediate previous utterances; i.e., for u_i, its parent utterance is u_{i-1}.

Featurized. Zhu et al. (2021) showed that manually selected features could offer a robust baseline; we take inspiration from Kummerfeld et al. (2019) and selected 8 features to train a featurized model:

- each utterance:
  1. The number of other speakers this character speaks after
  2. The number of utterances ago this character last spoke
  3. Whether the next utterance is spoken by the same character
- pairwise: between u_i and candidate parent utterance u_j
  4. The number of WordPiece tokens u_i and u_j have in common
  5. The distance between the two utterances $|i - j|$
  6. Whether there are utterances from either speaker between u_i and u_j
  7. Whether u_i and u_j are in the same turn
  8. Whether u_i and u_j are from the same speaker

BERT baseline. We adapt the Siamese encoders used in previous work on conversation analysis (Jiang et al., 2018; Henderson et al., 2020; Wu et al., 2020) to independently encode representations of the utterance of interest, parent utterance, and their associated scene context. We used two classes of embeddings: contextual and feature-based.

Contextual embeddings. Given a pre-trained model F like BERT, each utterance u spoken by speaker k is stringed together with special tokens as [CLS] k [SEP] u [LINE], where [SEP] separates a speaker label and the associated line, and [LINE] marks the end of the line. Here, [LINE] is a custom token whose representation is learned during training. The contextual embeddings are derived by

$$\boldsymbol{e} = F(\texttt{[CLS]}k\texttt{[SEP]}u\texttt{[LINE]}) \tag{2}$$

We extract the [CLS] token, denoted $\boldsymbol{e}^{\texttt{[CLS]}}$.

Feature-based embeddings. To enhance the expressivity of our models, we introduced additional embedding layers, randomly initialized, to encode information pertinent to the conversation structure in the scene. Each of the following features is assigned an embedding vector: utterance distance $\boldsymbol{f}_{d_i}$ (feature 5 from the featurized model), turn $\boldsymbol{f}_{t_i}$ (whether this line in the same turn as last, feature 7), scene speaker $\boldsymbol{f}_{k_i}$ (whether two speakers are the same, feature 8). All $\boldsymbol{f} \in \mathbb{R}^{250\times 2}$. We can then represent each utterance pair as the concatenation of all embeddings:

$$\begin{aligned}\boldsymbol{u}_{i,j} = [&\boldsymbol{e}^{\texttt{[CLS]}}_{c_j}; \boldsymbol{e}^{\texttt{[CLS]}}_{u_j}; \boldsymbol{e}^{\texttt{[CLS]}}_{u_i};\\ &\boldsymbol{e}^{\texttt{[CLS]}}_{u_i} - \boldsymbol{e}^{\texttt{[CLS]}}_{u_j}; \boldsymbol{e}^{\texttt{[CLS]}}_{u_i} \odot \boldsymbol{e}^{\texttt{[CLS]}}_{u_j};\\ &\boldsymbol{f}_{k_i}; \boldsymbol{f}_{t_i}; \boldsymbol{f}_{d_{i,j}}]\end{aligned} \tag{3}$$

Finally, we pass $\boldsymbol{u}_{i,j}$ through a non-linearity before the sigmoid output layer to compute the matching score $m_{i,j}$:

$$\boldsymbol{h}_{i,j} = \tanh(\boldsymbol{w}^{(0)\top}\boldsymbol{u}_{i,j}) \tag{4}$$

$$m_{i,j} = \sigma(\boldsymbol{w}^{(1)\top}\boldsymbol{h}_{i,j}) \tag{5}$$

For training, we used a self-link token [SELF] as a parent candidate for every u_i, which is assigned as the true parent if u_i is the start of a thread. For each (u_j^+, u_i) pair in our annotation, we sample five u_j^-. The objective is to minimize the binary cross-entropy loss, $\mathcal{L}_{\text{link}}$.

At inference time, for each u_i we have a candidate pool $\mathcal{P}_i = \{u_i, u_{i-1}, \ldots, u_{i-C-1}\}$ to consider u_i itself (self-pointing) along with $C-1$ previous candidates. Since in 90% of our annotation the true parent is within 5 utterances, we picked candidate pool size $C = 6$. We calculate the matching score between u_i and all $u_c \in \mathcal{P}_i$ and select $\text{argmax}_{\mathcal{P}_i}\, m_{i,c}$ as parent.

In addition to the BERT baseline, we adapt three recent architectures for dramatic conversation disentanglement. They are designed with Kummerfeld et al.'s (2019) IRC chat log in mind, and while many textual features do not have equivalents in our dramatic domain, we incorporate some designs as we saw fit, described below:

BERT with soft attention alignment. We adapt the soft alignment mechanism in the pointer module from Yu and Joty (2020) to emphasize the textual similarity between u_i and u_j:

$$\boldsymbol{H}_i^{'} = \text{softmax}(\boldsymbol{H}_i\boldsymbol{H}_j^\top)\boldsymbol{H}_j \tag{6}$$

$$\boldsymbol{H}_j^{'} = \text{softmax}(\boldsymbol{H}_j\boldsymbol{H}_i^\top)\boldsymbol{H}_i \tag{7}$$

$$\boldsymbol{h}_i^f = [\boldsymbol{h}_i; \boldsymbol{h}_i^{'}; \boldsymbol{h}_i - \boldsymbol{h}_i^{'}; \boldsymbol{h}_i \odot \boldsymbol{h}_i^{'}] \tag{8}$$

$$\boldsymbol{h}_j^f = [\boldsymbol{h}_j; \boldsymbol{h}_j^{'}; \boldsymbol{h}_j - \boldsymbol{h}_j^{'}; \boldsymbol{h}_j \odot \boldsymbol{h}_j^{'}] \tag{9}$$

where $\boldsymbol{H}_i = (\boldsymbol{h}_{i,0}, \ldots, \boldsymbol{h}_{i,p})$ and $\boldsymbol{H}_j = (\boldsymbol{h}_{j,0}, \ldots, \boldsymbol{h}_{j,q})$ are the bidrectional LSTM representations for u_i and u_j. $\boldsymbol{H}_i$ is used as query vectors to compute attentions over the key/value vectors in $\boldsymbol{H}_j$ and the set of attended vectors $\boldsymbol{H}_i^{'}$, one for each $\boldsymbol{h}_i \in \boldsymbol{H}_i$. In Eq. 8–9 we enhance the interactions by applying difference and element-wise product between the original and attended vectors. Finally, we swap out BERT-based contextual embeddings for u_i and u_j with $\boldsymbol{h}_i^f$ and $\boldsymbol{h}_j^f$, with the following resultant representation of the two utterances:

$$\begin{aligned}\boldsymbol{u}_{i,j} = [&\boldsymbol{e}^{\texttt{[CLS]}}_{c_j}; \boldsymbol{h}_i^f; \boldsymbol{h}_j^f;\\ &\boldsymbol{f}_{k_i}; \boldsymbol{f}_{t_i}; \boldsymbol{f}_{d_{i,j}}]\end{aligned} \tag{10}$$

The matching score is calculated using Eq. 4–5.

6-way classifier. The structural characterization of conversation proposed by Ma et al. (2022) is the current state of the art. We use their architecture without reference dependency modeling, since we don't have mentions in our movie data in the same format as IRC, but retain the rest. Their goal is to train a C-way classifier: for each, u_i, pick one from candidates including u_i and u_{i-j} $(1 \leq j \leq C-1)$. The UOI and candidate pairs are stringed and fed into a pre-trained model as

$$\boldsymbol{H}_0 = F(\texttt{[CLS]}u_{i-j}\texttt{[SEP]}u_i), \tag{11}$$

$\boldsymbol{H}_0 \in \mathbb{R}^{C\times L\times |F|}$, where L is the input sequence length. To obtain aggregated contextualized representations, we extract the [CLS] token: $\boldsymbol{H}_1 = \boldsymbol{H}_0^{\texttt{[CLS]}}$, $\boldsymbol{H}_1 \in \mathbb{R}^{C\times|F|}$. For candidate window size, we chose $C = 6$ (including self-pointing as one candidate).

Their architecture features two components: speaker property modeling and the Syn-LSTM module. Speaker property modeling leverages the masked Multi-Head Self-Attention (MHSA, Liu et al., 2021) mechanism to account for utterances from the same speaker with a speaker-aware mask matrix M, which we include in our adaptation:

$$M[i,j] = \begin{cases} 0, & k_i = k_j \\ -\infty, & \text{otherwise} \end{cases} \tag{12}$$

Syn-LSTM (Xu et al., 2021) is a biLSTM with an additional input gate to retain the information of utterances within the candidate window, designed to make the model context-aware. In other words, we have $\boldsymbol{H}_2 = \text{MHSA}(\boldsymbol{H}_1, M)$ and $\boldsymbol{H}_3 = \text{Syn-LSTM}(\boldsymbol{H}_2)$, where $\boldsymbol{H}_2, \boldsymbol{H}_3 \in \mathbb{R}^{C\times|F|}$. In this structural characterization, the final representation between each $[u_i, u_{i-j}]$ pair $(1 \leq j \leq C-1)$ and the self-pointing $[u_i, u_i]$ pair is:

$$\boldsymbol{h}_{i,j} = [\boldsymbol{p}_{ii}, \boldsymbol{p}_{ij}, \boldsymbol{p}_{ii} \odot \boldsymbol{p}_{ij}, \boldsymbol{p}_{ii} - \boldsymbol{p}_{ij}], \tag{13}$$

where $\boldsymbol{p}_{ij}$ is the representation for the pair of $[u_i, u_{i-j}]$ from $\boldsymbol{H}_3$. $\boldsymbol{h}_{i,j}$ is then fed into the classification head to predict the parent. The training objective is to minimize the cross-entropy loss.

Multi-task learning. We follow Yu and Joty (2020); Zhu et al. (2021); Huang et al. (2022) and introduce an auxiliary task, a binary classifier to predict whether u_i and u_j belong to the same thread, the probability of which is:

$$t_{i,j} = \sigma(\boldsymbol{w}_t^\top \boldsymbol{u}_{i,j}) \tag{14}$$

where $\boldsymbol{u}_{i,j}$ is the representation of the utterance pair given. The objective is to minimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{thread}} = -y_{i,j} \log t_{i,j} - (1 - y_{i,j}) \log(1 - t_{i,j}) \tag{15}$$

where $y_{i,j} = 1$ if u_i and u_j are in the same thread, 0 otherwise. The total training loss $\mathcal{L}$ of this model class is:

$$\mathcal{L} = \mathcal{L}_{\text{link}} + \alpha\mathcal{L}_{\text{thread}} \tag{16}$$

where α is a hyper-parameter to control the impact of the auxiliary task. We experimented with $\alpha = \{0.1, 0.5, 1.0\}$ and 0.1 performed best.

BERT baseline trained with Kummerfeld et al.'s (2019) data. Lastly, to test the influence of domain difference between IRC chat logs and dramatic conversations, we trained our models instead on Kummerfeld et al.'s (2019) training data, to be evaluated on our drama data. We extracted all usernames as speaker labels and treated system messages as action statements. Since individual users can send multiple messages in what would be one dialogue turn in movies, we did not perform sentence segmentation on messages.

set	train	dev	test
# titles	563	127	141
# unique speakers	1,711	371	389
# dialogue lines	11,672	2,639	2,743
# turns	5,988	1,298	1,475
# action lines	8,756	2,059	1,980

Table 2: Data statistics.

5.3 Setup

We trained all our models for 10 epochs and used the dev set for early stopping with the learning rate 5×10^{-6} (BERT-based) and 10^{-3} (linear). Our train–test split is reported in Table 2. For our BERT implementation, we used `bert-base-cased` from HuggingFace 4.19.2 with PyTorch 1.10.0.[7] An epoch took 1 hour 8 minutes on average on two NVIDIA GeForce RTX 2080 Ti GPUs.

5.4 Results

Experimental results are presented in Table 3. We report the standard set of metrics (described in Appendix A), along with their 95% bootstrap confidence intervals. We first note that the clustering metrics are low while link prediction accuracy is high because the most reasonable parent utterance for most UOIs (90%) is the immediately previous utterance, which leads to high baseline accuracy for link prediction, but low clustering baselines when considering entire threads.

The performance of models trained on Kummerfeld et al.'s (2019) suggests that domain difference matters. While seven times larger, their dataset is in an entirely different domain, and intuitively, chatroom users interact differently from movie characters. Such differences might account for the inferior performance, especially on stricter cluster metrics like One-to-One and Exact Match.

Enhancements to the baseline lead to minor, statistically insignificant, improvements, and the 6-way classifier outperforms the rest model classes on most metrics. Therefore, for the analysis below, we use the 6-way classifier.

[7] https://huggingface.co/bert-base-cased; https://pytorch.org/.

Model	Link prediction Acc.	Clustering ARI	1−VI	Shen F_1	1-1	Exact match F_1
trained with Kummerfeld et al. data						
BERT baseline	51.10 [49.28-52.96]	48.64 [45.48-51.78]	83.67 [82.89-84.43]	62.05 [60.22-64.07]	54.60 [52.67-56.62]	6.42 [4.67-8.13]
6-way classifier	60.84 [59.35-62.35]	55.10 [51.16-59.46]	86.85 [85.88-87.90]	63.65 [60.77-66.97]	60.20 [57.20-63.56]	11.62 [8.91-14.37]
trained with our dataset						
Previous	90.26 [89.78-90.75]	46.69 [45.30-48.72]	85.29 [84.65-86.20]	54.80 [53.09-57.31]	51.89 [50.13-53.99]	14.95 [12.13-17.50]
Featurized	89.75 [88.86-90.65]	47.80 [44.53-52.28]	85.61 [84.54-86.88]	55.90 [52.77-59.70]	53.05 [49.64-57.13]	15.25 [11.66-19.14]
BERT baseline	89.44 [88.49-90.44]	57.98 [54.17-63.20]	88.78 [87.85-89.83]	66.55 [63.66-69.89]	63.71 [60.55-67.43]	25.25 [20.90-29.73]
+ attn. alignment	**90.28** [89.31-91.27]	57.28 [53.35-62.47]	88.62 [87.54-89.80]	65.37 [62.17-69.04]	62.78 [59.23-66.81]	**25.88** [21.11-30.63]
+ aux. task	90.12 [89.26-91.02]	53.32 [49.37-59.02]	87.63 [86.59-88.86]	62.25 [59.12-66.06]	59.81 [56.34-63.93]	21.20 [17.19-25.54]
+ both	90.24 [89.34-91.17]	57.63 [54.04-62.15]	88.60 [87.61-89.72]	65.54 [62.61-69.01]	63.21 [60.03-66.98]	25.27 [20.97-30.03]
6-way classifier	87.23 [86.24-88.27]	**64.81** [60.70-69.98]	**90.11** [89.28-91.05]	**72.20** [69.63-75.29]	**69.02** [66.30-72.19]	25.40 [21.67-29.27]

Table 3: Experimental results. All metrics are reported with 95% bootstrap confidence intervals.

6 Analysis

To illustrate the usefulness of conversation disentanglement in drama, we disentangled 808 movies from SCRIPTBASE-J and carried out two analyses enabled by this work to explore two questions that engage previous work in film studies:[8]

Are conversational threads in movies getting shorter over the years? In his analysis on visual style in contemporary films, film historian Bordwell (2002, p. 16) observed, "For many of us, today's popular American cinema is always fast": the average shot length is decreasing and cuts and camera movements have become more rapid over the course of the twentieth century, leading to an impression of "intensified continuity." A similar observation is made in Cutting et al.'s (2010) empirical work, which relates this trend to our natural fluctuation of attention. Notably, such work emphasizes the visual aspect of films, which reinforces the established hierarchy in film studies: the film is a visual medium, and image is more important than sound. This hierarchy is critiqued in studies on film dialogues in particular (Kozloff, 2000), since characters converse with one another only after the advent of sound films. It then leads us to ask: Are conversational threads in films also getting shorter over the years?

Since we have disentangled movie conversations into threads, we can calculate the average number of utterances there are in a thread in a given movie and in a given year. Our movie data, while spanning from 1930s to 2010s, is not evenly distributed. As a result, for this analysis, we aggregated movies

[8]The list of movies we used for analysis can be found in our GitHub repo: https://github.com/kentchang/dramatic-conversation-disentanglement/blob/bf3d2fbc00f9d64356c308a2c0ca6b2e73580c19/list/titles-for-analysis.txt

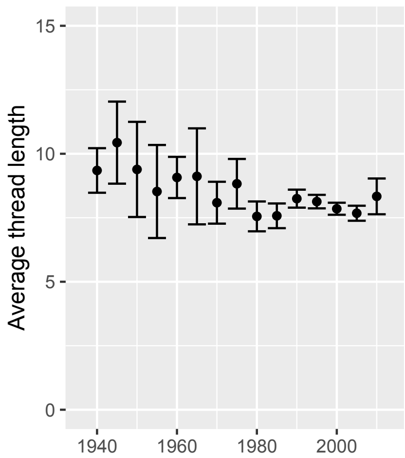

Figure 3: The average thread lengths of movies in a 5-year range, along with 95% confidence intervals.

in a 5-year range. While we would expect thread lengths to also decrease, Fig. 3 tells a different story. We can see the average thread length decreasing (although not statistically significantly so) until 1970, and the trend is flat since. Film dialogues seem to be resisting the broader trend associated with visual styles.

What is the pattern of floor claiming between men and women in movies? It has been pointed out that the film industry became dominated increasingly by men over the twentieth century (Boon, 2008). The Bechdel Test (Bechdel, 1986) is a popular and well-known measure for the representation of women in films, often used for advocating that women on screen should "speak up" (O'Meara, 2016) to encourage more diverse representation. Through this work, we would like to add an additional dimension to it: How often do characters who are women start a conversation in films?

In the tradition of continental philosophy, to initiate a conversation—or, to become a *speaking subject*—is a socially and ethically significant act (Foucault, 1972; Lacan, 2006). This influenced much of feminist film studies that considers the presence and absence of women's voices in films (Silverman, 1988; Lawrence, 1991; Sjogren, 2006). This work inspires us to frame and measure the agency of characters who are women in relation to the frequency with which they get to start a new conversational thread and claim the floor.

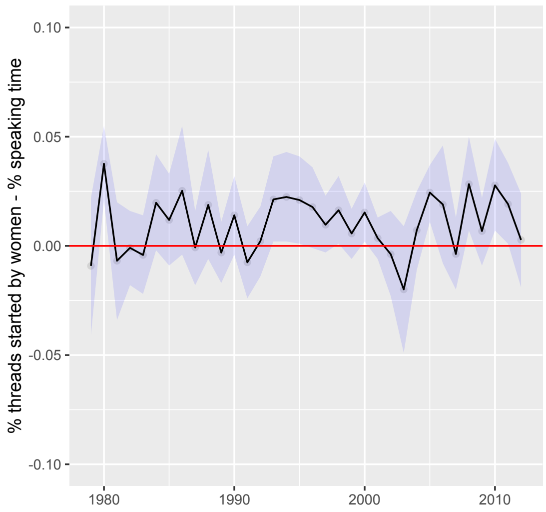

Figure 4: The percentage of threads started by women relative to their speaking time, along with 95% CIs.

We start by using TMDb's API to look up the gender of the actor portraying the character.[9] For this analysis, we only consider movies released after 1979, after which point we have at least five movies each year.In our data, 30.4% of threads are started by women. This is aligned with the oft-stated observation that men talk more than women in films (Ramakrishna et al., 2017; Lauzen, 2019), and the trend has not significantly changed for decades.

However, as we see in Fig. 4, when we subtract the percentage of *speaking time* by women (the number of lines they have) from that of *threads started* by women (e.g., in 2011, 32.7% of threads were started by women, and 30.8% of lines were spoken by women, so we see an absolute difference of $+1.9$), we see that women generally start more threads relative to their speaking time, and that is also relatively constant over time. In the figure, any year in which the 95% confidence interval does not overlap with 0 is significant at that level; while this is not significant for many individual years (given the limited number of movies per year), it is significant over all years ($+1.0$, $[0.07, 0.14]$). This finding is surprising because it suggests that despite their under-representation, women characters are written to initiate conversations more than their male counterparts.

7 Conclusion

We present in this work a new dataset for studying conversation disentanglement in movies and TV shows in order to enrich the landscape of this line of work in NLP. Movie and TV dialogues offer pragmatic patterns and interaction structures different from chat logs, on which standard benchmarks for this task are built. To ensure high quality of this dataset, we digitized teleplays written for TV pilots, so we have more screenplays in the standard format, which we find most useful for annotation. In addition, we draw on theoretical resources from sociolinguistics, sociology, and film studies to create a robust annotation scheme that considers topic and floor changes specifically in the context of drama, which we believe speaks to the needs of the wider scholarly community.

To the best of our knowledge, no previous technical or theoretical work has offered a working operationalization of conversational thread in the context of dramatic (scripted, spoken, face-to-face) conversations, or examined the significance of initiating a conversation or gaining the floor in this domain. While we do not claim that the results from our analysis are definitive, our work has demonstrated a new method to further investigate the sound-image hierarchy, gendered power dynamics in films, and communication behaviors in cultural representations on screen. We hope this will encourage and facilitate future research on drama and conversation in NLP, film studies, and the computational humanities.

Acknowledgments

We thank Timea Ryan for feedback on our annotation guidelines, and our anonymous reviewers for their constructive comments. The research reported in this article was supported by funding from the National Science Foundation (IIS-1942591) and the National Endowment for the Humanities (HAA-271654-20).

[9]https://developers.themoviedb.org/3/people/get-person-details

Limitations

This work is first limited by the availability of screenplays in the standard format. While movie or TV show transcripts are more readily available (subject to permission to use), they are less ideal for annotation due to unreliable scene headers and speaker labels. This therefore limits the size of our corpus, as digitization and correction are labor-intensive. In our analysis, we relied on metadata from SCRIPTBASE-J (each movie has a Jinni[10] profile that includes its corresponding IMDb[11] ID) and TMDb.[12] Given it scale, we weren't able to check individually whether the release year or the gender of actors in those community-built resources is correct or up-to-date. This work is also unimodal, while movies and TV shows are multimodal, which meant we did not have access to the video for annotation, and we could not compare thread length and shot length, among other things.

Ethics Statement

We are aware that this dataset and the analytical work that follows only represent a limited set of cultural and ethnic groups as well as language uses. The dataset we're annotating highlights US movies (and not e.g., Bollywood, Nollywood or the global film industry more generally), and so one risk is the centering that culture (and conversational norms) within that dataset at the expense of others. There have been documented allocational and distributional biases in the film industry (Baker and Faulkner, 1991; Ravid, 1999; O'Brien, 2014; Khadilkar et al., 2022), and we encourage those interested in furthering this line of work to acquaint themselves with relevant discourses. We are also aware that the dataset contains potentially problematic content, such as vulgar, violent, or offensive language in screenplays, or other biases held by individual screenwriters.

References

Mahmoud Azab, Noriyuki Kojima, Jia Deng, and Rada Mihalcea. 2019. Representing movie characters in dialogues. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 99–109, Hong Kong, China. Association for Computational Linguistics.

[10] https://en.wikipedia.org/wiki/Jinni_(search_engine)

[11] https://www.imdb.com/

[12] https://www.themoviedb.org/

Wayne E Baker and Robert R Faulkner. 1991. Role as Resource in the Hollywood Film Industry. *The American journal of sociology*, 97(2):279–309.

Alison Bechdel. 1986. *Dykes to Watch Out for*. Firebrand Books.

Gayatri Bhat, Avneesh Saluja, Melody Dye, and Jan Florjanczyk. 2021. Hierarchical encoders for modeling and interpreting screenplays. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 1–12, Virtual. Association for Computational Linguistics.

Kevin Alexander Boon. 2008. *Script Culture and the American Screenplay*. Wayne State University Press.

David Bordwell. 2002. Intensified continuity visual style in contemporary American film. *Film quarterly*, 55(3):16–28.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A Dataset for Abstractive Screenplay Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

James E Cutting, Jordan E DeLong, and Christine E Nothelfer. 2010. Attention and the evolution of Hollywood film. *Psychological science*, 21(3):432–439.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2010. Disentangling Chat. *Computational Linguistics*, 36(3):389–409.

Susan Ervin-Tripp. 1964. An analysis of the interaction of language, topic, and listener. *American anthropologist*, 66(6_PART2):86–102.

Michel Foucault. 1972. *The Archaeology of Knowledge*. Pantheon Books, New York.

Erving Goffman. 1963. *Behavior in Public Places*. The Free Press, New York.

Charles Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2018. What's This Movie About? A Joint Neural Network Architecture for Movie Content Analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1770–1781, New Orleans, Louisiana. Association for Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022. Who says what to whom: A survey of multi-party conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, California. International Joint Conferences on Artificial Intelligence Organization.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. 2002. Cluster validity methods: part I. *ACM Sigmod Record*.

Agnes Weiyun He and Vimala Herman. 1998. Dramatic discourse: Dialogue as interaction in plays. *Language*, 74(2):384.

Caroline Heldman. 2016. Hitting the bullseye: Reel girl archers inspire real girl archers. Technical report, Geena Davis Institute.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and Accurate Conversational Representations from Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

bell hooks. 1992. *Black Looks: Race and Representation*. South End Press.

Chengyu Huang, Zheng Zhang, Hao Fei, and Lizi Liao. 2022. Conversation Disentanglement with Bi-Level Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2985–2996, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to Disentangle Interleaved Conversational Threads with a Siamese Hierarchical Network and Similarity Ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Kunal Khadilkar, Ashiqur R KhudaBukhsh, and Tom M Mitchell. 2022. Gender bias, social bias, and representation: 70 years of Bollywood and Hollywood. *Patterns (New York, N.Y.)*, 3(2):100409.

Sarah Kozloff. 2000. *Overhearing Film Dialogue*. University of California Press.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Jacques Lacan. 2006. *Ecrits: The First Complete Edition In English*. W. W. Norton & Company.

Robin Tolmach Lakoff and Deborah Tannen. 1984. Conversational strategy and metastrategy in a pragmatic theory: The example of Scenes from a Marriage. *Semiotica*, 49(3-4):323–346.

Martha M Lauzen. 2019. It'sa man's (celluloid) world: Portrayals of female characters in the top grossing films of 2018. *Center for the Study of Women in Television and Film, San Diego State University*, pages 411–428.

Amy Lawrence. 1991. *Echo and Narcissus: Women's Voices in Classical Hollywood Cinema*. University of California Press.

Hui Liu, Zhan Shi, Jia-Chen Gu, Quan Liu, Si Wei, and Xiaodan Zhu. 2020. End-to-End Transition-Based Online Dialogue Disentanglement. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3868–3874, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13406–13414.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural characterization for dialogue disentanglement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

Khyati Mahajan and Samira Shaikh. 2021. On the Need for Thoughtful Data Collection for Multi-Party Dialogue: A Survey of Available Corpora and Collection Methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online. Association for Computational Linguistics.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2012. Hierarchical Conversation Structure Prediction in Multi-Party Chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 60–69, Seoul, South Korea. Association for Computational Linguistics.

Susan E McDaniel, Gary M Olson, and Joseph C Magee. 1996. Identifying and analyzing multiple threads in computer-mediated and face-to-face conversations. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, CSCW '96, pages 39–47, New York, NY, USA. Association for Computing Machinery.

Robert McKee. 2016. *Dialogue: The art of verbal action for page, stage, and screen*. Hachette UK.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895.

Jill Nelmes, editor. 2010. *Analysing the Screenplay*. Routledge, London, England.

Sik H Ng and James J Bradac. 1993. *Power in Language: Verbal Communication and Social Influence*. SAGE Publications.

Anne O'Brien. 2014. 'Men own television': why women leave media work. *Media Culture & Society*, 36(8):1207–1218.

Jennifer O'Meara. 2016. What "The Bechdel Test" doesn't tell us: examining women's verbal and vocal (dis)empowerment in cinema. *Feminist Media Studies*, 16(6):1120–1123.

Jennifer O'Meara. 2019. *Engaging Dialogue: Cinematic Verbalism in American Independent Cinema*. Edinburgh University Press.

Jacki O'Neill and David Martin. 2003. Text chat in action. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, GROUP '03, pages 40–49, New York, NY, USA. Association for Computing Machinery.

Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678, Vancouver, Canada. Association for Computational Linguistics.

S Abraham Ravid. 1999. Information, Blockbusters, and Stars: A Study of the Film Industry. *The Journal of Business*, 72(4):463–492.

Kay Richardson. 2010. *Television Dramatic Dialogue: A Sociolinguistic Study*. Oxford University Press.

Craige Roberts. 1996. Information structure in discourse: Toward a unified theory of formal pragmatics. *Ohio State University Working Papers in Linguistics*, 49:91–136.

Marjorie Rosen. 1973. *Popcorn Venus; Women, Movies and the American Dream*. Coward, McCann and Geoghegan.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character Comprehension in Stories as Speaker Guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 35–42, New York, NY, USA. Association for Computing Machinery.

Kaja Silverman. 1988. *The Acoustic Mirror: The Female Voice in Psychoanalysis and Cinema*. Indiana University Press.

Britta H Sjogren. 2006. *Into the vortex female voice and paradox in film*. University of Illinois Press, Urbana.

Weishi Wang, Steven C H Hoi, and Shafiq Joty. 2020. Response Selection for Multi-Party Conversations with Dynamic Topic Tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6581–6591, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C H Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3469, Online. Association for Computational Linguistics.

Tao Yu and Shafiq Joty. 2020. Online conversation disentanglement with pointer networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6321–6330, Online. Association for Computational Linguistics.

Henghui Zhu, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Who Did They Respond to? Conversation Structure Modeling Using Masked Hierarchical Transformer. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 34(05):9741–9748.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on Conversation Disentanglement. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 1–11, Online. Australasian Language Technology Association.

A Evaluation metrics

- **Adjusted Random Index (ARI)** (Halkidi et al., 2002) is defined as:

$$\frac{\sum_{i,j}\binom{n_{i,j}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{17}$$

- **Variation of Information (VI)** is the information gain or loss when going from one clustering to another (Meilă, 2007). It is the sum of conditional entropies $H(Y|X)+H(X|Y)$, where X and Y are clusters of the same set of items. We report $1-$VI, so the larger the value the better.

- **Shen $\mathbf{F_1}$** (Shen et al., 2006): Given a detected thread j and true thread i:

$$F(i,j) = \frac{2 \times \text{Precision}(i,j) \times \text{Recall}(i,j)}{\text{Precision}(i,j) + \text{Recall}(i,j)} \tag{18}$$

$$F = \sum_i \frac{n_i}{n} \max_j F(i,j) \tag{19}$$

  where $\text{Recall}(i,j) = \frac{n_{i,j}}{n_i}$, $\text{Precision}(i,j) = \frac{n_{i,j}}{n_j}$, $n_{i,j}$ is the number of messages of thread i in j, n_j the number of messages in detected thread j, n_i thread i.

- **One-to-One Overlap** (Elsner and Charniak, 2008) calculates the percentage of overlap between two sets of conversational threads paired up with the max-flow algorithm.

- **Exact Match $\mathbf{F_1}$** (Kummerfeld et al., 2019) calculates the number of perfectly matched conversational threads between two sets. During our annotation process, we did see threads with only one dialogue line, which functions quite differently from system messages in the context of IRC, so we did not exclude conversations with only one dialogue line.

B Annotation guidelines

B.1 Building intuitions

This is a study of conversational behaviors of characters in drama: here, we consider TV shows and movies and the specific task of conversation disentanglement. On the highest level, we want to identify *threads* in a conversation between multiple characters in a scene of a TV show or movie. In the same scene, some characters can change the subject of a conversation, or redirect other characters' attention to themselves, while others might never do so. Characters in a closer relationship might converse more frequently with each other. We annotate to build a dataset that can help us investigate those inquiries. To (hopefully) ease understanding, all examples below are drawn from *Gilmore Girls*, which follows the story of Lorelai Gilmore and her daughter, Rory, in a small town in Connecticut. In the excerpts below, we see Luke, Lorelai's will-they-or-won't-they love interest throughout the series, and Emily, Lorelai's mother. *Gilmore Girls* is famous for its fast-paced dialogue and offers illustrations of conversational behaviors that we wish to study.

B.1.1 Dialogue, conversation, and reply-to

Dialogue as speech act. Our definition of dialogue is an all-encompassing one taken from McKee (2016, p. 2): "Any words said by any character to anyone." A line of dialogue is, then, a sequence of words uttered (or, an *utterance*) by a character to themselves, another character, or a few other characters. This view, unlike a narrower one, where dialogue is a conversation held between characters, sees dialogue as a verbal tactic initialized by a character to achieve a certain goal: "All talk responds to a need, engages a purpose, and performs an action" McKee (2016, p. 3). In other words, a dialogue is a *speech act*. Characters use dialogue to inform us of an event (exposition), tell us something about themselves (characterization), or try to make something happen (action).

Conversation and conversational thread. We adopt a broad definition of a conversation: conversation is a talk between characters. In defining thread, we first adapt Goffman's (1963)'s definition of *conversation*[13]: a thread is a kind of *focused interaction*: one where "persons gather close together and openly cooperate to sustain a single focus of attention, typically by taking turns at talking" (Elsner and Charniak, 2008, p. 24). Second, we assume conversations in a given scene in drama are *entangled* and there are often more than one *thread* in any given conversation. Taken together, here is our operative definition:

[13]We note that thread in other work might be called *sub-conversation*, or simply *conversation*, as opposed to a stream of messages described in e.g., Elsner and Charniak, 2008. Given the complexity of drama, we chose not to overload the term of *conversation* and preferred *thread*, which also captures their interleaving nature.

> A thread is a cluster of semantically and pragmatically coherent utterances that are part of a conversation. Those utterances share a single, sustainable focus of attention (Goffman, 1963; Ervin-Tripp, 1964; Sacks et al., 1974), either on a character (who has other characters' attention) or a topic (often related to the wants and needs of a character), as well as other observable contextual relations (O'Neill and Martin, 2003).

In a conversation, attention can be paid to a character (who has the floor) or a topic (why they are having this conversation). In the context of drama, we can often relate the topic of a thread to **the desire or intent of the character who started the conversation** (this loops back to McKee's idea that dialogues are speech acts): characters can express their own needs, or respond to someone else's needs, and the need acts as the driving passion of the dramatic conversation (and subsequently, plot, characterization, etc.).

Attention can be paid to a character or a topic, but such attention should be *sustainable* over multiple utterances to form a thread. In other words, when an utterance redistributes the focus of other characters (or, a change of *floor*) or shows us new wants and needs of a character, and such distribution or attention or topical focus is carried over to the next couple of utterances, it usually marks the start of a new conversational thread. However, there are occasions where threads can be short, which we will describe at the end of this section.

It has been noted that the exact start of a conversational thread is not easy to determine (McDaniel et al., 1996; Elsner and Charniak, 2010), so we will dedicate the second section (§A.1.2) to this topic. There, we try to unpack our operational definition with more examples.

Reply-to relationship. Following conventions in NLP (Zhu et al., 2021), we understand individual dialogues in a multi-party conversation in terms of *parent utterance* and *utterance of interest* (UOI), and *utterance* is roughly synonymous with a dialogue line. The following is the first two utterances in the entire series of *Gilmore Girls*:

LORELAI Please, Luke. Please, please, please.

LUKE How many cups have you had this morning?

(I.i)[14]

In practice, UOI means the line you are currently annotating, and its parent utterance is the previous line it most logically *replies to*. In the example above, if Luke's utterance is the UOI, its parent utterance is the immediately previous utterance by Lorelai ("Please, Luke ..."). In fact, the *default* parent utterance is usually the previous line. Perhaps we can think of conversations in drama as sequences of sentences, where one *triggers* the next. Given our qualitative observations of drama, we make the following remarks on UOI, parent utterance, and thread:

- If an UOI does not have any parent utterance, it's the start of a new thread.
- One UOI can only have ONE parent utterance. An utterance can have multiple *children* (utterances that point to it as parent utterance).
- The default parent utterance is the previous utterance.

Nuances of reply-to and sentences in one dialogue turn. Those preliminary remarks don't always apply. Often one dialogue line is too large a unit for us to fully understand conversational behaviors, and more often the next line isn't the response to the current:

EMILY You're being stubborn, as usual.

LORELAI No, Mom, I'm not being stubborn. I'm being me! The same person who always needed to work out her own problems and take care of herself. Because that's the way I was born. That's how I am!

EMILY Florence, I'm dripping.

LORELAI I appreciate what you have done for Rory in paying for this school. That will not be forgotten. You won't let it. But she is my daughter. And I decide how we live, not you. Now then, do they validate parking here?

(I.ii)

Here, "Florence, I'm dripping." is certainly not a reply to the previous line. And even within one

[14]Season 1, episode 1 from *Gilmore Girls*. We are adapting the olden MLA convention for Shakespearean plays, where uppercase Roman numerals denote season, and lowercase ones, episode.

dialogue turn, "Now then, do they validate parking here?" has nothing to do with whether Lorelai is stubborn or not (which we tentatively call a *topic*). To reflect this, in this work, we study utterances on the *sentence* level. In other words, a screenplay/teleplay exhibits the following hierarchy:

> title $>$ scene $>$ conversational thread $>$ dialogue turn $>$ dialogue line/utterance $>$ sentence

Continuation as reply-to. Lorelai's long speech has multiple sentences. Given our design choice, we will say that the default parent utterance is still mostly *previous* sentence. The notable exception is, of course, "Now then, do they validate parking here?"—it *does not* have a parent utterance; it's starting a new thread. To distinguish between a true reply and continuation, we will encode speaker and turn information later in the model. Qualitatively, though, it is reasonable to treat continuation as a special case of reply-to. Consider:

SOOKIE And if we go down after two years . . .
LORELAI It'll be the most exciting two years of our lives!

(II.viii)

In conversations, one person can *finish* each other's lines. So, we say, for multiple sentences in one dialogue turn, the parent utterance of sentence n is, by default, sentence $n-1$.

Look for parent utterances, not addressee. In determining the reply-to relationship between utterances, the speaker/addressee information can help you find the parent utterance of an UOI, but it shouldn't be your sole basis of judgment, because they are fundamentally different task: determine the relationship between utterances vs. the person to whom a speaker addresses. Consider the following example:

LORELAI Michel, come on, we've got to get into these budgets.
SOOKIE Now.
MICHEL Does the red light mean it's programmed?
SOOKIE [*to Lorelai*] I explained it a hundred times.
LORELAI Michel, you've been setting that machine for 20 minutes now.

(IV.xvi)

According to the action statement, Sookie says "I explained it a hundred times." to Lorelai. But that line is *triggered* by Michel's "Does the red light mean it's programmed?" So, we say Sookie's "I explained it a hundred times" is the parent utterance of Michel's "Does the red light mean it's programmed?"—even though she says that directly to Lorelai. If this seems odd, remember we cannot assume Michel does not hear Sookie's line, and that if we were to say Sookie replied to Lorelai, there's no line from Lorelai that can reasonably be the parent utterance.

Intuition: thread, topic, and floor Before we define a thread more thoroughly, we use two examples to drive the intuition. If the next section (§A.1.2) gets confusing, revisit those two examples.

A *topic* is a semantically and pragmatically coherent unit:

LORELAI Who is that?
RORY I don't know. She just followed me in here like a puppy dog without saying a word.
LORELAI Maybe she's lost.
RORY Or, maybe she's one of my new suitemates who I'm already off to a swell start with.
LORELAI Do you know how vulnerable you are to venereal disease?
RORY All hail to the queen of the nonsequiturs.
LORELAI This parent orientation I went to was a nonstop litany of the horrors awaiting college freshman. You're supposed to carry a whistle, a flashlight, a crucifix, and a loaded Glock with you at all times.
RORY We should go out there. She'll think we're hiding.
LORELAI Okay, just don't shake hands with her. Bacteria.
RORY Mom.
LORELAI Or tell her where you live.
RORY Too late.
LORELAI Oh, you touched the doorknob.
RORY Good grief.

(IV.ii)

We first note that topic is not entirely about coherence from the semantic point of view: we expect to see metaphors or jokes in TV shows and movies, and there's no puppy dog, queen, or crucifix really present in the scene; and just like in real life, we might bring up something that appears entirely random in any conversation ("You have to be in my brain to see the connection!") Still, we can intuitively tell there are two threads of conversation going on, and we aren't perplexed when Rory says "All hail to the queen of nonsequiturs." We can perhaps look at this previous exchange from a pragmatic perspective: Lorelai starts that "Who is this" conversation because she wants to find out who this girl in Rory's dorm room is. Venereal disease has nothing to do with that. It *does not follow*.

On the other hand, *floor* has to do with *ownership* and *attention*. Who has started and owns the conversation? Who controls our (and other characters') attention? Who do we direct our gaze to if we were also in the scene?

LORELAI You could tape the movies, or get a DVD player.

EMILY I don't need a DVD player.

LORELAI Well, why not? Then you could buy all those musicals you love and watch them whenever you felt like it.

EMILY I'm not an invalid, Lorelai.

LORELAI Well, of course you are, Mother. Why else would I suggest a DVD player?

EMILY I can fill my time all by myself and I'd like you to drop this conversation right now.

LORELAI Where are you going?

EMILY We're going to eat.

LORELAI Just because you leave the room doesn't mean the conversation's over. I started the conversation. The conversation's in me. Therefore, when I get over there, the conversation's just gonna start up again!

(III.xiii)

Lorelai's last line can help us understand the nature of a conversational thread: a specific character first *started* and *owned* the thread, and then they let someone else do so. One character is ready to speak or stays speaking, and others listen and reply. Here, Lorelai wants to buy Emily a DVD player, and the latter refuses. Emily leaving the room can be understood as her unwilling to pay more attention to Lorelai: Emily doesn't want to stay in the conversation anymore. Also, Lorelai asking "Where are you going" is a shift in topic: it has nothing to do with DVD players.

We will emphasize this again, but floor change tends to be more common when the thread involves at least three characters. Floor still exists between two-party conversations: if a high school student finds herself in the principal's office, we can expect the principal might be the one doing most of the talking and *has the floor*. But floor change should be more frequent when we have more than two characters.

This exchange between Emily and Lorelai is also an interesting instance where Emily tries to gain control over the conversation (or floor) and Lorelai doesn't let her, so she just leaves. This behavior of floor gaining and changing is a particularly interesting aspect we want to consider for analysis. In the world of *Gilmore Girls*, Emily is the matriarch of the family, and she does have the most *power* in her conversations with Lorelai (her daughter) and Rory (her granddaughter). Who tends to gain the floor? How often does a new thread start? How long is a thread? Those are interesting empirical questions we hope this work can eventually help us answer.

Let's try another one and revisit this example:

LORELAI (D1) Michel, come on, we've got to get into these budgets.

SOOKIE (D2) Now.

MICHEL (D3) Does the red light mean it's programmed?

SOOKIE (D4) [*to Lorelai*] I explained it a hundred times.

LORELAI (D5) Michel, you've been setting that machine for 20 minutes now.

(IV.xvi)

Here, we have two threads: thread one has D1, D2, D5 (topic: budget meeting); thread two has D3 and D4 (topic: Michel's recording device). D5 relates most strongly to Lorelai's desire to get Michel to join the meeting, which is expressed already in D1, so it takes precedence over the weak semantic relation of *it* (D3) and *that machine* (D5) and makes D5 part of the first thread.

Here's one last example for intuition. This is a family dinner scene. Rory, the granddaughter, brings her boyfriend, Dean, along, who meets Richard, the grandfather, for the first time. Interrogation ensues. Pay attention to the intention and desire of each character, and who has the control of where the conversation is going.

EMILY Antonia, please bring out the Twinkies.

LORELAI I can't believe I just heard you say those words.

EMILY Well, don't get used to it.

RICHARD So, Dean, where are you planning to go to college?

DEAN Oh, uh, well I. . .

LORELAI Geez Dad, start off with "what's your favorite baseball team" or something.

RICHARD I'm talking to Dean.

DEAN I don't know yet.

RICHARD You don't?

DEAN No, not yet.

RICHARD Well, what kind of grades do you get?

EMILY Richard please, don't grill the boy.

RICHARD I'm not grilling the boy Emily. It's an easy question. A's, B's, C's?

DEAN I get a mixture actually.

RICHARD Mixture? [*laughs*] What's the ratio?

EMILY Richard.

RICHARD I'm just trying to get to know the boy Emily. After all, Rory brings home a young man to dinner, the least we can do is learn something about him.

LORELAI He changes a mean water bottle.

DEAN I get a couple A's, couple B's, few C's.

RICHARD Really?

DEAN I'm not great in math.

LORELAI Yeah, except who is really? You know, except mathematicians or the blackjack dealers, or I guess Stephen Hawking doesn't suck, but you know. You know what else is good though Mom, is a Ho-Ho. Because if you can't find a Twinkie, you know, treat yourself to a nice Ho-Ho. How long does it take to open a box?

EMILY She's making them.

LORELAI She's making the Twinkies? You're kidding.

EMILY Oh Richard, wasn't there a book you wanted to give Rory?

RICHARD In a minute. So Dean . . .

RORY Uh, Grandpa?

RICHARD You do know that Rory is going to an Ivy League school?

DEAN I know.

RICHARD Harvard, Princeton, Yale.

LORELAI He said he knew, Dad.

RICHARD You need top grades to get into a top school.

DEAN Yeah, well, Rory's really smart.

RICHARD Yeah, she is really smart.

RORY Mom?

LORELAI Yeah, why don't we all go sit in the uh . . .

RICHARD So, how are you planning to make a living once you graduate from this college you haven't thought anything about yet?

RORY Grandpa, can we talk about something else?

EMILY I'm going to get that book.

(II.i)

There are two threads: one about Twinkie, the other about Richard's approval (or the lack thereof) of Dean being his granddaughter's boyfriend. For most of this excerpt, Richard has the floor. Lorelai, Emily, and Rory all try to take over (or, gain the floor) and stop Richard, and all fail. Semantically, you can say that Richard asks questions that span through a few topics—his college plans, his grades at school, Rory's college plans, his career plans (or whether he could provide for Rory). But pragmatically, all those utterances are for Richard to get to know Dean. Each question, then, will not start a new thread.

Threads can be short. While they are mostly organized by a sustainable distribution of attention, threads don't have to be long. There can be *hanging threads*, where one character tries to switch subject or gain floor but failed. If an utterance is a reply to an action, the thread might be short as well. Here are two examples that illustrate that:

RORY Was anything resolved? Are she and grandpa gonna be all right?

LORELAI Don't worry about it. They're a team. They'll be okay.

RORY Good. I like them. [*begins eating with fingers*]

LORELAI I know. [*takes a bite*]

[*Luke brings plates and forks and transfers napkins to plates*]

RORY Thanks.

[*Lorelai pulls out her rosebud and hands it to Luke. While they resume eating, Luke walks away sniffing the rose.*]

(III.xiii)

Rory and Lorelai were talking about Rory's grandparents at first. From the action statements, we see Luke brings plates and forks to the table, and Rory's "Thanks" at the end replies to that. It does not belong in the previous thread about her grandparents, and here the only logical annotation would indicate that Rory starts a new thread, which has only one sentence, before the scene ends.

EMILY We were buying the two of you a house. Doesn't the fact that we were willing to spend an enormous amount of money on a wedding present count for anything?

LORELAI So, that's what you're mad about? Your mad about the enormous amount of money you might have wasted?

RORY Mom.

EMILY That's not what I was saying.

LORELAI Well, you implied it.

EMILY Lorelai, that's—

RORY Bangalore! Bangalore! Bangalore!

(VII.iii)

In this scene, Emily and Lorelai are having a quarrel. Rory tries to distract everyone and change topic, yelling "Bangalore" entirely out of the blue. Rory's "Bangalore" also starts a new thread.

In both examples, both threads have only one utterance, which might be shorter than usual, but contextually, it makes sense.

In the next section, we will talk in detail about how to determine topic/floor change or thread start in a more principled manner. But ultimately you need to use your judgment, and hopefully you have a rough sense of what a topic or a floor is.

B.1.2 Threading drama

This section offers more instructions on how to decide when a conversational thread starts in drama. We already noted the two layers at work in the previous section:

1. Conceptually, the start of a thread introduces a new floor ("focus of attention", Goffman, 1963) or an observable change in topic (contextual relations, O'Neill and Martin, 2003).

2. Practically, the start of a thread has no parent utterance.

The rest of this section further elaborates our (perhaps idiosyncratic) definitions of floor and topic as they apply to our conversational analysis of drama.

Floor: Who are we paying attention to? Elsner and Charniak (2010, p. 392) describe the start of a new conversational thread as the process of participants (or in our case, characters) "hav[ing] refocused their attention . . . away from whoever held the floor in the parent conversation." Like in Goffman (1963), *attention* is the operative word: the character holding the floor can safely assume that they have attention from others. Such attention is singular and must be sustained throughout the thread; or someone else has the floor. A test we could use is whether we could logically insert "Now, everyone listen to me!" or "Now, everyone look at me!" before the UOI that we see as a potential start of a new thread.

Consider the following example:

TAYLOR This goes well beyond a head of lettuce, young man. The charges against your nephew are numerous. He stole the "save the bridge" money!

LUKE He gave that back.

TAYLOR He stole a gnome from Babette's garden.

LUKE Pierpont was also returned.

MISS PATTY He hooted at one of my dance classes.

FRAN He took a garden hose from my yard.

ANDREW My son said he set off the fire alarms at school last week.

LORELAI I heard he controls the weather and wrote the screenplay to *Glitter*.

BOOTSY [Now, everyone listen to me:] I think it's time for me to pipe up here.

LUKE Oh yeah, that'll be good.

BOOTSY I have every right to pipe in here, Luke. I'm a local entrepreneur.

LUKE You took over your father's newsstand, Bootsy. It doesn't make you an entrepreneur.

BOOTSY And you took over your old man's hardware store.

LUKE And turned it into a diner!

BOOTSY Big whoop. Who can't fry an egg?

TAYLOR Let's keep it moving here boys, huh?

BOOTSY I never liked the look of that kid from the second I saw him.

LUKE Unbelievable.

BOOTSY Excuse me, but I've got the floor.

LUKE You don't have the floor.

BOOTSY I'm standing, aren't I?

LUKE Well, I was standing first, which means I have the floor and I'm not giving it to you.

TAYLOR What is with you two?

(II.viii)

It should be clear that by saying "I think it's time for me to pipe up here," Bootsy is trying to get everyone else's attention. And since he is able to keep talking, we see he manages to sustain the attention and (unfortunately, *contra* Luke) hold the floor, which is very different from the utterances of Miss Patty, Fran, and Andrew: they *interject*, but they do not gain the floor. In this light, threads are usually composed of several dialogue turns because that is the only way for us to tell whether such attention lasts (unless, of course, the scene ends abruptly).

Topic: What does the character want? As we've seen in McKee, 2016: "All talk responds to a need, engages a purpose, and performs an action." With that in mind, when you see the first utterance in the scene, try to identify that need, purpose, or action it speaks to, which you can describe in a *verb* phrase. From there, extrapolate a broader, general topic, which you can describe in a *noun* phrase. Remember, you might not be able to do such extrapolation without knowing what's happening in the scene. So read ahead and skim a little bit. A new conversational thread begins when the UOI has nothing to do with that previous topic. Otherwise, it continues the current thread.

A simple test to see if this UOI starts a new topic is to insert such parenthetical statements as "Changing subject" at the beginning, and see if the conversation still makes sense. On the flip side, we, adapting McDaniel et al. (1996), introduce some basic semantic and pragmatic mechanisms that signal the *continuation* of a thread below:

- *semantically and pragmatically coherent response*: By definition, if the UOI makes a coherent response to any previous lines, the closest candidate to it is its parent utterance.
- *semantically non-topical speech*: Expressive speeches ("Ouch") or greetings ("Hi") do not have a "manifest topic" (Ervin-Tripp, 1964), so we consider them as non-topical. Unless they are used to gain floor, they will always be regarded as a continuation of the same thread.
- *successive greetings*: If there are other things going on in the scene, "Hi" and "Goodbye" *alone* do not form a thread.
- *co-reference*: Pronouns whose referent would be less ambiguous if we determine that the UOI continues the current thread.
- *term of address*: If one character addresses another directly, there's a higher chance that they are in the same thread.
- *acknowledgement*: If one character acknowledges, in their utterance, the presence of another character in the same scene, they are more likely in the same thread.
- *same physical location*: If all characters stay where they are at the start of the thread, they might be in the same thread. **In contrast**: of course, hanging up the phone, someone leaving or entering the room/scene, are useful signals that indicate a new thread might be about to start.

B.1.3 Examples

We offer three extended examples to motivate threading in drama. Parenthetical statements are in brackets; additional comments are prefixed by the pound (#) sign:

Example 1. Focus on how Emily keeps changing subject.

LORELAI [Now, tell me:] Why are you smiling like that?

Lorelai's purpose: find out why Emily "smiles like that" → topic: Emily's smile

EMILY What are you talking about?

LORELAI You're smiling.

EMILY I'm happy.

LORELAI That's not your "I'm happy" smile.

EMILY Well, what smile is it, Loerlai?

LORELAI That's your "I've got something on Lorelai" smile.

EMILY [Changing subject:] Rory, your mother must be very tired.

RORY She works a lot.

LORELAI I grew up with that smile—I know that smile.

EMILY [Changing subject:] Tell me about school.

RORY Well, my French final went pretty well.

LORELAI You can change the subject. I know the smile.

EMILY Whatever you say, dear.

cotinuation: acknowledgement

LORELAI I've used it a few times myself.

RORY Mom.

continuation: direct address

EMILY [Changing subject:] So, tell me about parents' day.

LORELAI What?

EMILY Parent's day? Next Wednesday? When all the parents are supposed to go to the classes with their children all day long?

LORELAI The Chilton newsletter came out today!

Chilton is the name of Rory's school, which might not be immediately obvious without prior knowledge of the show. But if you read the rest of the scene, or the excerpt here, you should be able to see Chilton is related to the "smile" and "school" threads and Lorelai is not bringing up something entirely random.

RORY Yup.

LORELAI Right.

EMILY You didn't read yours?

LORELAI Not yet.

EMILY Ah.

LORELAI But you knew that.

EMILY Well.

LORELAI Hence the smile.

EMILY Lorelai, you're really being silly. There's no evil plan afoot here. I simply brought up a subject I thought we could all talk about.

LORELAI Oh right.

EMILY [Changing subject:] I'll try another subject—the colour blue is very pleasant, isn't it?

LORELAI Mom! Not everybody can wait outside the mailbox for the Chilton newsletter to arrive and then instantly memorize the contents in three seconds.

Lorelai's goal: explain why she hasn't read the Chilton newsletter yet → topic: Lorelai's self-defense

RORY I'd like to weigh in on the blue colour subject, please.

EMILY You have your priorities. Far be it from me to question them.

LORELAI Just because I don't read the newsletter doesn't mean I don't care about my daughter!

(I.xi)

Example 2. Now, we revisit our first real example, this time at sentence level (with line break added at the end of each sentence) with attention paid to floor and topic change.

EMILY You're being stubborn, as usual.

Emily's goal: convince Lorelai to do something → topic: Lorelai's stubbornness.

LORELAI No, Mom, I'm not being stubborn.
I'm being me!
The same person who always needed to work out her own problems and take care of herself.
Because that's the way I was born.
That's how I am!

EMILY Florence, I'm dripping.

Emily's goal: have Florence deal with the dripping → topic: perm

Emily engages a new character but she does not grab any attention from Lorelai, nor does Florence join the conversation, so no floor change.

LORELAI I appreciate what you have done for Rory in paying for this school.

continuing the previous topic

That will not be forgotten.
You won't let it.
But she is my daughter.
And I decide how we live, not you.
[Changing subject:] Now then, do they validate parking here?

Lorelai's goal: find out about parking validation on the premise → parking

(I.ii)

Example 3. Pay attention to how threads can intervene, and topic is a pragmatic phenomenon.

KIRK This doesn't smell right.

action: fix the egg situation → theme: his dining experience

LANE Smells fine, Kirk.

KIRK I think the eggs were bad.

LANE The eggs are fine, Kirk.

KIRK Were they cooked in the fish pan? They smell like they were cooked in the fish pan.

LANE No, the eggs were not cooked in the fish pan. They were cooked in the egg pan.

KIRK Was the fish pan sitting next to the egg pan? Because perhaps—

[*Lorelai walks in the door*.]

LORELAI [Now, everyone look at me:] I need something with cheese!

reading ahead will suggest Lorelai's goal is to find Luke. "I need something with cheese" is just like "Hey guys", an expression to grab everyone's attention.

KIRK Lorelai, smell my eggs.

continuing his egg topic

LORELAI Not today, Kirk.

replies to Kirk after she started her own thread

Hey, where's Luke? I want him to make that breakfast quesadilla thing he made yesterday.

LANE Luke's not here.

LORELAI Where is he? He knows the exact right jack-to-cheddar ratio.

KIRK He's out there. [*Kirk points out the window.*]

LORELAI Where?

KIRK Over there with Nicole.

(IV.xvi)

Example 4. This one shows how thread membership can be hard to determine at times.

[*Jess enters.*]

JESS Morning.

LUKE You're up early.

JESS Gotta catch me that worm.
See ya.

LUKE Where you off to?

JESS School.

LUKE This early?

JESS I got a lab project going on.
Me and my team are meeting early.

LUKE Well, have a good day.

JESS If I have a choice.
[*to Rory and Lorelai*] Hey.

Floor change: Jess joins the conversation and attracts the attention from Lorelai and Rory.

LORELAI Good morning.

Lorelai replies to Jess's "Hey."

JESS Talk to you later. [*He and Rory kiss*]

Jess still has the floor, and there's no obvious change in topic. We have a series of non-topical statements here, and this does not reasonably start a new thread. He, however, does not reply to Lorelai's "Good morning" either. So, the most logical reply-to here is his own "Hey."

RORY Later.

[*Jess exits. Lorelai's cell phone rings.*]

LORELAI [*to Rory*] By the way, your boyfriend snores.

Tricky! See below.

RORY Didn't need to know that.

(III.xvii)

It's tricky to determine whether "By the way, your boyfriend snores." is the start of a new thread. Jess's snoring is a new subject. Jess left the room, and no one can pay attention to him anymore, so there's a redistribution of attention. It is also true, however, that most of the thread is composed of semantically non-topical utterances. Since we're reaching the end of scene, this thread would have only two utterances, making it very short. In general, **we want floor/topic change (the way we define it) to take precedence**, despite some caveats and potential disagreements. Given our definition, we have both topic and floor changes here, so we would like to make "By the way, your boyfriend snores." be the start of a new thread.

This last example is also an important reminder that ultimately this project is about finding the reply-to relationship (and from there, threads of conversations). It's not about who replies to whom or who is listening. Addressees or participant roles should not be your primary judgement in deciding the reply-to relationship.

Lastly, many social media and instant messaging apps have this notion of "thread" built in: a Twitter or Slack thread usually explores the same topic; if one person explicitly indicates to which previous message they are replying, and then another one replies to that message, those messages naturally form a thread. Threads work in a similar way here.

B.2 Annotation in action

Data disclaimer. Please be aware that our sampled scenes may contain potentially problematic content, such as vulgar, violent, or offensive language in screenplays, or other biases held by individual screenwriters.

Annotation principles. Here are the general rules for annotation:

1. Intuitively, dialogues follow the *basic economic principle*, where D_n replies to D_{n-1}.
2. A new thread starts when a speaker refocuses other characters' attention or starts a new topic.
3. Use speaker labels, action lines, dialogue turn information to enhance your understanding of the scene.
4. Always quickly skim through a couple of lines and get a sense of what's happening in the scene before starting to annotate.

Summary of symbols. For each sentence in a dialogue line, annotate with the following symbols. Next section is on how to use those symbols.

- this line is the start of a new thread:
  - `T`: *both* floor and topic changes occur, or any signals that indicate the previous conversation was over. Also use this symbol at the start of any scene.
  - `F`: you're certain only Floor change occurs/can add "now, everyone listen to me" at the beginning (or a phrase that serves the same function is actually part of the line)
  - `P`: you're certain only toPic change occurs/can add "switching subject" at the beginnin (or a phrase that serves the same function is actually part of the line)
- `-`: this line replies to the preceding line
- `D`x: this line replies to sentence D_x
- symbols for editorial convenience (should be used very sparingly):
  - `S`: skip the current sentence, due to significant OCR/parsing errors
  - `X`: this line requires further discussion for adjudication

Handling parsing/OCR errors Fig. 5 shows you what kind of parsing/OCR related errors you might encounter and why they are there. If you spot an obvious/easily fixable OCR or parsing error, please correct it. If you suspect you've spotted an error of any kind, you could take a look at the original `txt` or `pdf` file. It also comes with experience. After you're sure there are errors, here's how you fix them:

First, if you are turning an action line into a dialogue line, supply dummy dialogue turn ID and dialogue ID. Our suggestion is something like `La` and `Da`. We do so because those lines you are rescuing might become parent utterances of UOIs to come, in which case you can annotate with `Da`.

Second, you might need to turn a dialouge line into an action line:

turn id	line id	speaker	line
L371	D470	THE BATHROOM	Where ...

Many entities are singled out printed in uppercase, such as THE BATHROOM here. When they appear alone in the line, there's no way to distinguish them from a regular speaker label (we also don't want to simply exclude locations, because e.g., MAN IN THE STREET is a well-formed speaker label). To correct this, simply change the dialogue ID to **A** Changing line type is NOT necessary (saves you two seconds, which add up). Since we don't really use action IDs for annotation, it's not necessary to add them. Removing dialogue turn ID or speaker label is optional. This is what that row should be:

turn id	line id	speaker	line
L371	**A**	THE BATHROOM	Where ...

Third, if a line is now empty after your correction, or if you spot a line that does not contain

```
EXT. WEDDING LAWN - BAND STAGE - DAY

The band sets up.  Eddie hands Zoe a YO GABBA CONCERT DVD.

                    EDDIE
          I paid top dollar for this bootleg.
          So make sure the kids study it.
          Ingrid asks very detailed questions.

As Stevie pours water in his FOG MACHINE he sees Barry's FISH.

                    STEVIE
          What's with the goldfish?

                    BARRY
               (laughs; to Gardener)
          Hear that?  He thinks these are goldfish!

The GARDENER just keeps on blowing with his blower.
```

(a) the original pilot script for *The Wedding Band*

```
EXT. WEDDING LAWN - BAND STAGE - DAY

The band sets up. Eddie hands Zoe a YO GABBA CONCERT DVD.

EDDIE

I paid top dollar for this bootleg.

So make sure the kids study it.

Ingrid asks very detailed questions.

As Stevie pours water in his FOG MACHINE he sees Barry's FISH.

STEVIE

What's with the goldfish?

BARRY

(laughs; to Gardener)

Hear that? He thinks these are goldfish!

The GARDENER just keeps on blowing with his blower.
```

(b) script after OCR

SCENE	S37				EXT. WEDDING LAWN - BAND STAGE - DAY
ACTION	S37		A462		The band sets up.
ACTION	S37		A463		Eddie hands Zoe a YO GABBA CONCERT DVD.
DIALOGUE	S37	L334	D584	EDDIE	I paid top dollar for this bootleg.
ACTION	S37		A464		So make sure the kids study it.
ACTION	S37		A465		Ingrid asks very detailed questions.
ACTION	S37		A466		As Stevie pours water in his FOG MACHINE he sees Barry's FISH.
DIALOGUE	S37	L335	D585	STEVIE	What's with the goldfish?
ACTION	S37		A467		(laughs; to Gardener)
DIALOGUE	S37	L336	D586	BARRY	Hear that?
DIALOGUE	S37	L336	D587	BARRY	He thinks these are goldfish!
ACTION	S37		A468		The GARDENER just keeps on blowing with his blower.

(c) Script after parsing/pre-processing

Figure 5: This is an example of typical OCR/parsing errors. It's unclear why Eddie's line is broken into three paragraphs (which is not normal), but since our pipeline relied heavily on line breaks, our parser wrongly recognized Eddie's lines as action statements because you can't really distinguish between a true action statement ("As Steve pours water …") and a dialogue line just by using line breaks. Errors like these are easily fixed, however: supply dummy line turn ID and dialogue IDs, and annotate accordingly.

an action statement or dialogue line but has information about the script or purely logistical information (like `Untitled Project (04/12/22)`, `CBS Studio Production`), put `S` (skip). Or long-tail errors: screenplays to *Star Trek* movies put all dialogues in Klingon in parentheses, and they will be parsed as action lines. It's impossible to find all of them through regular expressions, and we don't need our model to see them.

Fourth, pay attention to ellipses and em dashes: IMSDB/Scriptbase-J can use '[sentence] . . .' (with leading space), '[sentence]. . .' (without), and there can be space between each dot ('. . .' vs '...'). Em-dashes can be ' - ' (dash separated by space), '–' (two dashes, no space surrounding), or '-' (that looks just like two words being linked together. There's no easy way for us can clean and normalize that in our pipeline, and in some cases they interfere with the semantics. So correct those too.

B.2.1 Questions to ask while annotating

1. This is the beginning of a scene.
    - Put `T`.
    - Read a couple of lines ahead and gain a sense of:
        - Why does the character speak at all?
        - What do they want?
        - Who has the floor?

2. This is a new sentence:
    (a) Can D_{n-1} be the sensible reply to D_n?
        - If so, put `-`.

    (b) If not, what previous line leads to this line? Is there any previous line that triggers (or, gives the necessary context for us to understand) this UOI?
        - If there's one, put the utterance ID.

    (c) If not, is there a topic/floor change?
        i. Is there any new intent or desire being expressed? Can I insert "switching subject" at the beginning of UOI?
            - If so, put `P` (toPic).

        ii. Is there a character replacing another one as the center of attention? Can I insert "Now, look at me/listen to me" at the beginning of UOI?
            - If so, put `F` (Floor).

        iii. Do floor and topic changes happen at the same time? Or are there other

signals that make you think the previous conversation is over (someone has left the room, or other observable contextual changes)?
  - If so, put `T` (Thread).

(d) If none applies, this seems to be an edge case. Follow the following steps:
  i. Is this a dialogue line wrongly parsed as action line?
    A. If so, change the line ID from Ax to Dx.
    B. Correct it. Supply dialogue turn ID and speaker label to match the format of a regular dialogue line. Use the original `txt` or `pdf` file if necessary.
    C. If there are any empty lines (for example: you moved an obvious speaker label from one line to the "speaker" column of the current line) or lines that contain redundant information (for example: the content of this line and is copied to the previous line, which made the previous line a complete dialogue line) or information irrelevant to the task at hand (for example: "PILOT DRAFT #4"), put `S` (Skip)
  ii. Is this an action line wrongly parsed as a dialogue line?
    A. If so, change the line ID from Dx to `A`.
    B. Correct it.
    C. Put `S` (Skip) when you see fit.
  iii. Are there multiple wrongly parsed lines, or any OCR/parsing errors that takes more than **TWO MINUTES** to fix?
    - If so, don't spend time fixing all that.
    - Put `S`.
    - Take a note of the corpus name and title slug (e.g., `[tvpilot, the-wedding-band]`). We'll see if we want to remove that title from the corpus altogether later.
  iv. Is there no logical parent utterance, but you can't say there's any change in distribution of attention or switch in subject?
    - If so, put `X` and move on. We can discuss later.

B.2.2 Post-processing

We will be post-processing all annotations to:

- convert `-`'s into the correct ID's
- change `{T, P, F}` into `Tn`: We will NOT distinguish between `T`, `P`, or `F` for modeling, and they will be converted to `T1`, `T2`, etc.
- delete any lines tagged `S` (Skip)
- correct any non-conventional dialogue/action line to account for individual fixes: We will create a mapping between old and new line ID's: e.g., `D45, Dg, D46`, where `Dg` was added during annotation to fix a parse/OCR error, will become `D45, D46, D47`.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Discussed in the Limitation section.

☑ A2. Did you discuss any potential risks of your work?
Discussed in the Ethics Statement section.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Main claims are summarized in the Abstract and Introduction sections.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
1. TV Writing does not impose terms or licensing restrictions on the use of its screenplays or teleplays; we annotate and re-publish short snippets of those materials with a Creative Commons license under an argument of fair use.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We mentioned the occurrence of potentially offensive language (such as swear words used for dramatic effects) in screenplays in the Ethics Statement section. Names of individual actors are public data and not distributed along with this work or referenced anywhere.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
1 and Ethics Statement

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.3 and 5.3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

4.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.