Identifying Collaborative Conversations using Latent Discourse Behaviors

Ayush Jain, Maria Leonor Pacheco, Steven Lancette,
Mahak Goindani and **Dan Goldwasser**
Department of Computer Science, Purdue University
{jain207, pachecog, slancett, mgoindan, dgoldwas}@purdue.edu

Abstract

In this work, we study collaborative online conversations. Such conversations are rich in content, constructive and motivated by a shared goal. Automatically identifying such conversations requires modeling complex discourse behaviors, which characterize the flow of information, sentiment and community structure within discussions. To help capture these behaviors, we define a hybrid relational model in which relevant discourse behaviors are formulated as discrete latent variables and scored using neural networks. These variables provide the information needed for predicting the overall collaborative characterization of the entire conversational thread. We show that adding inductive bias in the form of latent variables results in performance improvement, while providing a natural way to explain the decision.

1 Introduction

Online conversations are rampant on social media channels, news forums, course websites and various other discussion websites consisting of diverse groups of participants. While most efforts have been directed towards identifying and filtering negative and abusive content (Wang and Cardie, 2014; Wulczyn et al., 2017; Zhang et al., 2018), in this paper we focus on characterizing and automatically identifying the positive aspects of online conversations (Jurafsky et al., 2009; Niculae and Danescu-Niculescu-Mizil, 2016; Napoles et al., 2017a). We specifically focus on *collaborative conversations*, which help achieve a shared goal such as gaining new insights about the discussion topic like response informativeness, engagement etc.

Rather than looking at the outcomes of such conversations (e.g., task completion (Niculae and Danescu-Niculescu-Mizil, 2016)), we analyze conversational behaviors, specifically looking at indications of *collaborative* behavior that is conducive to group learning and problem-solving. These include purposeful interactions centered around a specific topic, as well as open and respectful exchanges that encourage participants to elaborate on previous ideas. To help clarify these concepts, consider the following conversation snippet.

> **User A** : We should invest in more resources to encourage young people to be responsible citizens.
>
> **Response Option 1** : I wonder if more initiatives at grassroots level can help them to identify and understand issues of their local community more deeply.
>
> **Response Option 2** : Good point, I agree.

We compare the two possible responses to User A's post. Option 1 offers a balanced contribution, developing the idea presented in the original post and allowing the conversation to proceed. Option 2, while polite and positive, is *not* collaborative as the initial idea is not expanded on. In fact, agreement is often used as a polite way to end conversations without contributing additional content. Despite the positive sentiment, capturing the absence of balanced content contribution and the absence of idea development as different discourse behaviors, one can infer that it is not a collaborative conversation.

While humans could tell the two apart by detecting constructive discourse behaviors, automatically capturing these behaviors is highly challenging. Anecdotal evidence, collected by extracting features from conversation transcripts, can lead to conflicting information, as identifying collaborative behavior relies on complex interactions between posts. Our main intuition in this paper is that reasoning and enforcing consistency over these behaviors can help capture the conversational dynamics and lead to more accurate predictions.

Our technical approach follows this intuition. We design a hybrid relational model that combines neural networks and declarative inference

, pages 59 - 63
July 1-3, 2020 ©2020 Association for Computational Linguistics

to capture high-level discourse behaviors. Since we only have access to the raw conversational text, we model these behaviors as discrete latent variables, used to support and justify the final decision – whether the conversation is collaborative or not.

Explicitly modeling discourse behaviors as latent variables allows us to add inductive bias, constraining the representation learned by the neural model. It also provides a natural way to "debug" the learning process, by evaluating the latent variables activation. Our experiments show that the joint model involving global learning of different latent discourse behaviors improves performance. We use the Yahoo News Annotated Comments Corpus (Napoles et al., 2017b), and expanded the annotation for the collaborative task.[1]

2 Task Definition

Collaborative conversations are purposeful interactions, often revolving around a desired outcome, in which interlocutors build on each others' ideas to help move the discussion forward. Collaborative conversations are an important tool in collaborative problem solving (Greiff, 2012) and require collaboration skills (Flor et al., 2016; Hao et al., 2016). We focus on identifying indicators of successful collaboration. We build on the work of Napoles et al. 2017a, who released a dataset annotated for engaging, respectful and informative conversations, and annotate it for collaborative conversations, in which participants build on each other's words, provide constructive critique, elaborate on suggested ideas, generalizing them and synthesizing new ideas and knowledge in the process.

During the annotation process, we identified several repeating behaviors (detailed below) that helped characterize and separate between collaborative and non-collaborative conversations.

2.1 Non-Collaborative Discourse Behaviors

(A) Low Idea Development users who: (1) deviate from the thread topic and change the topic, (2) ignore previously raised ideas and give preference to their own, (3) repeat or reinforce previous viewpoints. **(B) Low User Engagement** users who: (1) show little interest, (2) add shallow contributions, such as jokes or links. **(C) Negative Sentiment** relevant when disagreements are not resolved politely and respectfully. **(D) Rudeness** use of abusive, rude or impolite words.

2.2 Collaborative Discourse Behaviors

(A) High Idea Development when users stay on topic (with respect to the original post) and new ideas are formed and developed based on preceding turns. **(B) Reference to Previous Posts** users refer to the previous post to advance the conversation. **(C) Back and Forth** users support and appreciate the ideas shared by others, and are polite when expressing disagreements. **(D) Positive Sentiment** resulting in positive interactions among users, expressed through polite conversation or informal emoticons. **(E) High User Engagement** leading to insightful discussions, meaningful to its participants. **(F) Balanced Content Distribution** between all members in the group. **(G) Questions** raised by participants to advance the conversation.

Annotation Process Two annotators labeled the conversations based on these guidelines, with an accuracy in inter-annotator agreement of 81%.

3 Modeling Collaborative Behaviors

Identifying collaborative conversations requires characterizing nuanced behaviors. In previous work, this analysis was defined by extracting social and discourse features directly from the raw data. In contrast, we view this decision as a probabilistic reasoning process over the relevant conversational behaviors that were identified during the annotation process (Sec. 2.1 and 2.2). Since these behaviors are not directly observed, and have to be inferred from the raw conversational features, we treat them as discrete latent variables which are assigned together-with, and consistent-with, the final classification task.

Each behavior is captured by a binary latent variable, denoted as $\mathbf{h} = \langle h_1, ..., h_k \rangle$, indicating if it's active or not in the given thread. These decisions are then connected with the final prediction, denoted y, a binary output value. This results in a factor graph (Figure 1). Each individual decision is scored by a neural net, and uses a set of features capturing relevant properties in the input conversation. To learn this model, we extend DRaiL (Zhang et al., 2016), a recently introduced framework for combining declarative inference with neural networks, described briefly in the following section. Our extension allows for the introduction of discrete latent predicates into the model.

[1] Annotated dataset available at https://gitlab.com/ayush-jain/collaborative-yahoo-discourse

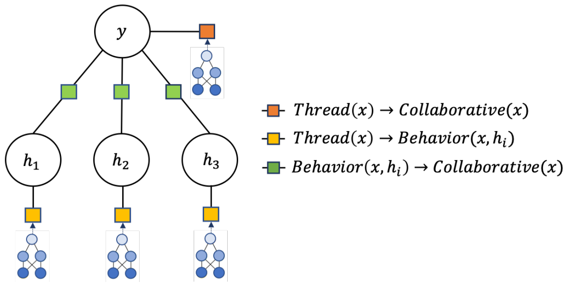
y
h₁
h₂
h₃
Thread(x) → Collaborative(x)
Thread(x) → Behavior(x, hᵢ)
Behavior(x, hᵢ) → Collaborative(x)

Figure 1: Factor Graph for Collaborative Conversations

3.1 Learning and Inference with DRaiL

DRaiL uses a first-order logic template language to define structured prediction problems. A task is defined by specifying a finite set of *predicates*, corresponding to observed or output information. Decisions are modeled using rule templates, formatted as horn clauses: A $\Rightarrow$ B, where A (*body*) is a conjunction of observations and predicted values, and B (*head*) is the output variable to be predicted. The collection of rules represents the global decision, taking into account the dependencies between the rules using a set of constraints C. Rule instances are represented by variables r_i, and they are scored using neural nets, defined over a parameter set $\mathbf{w}$.

$$\mathbf{y} = \arg\max_{r_i} \sum_i r_i \cdot score(x, \mathbf{w}, r_i)$$
$$\text{subject to C}, \quad \forall i; r_i \in \{0,1\} \qquad (1)$$

We define two models using this representation. The first, **DRaiL Local**, trains a single neural net, represented by the rule: THREAD(T)$\Rightarrow$ ISCOLLABORATIVE(T), mapping the thread to the predicted value directly. The input layer to the neural net is the union of word indicators and all the features used to capture conversational behavior (Table 1). This approach is similar in spirit to previous works, classifying conversational threads using aggregated features.

The second, **DRaiL Global**, builds on the previous model, augmenting it with rules capturing individual discourse behaviors, and then associating the predictions of these rules with the final prediction task. We define the set of latent conversational behaviors B $\in$ {*Idea Development, Reference to Previous Post, Sentiment, Balanced Content, Back and Forth, Questioning Activity, User Engagement, Rudeness and Controversial*}.

We define two rules for each behavior in B, as follows: THREAD(T)$\Rightarrow$ LATENTBEHAVIOR(T,B), corresponding to a neural net predicting the occurrence of the specific behavior B in conversational thread T. We also add the rule: LATENTBEHAVIOR(T,B)$\Rightarrow$ ISCOLLABORATIVE(T), capturing the relationship between the latent behavior and the collaborative prediction.

Behavior	Features
S	Degree of sentiment and intensity
B.C	Sentences per post, words per posts, post depth
C	Upvote/downvote ratio, $u-d$, $u+d$, $u/(u+d)$
R.P.P	2 per. pronouns, quotes of prev. posts, @username tags
B.F	(Dis)agreement markers, content indicators, post references
I.F	Lexical chains (Barzilay and Elhadad, 1997)
R	Profanity, bad words, short posts indicators
U.A	Number of posts, number of threads
Q.A	Question marks, question forms, question types

Table 1: Features per Behavior. Sentiment (S), Balanced Content (B.C), Controversial (C), Reference to Previous Posts (R.P.P) Back and Forth (B.F), Idea Flow (I.F), Rudeness (R), User Activity (U.A), Questioning activity (Q.A)

Each rule template is associated with an initial feature representation and a neural architecture to learn its scoring function. After scoring factors, values are assigned to the output variables by running an inference procedure. DRaiL uses Integer Linear Programming (ILP) to solve the inference problem. In our setup, we compare two models, with and without inference, corresponding to the global and local models.

Global Learning When multiple rules are defined in DRaiL, each has its own neural architecture and parameters. Since these rules are interconnected, DRaiL learns a globally normalized model which uses inference to ensure that the scoring functions for all rules result in a globally consistent decision. We adapted the structured hinge loss used in DRaiL to handle latent predicates. The loss function is defined over all neural parameters $\mathbf{w}$, and the error is back-propagated to update all networks.

$$L_D(\mathbf{w}) = \min_w \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i \qquad (2)$$

Where ξ_i is the slack variable, capturing the margin violation penalty for a given training example, and defined as follows:

$$\xi_i = \max_{y,\mathbf{h}}(f(\mathbf{x_i}, \mathbf{h}, y, \mathbf{w}) + \Delta(y, y_i)) - \max_{\mathbf{h}} f(\mathbf{x_i}, \mathbf{h}, y_i, \mathbf{w})$$

Here, $\mathbf{x_i}$ and y_i are the inputs and gold labels for the i-th instance and $\mathbf{h}$ denotes the active DRaiL rules corresponding to latent discourse behaviors.

4 Empirical Evaluation

4.1 Dataset and Experimental Settings

We annotate conversations on the Yahoo News Annotated Comments Corpus (Napoles et al., 2017b) following the guidelines specified in section 2, with 81% inter-annotator accuracy. The dataset consists of 2130 conversations for training, 97 for validation and 100 for testing. The data is imbalanced, with more conversations being non-collaborative (64%, 69% and 67% for training, validation and testing, respectively). Additionally, we annotated the fine-grained discourse behaviors for a sample set of 103 conversations.

We used feedforward networks for all rules, with one hidden layer and a softmax on top. All hidden layers use sigmoid activations. The number of hidden units are: 400 for the local rule, 50 for idea flow and 100 for all remaining behaviors. Rules that map a latent behavior to a final decision did not have a hidden layer. We used a learning rate of 0.01. All of these parameters, as well as the weights for the different rules, were tuned using the validation set.

4.2 Experiments

We compare the model that explicitly reasons about conversational behaviors and their relationships (*DRaiL Global*), with a local model that predicts whether a conversation is collaborative or not by using all discourse features as inputs to a single rule (*DRaiL Local*). To motivate the use of neural networks, we include two *Linear SVM* baselines, using bag-of-words and the set of all discourse features (Table 1). These results (Table 2) demonstrate the advantage of modeling competing discourse behaviors as latent variables and making a joint decision using inference, as opposed to just representing them using input features.

Model	Prec.	Rec.	F1
Linear SVM(BoW)	0.60	0.58	0.59
Linear SVM(BoW + disc.)	0.63	0.61	0.62
DRaiL Local(single NN)	0.65	0.64	0.64
DRaiL Global (latent vars.)	0.69	0.68	0.69

Table 2: Predicting Collaborative Conversations (Fixed splits)

We conduct an additional experiment to evaluate the quality of the predicted latent behaviors. To do this, we annotated the discourse behaviors based on the definitions provided in section 2, and evaluate the activations produced by our global model. We compare their correctness **before** learning (based on initialization parameters) and **after** global learning. Inference is used in both cases. Table 3 describes the results. We can see that performance consistently improved after global training compared to the initialization point, a clear indication of the connection between the latent information and the predicted conversational outcome. Identifying rude behaviors yields the highest F1 score (0.62), which can be expected as the decision relies on lexical information (negative and abusive words). Similarly, it is relatively easy to identify balanced content behavior, given that structural features (outlined in table 1) are very informative. Lexical chains, representing the repeated occurrence of a single word or of several closely related words over the course of a post (Barzilay and Elhadad, 1997), are also successful at capturing idea flow behaviors. However, controversial and back and forth behaviors are more challenging.

Individual Behavior	F1 (**before**)	F1 (**after**)
Idea Flow	0.371	0.574
Controversial	0.390	0.420
Balanced Content	0.541	0.610
Sentiment	0.462	0.548
User Activity	0.521	0.570
Reference to Previous Posts	0.299	0.427
Questioning Activity	0.427	0.511
Rudeness	0.514	0.620
Back and Forth	0.470	0.520

Table 3: Predicting Individual Latent Behaviors on Annotated Sample Set Before and After Global Learning

We performed an ablation study to see if the global model is driven by any particular discourse behavior (Table 4). We observe that performance drops significantly if the sentiment behavior is removed. Just using rules related to idea flow, sentiment and balanced content behaviors leads to an F1 score of 0.62.

Model	Precision	Recall	F1
All	0.690	0.680	0.687
All except S	0.483	0.495	0.489
All except I.F	0.635	0.554	0.591
All except B.C	0.581	0.593	0.586
All except QA	0.578	0.588	0.582
I.F + S + B.C	0.645	0.607	0.625
I.F + S + U.A	0.665	0.404	0.502
S + B.C + C + Q.A	0.693	0.546	0.610

Table 4: Ablation Study. Sentiment (S), Idea Flow (I.F), Balanced Content (B.C), Questioning Activity (Q.A), User Activity (U.A), Controversial (C)

5 Summary and Future Work

In this paper, we introduce the task of identifying collaborative conversations and provide annotations for a subset of the Yahoo News Annotated Comments Corpus. We suggest an approach that combines neural networks with constrained inference for identifying collaborative conversations, and showed how adding additional inductive bias in the form of discrete latent variables can improve learning. Moreover, we show that we are able to capture and explain individual discourse behaviors without additional supervision, which in turn allows us to gain insight into the final decision made by the model. Collaborative interactions help leverage the synergy between team members tackling complex problems, we hope to contribute in the development of automated systems supporting such processes.

References

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 31–41.

Samuel Greiff. 2012. From interactive to collaborative problem solving: Current issues in the programme for international student assessment. *Review of psychology*, 19(2):111–121.

Jiangang Hao, Lei Liu, Alina von Davier, Patrick Kyllonen, and Christopher Kitchen. 2016. Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In *Proceedings of the 9th International Conference on Educational Data Mining*.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646. Association for Computational Linguistics.

Courtney Napoles, Aasish Pappu, and Joel Tetreault. 2017a. Automatically identifying good conversations online (yes, they do exist!). In *Proceedings of the International AAAI Conference on Web and Social Media*.

Courtney Napoles, Joel Tetreault, Enrica Rosata, Brian Provenzale, and Aasish Pappu. 2017b. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of The 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain. Association for Computational Linguistics.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361. Association for Computational Linguistics.

Xiao Zhang, Maria Leonor Pacheco, Chang Li, and Dan Goldwasser. 2016. Introducing DRAIL – a step towards declarative deep relational learning. In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 54–62, Austin, TX. Association for Computational Linguistics.