USING LATENT DISCOURSE INDICATORS TO IDENTIFY GOODNESS IN ONLINE CONVERSATIONS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Ayush Jain

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

December 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF THESIS APPROVAL

Dr. Dan Goldwasser, Chair
    Department of Computer Science
Dr. Chris Clifton
    Department of Computer Science
Dr. Jennifer Neville
    Department of Computer Science
Dr. Elias Bareinboim
    Department of Computer Science

Approved by:

Dr. Voicu Popescu by Dr. William Gorman
    Head of the School Graduate Program

This is dedicated to the most kind-hearted, caring and crazy people in my life: my family.

ACKNOWLEDGMENTS

Firstly, I want to express my sincere gratitude to Professor Dan Goldwasser, my advisor, for introducing me to interesting and challenging problems within Natural Language Processing and giving me the opportunity to be a part of this project even though I had no prior exposure in the field. His genuine passion for the field and research in general ignited my interest and has been a great motivation throughout the course of this project. Research meetings with him involving interesting conversations about different topics were some of the best learning experiences at Purdue University and played a pivotal role in shaping the task of this project. His calmness and patience has been deeply inspiring and made it easier to embrace failures as part of this research work. I am highly grateful for his encouragement and support in the project even during academic vacations.

I also want to thank Steven Lancett from Purdue University for his contribution in annotation work, formulation of annotation guidelines and fruitful discussions about field of education in general. I would also like to thank Maria Leonor Pacheco from NLP lab at Purdue University for taking the time out to explain DRaiL platform and helping in paper writing as well. I also want to thank Mahak Goindani at Purdue University for her help in annotation work, Xiao Zhang from NLP lab at Purdue University for an interesting ride to my first NLP conference along with other members from the lab for insightful discussions during weekly reading group sessions. I also want to thank my committee members for their valuable feedback on this project.

Finally, I would like to thank Computer Science Department at Purdue University for offering me the position of Teaching Assistant in variety of courses and providing me financial support. It allowed me to experience teaching from the other side, study interesting courses and pursue my research work without any financial burden.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Ayush, Jain Master of Science, Purdue University, December 2018. Using Latent Discourse Indicators to Identify Goodness in Online Conversations. Major Professor: Dan Goldwasser.

In this work, we model latent discourse indicators to classify constructive and collaborative conversations online. Such conversations are considered *good* as they are rich in content and have a sense of direction to resolve an issue, solve a problem or gain new insights and knowledge. These unique discourse indicators are able to characterize flow of information, sentiment and community structure within discussions. We build a deep relational model that captures these complex discourse behaviours as latent variables and make a global prediction about overall conversation based on these higher level discourse behaviors.

DRaiL, a Declarative Deep Relational Learning platform built on PyTorch, is used for our task in which relevant discourse behaviors are formulated as discrete latent variables and scored using a deep model. These variables capture the nuances involved in online conversations and provide the information needed for predicting the presence or absence of collaborative and constructive characterization in the entire conversational thread. We show that the joint modeling of such competing latent behaviors results in a performance improvement over the traditional direct classification methods in which all the raw features are just combined together to predict the final decision. The Yahoo News Annotated Comments Corpus is used as a dataset containing discussions on Yahoo news forums and final labels are annotated based on our precise and restricted definitions of positively labeled conversations. We formulated our annotation guidelines based on a sample set of conversations and resolved any conflict in specific annotation by revisiting those examples again.

1 INTRODUCTION

In today's world, online conversations are prevalent on news forums, social media platforms and various other discussion websites involving a large number of people. Although many of these conversations contain sarcasm, abuse and personal insults, there are other conversations that are respectful and constructive in nature. Often, they are informative, smooth and effortless which lead to synthesis of new ideas. To date, most of the effort was directed towards identifying and filtering negative and abusive content [1–3]. Unlike these works, the focus is on characterizing and automatically identifying the positive aspects of online conversations [4–6].

Identifying the positive aspects of conversations can be helpful, not only in filtering bad content online, but also in initiating or moderating online conversations that promote group learning and forming stronger social bonds. Identifying these conversational behaviors can be used for estimating student engagement in classroom learning, making meetings more productive and many other cases involving purposeful conversations in a group setting.

We specifically focus on intentional conversations, which help in achieving a shared goal such as completing a task, group problem-solving or gaining new insights about the discussion topic. Rather than looking at the outcomes of such conversations (e.g., task completion [5]), we analyze conversational behaviors, specifically looking at indications of *collaborative* conversations. These types of conversations are conducive to group learning and problem-solving, they are characterized by purposeful interactions, centered around a specific topic. Collaborative conversations encourage an open and respectful exchange, allowing participants to present their ideas, respond to ideas by others and elaborate on them.

It is very easy for humans to identify elements of goodness in conversations, by detecting positive and constructive discourse behaviors as well as subtle negative and rude behavior by the participants. These can include an open flow of ideas, a high degree of engagement from all participants, politeness (or at least the lack of overtly rude or hostile behavior) especially when disagreements rise to the surface. These high level behaviors provide a strong indication for an overall collaborative conversation. However, while easy for humans to identify, capturing these behaviors in an automated way is highly challenging. Anecdotal evidence, collected by extracting features from conversation transcripts can lead to conflicting information, as identifying collaborative behavior relies on complex inter-post interactions. To further motivate this observation, an example of a subtle distinction between presence and absence of collaborative and constructive characteristics in conversations is shown below:

Non-collaborative and non-constructive conversation:

User A : I think that software piracy should not be tolerated in the developing nations as it leads to huge losses for the companies

User B : Good point, I agree.

User C : I am glad that you agree.

Collaborative and constructive conversation:

User A : I think that software piracy should not be tolerated in the developing nations as it leads to huge losses for the companies

User B : I wonder if we can ask only the professionals and upper classes in a nation to pay reasonable prices for software

User A : That actually makes sense. It would ensure that there are no losses and allow these countries to move into the information age resulting in widespread adoption of software in emerging markets.

In the first conversation, there is politeness and agreement but lacks enough content that could lead to development of new ideas. The conversation dies with a short reply from User B and C even if they agree with the User A's opinion and there is a overall positive sentiment. By capturing absence of balanced content contribution, absence of idea development and presence of positive sentiment as different discourse behaviors, one can easily infer that it is not a collaborative conversation. On the other hand, in the second conversation, User B makes a remark that not only leads to a new idea but allows User A to advance the idea further. The collaborative nature of this conversation is evident from the balanced engagement of both users and the development of relevant ideas in the conversation.

As seen in the examples above, polite disagreements can help promote collaborations, as opposed to rude disagreements which would prevent good insights from being developed further. However, politeness on its own is not enough, as it often can be used to avoid clarifying points of contention, which would hinder progress. Conversations can also drift from the main topic and lose their focus, or only serve a social objective and lack meaningful content. In order to capture such subtle differences, we define characteristics of collaborative and non-collaborative discussions (Section 3), and use these definitions as annotation guidelines to construct a dataset based on human judgments. Also, the focus is on conversation instead of a discussion. Since discussions usually have a predefined goal with participants trying to push their opinions and conversations are much more open-ended with participants listening to the perspective of others, conversations tend to be more collaborative with people respecting the viewpoints of others resulting in more chances for new idea development.

Identifying such discourse patterns and their inter-dependency among them to make a comprehensive judgment about the whole conversation is a challenging task. Distinguishing between distinctively good or bad conversations can be done using structural and content-based features. However, the nuanced task of identifying collaborative conversations is more challenging, as it often has the characteristics of both. Collaborative conversations can be long or short and non-collaborative conversations can be highly polite. Instead of relying on keyword-based analysis, it requires modeling the conversational flow.

Our technical approach follows this intuition. We design a global relational model 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. We use DRaiL [7], a recent framework for modeling decision dependencies in a deep learning framework. In this framework, decisions are formulated as first-order logic rules and interpreted probabilistically as a graphical model. A DRaiL program consists of a set of rules, each defining a factor template over a set of variables. Each rule is associated with a neural architecture used to learn its scoring function, and a feature representation definition, which describes how the model variables are represented. Finally, a global decision is made by performing MAP inference.

Each rule describing a latent discourse behavior is composed of: (1) a template definition written in first order logic specifying structural dependencies, (2) the neural network architecture that will be used to learn the parameters of its scoring function, and (3) the set of relevant raw features to be extracted and used for learning. These higher level behaviors are mapped to the final prediction of goodness using additional rules. Such rules enable the expressivity and interpretability of the relational model simultaneously.

Using DRaiL achieves 2 goals. First, the declarative rule definition provides a convenient formalism for expressing the dependencies between the discourse behaviors. When making a prediction, the latent variable activations provide a way to explain the prediction and interpret the learned model. With this approach, it is easier to inject domain knowledge or impose constraints, and we end up with a model that is easier to interpret and debug. Second, it allows for modular learning of discourse behaviors, with each rule learning a non-linear mapping from the raw-text to a specific discourse behavior. In the end, the activation/deactivation of certain latent behaviors allows the model to make the global prediction.

Our experiments (Section 5) show that the joint model involving global learning of different latent discourse behaviors significantly outperforms a local model in which all the raw features corresponding to different discourse behaviors are just combined in a single neural network. We use the Yahoo News Annotated Comments Corpus [8] and final labels are annotated based on the definitions mentioned above. We perform additional experiments to see the performance of each individual latent discourse behavior and measure the effectiveness of joint learning in the global model as well.

2 RELATED WORK

Analyzing conversational data and identifying social and linguistic indicators for collaborative and anti-social interactions was previously studied in several works, including dispute identification [1], counseling conversations [9] and most relevant to this work, identifying constructive conversations [5, 6]. In this paper we adapt the conversational data provided by Napoles et al. 2017b [8] to accommodate a more restrictive definition of *good* conversation, focusing on collaborative behavior. Conversations that are polite and socially pleasant without much content are not considered as collaborative in our case. Also, conversations that do not include balanced engagement from all the participants or contain few off-topic, insulting and rude posts are not considered collaborative as well.

From a technical perspective these works attempt to characterize desired and undesired conversational behaviors using lexical and discourse features. For example, [10] make use of domain-independent lexical and syntactic features on Wikipedia edits to study the relationship between politeness and social power. Other works [11–14] focus on the persuasive power of arguments made during the conversational interactions.

Our technical approach is different, instead of directly building on the raw inputs, we formulate the decision over a set of latent variables designed to capture fine-grained behaviors. Reasoning over conversational interactions using latent variables was previously suggested by Chaturvedi et al. 2014 [15], for predicting instructors' intervention in MOOCs, our task aims to characterize the entire conversation, rather than the actions of a single participant. Our latent variable formulation is used to characterize the conversational style, engagement and information flow. Other works focused on similar analysis in the supervised settings. For example, discourse relations between posts in conversational threads [16], and agreement and disagreement in social media dialogs [17].

To formulate our decision over latent discourse behaviors, we define a global relational model. The trade-off between local and global learning was explored in traditional graphical models (e.g., MEMM vs. CRF), and more recently, specifically, for dependency parsing [18,19]. While local learning is significantly faster, as it does not require solving a combinatorial inference problem during training, the different scoring functions learned might not be consistent with the correct global prediction. For this reason, building complex global models over relational data has attracted considerable attention in the machine learning community, and several high level languages for specifying the structure of different graphical models have been suggested. For example, BLOG [20] and CHURCH [21] were suggested for generative models, and MLN [22], PSL [23], FACTORIE [24], and CCM [25,26] were suggested for conditional models. On the other hand, combining deep learning with structured models has been studied by several works, typically in the context of a specific task or a specific inference procedure. These include dependency parsing [18,27], transition systems [19], named entity recognition and sequence labeling systems [28,29], and models for combining deep learning and graphical models for vision tasks [30,31]. Using DRaiL, we can explore different modeling decisions, with the added benefit that the scoring function for each factor can be learned using highly expressive models; unlike the other declarative frameworks, which assume a fixed representation.

3 TECHNICAL APPROACH

3.1 Task Definition

Collaborative and constructive conversations are purposeful interactions, often revolving around a desired outcome, in which interlocutors build on each others' ideas to help move the discussion forward. These conversations are an important tool in collaborative problem solving [32] and require collaboration skills [33, 34]. To help make this concept concrete, we manually analyzed the collaborative behaviors found in two conversational datasets. The first, consisting of online students interactions which focused on topics discussed in class, and were graded by the instructors. The second dataset consisted of the Yahoo News Annotated Comments Corpus [6], which is less structured and discusses a diverse set of topics. Building on previous work characterizing collaborative interaction [33, 35], we identified repeating discourse behaviors in collaborative and constructive behaviors and conversations lacking these traits. These characteristics helped in operationalizing the definition of such discussions and were used in the annotation guidelines for labeling data. Similar characteristics were grouped together to represent a specific collaborative/non-collaborative discourse behaviors (as shown in bold below)

3.2 Model Overview

We build a probabilistic model mapping raw features from conversations to different higher level latent discourse behaviors and make global prediction about the goodness of the conversations based on these competing discourse behaviors. We defined different characteristics of collaborative and non-collaborative conversations above that were used to model high level latent discourse behaviors. These characteristics help in differentiating collaborative and non-collaborative behaviors allowing for precise human annotation along with feature extraction corresponding to each discourse behavior. This is followed by learning and inference in the joint model as specified in DRaiL. Finally, all the relevant features are extracted from conversation threads and are used to model different latent discourse behaviors are described in detail. These features are used as input to the neural network used as scoring function for each latent behavior.

3.2.1 Characteristics of Different types of conversations

It is important to specify our definition of collaborative and non-collaborative conversations along with their characteristics to ensure consistency in annotation and allow grouping of different characteristics to unique discourse behaviors.(shown in bold below).

Definition: A *collaborative* conversation is interesting, polite, rich in content and one in which everyone feels part of. On the other hand, a *non-collaborative* conversation is boring, lacks relevant content with respect to the original top-most post and does not seem like a constructive discussion. A *collaborative* conversation involves working together and coming up with a set of ideas relevant to the original post. It should have sense of direction to resolve something (such as answering a query or substantiating their individual arguments for the original post). A *non-collaborative* conversation should have a nice flow and continuity from one post to the next. The participants take sincere interest in others and their viewpoints even when they are in disagreement. In a *collaborative* conversation, the participants are able to relate to what others are saying and can ask clarifying questions if they feel like. In contrast, a *non-collaborative* conversation involves people not taking each other seriously. The overall tone of the conversation shows lack of respect (example: opinions are dismissed, talking over others, interrupting others etc.). The post may or may not be long but it definitely lacks relevance to the original post.

Now, we describe characteristics of different higher level good and bad discourse behaviors that help in the formulation of our annotation guidelines.

NEGATIVE DISCOURSE BEHAVIORS:

Low Idea Development

- Users deviate from the thread topic and discuss something else
- Users ignore ideas raised in previous turns and only care about their ideas
- The participants are just repeating each other viewpoints

Low User Engagement

- Users show little interest in the discussion topic
- Users involve in shallow discussion, consisting of mostly telling jokes or sharing links, or similar activities

Negative Sentiment

- Disagreements are not resolved politely and respectfully

Presence of Rudeness

- Abusive, impolite or rude content

POSITIVE DISCOURSE BEHAVIORS:

High Idea Development

- The users stay on topic with respect to the original top-most post

- New ideas formed and developed based on preceding turns

Reference to previous posts

- There is continuity in discussion and the users refer to the previous post

Back and Forth

- Users support and appreciate the ideas shared by others. Disagreements are generally polite

Positive sentiment

- Positive interaction between users. It can be informal through use of emoticons etc.

High User Engagement

- Discussion is insightful and/or meaningful to the participants of the conversation

Balanced content distribution

- Balanced contribution by the all the members in the group

High Questioning Activity

- People advance the conversations by asking interesting questions

Some of the interesting sample conversations representing the different sets of above mentioned nuanced behaviors (found during annotation phase) are shown below:

Example 1: Families struggling with teens' phone addiction

Behaviors present: High Idea development, Reference to previous posts, High user engagement, Balanced content distribution

User A : All the people I know that are under 35 say that we old people are out of touch and that things have changed. Part of that may be true, but we can see what this phone addiction is doing to people. I am just surprised to see that the research shows young people check their phones at least once an hour. It seems more like once every 10 minutes. Of course, the survey did not point out the rest of the hour is spent holding the phone, massaging the phone, using the phone as a pacifier, or cuddling the phone as if it is a helpless puppy. Does anyone but me think it might not be safe for an eight or nine year old kid to be distracted by a phone when they are around busy streets?

User B : Just look at the young people driving, they constantly have their phone in their hand. If they aren't looking at it while driving they're getting ready to as they slowly come to a stop at a red light. I saw a guy so engrossed in his phone at a stop light about 15 cars made their left turn, double lane, while he sat there all alone. I was actually moving to go through my green when he finally looked up to see he was the only complete idiot on the road at that time.

User C : Well said. I have been saying this for years but let's not put all the blame on the kids. We, the adults, go out to dinner with our kids and allow the kids to look at their phones at the table. Us adults do the same thing so the kids see this as ok.

Example 2: Watch This Tesla's Autopilot Save a Driver's Life

Behaviors present: Reference to previous posts, Low idea development, Presence of subtle rudeness, High user engagement, Balanced content distribution

User A : The only great thing to self driving cars is that hopefully traffic congestion will not be as bad. The reason there's traffic, not accident bound, is because everyone is driving at different speeds. There's always those that are speeding and those that speed. These inconsistent speeds cause traffic.

User B : I speed every time I touch the interstate. Normally doing about 85-90 in a 75 and honestly it's a bad habit of mine. Luckily, I only see wrecks on people who actually do the speed limit, usually due to the fact that I may be speeding, but I'm also paying attention to everything around me instead of thinking I'm doing fine and checking my phone or looking at the sights around me. Get rid of phones and such in the car, see a decrease in idiotic driving. (Disclaimer: Don't speed. I do it because I'm an idiot.)

User C : If you're speeding that's not anyone else's fault, the speed limit is just that a limit! The minimum speed on all interstate highways is 40 mph per the us regulation.

Example 3: What Is Batman V Superman's Connection To the Flashpoint Paradox?

Behaviors present: Negative Sentiment, Presence of Rudeness, Low idea development, High Questioning Activity, Imbalanced Content distribution

User A : I understood it for what it was, most comic book fans probably understood it for what it was too.. The average movie-goer was probably like BAD-KEYWORD is this BAD-KEYWORD. Just like so many of them are complaining about the mere fact that Batman was fighting Superman and how in RL we all know Superman would just destroy Batman. LOL people literally saying in real life about comic book characters. learn to take movies for what they are. A chance to suspend or ignore reality for 2 hours and just have fun

User B : Yes.. enjoy the movie.. will watch it again.

User C : Your criticism is short-sighted and narrow. I was perfectly willing to suspend by disbelief. However, when you are asked to suspend your disbelief to the point where the characters no longer act as themselves, it is poor storytelling. I refuse to lower my bar and accept anything that is thrown at me simply because it is a superhero story (that I've waited a half century to see). Why must we continue to be subjected to these ”re-imaginings” of beloved characters by hacks who were never creative enough or talented enough to have imagined them in the first place?

3.2.2 Annotation guidelines

Based on our definitions of positive and negative conversations along with their characteristics, our annotation guidelines are formulated to classify conversations in the dataset as positive or negative. Two human annotators labeled the conversations based on these guidelines and tie breaks in annotation are resolved with thorough discussion after one iteration of annotation. The higher level discourse behaviors allow annotators to look for different indicators and make a final decision about the overall conversation based on the weightage of different behaviors in a conversation.

We have a very restricted definition of goodness in our scenario as mentioned before. Conversations that are polite and socially pleasant without much content are not considered as good in our case. Also, conversations that doesn't include balanced engagement from all the participants or even contain few off-topic, insulting and rude posts are not considered good as well.

We also made sure that the topic of original post is not used in the determination of discussion quality. The length of a post was also not used in the determination of discussion quality. The respectfulness of posters was determined based on context in the discussion and not just finding good/bad words individually. If most of them have elements of goodness, then it is considered a positive conversation If the conversation has concluded based on the original post, still conversation is positive even if there is slight deviation from the theme. We ignored grammatical mistakes as long as we were able to identify the ideas/thoughts in the discussion

3.3 Probabilistic Model in DRaiL

Identifying collaborative conversations requires characterizing nuanced behaviors. These behaviors often are not the product of a single conversational turn, or are expressed directly in the specific word choice in it. These behaviors are defined at an aggregate level by combining multiple turns. Previously, this analysis was defined by extracting features directly from the raw data, each behavior associated with a set of features. Unfortunately, these features provide a rough characterization of social and discourse behaviors, and can include conflicting indications.

Instead, we view this decision as a probabilistic reasoning process, in which the raw features from conversations are associated with different higher level latent discourse behaviors and allowing us to make a globally consistent prediction.

This process is described in figure below. First, we define different characteristics of collaborative conversations that are used to model high level latent discourse behaviors. We denote these behaviors as $\mathbf{h} = \langle h_1, ..., h_k \rangle$, each captured by a binary latent variable, indicating if the high level pattern is active or not in the given thread. Each latent variable decision is scored by a Neural net, and uses a set of features capturing relevant properties in the input conversation. These indicators help

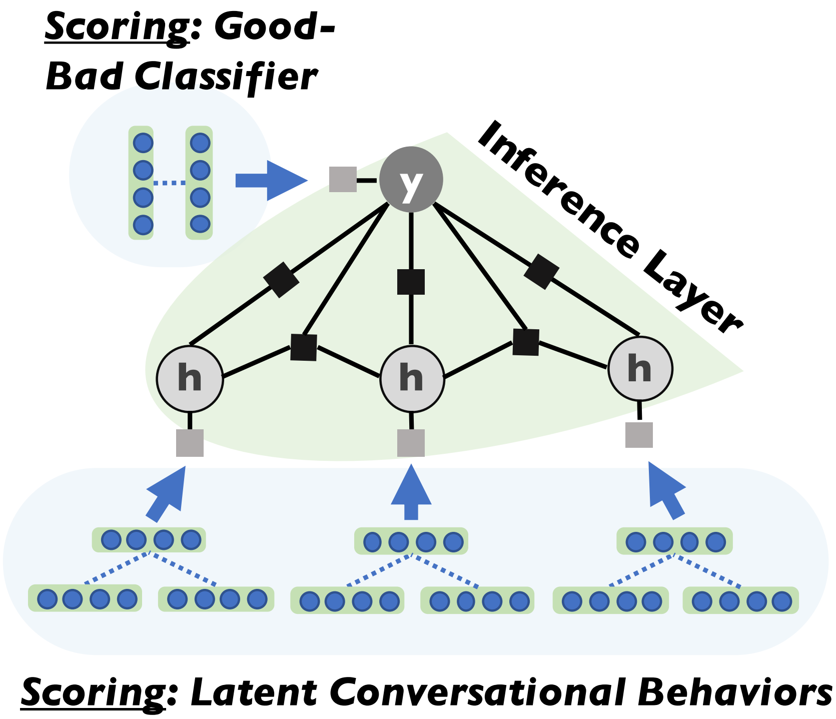
Scoring: Good-Bad Classifier
y
Inference Layer
h
h
h
Scoring: Latent Conversational Behaviors

Figure 3.1. Inference

in identifying collaborative behavior, captured by the higher order factors in Figure 3.3.1, connecting the discourse behaviors with the final prediction, denoted y, a binary output value. Finally, all the relevant features extracted from conversation are used to score the final prediction, in addition to the latent discourse indicators. We implemented this model in DRaiL [7], a recently introduced framework for combining declarative inference with deep learning, described briefly in the following section.

3.3.1 Inference in DRaiL

DRaiL uses a First-order logic template language to define structured prediction problems. A task in DRaiL is defined by specifying a finite set of *entities* and *predicates*. The predicates can correspond to hidden, latent or observed information and a specific input is defined by the instantiations of these elements. Decisions are defined using rule templates, formatted as horn clauses: A ⇒ 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 our global decision, taking into account the dependencies between the rules using a set of constraints C, defined over indicator variables r_i for each rule instance. Each rule grounded is scored using a neural net, defined over a parameter set w.

The inference procedure can then be expressed as shown below where r_i corresponds to a particular rule grounding along with it's particular score.:

$$y* = \arg\max_{\forall r_i} \sum_i r_i * score_i(x_i, \mathbf{h}, y, \mathbf{w}) \tag{3.1}$$

subject to set of constraints C, for all i: r_i belonging to [0,1] and x_i is a problem component tied to rule network i

We define rule activations over Boolean variables r_i for each rule grounding, indicating whether they are active or not. The final prediction y* corresponds to the collection of heads in active rule groundings.

3.3.2 Local vs Global learning in DRaiL

rule: Thread(T) ⇒ FinalLabel(T)

Local Learning: In this modality, each rule template is treated as an independent learning problem and their associated network parameters are optimized separately. We feed each network all observed instances of the rule in the training data. The architecture, learning choices and hyperparameters are configurable and can be tuned for each subproblem. At prediction time, we generate rule groundings by enumerating all possible values for the rules variables given its domain and score the factors using the neural nets. In order to enforce consistency between variable assignment and dependency among them, relevant constraints are taken into consideration in the

ILP formulation.

rule: Thread(T) ⇒ LatentBehavior(T,B)

rule: LatentBehavior(T,B) ⇒ FinalLabel(T)

where B ∈ Latent Behavior Set {Idea Development, Reference to previous post, Sentiment, Content Length, Back and Forth, Rudeness, High Questioning Activity, User Engagement}

Global Learning using latent variables: While prediction in DRAIL always uses global inference, we might fail to optimize for certain structural dependencies when we minimize the loss function for each rule independently. For this reason, we incorporate the option of using inference at training time and define a structural objective that promotes correct global predictions. For example, when training a collective classification task such as stance prediction on a debate network, each user and its friends can be considered to make a collective prediction. Then the structured hinge loss is used for updating the parameters in all networks

$$\min_{w} \frac{\lambda}{2} \left\|\mathbf{w}\right\|^2 + \delta \tag{3.2}$$

$$\delta = \sum_{i \epsilon rule} \max_{y \epsilon Y, \mathbf{h} \epsilon H} (score_i(x_i, \mathbf{h}, y, \mathbf{w}) + cost(y, t_i)) - \max_{\mathbf{h}} score_i(x_i, \mathbf{h}, t_i, \mathbf{w}) \tag{3.3}$$

where x_i and t_i are the problem components and gold predictions tied to rule network i, Y denotes all possible predictions, and $score_i()$ is the output of the last hidden layer of the corresponding network.

3.3.3 Constraints in DRaiL

In order to enforce consistency between variable assignments and dependencies among them, the following five types of constraints are taken into consideration in an

ILP formulation:

- Negation Constraints: The first type constraints ensure exclusive activation of a head predicate and its negation at the same time
- Implied Contraints: Each rule template defines the dependency between body and head. This dependency is reflected between the rule groundings variable and the head variables in the body
- Rule/head constraints: One head predicate can be associated with multiple rule grounding variables. Activation of any rules in ruleset(j) ensures the activation of the head variable. On the other hand, the activation of the head variable ensures the activation of at least one of its corresponding rule variables
- Binary/multi-class/multi-label constraints: In many problems, we are facing multi-class or multi-label decisions. DRAIL guarantees this by adding suitable constraints. For instance, in the multi-class case, among all head variables on the same entity, only one of them is activated while the others remain inactive, as a decision is made on which class to choose. Note that the constraints for binary predicates can be covered by the negation constraints mentioned above.
- Hard constraints from rule definitions: Users can define hard constraints in the rule templates, which usually infuse prior knowledge and thus improve the prediction capacity. Rule groundings of these templates are dealt differently as the activation of such a rule depends on the activation of all body predicates

Overall, a DRaiL program describes the interaction between inference, learning, and representation. It is defined over a set of predicates, which can represent either an observed value or an output prediction. From that perspective, each rule defines a factor template, and we learn the parameters of the scoring function for each rule using deep learning architectures. Architectures can be different for each rule and normalized into a probability distribution to allow global inference over all competing values representing different discourse behaviors

3.4 Modeling different discourse behaviors

We model a set of nine latent discourse behaviors in DRaiL. We associate one neural network with each latent behavior rule, and relevant features are extracted as inputs. The presence or absence of a latent behavior depends on extracting the correct features from the conversation threads. In this section, we describe the behaviors that we consider, as well as the set of features used to capture them.

Sentiment related behaviors

Capture the overall emotion and attitude of the conversation.

rule: Thread(T) ⇒ Sentiment(T,S)

Feature Representation: We use the degree of positive, negative and neutral sentiment, along with the degree of intensity for the top most post, and the mean of the degree of intensity for subsequent posts.

Balanced Content Distribution related behaviors

Capture the level of participation for all users in the conversation.

rule: Thread(T) ⇒ Balanced(T,B)

Feature Representation: We use the number of sentences per post, the number of words per posts, the depth of the post, and indicators for the distribution of content among the participants. For content distribution, we use the average of the ratio between the length of each post in a thread and the length of the main post.

Controversial behaviors

Capture the level of disagreement and heated discussion on specific topics.

rule: Thread(T) $\Rightarrow$ Controversial(T,C)

Feature Representation: We measure ratio between upvotes (u) and downvotes (d), as well as other measurements: $u - d$, $u + d$, and $u/(u + d)$. This way we gauge post popularity as well as disagreement among people.

Reference to Previous Posts related behaviors

Capture the degree to which users follow up on previous statements.

rule: Thread(T) $\Rightarrow$ PrevRef(T,R)

Feature Representation: We use the presence of second person pronouns (e.g. you, yours, yourself), quotes of previous posts, and the use of the @username tag to refer to another person.

Back and Forth behaviors

Capture the exchange of ideas with a lot of competing arguments.

rule: Thread(T) $\Rightarrow$ BacknForth(T,B)

Feature Representation: We use agreement and disagreement markers, indicators of sufficient content, and references to previous posts

Idea flow related behaviors

Capture the advancement of ideas put forth by members in a conversation.

rule: Thread(T) ⇒ IdeaFlow(T,I)

Feature Representation: We use lexical chains [36] to link related words across the conversations representing ideas. This is followed by ranking the chains using different criteria: 1) Length, measured by counting the number of occurrences of members in the chain, 2) Homogeneity index, measured as 1 - the number of distinct occurrences divided by the length.

Rude behaviors

Capture offensive or ill-mannered speech that obstructs meaningful discussions.

rule: Thread(T) ⇒ Rude(T,R)

Feature Representation: We model the presence of profanity, use of bad words, and indicators for posts that are too short.

User Engagement related behaviors

Capture the general response of users to a conversation.

rule: Thread(T) ⇒ UserEng(T,E)

Feature Representation: We use the number of posts, as well as the number of threads initiated by the user

Questioning behaviors

Capture the different questions asked by the participants.

rule: Thread(T) ⇒ Ques(T,Q)

Feature Representation: We model the presence of question marks, indicators for different types of question (who, how, what, why, when etc.), and indicators of whether it is a descriptive, relational or causal question.

Finally, as explained in section above, we add a rule of the form LatentBehavior(T,B) ⇒ FinalLabel(T) for each described behavior to condition the final decision on all the latent behaviors. For each conversation thread and behavior, we generate one instance of the rule for the case of an active latent behavior, and one for the case of an inactive latent behavior. We use a bias term as an input to these networks (i.e. 1) and learn a single parameter to capture whether a latent variable is active or not. The inference procedure in DRaiL makes the final decision and guarantees that each behavior is either active or inactive for a given thread.

4 EXPERIMENTS AND EVALUATION

In this section, we describe the dataset and evaluate our two DRaiL learning models. First, a **local model** that predicts whether a conversation is collaborative or not by using all discourse features as inputs to a single neural net. Then, a **global model** extending the local version by modeling the latent discourse behaviors defined in previous section. We also conduct additional experiments evaluating the contribution of different discourse behaviors, as well as our model's ability to capture these behaviors using latent variable training.

4.1 Dataset and Experimental Settings

We annotate the conversations on the Yahoo News Annotated Comments Corpus [8] by following the guidelines specified in previous section. The resulting dataset consists of 2130 conversations for training, 97 for validation and 100 for testing. The data is imbalanced with more conversations labeled as non-collaborative. In addition, to evaluate our latent variable formulation, we annotated the fine-grained discourse behaviors for a sample set of 103 conversations from the training data.

We used feedforward networks for all rules, with one hidden layer and a softmax layer on top. All hidden layers use sigmoid activation functions. The number of hidden units are: 400 for the local rule, 50 for idea flow and 100 for all remaining behaviors. We didn't use a hidden layer for rules that map a latent behavior to a final decision. We used SGD and a learning rate of 0.01 for training. After training, we tuned weights for the different rules using the validation data.

As mentioned before, the conversations are annotated based on our strict definition of goodness which resulted in majority conversations (greater than 60%) being labeled as bad conversation in the dataset.

We use Macro precision, recall and F1 score as our evaluation measure because the dataset is imbalanced and we want to focus our model's performance on minority class (good conversations) as well. We also use validation set for testing our trained model and prevent overfitting. In DRaiL, all the implemented models (global or local) use early stop approach to terminate training of neural network(s) after specific number of epochs and select the best estimated trained model based on performance on validation set. This trained model is then evaluated on our test set and results are shown below.

4.2 Experimental results

For evaluation, we first compare our global model involving latent discourse behaviors with different baselines in order to measure the effectiveness of joint learning as shown in Table 4.1. We compare the two DRaiL models described above with two linear baselines: a linear model using BoW features, and an additional model enhanced with discourse markers features. Both models use a SVM classifier. Results can be observed in table 4.1. Since our DRaiL models use non-linear neural networks, both perform better than the linear models. The global model outperforms the local version, increasing the F1 score by 4.7 points. These results demonstrate the advantage of modeling competing discourse behaviors jointly and performing inference, as opposed to just representing them as features to a neural model. Our experiments show that the global model results in improved performance, as a result of modeling discourse behaviors. However these behaviors are learned as latent variables, which may not capture valid patterns. To evaluate the latent model correctness, we conducted an additional experiment below.

Model	Precision	Recall	F1
Baseline Linear SVM(BoW)	0.600	0.580	0.590
Baseline Linear SVM(BoW + discourse indicators)	0.630	0.610	0.620
DRaiL Local model (Non-linear - single neural net)	0.650	0.640	0.640
DRaiL Global Model (with all latent behaviors)	0.690	0.680	0.687

Table 4.1.
Global Model compared with different baselines

4.3 Latent discourse behavior analysis

A sample set of 103 conversations annotated by us is used from training set to analyze individual latent behavior to evaluate how well we capture the latent behaviors described in previous section. For this purpose, we used a sample set, annotated for discourse behaviors based on the definitions provided in above section, and evaluate the latent variables activations produced by our global model. Table 4.2 describes the results. We can observe that identifying rude behaviors yields the highest F1 score (0.62), this is because it is straightforward to identify negative and abusive words in a conversation. The same can be said for the balanced content behavior, given that structural features are very informative. Lexical chains are also successful at capturing idea flow behaviors. However, controversial and back and forth behaviors are more challenging and thus exhibit lower performance.

These results are obtained without explicit supervision, and depend on the initialization point of the model and the learning process connecting the latent model with observed outcomes. Table 4.3 captures the performance on sample set conversations before and after global learning i.e. initial epoch to the best epoch. This experiment is just to validate that performance on sample set is better than evaluation set since sample set is a part of training set. Table 4.4 captures the impact of the learning process, by comparing the performance of latent behaviors prediction before (i.e., using the model initialization point) and after global training. This is done to measure the bias introduced by initializing the model and tuning the rule weights, against the outcome of the global procedure. Performance consistently improves for all discourse behaviors, a clear indication that we are learning meaningful latent information with our model.

Individual Behavior	Precision	Recall	F1
Idea Flow	0.627	0.609	0.574
Controversial	0.364	0.50	0.420
Balanced content	0.520	0.720	0.610
Sentiment	0.547	0.558	0.548
User Activity	0.568	0.592	0.570
Reference to previous posts	0.545	0.340	0.427
Questioning Activity	0.515	0.520	0.511
Rudeness specific	0.650	0.590	0.620
Back and Forth	0.530	0.510	0.520

Table 4.2.
Individual Latent Behavior performance on sample set after global learning

Evaluation metric	Before Learning	After Learning
Precision	0.45	0.71
Recall	0.49	0.69
F1 score	0.43	0.70

Table 4.3.
Comparison of performance on sample set conversations before and after global learning

Individual Behavior	F1 score before learning	F1 score after learning
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 specific	0.514	0.620
Back and Forth	0.470	0.520

Table 4.4.
Latent discourse behavior comparison of sample set before and after global learning

Model	Precision	Recall	F1
All behaviors except Sentiment	0.495	0.495	0.490
All behaviors except Idea flow	0.620	0.584	0.580
All behaviors except Balanced Content	0.579	0.587	0.546
All behaviors except Questioning activity	0.592	0.569	0.568
Idea flow + Sentiment + Balanced content	0.675	0.61	0.608
Idea flow + Sentiment + User Activity	0.66	0.507	0.264
Sentiment + Balanced Content + Controversial + Questioning activity	0.681	0.599	0.596

Table 4.5.
Discourse Behavior Ablation

4.4 Latent Discourse Behavior Ablation study

We also performed an ablation study to see if the global model is driven by any particular discourse behavior or group of discourse behaviors. It also gives an indica-

Model Type	F1 without gold latent labels	F1 with gold latent labels
Type I (0.653 F1 on test set)	0.620	0.640
Type II (0.666 F1 on test set)	0.689	0.630
Type III (0.687 F1 on test set)	0.700	0.640

Table 4.6.
Classification performance on sample set with and without gold latent labels

tion about the impact of a particular behavior in case performance drops significantly by removing one particular behavior. As we can see in Table 4.5, capturing sentiment behavior plays an important role as the performance significantly drops without it. Also, just using rules related to idea flow, sentiment and balanced content behaviors leads to an F1 score of 0.61.

4.5 Effect of gold values of latent behaviors on global prediction

In table 4.6, we compare 3 different types of global models to see if providing gold labels for latent behaviors help in boosting the performance of overall global prediction. Although it shows a performance increase in Type I model but not in others which can be due to the fact that different latent behaviors have not been captured correctly.

4.6 Different weightage of local classifier in the global model

Finally, we check the imapct of different weightages of local classifier i.e. local rule in the global model to see the extent to which global model is driven by latent rules and local classifier. Suprisingly, the performance of global model inclreases with the increase in weightage of local classifier upto 0.5 and then decreases. This maybe due to the fact that the features corresponding to local classifier might be conflicting with features corresponding to latent rules in the global model.

Weightage of local classifier in global model	Precision	Recall	F1
0.1	0.509	0.510	0.508
0.3	0.652	0.614	0.619
0.5	0.679	0.660	0.687
0.7	0.660	0.637	0.644
0.9	0.653	0.606	0.610

Table 4.7.
Final global classification performance using different weights of Local classifier in the global model

4.7 Additional Experiments

We also performed 5-fold cross validation in which training, testing and validation data were combined and reshuffled to test the robustness of the global model and sensitivity of test data. The results were in the range of 0.50-0.69 for same configuration of initial weights. It is possible that these weights can be tuned more to give better performance for different sets.

Also, random feature experiment was performed in which different latent rules were associated with random sets of features to see if the semantic interpretation of latent rules plays a role in overall global decision for the conversation. It was seen that the random latent rules performed much worse around 0.50-0.55 for the same test data.

One more important experiment was to see the drastic difference between linear and non-linear neural networks used for the latent rules. If there was no hidden layer used for each of the latent rules, the performance dropped drastically to F1 score of around 0.50 from 0.687.

Finally, different other configurations like hot start in which local learning for the local rule was done initially followed by global learning involving all the latent rules was done which decreased the performance of the global model. Also, few hard constraints were tested in the global model involving some values of latent rules to see if it helps in the global prediction, but it didn't make any different in the final prediction. It can be due to inappropriate choice of latent rules for hard constraints or incomplete features used for those specific latent rules.

5 SUMMARY AND FUTURE WORK

A deep latent-variable approach for the problem of identifying collaborative conversations online was shown above. The beauty of this approach is that it allows us to accommodate more nuanced discourse behaviors in the future . Each nuanced discourse behavior can be considered as a separate line of research. Capturing such complex behaviors (almost) perfectly and their dependencies allows us to make global prediction about various kinds of decisions including overall collaborative/non-collaborative behavior. We used DRaiL, a framework for combining declarative structural modeling with deep learning, and showed that both aspect contribute to better performance on Yahoo forums dataset, demonstrating how adding additional inductive bias through constrained latent variable models can improve learning.

We can have more refined annotation guidelines to capture diverse set of conversations, complex topic ideas and resolve tie-breaks for ambiguous/interesting examples. Similarly, it can be used to test the learnt global model on a completely different sets of conversations in out of domain settings to see the effectiveness of latent rules used to represent semantic behaviors.

Out of various discourse behaviors identified during annotation process, novelty and off-topic behavior in conversations were most difficult to identify as the conversation has to be off-topic to a certain extent in order to be classified as novel but not deviate too much to derail the conversation.

Also, we can have more complex rules involving latent variables to capture more nuanced behaviors identified during the annotation process and initial weights can be assigned based on certain heuristics developed during the annotation process.

Finally, collaborative and constructive interactions can help leverage the synergy between team members when tackling complex problems, we hope that this work will help contribute to the efforts of developing automated systems supporting such processes.

REFERENCES

REFERENCES

[1] Lu Wang and Claire Cardie. 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)*, volume 2, pages 693–699, 2014.

[2] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 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, 2017.

[3] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 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, 2018.

[4] Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 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, 2009.

[5] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational markers of constructive discussions. In *Proceedings of NAACL-HLT*, pages 568–578, 2016.

[6] Courtney Napoles, Aasish Pappu, and Joel R Tetreault. Automatically identifying good conversations online (yes, they do exist!). In *ICWSM*, pages 628–631, 2017.

[7] Maria Leonor Pacheco, Xiao Zhang, Chang Li, and Dan Goldwasser. Introducing DRAIL - a step towards declarative deep relational learning. In *Proceedings of the Workshop on Structured Prediction for NLP@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 54–62, 2016.

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

[9] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463, 2016.

[10] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 250–259, 2013.

[11] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

[12] Zhongyu Wei, Yang Liu, and Yi Li. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 195–200, 2016.

[13] Ivan Habernal and Iryna Gurevych. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas, 2016. Association for Computational Linguistics.

[14] Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 742–753, 2017.

[15] Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1501–1511, 2014.

[16] Amy X Zhang, Bryan Culbertson, and Praveen Paritosh. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the Eleventh International Conference on Web and Social Media*, 2017.

[17] Amita Misra and Marilyn Walker. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, 2013.

[18] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[19] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *ACL*, pages 2442–2452, 2016.

[20] Brian Milch, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L Ong, and Andrey Kolobov. Blog: probabilistic models with unknown objects. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK*, 2005.

[21] Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. In *UAI*, 2012.

[22] Pedro M Domingos, Stanley Kok, Hoifung Poon, Matthew Richardson, and Parag Singla. Unifying logical and statistical ai. In *AAAI*, 2006.

[23] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *arXiv:1505.04406 [cs.LG]*, 2015.

[24] Andrew McCallum, Karl Schultz, and Sameer Singh. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *NIPS*, 2009.

[25] N. Rizzolo and D. Roth. Learning based java for rapid development of nlp systems. In *LREC Proceedings,Malta*, 2010.

[26] Parisa Kordjamshidi, Dan Roth, and Hao Wu. Saul: Towards declarative learning based programming. In *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI-15, Buenos Aires, Argentina*, 2015.

[27] David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. Structured training for neural network transition-based parsing. In *ACL*, 2015.

[28] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[29] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[30] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[31] L.-C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun. Learning deep structured models. In *ICML*, 2015.

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

[33] Michael Flor, Su-Youn Yoon, Jiangang Hao, Lei Liu, and Alina von Davier. 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, 2016.

[34] Jiangang Hao, Lei Liu, Alina von Davier, Patrick Kyllonen, and Christopher Kitchen. Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In *EDM*, 2016.

[35] Jingyan Lu, Ming Ming Chiu, and Nancy WaiYing Law. Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior*, 27(2):946–955, 2011.

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