Predicting Dialogue Outcomes over Structured Latent Representations

Dan Goldwasser
University of Maryland
College Park, MD
goldwas1@umiacs.umd.edu

Hal Daume III
University of Maryland
College Park, MD
hal@umiacs.umd.edu

1 Introduction

Many dialogues lead to decisions and actions. The participants in such dialogues each come with their own goals and agendas, their own perspectives on dialogue topics and ways of interacting with others. *Understanding natural language dialogue can be considered as the process of identifying how relevant dialogue content combined with its participants features, lead to the dialogue outcome.*

Studying dialogue relations as a complex structured output prediction problem, independently of the dialogue outcome, has been the focus of significant research efforts, such as discourse relations [9], rhetorical structure [8, 1] and dialogue act modeling [13]. While these works capture linguistically-motivated fine-grained dialogue relations, they require considerable effort and linguistic knowledge. Furthermore, the subjective nature of our task, focusing on dialogue relations relevant to the dialogue outcome, allow for many different interpretations of the interactions leading to the observed outcome.

In this work, we devise a discriminative latent variable model that captures the overall structure of a dialogue as relevant to specific acts that occur as a result of that dialogue. We aim to model both the *relevance* of preceding dialogue to particular real-world action, as well as a binary structured relationship among utterances, while taking into account the pragmatic effect introduced by the different speakers' perspectives. We focus on a particular domain of dialogue: courtroom transcripts. This domain has the advantage that while its range of topics can be broad, the roles of participants are relatively well-defined. Courtroom dialogues also contain a specialized speech act: **the objection**.

Our technical goal is to drive latent learning of dialogue structure based on a combination of raw input and pragmatic binary supervision, derived from *objection* speech acts naturally appearing in the dialogue. Our model makes use of three conceptually different components capturing linguistic and pragmatic considerations and their relevance in the context of the dialogue structure.

Our linguistic model focuses on enriching a lexical representation of the dialogue utterances (complex tools fell apart on this data), using linguistic resources capturing biased language use, such as subjective speech, expressions of sentiment, intensifiers and hedges. For example, the phrase *"So he was driving negligently?"* is an *argumentative* expression, as it requires the witness to draw inferences, rather than describe facts. Identifying the use of biased language in this phrase can help capture this objectionable aspect. In addition, we use a named entity recognizer, as we observe that relevant entity mentions provide a good indication of the dialogue focus.

Since turn utterances are *situated* in the context of the court case, the relevance and information flow properties captured by the latent variables can be conditioned on the speaker (or the side in trial), thus allowing the model to focus the learning on relevant pragmatic influences.

Next, a discriminative latent variable model learns a structured representation of the dialogue, useful in making the high-level prediction. The model consists of two variable types. *The first* aims to identify relevant content for the objection decision, based on properties such as lexical items and expressions of subjectivity combined with speaker's features. *The second type* focuses on the infor-

mation flow between speakers, and identifies relevant dialogue relations between turns by constructing a joint representation of the turns, capturing responses to questions and combining lexical items appearing in different turns. Both of these aspects of the dialogue are formalized as latent variables, trained jointly with the final classification task using the automatically extract binary supervision.

We evaluate our approach over short dialogue snippets extracted from the O.J. Simpson murder trial. Our experiments evaluate the contribution of the different aspects of our system, showing that the dialogue representation determined by our latent model results in considerable improvements.

2 Dialogue Structure Modeling

Predicting dialogue outcomes requires capturing the interactions between different participants, the tone of conversation, understanding of controversial issues presented during the trial, and their different interpretations. Instead of treating this information as part of the output space, we treat it as set of latent variables determining the relevant parts of the dialogue and the relations between them. We inform these decisions using generic resources providing *linguistic* and *pragmatic* information, situating the dialogue in the context of the trial.

2.1 Linguistic and Pragmatic Information

Linguistic Resources **(1)** *Named Entities (NE)*: provide strong indications of the topics discussed in the dialogue and help uncover relevant utterances. We use the tagger described in [5].
(2) *Subjective and Biased Language*: Equally important to understanding the topics of conversation is the way they are discussed. Expressions of subjectivity and sentiment are useful linguistic tools for changing the tone of the dialogue and are likely to attract opposition. We use a lexicon of subjective and positive/negative sentiment expressions [12]. We use a list of hedges and boosters [7], which can help identify evasive and (overly) confident responses. We use a lexicon of biased language provided by [11], and finally we use a Patient Polarity Verbs lexicon [6]. This lexicon consists of verbs in which the agent performs an action with a positive (*"He **donated** money to the foundation"*) or negative (*"He **stole** money from the foundation"*) consequence to the patient.
(3) *Sentence Segmentation*: Since turns discuss multiple topics, we segment each turn into sentences and associate sentences with a label, such as FORMALITY (e.g., a witness being sworn in), QUESTION, RESPONSE (POSITIVE or NEGATIVE) and a general STATEMENT[1].

Capturing Pragmatic Effects We observe that in the context of a courtroom discussion, utterance interpretation (and subsequent dialogue actions) is conditioned to a large extent on the speaker's motivation and goals rather than in isolation. We capture this information by explicitly associating relevant characteristics of the speakers involved in the dialogue with their utterances (denoted as $\mathbf{x}_{Sit}$when this information is used). We use the list of *actors* which appear in the trial transcripts, and associate each turn with a speaker, their role in the trial and the side they represent.

2.2 Identifying Relevant Interactions using Constrained Optimization

We use the information sources described above to enrich our dialogue model, since the dialogue relations are not pre-annotated, we allow it to be learned as a latent variable. Uncovering the dialogue structure is formalized as ILP optimization problem over the components of the dialogue representation. In our experiments, use the highly optimized Gurobi toolkit[2]. More formally, given an input $\mathbf{x}$, we denote the space of all possible dialogue relevance and sentence-connections substructures as $\Gamma(\mathbf{x})$. Assuming that $\Gamma(\mathbf{x})$ is of size N, we denote latent representation decisions as $\mathbf{h} \in \{0,1\}^N$, a set of indicator variables, that selects a subset of the possible substructures that constitute the intermediate representation. For a given dialogue input $\mathbf{x}$ and a substructure $s \in \Gamma(\mathbf{x})$, we denote $\phi_s(\mathbf{x})$ as the feature vector of s [3]. Given a fixed weight vector $\mathbf{w}$ we score the intermediate representations for the final classification task, using the following linear objective.

[1] Determined by lexical information (question marks, dis/agreement indications and sentence length)
[2] http://www.gurobi.com/
[3] For brevity, we omit the description of the feature functions used in each decision

$$f_{\mathbf{w}}(\mathbf{x}) = \max_{\mathbf{h}} \quad \sum_{s} h_s \mathbf{w}^T \phi_s(\mathbf{x}) \tag{1}$$

2.2.1 Relevance Decisions

Each dialogue input consists of six turns, which occurred prior to the objection speech act. In practice, fewer turns might be needed to capture the offending interaction. Given a dialogue consisting of $(t_1, .., t_n)$ turns, each consisting of $(t_i.s_1, .., t_i.s_k)$ sentences, we associate with each sentence. (1) **Relevance** variables (denoted $h^r_{i,j}$) indicating the relevance of the j-th sentence in the i-th turn, for the classification decision, and (2) **Irrelevance** variables (denoted $h^{ir}_{i,j}$). The values of these variables is mutually exclusive, we add this constraint to our ILP formulation. These decisions are represented using features capturing occurrences of words, NEs, and biased language in conjunction with speaker's information.

2.2.2 Dialogue Structure Decisions

The information required to make the classification may not contained in a single turn, it might be the product of the information flow between dialogue participants. Given a dialogue consisting of $(t_1, .., t_n)$ turns, each consisting of $(t_i.s_1, .., t_i.s_k)$ sentences, we associate with every two sentences, $s_j \in t_i$, $s_k \in t_l$, such that $(i \neq l)$: (1) **Sentences-Connected** variables (denoted $h^c_{(i,j),(k,l)}$) indicating that the combination of the two sentences is relevant for the classification decision and (2) **Sentences-not-Connected** variable (denoted by $h^n_{(i,j),(k,l)}$) indicating the opposite. Again, the activation of these variables is mutually exclusive. In addition, we require that connected sentences should be *relevant*. This decision is represented using features combining the properties of the two sentences and their speakers.

3 Learning and Inference

Unlike the traditional classification settings, the learning process, defined over a set of latent variables, is formalized as an optimization problem that selects the dialogue elements and associated features that best contribute to successful classification.

$$\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i \ell \left(-y_i \max_{\mathbf{h} \in \mathcal{C}} \mathbf{w}^T \sum_{s \in \Gamma(\mathbf{x})} h_s \phi_s(\mathbf{x}_i) \right) \tag{2}$$

This formulation is not a convex optimization problem and care must be taken to find a good optimum. In our experiments, we use the algorithm presented in [2] to solve this problem. The algorithm solves this non-convex optimization function iteratively, decreasing the value of the objective in each iteration until convergence. In each iteration, the algorithm determines the values of the latent variables of positive examples, and optimizes the modified objective function using a cutting plane algorithm. This algorithmic approach is conceptually related to the algorithm suggested by [18, 4].

Discriminative latent variables models have seen a surge of interest in recent years, both in the machine learning community [18, 10] as well as various application domains such as NLP [3, 17, 14, 15, 14, 2, 16] and computer vision [4].

4 Empirical Study

Evaluated Systems In order to understand the different components of our system, we construct several variations. We compare our latent model with and without using pragmatic information (denoted DIAL($\mathbf{x}_{Sit}$) and DIAL($\mathbf{x}$), respectively). We also compare two baseline systems, which do not use the latent variable formulation, with and without pragmatic information (denoted ALL($\mathbf{x}_{Sit}$) and ALL($\mathbf{x}$), respectively).

System	ALLOBJ	OVERRULEDOBJ	SUSTAINEDOBJ
ALL($\mathbf{x}$)	64.9	63.7	66.9
ALL($\mathbf{x}_{Sit}$)	65.1	63.7	67.9
DIAL($\mathbf{x}$)	65.4	65.1	66.7
DIAL($\mathbf{x}_{Sit}$)	**69.1**	**66.3**	**70.2**

Table 1: **Overall results** Accuracy results of the different variations of our systems. Results show considerable improvement when using our latent learning framework with pragmatic information.

Datasets Our dataset consists of dialogue snippets collected from the transcripts of the famous O.J. Simpson murder trial[4], we also extracted a list of all trial participants and their roles in the murder case. The collected dataset consists of 4981 dialogue snippets resulting in an objection (2153 *sustained*). We also mined the trial transcript for negative examples, collecting 6269 examples. We randomly select 20% as test data. We refer to this dataset as ALLOBJ. In addition, we created two additional dataset, consisting only of sustained/overruled objections. We denote the dataset consisting only of sustained/overruled objections as SUSTAINEDOBJ and OVERRULEDOBJ, respectively.

Overall results The most striking observation emerging from the results in Table 1 is the combined contribution of capturing relevant dialogue content and interaction (using latent variables), combined with pragmatic information. For example in the ALLOBJ, when used in conjunction, their joint contribution pushed performance to 69.1 accuracy, significantly better when each is used in isolation - 65.1 for the deterministic system using pragmatic information, and 65.4 of the latent-variable formulation which does not use this information. These results are consistent in all of our experiments. We also observe that sustained objections are easier to predict than overruled objections. This is not surprising since objections raised for unjustified reasons are harder to detect.

Pragmatic Considerations Pragmatic information (i.e., $\mathbf{x}_{Sit}$ representation) associates all decisions with a speaker identity and role. Table 1 shows that this information typically results in better quality predictions, regardless of which system was used.

An interesting side effect of using pragmatic information is its impact on the dialogue structure predictions. We quantify the effect by looking at the number of latent variables activated for each model. When pragmatic information is used, 5.6 relevance variables are used on average (per dialogue snipped). In contrast, when it is not, this number rises to 6.3[5]. In addition, the average number of sentence-connection variables active when pragmatic information is used is 3.44, this number drops to 2.53 when it is not. This suggests that dialogue pragmatics allows the model to better to take advantage of the dialogue structure , focusing the learner on higher level information,(i.e., relation between turns), and less on low level, lexical information.

5 Discussion

In this work we tackle the problem of identifying the actionable result of a dialogue. Unlike most of the relevant work in NLP, our approach requires only very lightweight annotation coming for "free" in the form of courtroom objections, and use a latent variable model to provide judgements of relevant linguistic and dialogue relations. We enhance this model using pragmatic information, capturing speakers' identity and role in the dialogue, and show empirically the relevance of this information when making predictions, and its impact on the resulting dialogue structures.

It is important to recognize that courtroom objections are not the only actionable result of dialogues. Many discussions that occur on online forums, in social media, and by email (in addition to in person) result in measurable *real-world* outcomes, which provide an easy and cheap source of supervision. Using a latent variable formulation allows to take advantage of this form of supervision, without the prohibitively high cost of learning structured dialogue models from dedicated data.

[4] http://en.wikipedia.org/wiki/O._J._Simpson_murder_case

[5] The average number of sentences per dialogue is 8.6

References

[1] Jason Baldridge and Alex Lascarides. Probabilistic head-driven parsing for discourse structure. In *CoNLL*, 2005.

[2] Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. Discriminative learning over constrained latent representations. In *NAACL*, 2010.

[3] M. Connor, C. Fisher, and D. Roth. Online latent structure training for language acquisition. In *IJCAI*, 2011.

[4] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[6] Amit Goyal, Ellen Riloff, and Hal Daumé III. Automatically producing plot unit representations for narrative text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2010.

[7] K. Hyland. Metadiscourse: Exploring interaction in writing. In *Continuum, London and New York.*, 2005.

[8] Daniel Marcu. The rhetorical parsing of natural language texts. In *ACL*, 1997.

[9] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *LREC*, 2008.

[10] Ariadna Quattoni, Sybor Wang, L-P Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2007.

[11] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*, 2013.

[12] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *NAACL*, 2003.

[13] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *COMPUTATIONAL LINGUISTICS*, 26:339–373, 2000.

[14] Oscar Täckström and Ryan T. McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *ECIR*, 2011.

[15] Rakshit Trivedi and Jacob Eisenstein. Discourse connectors for latent subjectivity in sentiment analysis. classification. In *NAACL*, 2013.

[16] Mengqiu Wang and Christopher D. Manning. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

[17] Ainur Yessenalina, Yisong Yue, and Claire Cardie. Multi-level structured models for document-level sentiment classification. In *EMNLP*, 2010.

[18] C. Yu and T. Joachims. Learning structural svms with latent variables. In *Proc. of the International Conference on Machine Learning (ICML)*, 2009.