Contextualized Representations for Low-resource Utterance Tagging

Bhargavi Paranjape
Carnegie Mellon University
bvp@cs.cmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

Utterance-level analysis of the speaker's intentions and emotions is a core task in conversational understanding. Depending on the end objective of the conversational understanding task, different categorical dialog-act or affect labels are expertly designed to cover specific aspects of the speakers' intentions or emotions respectively. Accurately annotating with these labels requires a high level of human expertise, and thus applying this process to a large conversation corpus or new domains is prohibitively expensive. The resulting paucity of data limits the use of sophisticated neural models. In this paper, we tackle these limitations by performing unsupervised training of utterance representations from a large corpus of spontaneous dialogue data. Models initialized with these representations achieve competitive performance on utterance-level dialogue-act recognition and emotion classification, especially in low-resource settings encountered when analyzing conversations in new domains.

1 Introduction

Spontaneous human conversations have been collected in different domains to support research in data-driven dialogue systems (Serban et al., 2015), affective computing (Zadeh et al., 2018; Busso et al., 2008; Park et al., 2014), clinical psychology (Althoff et al., 2016) and tutoring systems (Sinha et al., 2015). These conversations are analyzed by segmenting transcriptions into each speaker's utterances (Traum and Heeman, 1996), which are often labeled with different types of information. The exact type of label to be used depends on the downstream task or research questions to be answered, and thus the tagging paradigms are varied and numerous. For example, the speaker's intention can be specified using a dialogue acts (DAs) or speech acts (Searle and Searle, 1969), which capture the pragmatic or semantic function of the utterance.

Utterance	DA
A: Hi	Greeting
B: Hi, How are you?	Greeting
A: Are you done with your homework?	Question
B: Yeah	Yes Answer
B: How about you?	Question
A: I'm having trouble with Q4	Statement
B: Yeah	Backchannel
A: so it's going to take some time	Statement

Table 1: Snippets of conversation with dialogue act tags. "Yeah" is tagged differently in different contexts.

Utterances may also be tagged with traits such as sentiment, emotion and valence labels (Busso et al., 2008; Zadeh et al., 2018), speaker persuasiveness (Park et al., 2014), speaker dominance(Busso et al., 2008) and other characteristics at the utterance and conversational level.

While these labels vary greatly, one constant is that they are often ambiguous and context-dependent (Table 1), making it challenging for humans to annotate efficiently and accurately. Thus, curating large corpora is labor-intensive, and we are always faced with a paucity of data in new domains and labeling paradigms of interest.

Moreover, the label assigned to an utterance depends on the current state of the dialogue (Stone, 2005) and prediction of an utterance's label benefits from referring to other utterances in context and their labels (Jaiswal et al., 2019). Deep learning models like RNNs and CNNs have proven effective tools to encode neighbouring utterances (Chen et al., 2018; Liu et al., 2017; Blunsom and Kalchbrenner, 2013; Bothe et al., 2018; Kumar et al., 2017). However such models rely on large annotated corpora that are prohibitively expensive to procure, especially for niche domains.

One recently popular method to overcome the dearth of supervised data in NLP is unsupervised pretraining over large unlabeled corpora. For ex-

Proceedings of the SIGDial 2019 Conference, pages 68–74
Stockholm, Sweden, 11-13 September 2019.

ample, Melamud et al. (2016); Peters et al. (2018); Devlin et al. (2018) use language modeling as an unsupervised task to learn word embeddings in context, and demonstrate remarkable improvements on a number of downstream NLP tasks. However, these methods learn representations for individual words, whereas for dialog analysis tasks, we need representations for *utterances* in the context of the entire dialog.

In this paper, we adapt the technique of learning contextualized representations using unsupervised pretraining to learn representations for utterances in the context of the dialogue. We first introduce a general model architecture consisting of a token, utterance, and conversation encoder. We then present a method to efficiently train this model by predicting the *bag-of-word* vectors of previous and next utterances over a large heterogeneous corpus of spoken dialogue transcripts. We quantify the effectiveness of learnt contextual utterance representations on two downstream utterance-labeling tasks: DA tagging and emotion recognition. We obtain competitive performance on two popular DA tagging tasks (SwitchBoard and ICSI Meeting Recorder) and an emotion labeling task (IEMOCAP). Particularly, we observe significant improvements over training complex utterance tagging models from scratch for simulated low-resource settings for these tasks as well as for considerably smaller DA datasets such as LEGO and Map Task.

2 Methodology

We consider a large collection of conversations, where each conversation $\mathcal{C}$ is an ordered list of N utterances $\mathcal{C} = \{u_i, u_2, ..., u_N\}$ and each utterance is a list of tokens, $u_i = \{w_1, w_2, ...w_{|u_i|}\}$. Conversations may also have labels for every utterance: $Y = \{y_1, y_2, ..., y_N\}$ where each $y_i \in \mathcal{T}$, a finite set of labels expertly defined for a domain.

2.1 Unsupervised Pretraining

Contextualized Utterance Representations We adopt a hierarchical encoder model consisting of a token encoder, an utterance encoder and a conversation encoder, followed by an output layer. The token encoder layer ENC_{tok} encodes every token w_j in utterance u_i into a fixed-size embedding $e^{tok}_{w_j}$, while the utterance encoder ENC_{utt} encodes token embeddings of an utterance u_i into a fixed-sized utterance representation $e^{utt}_{u_i}$. For our specific instantiation, we combine both encoders: we use the pretrained ELMo (Peters et al., 2018) model to encode the sequence of tokens in an utterance u_i and take the final state of the forward and backward LSTMs (concatenated) as our utterance representation $e^{utt}_{u_i}, \quad i = 1, 2, ..., N$. We specifically choose ELMo because it is a strong general-purpose encoder and its character-based representations may be more robust to noise and OOV words in spontaneous conversations. This is followed by a conversation encoder ENC_{conv}, which further converts this sequence of context-independent representations of utterances to a context-dependent sequence of utterance representations. For ENC_{conv}, we use an architecture identical to the decoder variant of the Transformer (Vaswani et al., 2017) with $N = 2$ layers. We specifically choose the self-attentional Transformer for this purpose, as it is efficient to train, can easily capture long-distance dependencies over the entire conversation, and empirically outperformed other architectures such as LSTMs in preliminary experiments. The outputs, $h_{u_i}, \quad i = 1, 2, ..., N$, of this hierarchical encoder of Figure 1 can be used as contextualized representations for utterances.

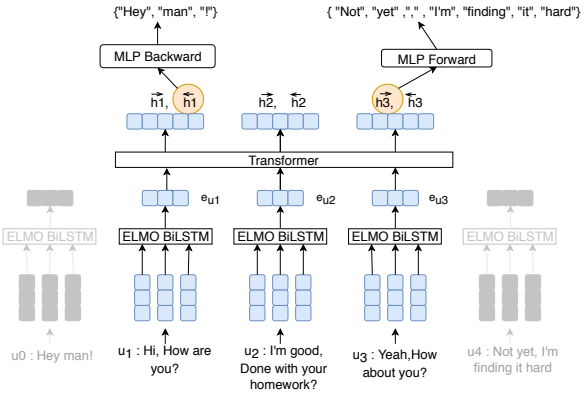

Figure 1: Hierarchical conversation encoder model

Predicting Utterance Bag-of-words In order to learn contextualized representations, the hierarchical encoder is trained to predict the bag-of-words of the previous and next utterances in the conversation using these representations. This training is done in the forward and backward direction respectively by allowing the self-attention layer of the transformer to only attend to earlier positions and later positions in the utterance sequence respectively (Figure 1). Hence, we learn *contextual utterance embeddings* in both directions: $\overleftarrow{h_{u_i}}, \overrightarrow{h_{u_i}}; \quad i = 1, 2, ..., N$. We use an MLP followed by sigmoid function as the output layer over

Corpus	# Utterances	# Tokens
SwitchBoard	460K	3M
Meeting Recorder	105K	11K
CALLHOME	27K	1M
AMI Meeting Corpus	150K	1M
BNC	1M	10M

Table 2: List of dialogue corpora for pretraining contextualized utterance representations

h_{u_i} to predict the set of words in the neighboring utterance. u_{i-1} is reconstructed from $\overleftarrow{h_{u_i}}$ and u_{i+1} from $\overrightarrow{h_{u_i}}$. We use binary cross entropy (BCE) loss, where the target is a vocabulary-sized binary vector with words present in the utterance marked 1 and others 0. Notably this formulation reduces training time by relaxing word-order in the reconstruction loss, unlike other methods that predict words in order for surrounding utterances (Kiros et al., 2015). For utterances u_{i-1} and u_{i+1} with vocabulary vectors U_{i-1} and $U_{i+1} \in \{0,1\}^{|V|}$ respectively, the *bag-of-word* loss for utterance u_i is given by:

$$\mathcal{L}_{BOW}(u_i) = BCE(\text{MLP}(\overleftarrow{h_{u_i}}), U_{i-1}) + BCE(\text{MLP}(\overrightarrow{h_{u_i}}), U_{i+1}). \quad (1)$$

where,

$$BCE(\mathbf{x}, \mathbf{y}) = \sum_{n}^{|V|} [y_n log(x_n) + (1-y_n) log(1-x_n)]$$

For conversation $\mathcal{C} = \{u_1, u_2, ..., u_N\}$,

$$\mathcal{L}_{BOW}(\mathcal{C}) = \frac{1}{N} \sum_{i=0}^{N} \mathcal{L}_{BOW}(u_i). \quad (2)$$

2.2 Utterance Tagging

Once we have learned contextualized utterance representations, we can use them to predict the sequence of labels $Y = \{y_1, y_2, ..., y_N\}$, such as dialogue acts, for utterances in the conversation. In this work we use a linear-chain conditional random field (Lafferty et al., 2001) as used in previous state-of-the-art models for DA tagging (Kumar et al., 2017; Chen et al., 2018) to predict one of the $|\mathcal{T}|$ tags for each u_i, where the utterance is represented as the concatenation of the forward and backward contextualized vectors: $\overleftarrow{h_{u_i}}, \overrightarrow{h_{u_i}}$.

3 Experiments

Pretraining Datasets and Hyperparameters We train contextualized utterance representations on transcriptions of spontaneous human-human conversation corpora (Serban et al., 2015). We choose the corpora presented in Table 2 for this work. A majority of the conversations are dialogues, and utterances across all corpora are 10 words long on average. However, the chosen corpora have conversations of widely varying lengths (no. of utterances/conversation). For computation/memory efficiency, and also because more distant utterances likely have diminishing influence on discourse modeling, we divide each conversation into conversational snippets of length 64 [1] by moving a 64-length window over the conversation with stride 1 and train the bag-of-word loss on each snippet thus obtained. For the conversational encoder, we use 2 layers of the transformer with 8 attention heads of 64 dimensions each. All feed-forward networks use 2 layers with hidden size of 512. For training and fine-tuning, we use the Adam (Kingma and Ba, 2014) with learning rate 0.0001.

Tasks We evaluate performance of our model on these utterance-level tagging tasks:

SwDA, the Switchboard Dialogue Act Corpus, annotates 1,155 telephonic conversations (224K utterances) with one of the 42 DAs in the DAMSL (Jurafsky, 1997) taxonomy.

MRDA, the ICSI Meeting Recorder Dialogue Act corpus annotates 75 multi-party meetings (105K utterances) with DAs according to 5 domain-specific tags (Dhillon et al., 2004).

IEMOCAP, an emotion recognition dataset of 12 hours of dyadic improvisations or scripted scenarios, with eight categorical emotion labels (Park et al., 2014) (10K utterances).

LEGO, a subset (14K utterances) of the Lets Go bus-information dialogue system corpus (Raux et al., 2006) annotated with the ISO 24617-2 standard for conversation functions of task by (Ribeiro et al., 2016).

Map Task, (Carletta et al., 1997; Anderson et al., 1991) is 18 hrs of dialogue where speakers collaborate to complete a map (5K utterances).

To simulate low-resource settings for the larger datasets like SWDA and MRDA, we experiment with different sizes of the training datasets and evaluate on the standard test set for these. For LEGO and MapTask, we use 10-fold cross validation.

Experimental Settings We use four different experimental settings to measure the efficacy of our

[1] tuned model hyper-parameter

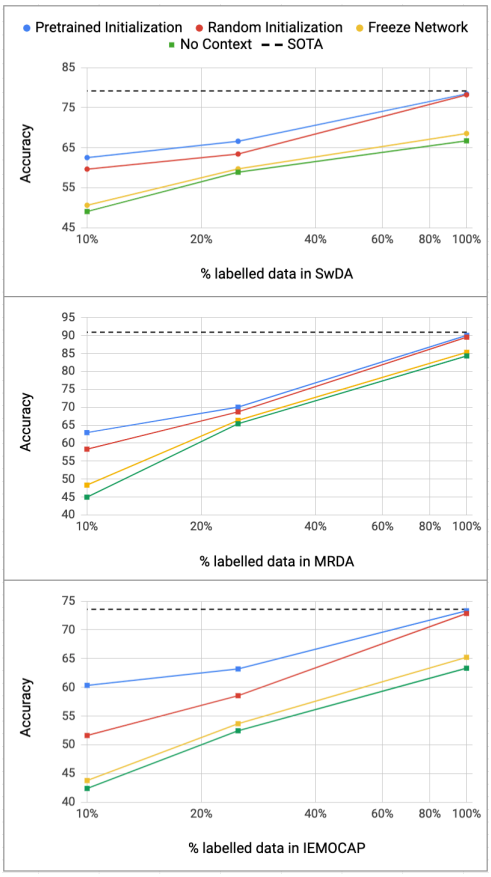

Figure 2: Performance by training data sizes. SOTA: comparable state-of-the-art model trained on tagging task for entire dataset.

pretrained utterance representations : *No Context* - With no conversational encoder (i.e. independently encoding every utterance using ELMo); *Random Initialization* - with the conversational encoder randomly initialized and trained on only the downstream tagging task; *Freeze Network* - the conversational encoder initialized using the model pretrained on our bag-of-word objective and kept fixed for downstream task; *Pre-trained Initialization* - the initialized conversation encoder fine-tuned on the downstream task. These settings are used to isolate the gains from using (1) contextualized representations, (2) pretraining them and then (3) fine-tuning them on the downstream task.

4 Result and Discussion

We observe that using pretrained utterance representations shows improved performance over random initialization and is competitive with existing state-of-the-art works by Kumar et al. (2017) for SwDA and MRDA, and Poria et al. (2017) for IEMOCAP that use similar hierarchical architectures but are only trained on the task (Random initialization setting). From Figure 2, we observe that the pretraining-based initialization is especially helpful when the amount of training data is significantly reduced for SWDA, MRDA and IEMOCAP, over other experimental settings. The improved performance of the random initialization setting over fixing the pretrained conversational encoder parameters underscores the need to fine-tune for downstream tasks. Our pretrained model also outperforms random initialization and existing best results (Ribeiro et al., 2015; Sridhar et al., 2009) for truly low-resource datasets like LEGO and Map Task, as shown in Table 4. We also analyze the gain in accuracy by dialogue act category for the pretrained model over other experimental settings. We find that the pretrained model shows improvements in the categories listed in Table 3 over random initialization. These acts typically requires models to keep track of longer contexts than other DAs like questions and back-channels. Dialogue examples in Table 5 further illustrate this.

DA Category	% Increase in accuracy	Example
Agree/Accept	43	That's exactly it.
Summarize/ Reformulate	180	Oh, you mean you switched schools..
Statement-Opinion	55	I think it's great.
Yes-Answer	33	Yes
Hold before answer or agreement	300	I'm drawing a blank

Table 3: SwDA DA categories that improve using pretrained utterance embeddings with % improvements in accuracy over other experimental settings.

DA Corpus	Pretrained	Random	SOTA
LEGO	93.70	92.98	88.75
Map Task	79.34	77.91	72.50

Table 4: Results on LEGO and Map Task

5 Conclusion

We show that using large dialogue corpora to train contextualized utterance embeddings using a bag-of-word reconstruction loss is beneficial for utterance-level tagging in the low-resource setting, indicating that these embeddings learn useful and generalizeable properties of conversational discourse. Future work involves incorporating speaker identity, utterance duration and speech/prosody features.

Utterance	Gold	Pre-trained	Random	No Context
B: where are you going to move to?	Wh-Question	Wh-Question	Wh-Question	Yes-No-Q
A: Uh, Maryland.	Statement	Statement	Statement	Hedge
B: Oh, are you?	Backchannel question	Backchannel question	Backchannel question	Yes-No-Q
A: Uh-huh.	**Yes answers**	**Yes answers**	**Yes answers**	Backchannel
B: Do you have friends there?	Yes-No-Q	Yes-No-Q	Yes-No-Q	Yes-No-Q
B: or,	**Abandoned**	**Abandoned**	**Abandoned**	Uninterpret.
A: My fiancee is down there ⟨laughter⟩.	Statement	Statement	Statement	Statement
B: Oh, I see.	Resp. Ack	Resp. Ack	Resp. Ack	Resp. Ack
B: So, does he work for a company down there?	Yes-No-Q	Yes-No-Q	Yes-No-Q	Yes-No-Q
A: Yeah,	Yes answers	Yes answers	Yes answers	Yes answers
A: he works for the government.	Statement	Statement	Statement	Statement
B: Oh, I see.	Resp. Ack	Resp. Ack	Resp. Ack	Resp. Ack
B: Oh, the big company.	**Summarize/ reformulate**	**Summarize/ reformulate**	Statement	Statement
A: Yeah	Agree/Accept	Yes Answer	Yes Answer	Yes Answer
A: and I said no, I'm just twenty-three,	Statement	Statement	Statement	Statement
B: Uh-huh.	Backchannel	Backchannel	Backchannel	Abandoned
A: you know, because I don't think of myself as needing to have children	Statement	Statement	Statement	Statement
A: but the first thing he says is, well, don't you miss that part of your life.	Statement	Statement	Statement	Statement
A: And I just,	**Abandoned**	**Abandoned**	**Abandoned**	Uninterpret.
A: my, my mind just went,	Statement	Statement	Statement	Statement
B: You didn't know what you're going to be missing.	**Collaborative Completion**	**Collaborative Completion**	Statement	Statement
A: I went, what.	Statement	Statement	Statement	Statement
B: ⟨Laughter⟩.	Non-verbal	Non-verbal	Non-verbal	Non-verbal

Table 5: Dialogue Examples from SwitchBoard with dialogue acts as labelled under different experimental settings. The pre-trained network performs better on categories like Summarizing and Collaborative Completion

Acknowledgments

We thank the reviewers for their insightful comments. This work was supported by the National Institute of Health (NIH) grant no. R01MH096951-08.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Phil Blunsom and Nal Kalchbrenner. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*. Proceedings of the 2013 Workshop on Continuous Vector Space Models and their .

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. *arXiv preprint arXiv:1805.06280*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C Kowtko, Gwyneth Doherty-Sneddon, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 225–234. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide.

Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. *arXiv preprint arXiv:1903.11672*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. *arXiv preprint arXiv:1709.04250*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57. ACM Press.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Ninth International Conference on Spoken Language Processing*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2015. The influence of context on dialog act recognition. *arXiv preprint arXiv:1506.00839*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2016. Mapping the dialog act annotations of the lego corpus into the communicative functions of iso 24617-2. *arXiv preprint arXiv:1612.01404*.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Tanmay Sinha, Ran Zhao, and Justine Cassell. 2015. Exploring socio-cognitive effects of conversational strategy congruence in peer tutoring. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 5–12. ACM.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422.

Matthew Stone. 2005. Communicative intentions and conversational processes in humanhuman and human-computer dialogue. *Approaches to studying world-situated language use*, pages 39–70.

David R Traum and Peter A Heeman. 1996. Utterance units in spoken dialogue. In *Workshop on Dialogue Processing in Spoken Language Systems*, pages 125–140. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246.