CONTEXTUAL TOPIC MODELING FOR DIALOG SYSTEMS

Chandra Khatri, Rahul Goel*, Behnam Hedayatnia**
Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, Arindam Mandal

ckhatri, goerahul, behnam, ametalli, anuvenk, raeferg, arindamm@amazon.com

ABSTRACT

Accurate prediction of conversation topics can be a valuable signal for creating coherent and engaging dialog systems. In this work, we focus on context-aware topic classification methods for identifying topics in free-form human-chatbot dialogs. We extend previous work on neural topic classification and unsupervised topic keyword detection by incorporating conversational context and dialog act features. On annotated data, we show that incorporating context and dialog acts leads to relative gains in topic classification accuracy by 35% and on unsupervised keyword detection recall by 11% for conversational interactions where topics frequently span multiple utterances. We show that topical metrics such as Topical Depth is highly correlated with dialog evaluation metrics such as Coherence and Engagement implying that conversational topic models can predict user satisfaction. Our work for detecting conversation topics and keywords can be used to guide chatbots towards coherent dialog.

Index Terms— dialog systems, dialog evaluation

1. INTRODUCTION

Most successful conversational interfaces like Amazon Alexa, Apple Siri, and Google assistant have been primarily designed for short task-oriented dialogs. Task-oriented dialogs follow template-like structures and have clear success criteria. Developing conversational agents that can hold long and natural free-form interactions continues to be a challenging problem [1]. Sustaining long free-form conversation opens the way for creating conversational agents or chatbots that feel natural, enjoyable, and human-like.

Elements of a good free-from conversation are hard to objectively define. However, over the years, there have been various attempts at defining frameworks on how free-form conversations can be rated [2, 3, 4, 5, 1]. Accurate tracking of the conversation topic can be a valuable signal for a system for dialog generation [6] and evaluation [1]. Previously, [7] have used topic models for evaluating open domain conversational chatbots showing that lengthy and on-topic conversations are a good proxy for assessing the users' satisfaction with the conversation and hence in this work we focus on improving supervised conversational topic models. The topic models proposed by [7] are non-contextual, which prevents them from using past utterances for more accurate predictions. This work augments the supervised topic models by incorporating features that capture conversational context. We apply new context-sensitive models to topic classification of free-form conversations. We also train an additional BiLSTM topic model with and without context as an another comparison point. The models were trained and evaluated on real user-chatbot interaction data collected during a large-scale chatbot competition. With the goal of improving the topic model in mind, we train a separate independent model to predict dialog acts [8] in a conversation and observe that incorporating dialog act as an additional feature improves topic model accuracy. We evaluate three flavors of models: (1) optimized for speed (deep average networks [9]), (2) optimized for accuracy (BiLSTMs), and (3) an interpretable model using unsupervised keyword detection (attentional deep average networks (ADAN)). We also evaluate the keywords produced by the ADAN model qualitatively by curating a manually-labeled dataset with keywords showing that incorporating context increases the recall of keyword detection. For this work, we annotated more than 100K utterances collected during the competition with topics, dialog acts, and topical keywords. We also annotated a similarly-sized corpus of chatbot responses for coherence and engagement. Using this data, we show that incorporating context and additional signals like predicted dialog act in topic models makes them robust to noise and reduces topic classification error. To the best of our knowledge, this is the first work that uses contextual topic models for open domain conversational agents. Our main contributions are: **1)** We annotated and analyzed conversational dialog data with topics, dialog acts, as well as conversational metrics like coherence and engagement. We show high correlation between topical depth and user satisfaction score. **2)** We show that including context and dialog acts in conversational topic models leads to improved accuracy. **3)** We provide quantitative analysis of keywords produced by the Attentional Topic models.

Equal Contribution

2. RELATED WORK

The work by [10] is an early example of topic modeling for dialogs. They define topic trees for conversational robustness. [11] propose using conversational topic models to model first encounter dialogs while [12] use Latent Dirichlet Allocation (LDA) to detect topics in conversational dialog systems. Topic tracking and detection for documents has been a long on-going research area [13]. [14] provide a good overview of classical approaches. Topic models such as pLSA [15] and LDA [16] provide a powerful framework for extracting latent topics in text. However, researchers have found that these unsupervised models may not produce topics that conform to users' existing knowledge [17] as the objective functions of these models often does not correlate well with human judgements [18]. This often results in nonsensical topics which cannot be used in downstream applications. There have been work on supervised topic models [19] as well as making them more coherent [17]. Human conversation containing speech act or dialog acts is a classic idea [8]. Over the years, there are extensive literatures on both supervised and unsupervised ways to classify dialog acts [20]. In this work, we perform supervised dialog act classification, which we use along with context as additional features for the topic classification.

The two tasks we have at hand are predicting dialog acts and topics. Similar to multi-task learning [21, 22], we want to exploit the correlations across these related tasks so in this work, we use one model to predict dialog acts and feed the predictions as an additional signal to the topic model.

A major hurdle for open-domain dialog systems is their evaluation [23] as there are *many* valid responses for any given situation. There has been a lot of recent work towards building better dialog evaluation systems [24, 25]. Some work include learning models for good dialog [5, 4], adversarial evaluation [26], and using crowd sourcing [2]. Inspired by [7] who use sentence topics as a proxy for dialog evaluation, we support claims about topical depth being predictive of user satisfaction and extend their models to incorporate context.

3. DATA

The data used in this study was collected during a large chatbot competition. Academic research teams participated and produced large-scale live conversational data from real users. Upon initiating the conversation users were paired with a randomly selected chatbot made by the competition participants. At the end of the conversation, the users were prompted to rate the chatbot quality and had the option to provide feedback to the teams that built the chatbot. We had access to over 100k such utterances containing interactions between user and chatbot collected during the 2017 competition which we annotated for topics, dialog acts, and keywords (using all available context).

3.1. Annotation

Upon reviewing a user-chat interaction, annotators:

- Classify the topic for each user utterance and chatbot response, using one of the available categories. Topics are organized into 12 distinct categories as shown in Table 2. It included the category **Other** for utterances or chatbot responses that either referenced multiple categories or do not fall into any of the available categories.

- Determine the goal of user or chatbot, which are categorized as 14 dialog acts in Table 2. Inspried by [8] we created a simplified set of dialog acts which were easy to understand and annotate. It includes the category **Other** for utterances which do not fall into any of the available categories and **Not Set**, which means annotators did not annotate the utterance because of potential speech recognition error or the utterance was too short or noisy. The goals are context-dependent, and therefore, the same utterance/Bot response can be evaluated in different ways, depending on the available context.

- Label the keywords that assist the analyst in determining the topic, e.g., in the sentence "the actor in that show is great," both the word "actor" and "show" will assist the analyst when classifying the topic for this utterance.

These topics and dialog acts were selected based on the overall parameters of the competition as well as observed patterns of user interactions. As topics of conversation are strongly related to human perceptions of coherent conversation, we hypothesize that these annotations would allow improved dialog modeling and more nuanced responses. We provide some example of less obvious topics like **Interactive** and **Phatic** in Table 4

In some cases, a user utterance or the chatbot response will imply multiple dialog acts. In all cases where a user utterance or chatbot response contains more than one dialog act, we default to **MultipleGoals**. If the request for more than one dialog act is within a topic, we categorize within the topic. Examples of annotations are shown in Table 3. The distribution of topics is shown in Figure 1 and for distribution of dialog acts is shown in Figure 2. Our inter-annotator agreement on the topic annotation task is 89% and is 70% on the dialog act annotation task. The Kappa measure on the topic annotations is 0.67(Good) and is 0.41 on dialog act annotation(Moderate). Our reference for quality is from [27].

In addition to these annotations, we also asked a separate set of annotators to rate each chatbot response as "yes" or "no" on these four questions: 1) The response is comprehensible: The information provided by the chatbot made sense with respect to the user utterance. 2) The response is relevant: If a user asks about a baseball player on the Boston Red Sox, then the chatbot should at least mention something about that baseball team. 3) The response is interesting: The chatbot

Turn	Agent	Sentence	Topic	DialogAct	Keywords
1	User	let's talk about politics	Politics	Information Request	Politics (KeywordPolitics)
1	Chatbot	ok sounds good would you like to talk about Gucci?	Fashion	TopicSwitch	Gucci (KeywordFashion)
2	User	Yes	Fashion	InformationRequest	
2	Chatbot	Sure! Gucci is a famous brand from Italy	Fashion		Gucci brand Italy (KeywordFashion)

Table 1: Typical Annotation workflow, here the keywords are marked with the topics which they helped identify

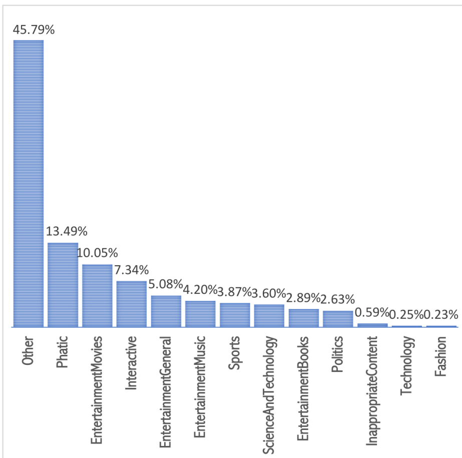

Fig. 1: Topic Distribution

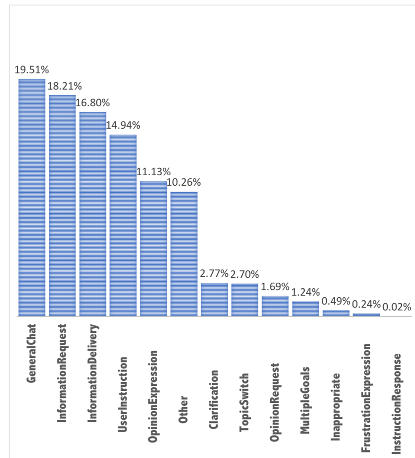

Fig. 2: Dialog Act Distribution

Topics	**Dialog Acts**
Politics	InformationRequest
Fashion	InformationDelivery
Sports	OpinionRequest
ScienceAndTechnology	OpinionExpression
EntertainmentMusic	GeneralChat
EntertainmentMovies	Clarification
EntertainmentBooks	TopicSwitch
EntertainmentGeneral	UserInstruction
Phatic	InstructionResponse
Interactive	Interactive
Other	Other
Inappropriate Content	FrustrationExpression
-	MultipleGoals
-	NotSet

Table 2: Topic and Dialog Act labels

Agent	Sentence	Dialog Act
User	what are your thoughts on yankees	Opinion Request
Chatbot	I think the new york yankees are great. Would you like to know about sports	MultipleGoals (OpinionExpression and request)
User	Yes	OpinionExpression

Table 3: Contextual Annotations and Multiple goals

would provide an answer about a baseball player and provide some additional information to create a fleshed out answer. 4)

Sentence	Topic	DialogAct
huh so far i am getting ready to go	Phatic	InformationDelivery
let's chat	Phatic	GeneralChat
who asked you to tell me anything	Other	FrustrationExpression
can we play a game	Interactive	UserInstruction

Table 4: Example user utterances with their topics and dialog acts

I want to continue the conversation: This could be through a question back to the user for more information about the subject. We use the sum of the first two metrics as a proxy for coherence, and the sum of last two as a proxy for engagement. We consider "yes" as a 1 and "no" as a 0 to convert these to scores.

3.2. Topical Metrics and Evaluation Metrics

Coherence and Engagement are thus measured on the scale of 0 to 2. They are defined in terms of the questions asked to the annotators, as follows:

- Coherence = Response is comprehensible + Response is relevant
- Engagement = Response is interesting + I want to continue this conversation

Metric	Value
Average Conversation Length	11.7
Median Conversation Length	10.5
Mean User Utterance Length	4.2
Median User Utterance Length	3
Mean Chatbot Utterance Length	24
Median Chatbot Utterance Length	17
User Vocab Size (words)	18k
ChatBot Vocab Size (words)	85k

Table 5: Some Data Statistics. Average Conversation Length is measured in turns (1 user and chatbot utterance). Utterance lengths defined as number of words in the sentence.

Metric	Correlation
Coherence	0.80
Engagement	0.77

Table 6: Pearson correlation of Topical Depth with Coherence and Engagement, p < 0.0001 for all rows

We annotated each chatbot response for Coherence and Engagement. We provide more statistics about our data in Table 5. We observe that user responses are very short and have limited vocabulary compared to chatbot responses which makes this task challenging as well as context crucial for effective topic identification.

Following [7], we define the following terms:

- **Topic-specific turn $\mathcal{T}$:** Defined as a pair of user utterance and chatbot's response where both utterances belong to the same topic. For example in Table 1 Turn 2 forms topic specific turn for fashion
- **Length of sub-conversation l_s:** Defined as the number of topic-specific turns it contains. In Table 1 there is a sub-conversation of length 1 for fashion

The correlation of topical depth of a conversation with Coherence, and Engagement is given in Table 6. We observe that the mean of Coherence given by the annotators,1.21 with a standard deviation of 0.75, is much higher than the mean of Engagement, 0.81 with a standard deviation of 0.62. Both of them are almost equally correlated with Topical Depth, which implies that by making conversational chatbots stay on topic there is room for improvement in user engagement. Hence, observing Table 6 we claim that the topical metrics as proposed by [7] can be used for automatically evaluating the ability of chatbots to engage users with a coherent conversations.

4. MODELS

We use DANs and ADANs as our topic classification baseline and explore various features and architectures to incorporate

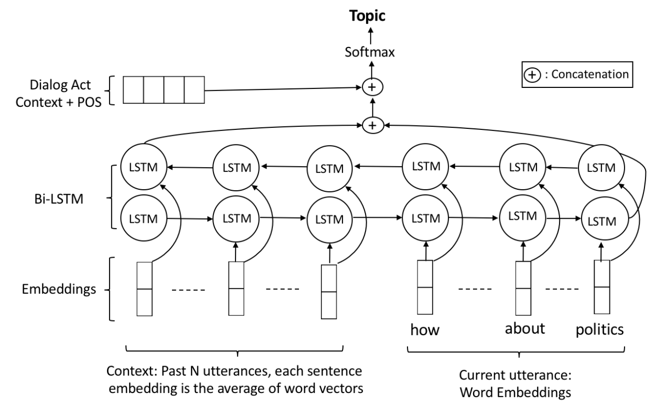

Fig. 3: Contextual Bi-LSTM for Classification

context. We also train BiLSTM classification models with and without context. We first describe the DAN, ADAN, and BiLSTM models. We then describe additional features for the models to incorporate context and dialog act information. All of our models are trained with 50% dropout for regularization.

4.1. DAN

DAN is a bag-of-words neural model (see Figure 4 for DAN with context) that averages the word embeddings in each input utterance as a vector representation s of a sentence. s is passed through a series of fully connected layers and fed into a softmax layer for classification along with the context. Assume an input utterance of length L and corresponding D dimensional word embeddings $e_i, i = 1, \cdots L$, then the utterance representation is: $s = \frac{1}{L}\sum_{i=1}^{L} e_i$. Figure 4 excluding the context layers illustrates the DAN model for topic classification. In Figure 4 we have output layer of size K which corresponds to the number of topics we want to classify between. The parameter H_1 is the hidden layer dimension and is a hyperparamter. Due to lack of recurrent connections and its simplicity, DAN provides a fast-to-train baseline for our system. The word embeddings are initialized with Glove [28] and fine-tuned.

4.2. ADAN

[7] extended the DAN model by attention mechanism to jointly learn the topic-wise word saliency for each input utterance while performing the classification. The keywords can thus be extracted in an unsupervised manner. Figure 5 depicts ADAN with context (CADAN). Figure 5 excluding the context layers illustrates the ADAN model for topic classification. As shown in the figure, ADAN models the distribution using a topic-word attention table of size $K \times |V|$, where $|V|$ is the vocabulary size and K the number of topics in the classification task. The attention table corresponds to the relevance-weights associated with each word-topic pair. The weights $w_{k,i}$ essentially measure the saliency of word x_i given topic k_i. The utterance representation $\boldsymbol{s}_k$ per topic is computed through weighted average:

$$[\alpha_{k,1}, \cdots, \alpha_{k,L}] = \mathrm{softmax}([w_{k,1}, \cdots, w_{k,L}]) \quad (1)$$

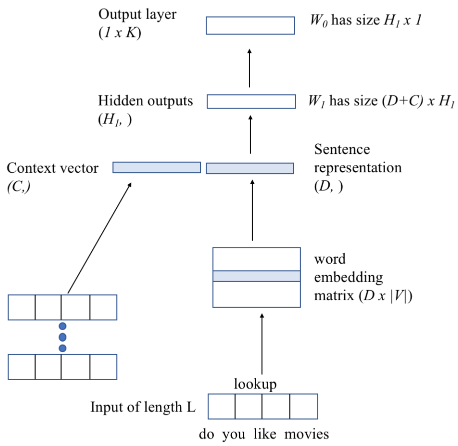

Fig. 4: The structure of CDAN, note that context is averaged and appended to the average sentence vector.

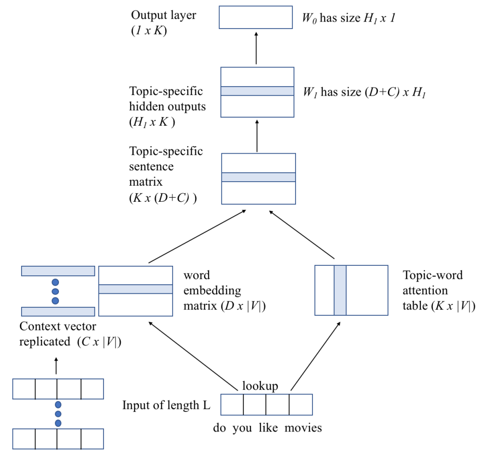

Fig. 5: The structure of CADAN, note that the same context is added to each word

$$\boldsymbol{s}_k = \frac{1}{L}\sum_{i=1}^{L} \alpha_{k,i}\boldsymbol{e}_i \quad (2)$$

ADAN provides an interpretable baseline for our system.

4.3. Bi-LSTM for Classification

We train a simple 1 layer BiLSTM model with word embeddings as an input for topic classification. The word embeddings are randomly initialized and are learnt during the task. For the final sentence representation, we use the concatenation of the final state of the forward and backward LSTM [29]. The contextual variation of the BiLSTM model is shown in Figure 3. The context is concatenated at each time step and is fed as an input to the BiLSTM.

4.4. Context as Input

We consider two variants of contextual features to augment the above-mentioned models for more accurate topic classification. We define a turn vector as concatenation of User utterance and Chatbot's response. In the conversation in Table 1, there are two turn vectors: [User: let's talk about politics; Chatbot: ok sounds good would you like to talk about Gucci?], [User: Yes Sure! Chatbot: Gucci is a famous brand from Italy]. We get a fixed-length representation of a turn vector by simply averaging the word embeddings of all the words in them. For current turn T_i of length L with words $s_i, i \in 1...L$, let S_i be the word embedding vector corresponding to the word s_i. Then the turn vector is $T_i = \frac{1}{L}\sum_{i=L}^{I} S_i$.

- Average turn vector as context: For the current turn I, the context vector is obtained by averaging the previous N turn vectors:
- Dialog Act as context: Dialog Act could serve as a powerful feature for determining the context. We train a separate DAN that predicts Dialog Acts given current and past utterances, which we use as an additional input to our models

We concatenate the feature vectors together and append them to each word vector in DANs and ADANs as an additional input to the model.

For BiLSTMs, we try two different ways to include the context of average word vector (1) Concatenating the context to the input of BiLSTM with word embeddings and (2) Adding context in sequential manner as an extension to the input sequence rather than concatenating the averaging embeddings. We append the dialog act to the output of BiLSTM before softmax. Additionally we extract Part of Speech(POS) features from current utterance using the SpaCY POS tagger [30]. We append the POS features to the output of BiLSTM before the softmax layer. We experiment with a combination of these features, the results are shown in Section 5.

4.5. Salient Keywords

We use our annotated keywords as a test set to evaluate our attention layer quantitatively. To choose the keywords produced by our attention layer, we first choose the topic as the topic produced by the attention model. From the attention table select the row w_k corresponding to the topic and then select top k keywords where k is equal to keywords in ground truth. We do this only for evaluation of the keywords produced by the ADAN model. For ground truth keywords, we use all the tokens marked with any topic by the annotators in a sentence.

5. EXPERIMENTS

Since our data set was highly imbalanced (see Figure 1), we down-sampled the **Other** class in our data set. We split our annotated data into 80 % train, 10 % development, and 10 % test sets. We used the dev set to roughly tune our hyper-parameters separately for all our experiments. We train our networks to convergence, which we define as validation accuracy not increasing for two epochs. Having multiple layers did not significantly increase accuracy on development set so we kept only 1 hidden layer for DAN, ADANs, and BiLSTMs. We

Sentence with ground truth keywords	ADAN	+C	+C+D
do you know hal | Sci&Tech	hal | Sci&Tech	hal | Sci&Tech	hal | Sci&Tech
i like puzzles | EntGeneral	i | Sci&Tech	puzzles|EntGeneral	puzzles | EntGeneral
can you make comments about music | EntMusic	comments | Phatic	you | EntBooks	music | EntMusic

Table 7: Keywords produced by models. C refers to context and D refers to dialog act. Only the words which are annotated are marked with | followed by the topic which they refer to. Labels are shortened to fit in the table.

choose a embedding size of 300 and the hidden layer size of 500 for both DAN and ADAN. For BiLSTM, we use 256 as the hidden layer size. We use ADAM as our optimizer with learning rate of 0.001 and other defaults. We noticed only marginal gains by including very long context windows, hence to speed up training, we only consider last 5 turns as context in our model. We measure our accuracy on the supervised topic classification task. Our results are shown in Table 8. For the salient key words detection task, we measure token level accuracy and the results are shown in Table 10.

Classifier	Baseline	$+C$	$+D$	$+C+D$
LSTM-Avg	0.55	0.56	0.59	0.68
LSTM-Seq	-	0.61	-	0.74
DAN	0.51	0.57	0.52	0.60
ADAN	0.38	0.39	0.42	0.40

Table 8: Classification accuracy for Bi-LSTM, DAN, and ADAN models using only current utterance (baseline) as well as adding context and dialog act features. ADAN+Context corresponds to the CADAN model. C refers to context and D refers to dialog act, LSTM-Avg is the model where average context is appended to word embeddings. LSTM-Seq refers to the model where context is fed sequentially to LSTM.

Classifier	Sentences	+POS	+C
DAN	0.50	0.61	0.69
Bi-LSTM	0.50	0.58	0.67

Table 9: Dialog Act classification accuracy. Since DAN with context gives us the best accuracy we use that model as an input to our topic classification models

Classifier	Precision	Recall
ADAN	0.37	0.36
CADAN	0.33	0.32
CADAN + DialogAct	0.40	0.40

Table 10: Keyword detection Metrics. CADAN corresponds to the ADAN model with context input features.

6. RESULTS AND ANALYSES

We show our main results in Table 8, we highlight a few key results. The BiLSTM baseline performs better than the DAN baseline. ADAN performs worse than DAN and BiLSTM across the board. We believe that this is because of the large number of parameters in the work-topic attention table of ADAN, which requires a large amount of data for robust training. Given our dataset of 100K utterances, the ADAN model may be overfitting. Adding contextual signals like past utterances and dialog acts help DAN and BiLSTM which can be seen in Table 8 by comparing the baseline with other models.

Adding context alone helps DAN but does not significantly improve BiLSTM performance. This could be because of the BiLSTM already modeling the sequential dependencies where the context alone does not add a lot of value. We observe that past context and dialog acts work complementarily where adding context makes the model sensitive to topic changes while adding in the dialog acts and make the model more robust to contextual noise.

The best performing model is a combination of all the input signals: Context and Dialog Acts. However, in the case of ADAN where the model is over-fitting because of insufficient data, adding both features degrades the classification performance further but improves the keyword detection metrics (Table 10). A few examples of keywords learned by the ADAN model are shown in Table 7.

7. CONCLUSION AND FUTURE WORK

We focus on context-aware topic classification models for detecting topics in non-goal-oriented human-chatbot dialogs. We extended previous work on topic modeling (DAN and ADAN models) by incorporating conversational context features to topic modeling and we introduce the Contextual DAN (CADAN) and Contextual BiLSTM models. We describe a fast topic model (DAN), an accurate topic model (BiLSTM), and an interpretable topic model (ADAN). We show that we can add conversational context in all of these models in a simple and extensible fashion. We also show that Dialog Act provides valuable input which helps improve the topic model accuracy. Furthermore, we depict that the topical evaluation metrics such as Topical Depth highly correlates with dialog evaluation metrics such as Coherence and Engagement, which implies that conversational topic models can play a critical role in building great conversational agents. Furthermore, topical evaluation metrics such as Topical Depth can be obtained automatically and can be used for evaluation of open-domain conversational agents, which has largely remained an unsolved problem. Unsupervised topic modeling is another future direction that we plan to explore. In the future, we plan to include more context like device state and user preferences in such models.

8. REFERENCES

[1] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, et al., "On evaluating and comparing conversational agents," *arXiv preprint arXiv:1801.03625*, 2018.

[2] Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel, "Can machine translation systems be evaluated by the crowd alone," *Natural Language Engineering*, vol. 23, no. 1, pp. 3–30, 2017.

[3] Dominic Espinosa, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant, "Further meta-evaluation of broad-coverage surface realization," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 564–574.

[4] Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo, "Evaluating coherence in open domain conversational systems," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," *arXiv preprint arXiv:1708.07149*, 2017.

[6] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[7] Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram, "Topic-based evaluation for conversational bots," *arXiv preprint arXiv:1801.03622*, 2017.

[8] Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al., "Dialog act modeling for conversational speech," in *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998, pp. 98–105.

[9] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III, "Deep unordered composition rivals syntactic methods for text classification," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2015.

[10] Kristiina Jokinen, Hideki Tanaka, and Akio Yokoo, "Context management with topics for spoken dialogue systems," in *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1998, pp. 631–637.

[11] Trung Ngo Trong and Kristiina Jokinen, "Conversational topic modelling in first encounter dialogues," .

[12] Jui-Feng Yeh, Chen-Hsien Lee, Yi-Shiuan Tan, and Liang-Chih Yu, "Topic model allocation of conversational dialogue records by latent dirichlet allocation," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–4.

[13] James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, "Topic detection and tracking pilot study final report," 1998.

[14] James Allan, "Introduction to topic detection and tracking," in *Topic detection and tracking*, pp. 1–16. Springer, 2002.

[15] Thomas Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[17] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[18] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[19] Jon D Mcauliffe and David M Blei, "Supervised topic models," in *Advances in neural information processing systems*, 2008, pp. 121–128.

[20] Aysu Ezen-Can and Kristy Elizabeth Boyer, "Understanding student language: An unsupervised dialogue act classification approach," *Journal of Educational Data Mining (JEDM)*, vol. 7, no. 1, pp. 51–78, 2015.

[21] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning.," in *Proc. of International Conference of Machine Learning (ICML) 2008*, 2008.

[22] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) 2017*, 2017.

[23] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[24] Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević, "Results of the wmt16 metrics shared task," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016, vol. 2, pp. 199–231.

[25] Rohit Gupta, Constantin Orasan, and Josef van Genabith, "Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks.," 2015.

[26] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky, "Adversarial learning for neural dialogue generation," *CoRR*, vol. abs/1701.06547, 2017.

[27] Douglas G. Altman, *Practical Statistics for Medical Research*, CRC Press, 1990.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.," in *Proceedings of EMNLP*, 2014.

[29] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[30] SpaCy, "Spacy," `https://spacy.io/`, 2018, [Accessed: 2018-07-01].