Dialogue Act Modeling in Information-seeking Conversations

Yue Yu
Georgetown University
Washington, D.C.
yy476@georgetown.edu

Siyao Peng
Georgetown University
Washington, D.C.
sp1184@georgetown.edu

Abstract

We describe a LSTM-based approach for modeling dialogue acts in conversations with complex dialogue structure and utterance context. Our model detects and predicts dialogue acts based on lexical context and structure knowledge automatically learned from data. Models are trained and evaluated using the MSDialog corpus with 2,199 multi-turn QA dialogues annotated with multiple dialogue acts on the utterance level. The model significantly outperforms the baseline model but still requires further tuning to better perform on non-dominant tags.

1 Introduction

Conversational assistants (CAs) such as Siri, Alexa and Cortana are now very popular. With CAs, users can issue simple queries and commands to conduct single-turn QA or goal-oriented tasks, such as asking for weather and setting timers. However, the current CAs are not yet capable of handling complex information requests which involve multiple turns of information exchange (Joho et al., 2018). Recognizing the intent of a user, i.e. dialogue act (DA), and meaningful structures of dialogue acts in these kinds of dialogue systems can be very useful and essential.

These conversations enforce multi-party scenarios and information-rich dialogues. To the best of our knowledge, there is a few prior research on dialogue act focusing on this type of conversations. In this paper, we applied Markov model to analyze dialogue act dynamics in information-seeking conversations and proposed a novel approach for context-aware DA modeling using deep neural models. We show that the proposed DAMIC model achieves the highest F1 score on collapsed utterances and significantly outperforms the baseline model. Meanwhile, the model requires further tuning to improve the performance on non-dominant tags.

To further this project, we plan to experiment with DA-driven response selection based on state-of-the-art deep neural matching networks (Yang et al., 2018).

2 Related Work

Much research has been done on dialogue act (DA) modelling and prediction for conversation models. There are three mainstreams for DA modeling: (1) DA models in literature frequently use a hidden Markov model (HMM) to structure a generative process of DA sequences (Stolcke et al., 2000; Ritter et al., 2010; Jo et al., 2017). This approaches works pretty well when we can safely assumed that each hidden state corresponds to a single DA. (2) DA modeling is also often cast as classification task (Oraby et al., 2017). Although dialogue turns annotated with multiple DAs will no longer be a problem, this approach cannot capture long term dependency between DAs across turns. (3) Encoder-decoder model is another popular approach for dialogue system (Mathur and Singh, 2018). Although many proposed neural approaches have incorporated DA as a feature for conversation modeling, the DA sequences in these models are often learned separately (Kumar et al., 2018) rather than learned together with the encoder-decoder model. As a result, these models require considerable amount of dialogue utterances annotated with the corresponding dialogue acts.

Therefore, contemporary DA models are most frequently applied to a very restricted set of conversations, most frequently customer-

Items	Min	Max	Mean	Median
# Turns Per Dialog	3	10	4.56	4
# Participants Per Dialog	2	4	2.79	3
Dialog Length (Words)	27	1,467	296.90	241
Utterance Length (Words)	1	939	65.16	47

Figure 1: Statistics of MSDialog dataset (figure from (Qu et al., 2018)

agent conversations, which usually have short utterances and each utterance with a single tag. In contrast, less attention has been paid to other forms of conversations with much more complex dialogue structure and content, for example question answering (QA) interactions on online forums. These conversations enforce less constraints on dialogue structure and content. Multi-party scenarios and information-rich utterances become possible. To the best of our knowledge, there is no prior research on dialogue act modeling focusing on this type of conversation. By adding DA annotations, we can now analyze these forum conversations to learn complex DA structures across dialogue turns and ultimately benefit the conversation modeling task.

3 Corpus

3.1 Design of Resource Corpus

In our project, we work with a recently released multi-turn information-seeking forum corpus, the MSDialog-Intent corpus[1] (Qu et al., 2018). It is a labeled dialog corpus of question answering (QA) interactions between information seekers and providers from an online forum on Microsoft products. The corpus contains more than 2,199 multi-turn QA dialogues with 10,020 utterances that are annotated with DAs on the utterance level. Basic statistics of the dataset are shown in Figure 1. Besides the utterance texts and DA tags, each utterance has also been annotated with several data fields including *actor_type*, *user_id*, etc. Table 3 (in Appendix A shows an example dialogue from the corpus. This corpus makes it possible to perform an in-depth analysis over dialogue acts in multi-turn human QA conversations.

The annotation scheme for MS corpus consists of 12 DA tags for different intent purposes, including greeting, questioning, answering, providing feedbacks, following-ups, requesting, etc[2]. As discussed in Qu et al. (2018), GG (Greetings/Gratitudes) and PA (Potential Answer) are the most frequent tags in this corpus, each accounts for 22% of the distribution (shown in Figure 2).

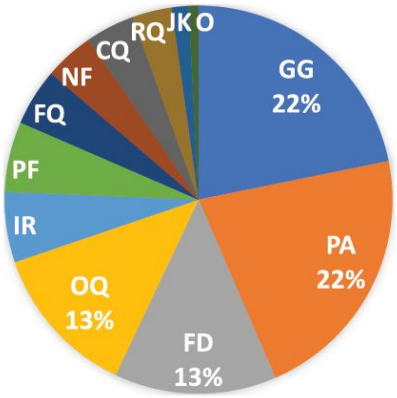


Figure 2: Distribution of DA tags in MSDialog-Intent Corpus (figure from Qu et al. (2018)

Another interesting design of the annotation scheme is that each utterance is annotated with all applicable DA tags without giving the segment boundaries. Due to the polite and friendly atmosphere of MS tech forum, many technically informative utterances are also labeled GG in addition to their more crucial DAs, e.g. PF (Positive Feedback).

3.2 Corpus analysis

We start the project by exploring the Markov chain of the flow pattern for DAs as in Qu et al. (2018). The Markov model starts with an INITIAL node and ends with a TERMINAL node. It includes all 12 DA tags but excluding co-occurring cases of GG. Specifically, only in this task[3], we follow Qu et al. (2018)'s method to ignore all GG tags when they co-occur with other DA tags. However, we leave GG tags that occur alone (i.e. when it is the only DA tag for an utterance) unchanged (instead of converting to JK (Junk)), to avoid the assimilation of GG and JK. Consequently, when constructing the Markov Chain, we attempted to prevent over-dominance of GG. In fact, it turns out that 90%

[1] https://ciir.cs.umass.edu/downloads/msdialog/

[2] See Appendix 8 for descriptions and examples of the 12 DAs

[3] We keep all DA annotations of GG unchanged in our following experiments since we are able to handle multi-tagged DA modeling.

of GG tags co-occur with other more informative DA tags, whereas only 10% occur by itself.

Moreover, since each utterance can be annotated with multiple DAs, we follow Qu et al. (2018)'s method which splits each multi-tagged utterance into separate paths, resulting in 12 different paths in Figure 3 throughout the conversation shown in Table 3 (in Appendix).

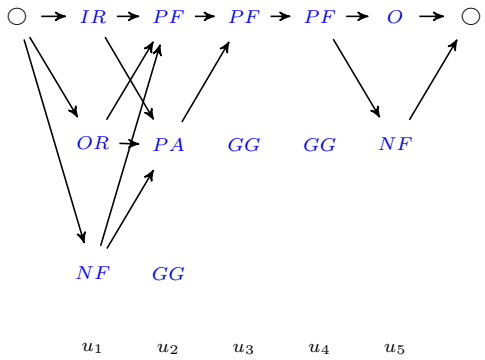


Figure 3: Example of Multi-tagged DA Flows following Qu et al. (2018)'s method

Based on this path-splitting design, we implemented a Markov model for the 10,020 utterances. Different from Qu et al. (2018), we did not exclude conversations with more than 100 paths. Because conversations differ in length and depth, there is no practical motivation to remove conversations with abundant DA tags, especially since utterances are quite long in MSDialog corpus and could include a variety of user intents.[4]

Echoing the resource paper, we observe that besides the TERMINAL node, PA (Potential Answer) and FD (Further Details) are the most frequent receiving nodes and the correlation between PA and FD is quite strong[5]. This arises a potential problem if we model DA tag prediction only using Markov model, the top transition probabilities PA $\rightarrow$ FD and FD $\rightarrow$ PA would result in an infinite loop between the two DA tags. As a result, the source paper only describes a general pattern of dialogue flow but does not use the Markov model for DA prediction tasks.

To take utterance contexts into consideration and to enable predicting multiple dialogue act labels for each utterance, we will turn to our LSTM-based models in the following sections.

3.3 Preprocessing

The MSDialog dataset has been preprocessed to better fit the LSTM model. Most crucially, due to the unusual length and complexity of these forum utterances, we cleaned miscellaneous non-words to reduce sparseness of the dataset. In addition, we extracted Bag-of-Words (BoW) information from this corpus to training our model, joined with utterance texts and DA tags. Preprocessing includes the following steps:

1. The original *MSdialog-Intent.json* is transformed into a tabulated csv file mirroring the design of Switchboard Dialog Act (SwDA) Corpus (Godfrey et al., 1992).[6] This step not only helps us better manipulate with the dataset using Python dataframes but also creates future potentials to extend our model to other similarly formatted datasets.

2. Email addresses, weblinks, numbers, etc. are replaced by their super-tags, EMAIL, URL, NUM respectively.

3. Named Entities are replaced by their corresponding types, using super-tags ORGANIZATION, PERSON, PLACE, etc using StanfordNERtagger (Finkel et al., 2005)[7] provided in nltk (Bird et al., 2009)[8].

4. Remaining words are stemmed (using Porter Stemmer (Porter, 1980))[9] and normalized to decrease variations.

In the meanwhile, we created a Bag-of-Words (BoW) model (Harris, 1954) consisting of 3311 stemmed and super-tagged word. We did not exclude stop-words since they could be an indicator for different dialogue intents. We believe the traditional BoW model is a better fit for our task compared to word embeddings since in these extraordinarily long utter-

[4] We include in Appendix Figure 9 which shows the reproduced Markov model with transition probabilities labeled on each edge; edges whose probability are less than 10% are filtered.

[5] Table 2 also shows this result in Appendix

[6] https://github.com/cgpotts/swda

[7] https://nlp.stanford.edu/software/CRF-NER.shtml

[8] https://www.nltk.org/

[9] https://tartarus.org/martin/PorterStemmer/

ances, keyword frequencies are more crucial than their semantic distributions.

3.4 Combination of Dialogue Utterances

Another factor we experimented is the combination of dialogue utterances in the dataset. Frequently, one dialogue turn signals the shift from one participant to another. Though not prominent, we observe 513 pairs of adjacent utterances that are generated from the same speaker[10]. We hypothesize that adjacent utterances generated by the same speaker have more similar dialogue intents thus it would be reasonable to collapse some of the adjacent same-speaker utterances into one by concatenating the utterances and obtaining a union set of DAs.

We used Sørensen–Dice coefficient (Sørensen, 1948; Dice, 1945) to calculate the similarity between two DA vectors from two adjacent utterances. Specifically, we utilized the dice distance from scipy to measure the distance of two binary vectors, 0 meaning exactly the same and 1 meaning totally different. The mean and standard deviation for all 7,821 adjacent utterances are 0.786 and 0.300 whereas adjacent utterances from the same user are of much shorter dice distance, with a mean of 0.552 and standard deviation of 0.381 for 'Users', 0.476 and 0.335 for 'Agents' respectively. As a result, in our experiments, we practice our model on two variations of the dataset: one without collapsing any utterance (consisting of 10,020 utterances) and one with adjacent utterances from the same speaker that has a dice distance less than the respective mean collapsed (consisting of 9,658 utterances). Results in section 6 show that collapsing relevant adjacent utterance moderately improves our LSTM model.

4 Approach

4.1 Baseline

A LSTM (Hochreiter and Schmidhuber, 1997) was used as baseline to predict DAs for each dialogue turn based on the utterance history. We applied the same pre-processing procedures, as described in Section 3.3, to the utterances for both the baseline model and the DAMIC model.

$$o_t = \text{LSTM}_\theta(BoW(u_1)...BoW(u_{t-1})) \quad (1)$$

4.2 Dialogue Act Modeling

Our model, DAMIC (Dialogue Act Modelling in Information-seeking Conversations) is based on LSTM and comprises three components: (1) Context Representation Construction; (2) Context-based DA Prediction; and (3) DA Sequence Modeling. Each part is implemented as a differentiable submodel and, therefore, the whole model can be learned simultaneously. Furthermore, we are not paying much attention to the DA structure within each turn and only the DA relations across dialogue turns are of interest,

Context Representation Construction In this work, the context for each turn is limited to the utterance in a single turn. Considering that the utterances in information-seeking scenario (e.g. forum post and email exchange) are usually long and information-rich, which contain images, URLs and computer languages (e.g. HTML and error logs), we converted each utterance u_t into a BoW representation as described in Section 3.3. The context representation c_t was then formed by encoding the BoW representation with a neural encoder.

$$c_t = \text{Encoder}_\theta(\text{BoW}(u_t)) \quad (2)$$

DA Sequence Modeling Inspired by the result from Section 3.2, this model is designated for learning the DA relations among dialogue turns. As Figure 5 shows, a LSTM was applied to learn DA structure from DA sequences. The last hidden state h_t^2 of the LSTM was then used as input for DA Prediction along with the context c_t.

$$h_t^2 = \text{LSTM}_\theta(a_1...a_{t-1}) \quad (3)$$

During testing, the reference DAs for prior steps are not available, we will use the predicted results instead (as Figure 6 shows).

$$h_t^2 = \text{LSTM}_\theta(o_1...o_{t-1}) \quad (4)$$

[10]Two utterances are from the same speaker if the *user_id* and *actor_type* are the same.

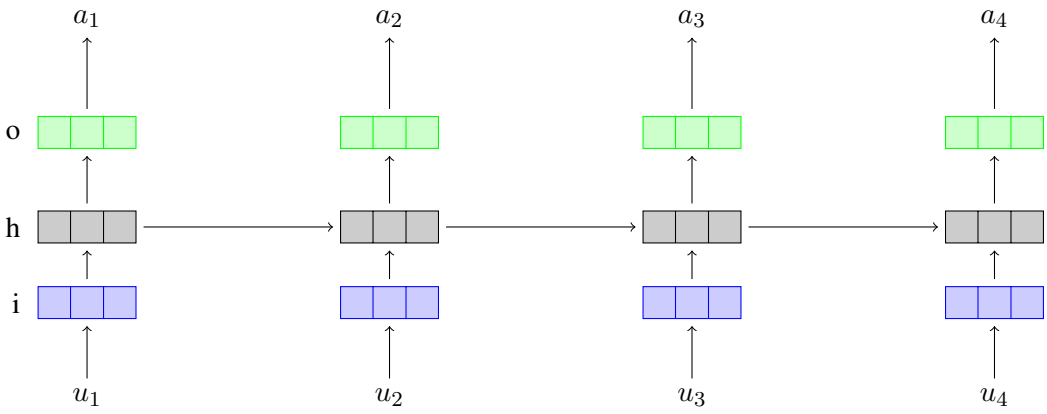


Figure 4: Baseline Model

Context-based DA Prediction Based on outputs from prior models, another neural network was trained to predict the appropriate DAs for each dialogue turn.

$$o_t = \mathrm{NN}_\theta(h_t^2, c_t) \tag{5}$$

4.3 DA-driven Response Selection

Going beyond dialogue act classification, we will try to apply the patterns, such as distribution, co-occurrence and flow patterns, learned from dialogue act dynamics to help improve the performance of neural matching networks (Yang et al., 2018) on the response selection task. In a nutshell, the automatic dialogue act tagger we trained on 2,000 dialogues will be used to preprocess 25,019 more dialogues so as to incorporate dialogue acts in addition to the wildly used lexical semantics feature in response selection task. As recent study shows, additional information available in the form of dialogue act significantly improves the performance of both generative and discriminative types of conversation models (Kumar et al., 2018). We want to test the hypothesis that meaningful patterns learned from dialogue act dynamics can help response selection for information-seeking conversations. The input of this system will be user posts on information-seeking forums, along with automatically tagged dialogue acts, and the output will be a ranked list of candidate responses.

Given the time restriction, this task goes beyond the scope of this project. Here we document the approach we plan to take for the completeness of the final report. We will also describe the reproduced baseline and results for this task in Section 5 and Section 6.

5 Experiment

5.1 Dialogue Act Modeling

The corpus was partitioned into training, validation, and test sets in the ratio 3:1:1. To make a direct comparison between the baseline model and the DAMIC model, the LSTM hidden layer number of baseline model was set to 4 and the total number of hidden layers in DAMIC Model was set to 4 as well (2 for prediction and 2 for LSTM). The size of all hidden layers were set to 128. The teacher forcing rate was set to 0.5.

For both models, the mean squared error (MSE) was used as loss function and stochastic gradient descent (SGD) was applied for optimization. The models were trained on a single GeForce RTX 2080 Ti GPU.

5.2 DA-driven Response Selection

In order to apply the patterns learned from dialogue act dynamics to help improve the performance of deep neural models for response ranking, we re-ran the experiments for Deep Matching Networks (DMN) proposed by Yang et al. (2018), which is the state-of-the-art work on response selection in multi-turn conversations.

We used the same experimental setup as Yang et al. (2018) stated in their work. Before model training, we preprocessed the datasets in following steps:

1. Perform word tokenization, stemming, lowercasing and indexing. Words ap-

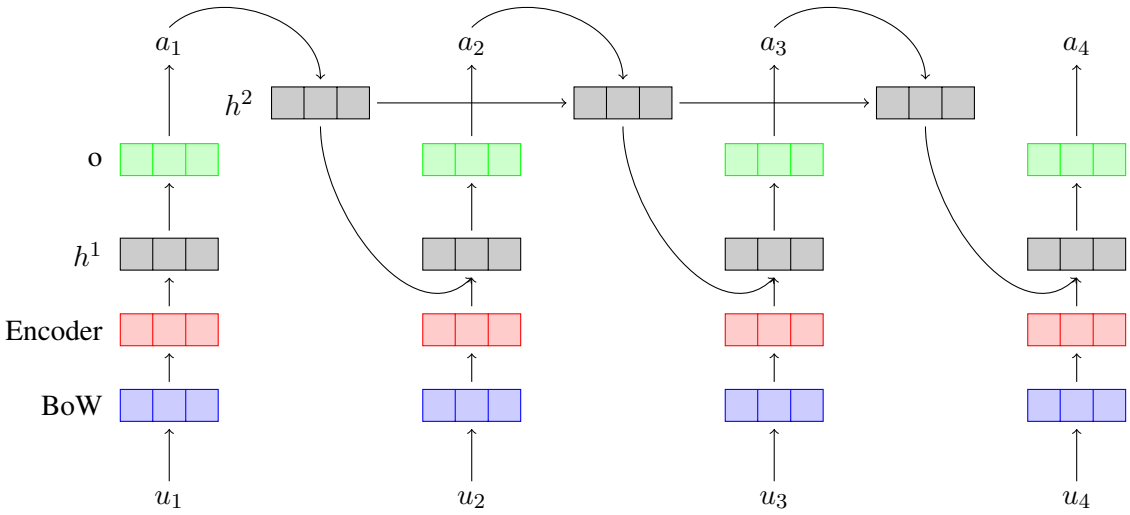


Figure 5: DAMIC Model - Training

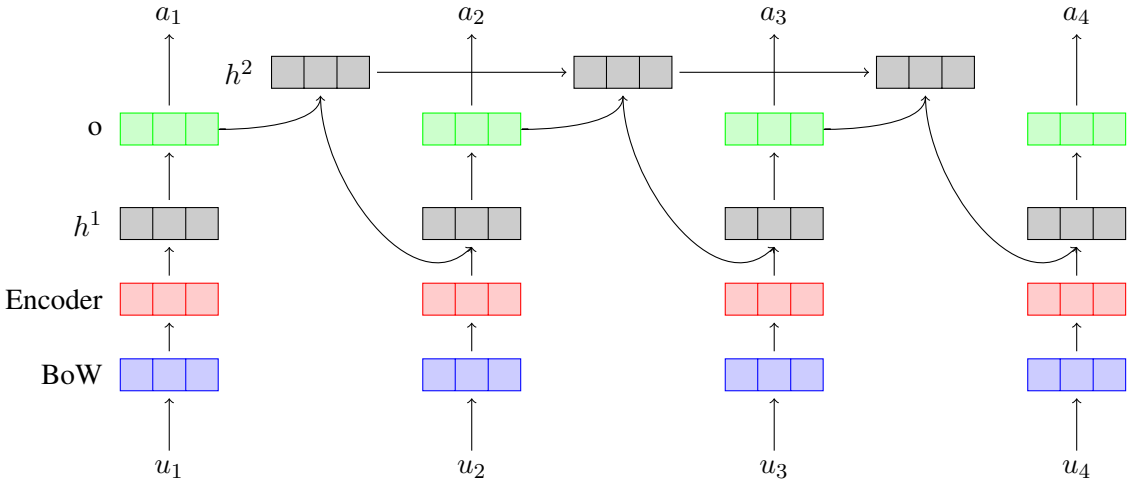


Figure 6: DAMIC Model - Testing

peared less than 5 times in the corpus are filtered.

2. Use Word2Vec[11] to pre-train the word embeddings and then update them during the model training process.

3. Filter the generated word embedding file by the words left after step 1 to save some memory.

The DMN model was trained on a single GeForce GTX 1080 GPU.

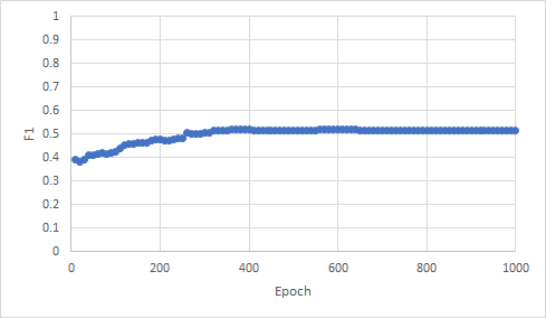
[11] https://github.com/dav/word2vec

6 Results

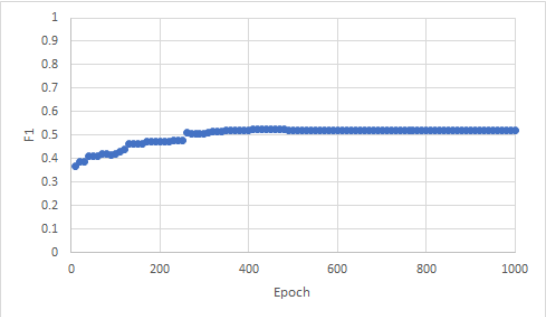
6.1 Dialogue Act Modeling

Results of the experiments in Section 5.1 are shown in Table 1. The table includes four sub-experiments: baseline LSTM and DAMIC model, with or without collapsing utterances. The results show that our DAMIC models significantly outperform the baseline models in both collapsed and non-collapsed experiments. Particularly, the DAMIC model with collapsed utterances achieves the highest F1 score 65.00. Figure 7 demonstrates the convergence of the four experiments. We also analyzed F1 for each of the 12 DAs in Table 1. For the baseline LSTM models, only OQ (Original Question) and PA (Potential Answer) receive high F1 scores. In contrast, our DAMIC model, es-

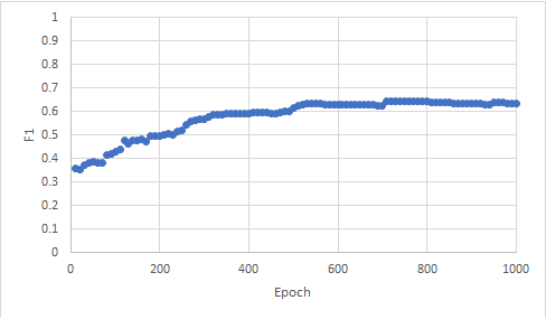
pecially when using the collapsed dataset, receives satisfiable results for the four most frequent DAs in Figure 2: OQ, PA, FD (Further Details) and GG (Greetings/Gratitude). Due to the unbalanced frequency of DA tags in the dataset, our model succeeds in identifying the most dominant tags. However, future work is still required to improve DA classification for the other DA tags.



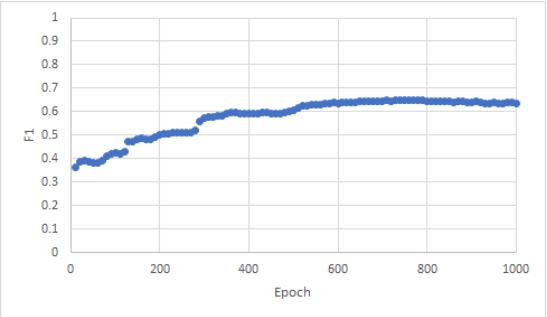
(a) Baseline model without collapsing same-speaker utterances



(b) Baseline model with collapsed same-speaker utterances



(c) DAMIC model without collapsing same-speaker utterances



(d) DAMIC model with collapsed same-speaker utterances

Figure 7: F1 vs. Epoch chart for four experiments

6.2 DA-driven Response Selection

As Table 2 shows, the result reported by Yang et al. (2018) is reproduced on both Ubuntu Dialog Corpus (UDC) (Lowe et al., 2015) and MS-Dialog corpus (Qu et al., 2018).

7 Discussion

The results in Section 6 show preliminary success in modeling multi-turn multi-tagged dialogues using LSTM-based DAMIC model. This model imposes less restrictions on the

		F1	F1 per DA											
			CQ	FD	FQ	GG	IR	JK	NF	O	OQ	PA	PF	RQ
Baseline	-collapsed	51.85	16.54	49.93	28.50	49.06	32.56	7.14	22.55	1.81	**85.35**	**72.76**	35.12	19.41
	+collapsed	52.48	19.04	47.81	23.73	**51.54**	27.60	6.64	25.52	1.13	**86.88**	**73.86**	34.68	17.56
DAMIC	-collapsed	64.39	23.83	**55.79**	27.87	**75.88**	34.50	8.06	32.31	3.10	**93.60**	**85.40**	**50.07**	34.42
	collapsed	**65.00**	21.22	**52.13**	27.76	**77.91**	34.86	8.93	32.54	2.09	**93.59**	**85.73**	**52.99**	32.14

Table 1: Baseline and DAMIC LSTM results with or without utterance collapsing

Data	UDC				MSDialog			
Methods	MAP	Recall@5	Recall@1	Recall@2	MAP	Recall@5	Recall@1	Recall@2
BM25	0.6504	0.8206	0.5138	0.6439	0.4387	0.6329	0.2626	0.3933
BM25-PRF	0.6620	0.8292	0.5289	0.6554	0.4419	0.6423	0.2652	0.3970
ARC-II	0.6855	0.8978	0.5350	0.6959	0.5398	0.8662	0.3189	0.5413
MV-LSTM	0.6611	0.8936	0.4973	0.6733	0.5059	0.8516	0.2768	0.5000
DRMM	0.6749	0.8776	0.5287	0.6773	0.5704	0.9003	0.3507	0.5854
Duet	0.5692	0.8272	0.4756	0.5592	0.5158	0.8481	0.2934	0.5046
SMN	0.7327	0.9273	0.5948	0.7523	0.6188	0.8374	0.4529	0.6195
DMN	**0.7363**	**0.9196**	**0.6056**	**0.7509**	**0.6415**	**0.9155**	**0.4521**	**0.6673**
Reproduced Result	**0.7403**	**0.9220**	**0.6106**	**0.7558**	**0.6493**	**0.9180**	**0.4645**	**0.6726**

Table 2: Reproduced Result of the DMN Baseline

structure of dialogue acts as well as utterance contexts. However, since we are aiming for a conference publication for a greater impact to the community, we would like to further develop this line of research to improve our DAMIC model (it is a pity that we did not have enough time to include them in this report).

In order to achieve better performance on less frequent DA tags, we would like to incorporate additional textual information and metadata from the corpus to refine our predication model. In addition, we believe that utilizing structured LSTM and attention mechanism can better enforce long-distance dependencies between DAs (Kim et al., 2017) (as Figure 10 shows). Furthermore, to accelerate our model and to increase efficiency, we will implement a batch learner, which allows us to run more epochs. These improvements can strengthen our model to produce a more promising result for DA prediction.

Furthermore, we can do a semi-supervised learning task on a much larger and unlabelled dialogue corpus[12] with the improved DAMIC model. Finally, we will be able to incorporate dialogue acts into state-of-the-art data-driven dialogue systems such as the deep neural matching networks (Yang et al., 2018) for the response selection task.

[12] https://ciir.cs.umass.edu/downloads/msdialog/

8 Conclusion

As dialogue act is very useful and essential for conversation modelling and very limited work has been done on applying DA-based model to information-seeking conversations on Internet forums. In this project, we applied Markov model to analyze dialogue act dynamics in information-seeking conversations and constructed our LSTM-based DAMIC model to predict DAs for a multi-tagged dialogue corpus. We have shown that the DAMIC model achieves the highest F1 score on collapsed utterances and significantly outperforms the baseline model. Meanwhile, the model requires further tuning to improve the performance on non-dominant tags.

For future work, it will be interesting to see if we can use transfer learning to leverage the models we build on MSDialog from other online forums, e.g. Reddit, Stack Overflow and phpBB. In addition, a more fine-grained (and possibly mutually exclusive) set of dialogue acts might be useful for a deeper analysis over dialogue acts in multi-turn conversations. It would also be useful to annotate *antecedents* for each utterance to indicate non-chronological dialogue flows (similar to the annotation practice conducted in our course Traum et al. (2018)). However, this will require more insights regarding the distribution of topics and language usage over the forum of interest.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yohan Jo, Michael Miller Yoder, Hyeju Jang, and Carolyn P Rosé. 2017. Modeling dialogue acts with content word filtering and speaker preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2017, page 2169. NIH Public Access.

Hideo Joho, Lawrence Cavedon, Jaime Arguello, Milad Shokouhi, and Filip Radlinski. 2018. Cair'17: First international workshop on conversational approaches to information retrieval at sigir 2017. In *ACM SIGIR Forum*, volume 51, pages 114–121. ACM.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. 2018. Dialogue-act-driven conversation model: An experimental study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1246–1256.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Vinayak Mathur and Arpit Singh. 2018. The rapidly changing landscape of conversational agents. *arXiv preprint arXiv:1803.08419*.

Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. 2017. How may i help you?: Modeling twitter customer serviceconversations using fine-grained dialogue acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 343–355. ACM.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. *arXiv preprint arXiv:1804.08759*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics.

Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, 5:1–34.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

David R Traum, Cassidy Henry, Stephanie M Lukin, Ron Artstein, Felix Gervits, Kimberly A Pollard, Claire Bonial, Su Lei, Clare R Voss, Matthew Marge, et al. 2018. Dialogue structure annotation for multi-floor interaction. In *LREC*.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. *arXiv preprint arXiv:1805.00188*.

A Appendix

Turn#	Actor Name	Actor Type	Utterance	Dialogue Act(s)
1	James	User	I had a couple of problems and contacted Microsoft support who did remote assistance. Resolved the problems but since then my computer has become increasingly slower, particularly when reading emails. I already felt Microsoft Edge was getting slower starting whereas at first it was a joy with its fast start. When explaining that to the assistant he went into settings and got rid of caches and cookies but since then it has got even slower. Any help Original title: Windows 10 Microsoft Edge	Information Request Original Question Negative Feedback
2	Faith	Agent	Thank you for posting in Microsoft Community. We understand that your Windows 10 computer and MS Edge has become slow. We2019ll certainly help you with this. Method 1: Run the System Maintenance troubleshooter. 1. Press Windows Key + X. 2. Click on Control Panel. 3. Select Troubleshooting. 4. On the left pane, click on View All. 5. Click on System Maintenance to run the troubleshooter. Method 2: Reset the Microsoft Edge I would suggest you to Reset the Microsoft Edge and see if it helps. Note: Resetting the Microsoft edge will remove your bookmarks and History.	Greetings/Gratitude Potential Answer Positive Feedback
3	James	User	Thank you so much for your help. I did option 1 and it solved the problem. Thanks again for your assistance. Much faster again.	Greetings/Gratitude Positive Feedback
4	Faith	Agent	Hi James, Thank you for posting back with the result. Glad to know the issue resolved. Feel free to post us if you need any assistance with Windows. We'll be glad to assist you. Thank you.	Greetings/Gratitude Positive Feedback
5	Juan	Agent	There is no c̈ontrol panelïn the (windows) X pop up list. Looks like MS themselves don't know what their system/commands are doing. This is a very frustrating problem!	Others Negative Feedback

Table 3: an Example Dialogue from the Corpus.

Code	Label	Description	Example
OQ	Original Question	The first question by a user that initiates the QA dialog.	If a computer is purchased with win 10 can it be downgraded to win 7?
RQ	Repeat Question	Posters other than the user repeat a previous question.	I am experiencing the same problem ...
CQ	Clarifying Question	Users or agents ask for clarification to get more details.	Your advice is not detailed enough. I'm not sure what you mean by ...
FD	Further Details	Users or agents provide more details.	Hi. Sorry for taking so long to reply. The information you need is ...
FQ	Follow Up Question	Users ask follow up questions about relevant issues.	Thanks. I really have one simple question – if I ...
IR	Information Request	Agents ask for information of users.	What is the make and model of the computer? Have you tried installing ...
PA	Potential Answer	A potential answer or solution provided by agents.	Hi. To change your PIN in Windows 10, you may follow the steps below: ...
PF	Positive Feedback	Users provide positive feedback for working solutions.	Hi. That was exactly the right fix. All set now. Tx!
NF	Negative Feedback	Users provide negative feedback for useless solutions.	Thank you for your help, but the steps below did not resolve the problem ...
GG	Greetings/Gratitude	Users or agents greet each others or express gratitude.	Thank you all for your responses to my question ...
JK	Junk	There is no useful information in the post.	Emojis. Sigh .... Thread closed by moderator ...
O	Others	Posts that cannot be categorized using other classes.	N/A

Figure 8: Description & example of 12 DA tags (figure from Qu et al. (2018)

from\to	CQ	FD	FQ	GG	IR	JK	NF	O	OQ	PA	PF	RQ	TERM.
CQ	5.3%	24.7%	4.4%	1.1%	7.1%	1.1%	3.6%	0.8%	0.5%	**31.4%**	6.6%	3.9%	9.8%
FD	4.2%	17.7%	5.2%	2.9%	6.9%	1.6%	4.1%	1.1%	0.7%	**31.6%**	6.3%	3.4%	14.4%
FQ	3.2%	17.6%	5.6%	0.8%	5.2%	1.5%	3.6%	0.8%	0.6%	**37.0%**	4.8%	3.2%	16.2%
GG	1.0%	6.6%	3.3%	8.8%	0.3%	2.5%	3.3%	0.3%	1.0%	8.3%	4.0%	5.6%	**55.1%**
IR	6.8%	**30.1%**	7.2%	0.7%	7.0%	1.3%	5.4%	1.2%	1.2%	16.3%	9.6%	4.3%	8.9%
JK	2.5%	8.3%	2.5%	2.5%	3.2%	5.7%	3.8%	1.6%	0.6%	17.8%	3.2%	4.8%	**43.5%**
NF	3.0%	15.3%	3.2%	1.4%	6.7%	2.5%	4.7%	1.4%	0.6%	35.6%	3.6%	4.0%	18.0%
O	1.6%	8.6%	3.2%	5.9%	3.2%	5.4%	5.9%	5.4%	0.5%	16.6%	3.7%	3.2%	**36.9%**
OQ	6.2%	9.5%	2.7%	0.1%	16.2%	0.7%	0.8%	0.8%	1.2%	**56.7%**	1.3%	2.6%	1.5%
PA	6.3%	**20.4%**	9.3%	1.9%	3.7%	1.8%	8.6%	0.9%	1.4%	12.2%	12.9%	5.6%	15.0%
PF	1.7%	10.5%	4.4%	12.7%	2.0%	2.1%	2.8%	1.4%	0.8%	14.0%	10.2%	3.7%	**34.0%**
RQ	3.6%	11.9%	2.4%	0.8%	5.5%	2.3%	2.9%	1.3%	1.0%	**29.2%**	2.5%	6.1%	30.6%
INIT.	0.9%	7.4%	1.2%		2.5%	0.5%	2.4%	0.1%	**84.0%**	0.5%	0.2%	0.4%	

Table 4: Transition probability (for each DA in the row, the probability of next DA in the column; each row sums up to 100% and top probabilities are bolded)

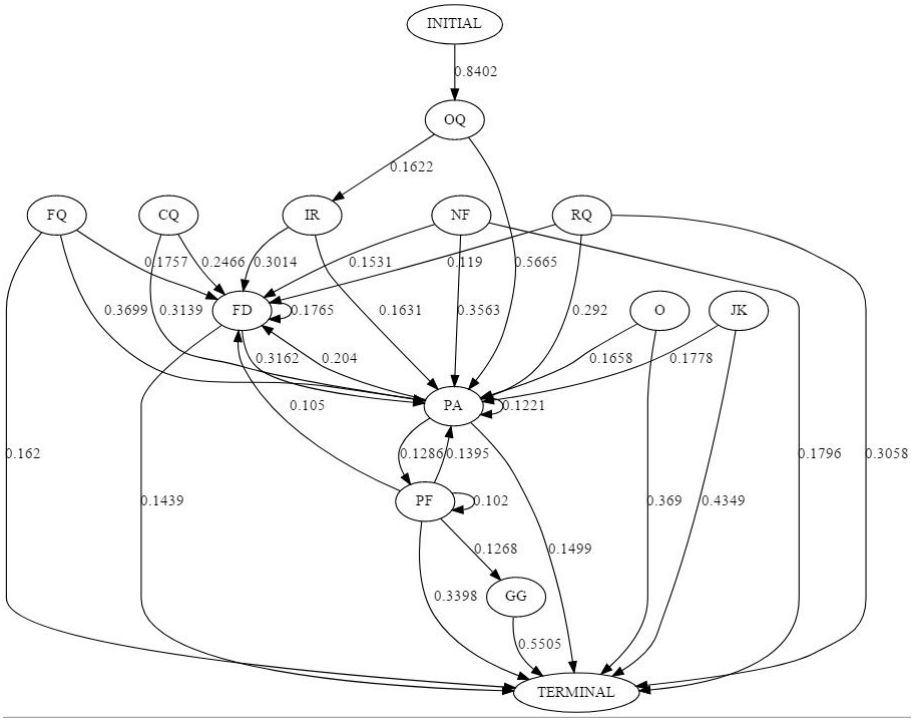


Figure 9: Reproduced Markov chain for Dialogue Acts in MSDialog-Intent

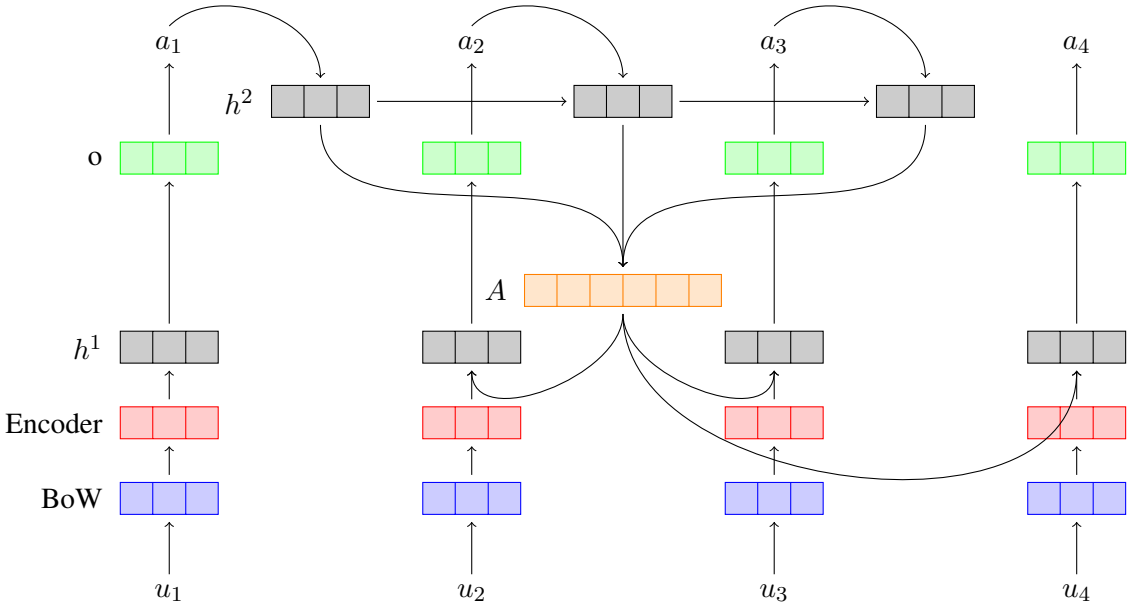


Figure 10: DAMIC Model with Attention Mechanism