Conversation Management in Task-oriented Turkish Dialogue Agents with Dialogue Act Classification

O. Fatih Kilic[a], Enes B. Dundar, Yusufcan Manav, Tolga Cekic and Onur Deniz

Natural Language Processing Department, Yapi Kredi Technology, Istanbul, Turkey

Keywords: Dialogue Act, Text Classification, Dialogue Policy.

Abstract: We study the problem of dialogue act classification to be used in conversation management of goal-oriented dialogue systems. Online chat behavior in human-machine dialogue systems differs from human-human spoken conversations. To this end, we develop 9 dialogue act classes by observing real-life human conversations from a banking domain Turkish dialogue agent. We then propose a dialogue policy based on these classes to correctly direct the users to their goals in a chatbot-human support hybrid dialogue system. To train a dialogue act classifier, we annotate a corpus of human-machine dialogues consisting of 426 conversations and 5020 sentences. Using the annotated corpus, we train a self-attentive bi-directional LSTM dialogue act classifier, which achieves 0.90 weighted F1-score on a sentence level classification performance. We deploy the trained model in the conversation manager to maintain the designed dialogue policy.

1 INTRODUCTION

Conversation behaviors of humans with machines in online dialogue systems (e.g. chatbots) differ from human-human social conversations. People tend to use split utterances, several non-informative statements before they convey their intents (Purver et al., 2009). Also, they tend to use several feedback statements to direct goal-oriented conversation agents towards a desired intent.

Dialogue acts address these different behaviors by capturing the intent of humans in generating an utterance (Austin, 1975). The main dialogue acts that people use in task oriented dialogue systems are giving orders, asking questions, making informational statements or making feedback statements (Jurafsky and Martin, 2014). Dialogue systems must detect these dialogue acts to understand the users better and maintain smoother conversations (Jurafsky and Martin, 2014). For instance, distinguishing a user statement from a directive can help agents collect information from the statements and act according to the correct intent at a directive state.

Although there exist several dialogue act class schemes, human-machine dialogue systems that are working on a specific domain and a language with different grammatical structures require labeling different classes of dialogue acts for better management of the conversation (Bunt, 2009; Jurafsky and Martin, 2014; Jurafsky et al., 1997).

To this end, we analyze the interlocutor behavior in real-life human-machine conversations collected from the banking domain, Turkish dialogue agents. We propose 9 core dialogue acts based on this behavior analysis and a hybrid conversation management architecture, where human support agents and bot agents work concurrently, for dialogue agents using these act classes. We also propose a mapping from ISO 24617-2 dialogue act annotation scheme to designed classes for consistency (Bunt et al., 2012). Finally, we manually label 462 conversations collected according to proposed classes and train a bi-directional self-attentive LSTM model for automated tagging of user utterances to be used in the conversation manager. We compare the proposed model with several dialogue act tagging architectures on our corpus, which shows the superior performance of our model and states that each specific domain and task should be handled explicitly.

We organize the paper as follows. In Section 2, we present the earlier studies that are related to our work. In Section 3, we show the development of 9 core dialogue act classes that we propose with relational examples from data collected from the banking domain Turkish dialogue agent. Then, in Section 4, we introduce a dialogue management architecture for a hybrid conversational agent based on the developed classes.

[a] https://orcid.org/0000-0003-2304-4658

Kilic, O., Dundar, E., Manav, Y., Cekic, T. and Deniz, O.
Conversation Management in Task-oriented Turkish Dialogue Agents with Dialogue Act Classification.
DOI: 10.5220/0010020700290035
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 1: KDIR*, pages 29-35
ISBN: 978-989-758-474-9

Finally, in Section 5, we present the model architecture we use for automated dialogue act tagging and we compare the performance of the proposed model with other baseline algorithms through experiments. We conclude the paper with final remarks in Section 6.

2 RELATED WORK

The dialogue act (DA) captures the generation intent of an utterance by the speaker in a conversation (Austin, 1975). Several studies have been conducted in analyzing these different type of intents and multitude of taxonomies classifying the type of DAs have been proposed (Jurafsky et al., 1997; Bunt, 2009; Bunt et al., 2012; Paul et al., 2019). Corpora of human-human social conversations have been annotated using these different DA taxonomies. The Switchboard-DAMSL corpus (Jurafsky et al., 1997), the ICSI meeting corpus (Janin et al., 2003), the AMI meeting corpus (Carletta et al., 2005), the Meeting Recorder Dialogue Act corpus (Shriberg et al., 2004) and the HCRC MapTask corpus (Anderson et al., 1991) present DA annotations for multi-human conversations in English.

In (Bunt et al., 2012), authors introduced ISO 24617-2 standards, which is further developed in (Bunt et al., 2017), for setting global standards on DA class definitions and annotation. Multiple human-human spoken corpora that are focusing on social conversations have since been annotated according to these standards and presented in DialogBank (Bunt et al., 2016).

In addition to human-human social conversation DA annotations, core DA classes proposed for task-oriented dialogue systems in recent works (Young, 2007; Shah et al., 2018). In (Paul et al., 2019), the authors proposed a universal DA annotation scheme in an effort to unify the DA classes for task-oriented dialogue systems. They aim to train a universal DA tagger using this scheme to later tag the human-human task-oriented dialogues.

For the development and the evaluation of goal-oriented DA tagging models and conversation tracking systems, large scale human-human goal-oriented conversation corpora such as Frames and Multi-WOZ are recently introduced in the literature (Asri et al., 2017; Ramadan et al., 2018). We note that while the released corpora for DA annotations and classification are mostly in English, DA annotation corpora is infrequently available in other languages as well. The Czech Railways dialogue corpus contains DA annotations for task-oriented human-human dialogues (Cerisara et al., 2018a). The French Emospeech corpus consists of dialogues in the context of a serious game between humans and machines along with the corresponding DA annotations (Barahona et al., 2012). The German VERBMOBIL corpus is especially interesting for DA classification since it is prepared for a morphologically rich language, where traditional statistical classification methods are harder to apply (Kay et al., 1992).

Although presenting these universal taxonomies for DA annotation is convenient, the specificity of domain and use case prevents the generalization of the DA class types in certain cases (Chowdhury et al., 2016). Thus, transferring the existing DA annotations of a corpus or annotating a new corpus with the generalized standards might hurt the performance of the dialogue system (Chowdhury et al., 2016). In this regard, we study a DA annotation taxonomy and classification for task-oriented dialogue systems along with a manually tagged corpus in the Turkish language, which also features different morphological aspects, for the first time in the literature.

The automated DA tagging of utterances in a conversation is also studied with mainly two different approaches. The first approach considers the traditional statistical methods, which includes the use of Hidden Markov Models (HMMs) (Stolcke et al., 2000), Maximum Entropy (Choi et al., 1999), Conditional Bayesian Networks (Ji and Bilmes, 2005) and Support Vector Machines (SVMs) (Quarteroni and Riccardi, 2010). The second approach, which is the widespread approach in recent works, considers Machine Learning (ML) based methods especially using deep learning techniques such as Convolutional Neural Networks (CNNs) (Kalchbrenner and Blunsom, 2013), Recurrent Neural Networks (RNNs) (Lee and Dernoncourt, 2016) and neural embedding (Cerisara et al., 2018b) models. Studies taking the second approach show superior performance compared to traditional statistical methods and present the current state-of-the-art in automated DA tagging.

3 DEVELOPMENT OF DIALOGUE ACT CLASSES

Different type of dialogue systems that are working on a particular domain or language requires labeling of different DA classes considering the requirements of the dialogue system (Jurafsky and Martin, 2014). The DA taxonomies presented in literature include 30 or more DA classes for recognition (Jurafsky et al., 1997; Bunt, 2009). However, these large DA tag-sets are often reduced to smaller subsets or re-designed with fewer classes that are convenient for the dialogue system management. This is mainly due to some DA classes occur rarely in certain domains (Cerisara et al., 2018a). Thus, they do not have a significant effect on the dialogue management system. Also, it is hard to collect samples for the rare classes to design a recognition system for all the DA classes presented in contemporary DA taxonomies.

In this regard, we analyze 426 conversations collected from the banking domain Turkish chatbot of Yapi Kredi that consists of 5079 user utterances. We omit the system-side utterances since the dialogue system was working only as a QA agent and did not follow a dialogue policy considering the DA classes. We develop 9 possible DA classes based on interlocutor behavior presented in the collected data.

We observe that the users tend to use several statement utterances or split sentences describing the context before conveying an intent that requires an action to be taken by the system. We tagged such utterances as a descriptive statement (SD). After describing the context, users either ask a question or gives an order. We grouped such utterances requiring action in three distinct DA classes: request statement (SR), yes/no question (QY), and open question (QW).

We also realize that the users generate feedback acts mainly in two distinct patterns. The one behavior of feedback occurs when the system detects the intent of the user incorrectly. The other behavior occurs when the users are not satisfied with the response provided by the system. Therefore, we generate 2 distinct DA classes for these feedback acts, feedback-incorrect (FI) and feedback-weak (FW) respectively. There were also opening and closing statements, which we grouped under another DA class (OC). We developed the remaining two DA classes for answer-accept (AA) and answer-reject (AR) behaviors. We present the list of developed DA classes and example utterances for each of the classes along with their English translations in Table 1. Moreover, we give further statistics about the class distributions of the corpus in Table 2.

To be consistent with universal DA annotation taxonomies, we also present a loose mapping from ISO 24617-2 annotation scheme to the developed and reduced DA tag-set in Table 3.

4 PROPOSED CONVERSATION MANAGEMENT ARCHITECTURE

We build a dialogue policy for conversation management using the presented DA tag-set. We design the dialogue policy for hybrid conversational systems, where a machine-agent and human-agent works concurrently such that when the capabilities of the machine agent are insufficient regarding the user query, the human-

Table 1: Example utterances for selected DA classes. We present examples collected from banking domain Turkish dialogue agents along with their English translations.

DA Class	Example Utterance	English Translation
SD	Bugün vade bitiş tarihimdi.	Today was the due for my statement.
SR	Bireysel ihtiyaç kredisi almak istiyorum.	I want to get a personal loan.
QW	Hesabımda ne kadar para var?	How much savings do I have in my account?
QY	Yıl sonunda puanlarımız siliniyor mu?	Will our points be invalidated at the end of the year?
FI	Ben bunu aramıyorum!	I was not asking this!
FW	Ordan sorunumu çözemedim!	I could not solve my problem there!
OC	Merhaba!	Hello!
AA	Evet, lütfen.	Yes, please.
AR	Yok, istemiyorum.	No, I do not want.

Table 2: DA class distribution in manually tagged corpus.

DA Class	Count	DA Class	Count
SD	1567	SR	634
QY	1020	QW	637
FI	506	FW	393
OC	292	AA	25
AR	5		

Table 3: Developed DA tagset mappings from ISO 24617-2 tagset.

DA Class	ISO 24617-2 Mappings
QW, QY	Question
SR	Request
SD	Inform
FW	Instruct
FI	Suggest
AA	Accept Offer, Accept Request, Accept Suggest
AR	Decilne Offer, Decline Request, Decline Suggest
OC	Initialize Greeting, Initialize Goodbye, Thanking, Return Greeting, Return Goodbye

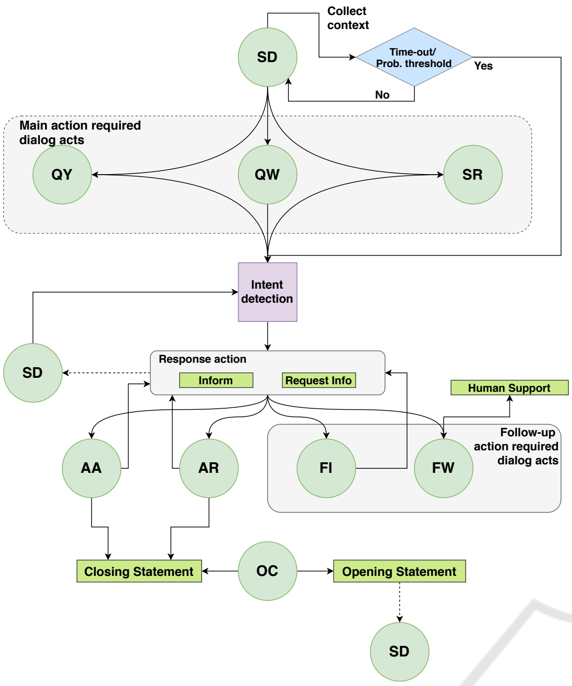

Figure 1: The dialogue policy for the conversation management with the use of DA classes is presented as a FSM.

agent takes over the conversation and completes it.

We present the dialogue policy in Figure 1 as a finite state machine (FSM). We envision in this architecture, the user either starts with an OC act or an SD act. This can be immediately followed with AD, QY, or QW act where the user intent with its context is conveyed. These acts can occur either in the same utterance or in split utterances. When one of the main-action requiring act occurs, the language understanding part of the conversation manager takes place and detects the user intent. The manager then takes the appropriate action, which can be generating an answer, directing the user into a defined flow, or collecting information from the user for using an external service. We note that if the probability of the detected SD act is low or time-threshold is passed, the manager takes action with whatever context is collected so far.

The user might provide feedback about the given answer by following the FI and the FW acts. If the user states that the generated answer is not related to their intent (FI state), then the dialogue system asks for further clarification for their intent. On the other hand, if the generated answer is correct, but does not fully satisfy the user (FW state), then the human-agent in the hybrid conversation system takes over. The human-agent further helps the user considering their intent. The generated answer might require additional approval from the user. In such a case, the user generates an utterance with the AA or the AR act. According to the class of the act, the dialogue system either ends the conversation or continues to generate utterances according to the initial intent.

Table 4: An example conversation between a user and the conversation agent. The agent handles context collection by detecting the statement utterances. It also detects the user feedback DA and asks user to direct it to the correct intent. System side utterances are not classified into any DA class, thus they are tagged as '*na*'.

Turn	Utterance	Translation	DA Class
User	Hemen her ay dọnem faizi olarak kartıma ek ücret yansıtılıyor.	Almost every month, a monthly interest is being reflected to my statemet.	SD
User	Sebebi nedir?	What is the reason for that?	QW
Agent	Yıllık üyelik ücretleri hakkında detaylı bilgi için linki takip ediniz.	You can learn about the yearly subscription rates from the link provided.	na
User	Bahsettiğim şey bu değil!	I was not talking about this!	FI
Agent	Size başka nasıl yardımcı olabilirim?	What else can I help you with?	na

The dialogue policy presented in Figure 1 is designed for task-oriented conversation systems. It does not consider the system side DAs and generates predefined utterances based on user intents. It takes the user from context defining states to fulfill their intents with feedback loops to improve their experience with the detection of DA classes at every utterance. Therefore in Section 5, we present a DA classifier to conduct the dialogue policy presented in this section.

We present an example dialogues collected from the real-life human-machine conversations in Table 4 that are managed according to the dialogue policy presented in 1. We collected this example after we deployed the automated DA classifier and employed the presented dialogue policy on banking domain conversational agents.

We observe the dialogue agent can handle several informative statements and split utterances as context collection. We note that the agent can also handle the feedback acts of users considering the presented dialogue policy by either directing them to the human customer representative for further help or by asking the user for clarification about their correct intent.

5 DIALOGUE ACT CLASSIFIER MODEL DEVELOPMENT AND EXPERIMENTS

In this section, we propose a model for automated DA tagging of user utterances. In this regard, we use a deep learning approach for the multi-class sentence classification task that is DA tagging. We compare different architectures and propose the bi-directional LSTM with the attention network model in detail. We choose the proposed model because Turkish being an agglutinative language the proposed DA classes can be identified according the use of certain morphemes and words. In this regard, the proposed self attentive bi-LSTM model can focus on the characters and words that can properly discriminate the certain DA classes. Therefore, we expect this architecture to work well with an agglutinative language on DA classification task, e.g. Turkish. We observe the proposed model is superior compared to other architectures in our experiments.

The proposed model is composed of the following components, as shown in Figure 2:

- Input Layer; inputs sentences to the model.
- Embedding Layer; maps each word into low dimensional vectors. We use pre-trained ELMo embeddings for this layer.
- bi-LSTM Layer; produces forward and backward LSTM features extracted from the embedding sequence.
- Attention Layer; learns weight vectors to extract sentence level features from word level sequence features.
- Output Layer; concatenates side information about the input sentence with attention layer features and produces the classification output of the model through a dense network with softmax activation.

The input layer takes sequence of word tokens $\{x_1, x_2, \cdots, x_n\}$, where $n = 20$ and left-zero padding is used for the sentences with less than 20 token length.

We trained ELMo language model with BOUN Turkish web corpus and Turkish Wikipedia dump extended with human-human chat conversations collected from the banking domain (Sak et al., 2008; Peters et al., 2018). BOUN web corpus is created by crawling three Turkish newspapers and web pages which contains 491 million tokens. Moreover, the Turkish Wikipedia dump contains more than 50 million tokens and the human-human chat conversations are collected from customer service conversations of the bank which contains more than 100 million tokens collected from 10 million conversations. Therefore, we use a corpus of 650 million tokens in total for the training of ELMo language model.

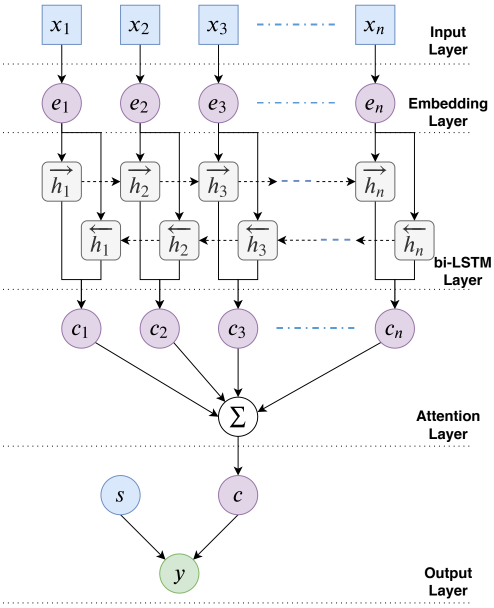

Figure 2: DA Classifier Structure.

The embedding layer produces a 1536 dimensional feature vector for each token following the default ELMo embedding architecture (Peters et al., 2018). The bi-LSTM layer contains 128 hidden units in each direction and the attention weights are calculated for both directions. We employ a dense layer with softmax activation for the output layer. To increase the generalization performance of the model, we also employ dropout at the embedding layer and the output layer with rates 0.1 and 0.24 respectively.

We perform the weight updates with Adam optimizer with 0.02 learning rate and 0.002 decay (Kingma and Ba, 2014). We select $\beta_1 = 0.83$ and $\beta_2 = 0.92$ for the optimizer and we use categorical-cross entropy as the objective function.

The side information used in the output layer consists of the character length of the utterance, the number of tokens used, and one-hot encoded representations of the punctuation used in the sentence.

Considering the class distributions, we randomly split the collected data into train, validation, and test sets with 0.75, 0.125, and 0.125 proportional amounts respectively. We use the validation set for hyper-parameter selection and early-stopping in the training phase. We compare the performance of different models using the held-out test data.

During the comparison, we experiment with CNN, CNN+LSTM, and contextual LSTM (Ctx. LSTM) models. CNN model follows the method presented in (Kim, 2014) considering the task as a sentence classifi-

Table 5: Performance comparison of different DA classifier models with the proposed model.

Method	Train			Test		
	Prec.	Rec.	AUC	Prec.	Rec.	AUC
bi-LSTM Atn.	0.984	0.984	0.999	**0.899**	**0.899**	**0.984**
CNN	0.711	0.696	0.916	0.635	0.645	0.87
CNN+LSTM	0.96	0.96	0.996	0.88	0.879	0.982
Ctx. LSTM	0.761	0.65	0.894	0.707	0.607	0.84

cation. CNN+LSTM model uses a CNN and an LSTM auto-encoder for feature extraction on top of an embedding layer. It uses a dense layer with both features concatenated as input for the classification task. The contextual LSTM model uses the utterance sequences in conversations as inputs. It takes a window of utterances from the same conversation sequentially. It then generates sentence representation for each utterance in the window using the embedding layer. It then further employs an LSTM classifier using the sentence representations in a conversation sequentially.

We present the train and test performance of the models considering Precision, Recall, and Area-under-Curve (AUC) metrics in Table 5. We observe that the proposed bi-LSTM Attention model shows superior performance both during train and test phases compared to the other architectures and selected this architecture for the deployment.

This is mainly due to the use of the contextual embedding layer and sequential processing of word features using the bi-LSTM layer. We also observe that the CNN+LSTM model shows comparable performance with the proposed model again due to the aforementioned reasons.

In light of these experiments, we decide to deploy bi-LSTM with Attention model as the automated DA tagger for implementing the dialogue policy presented in Figure 1. We deploy the DA classifier such that every user utterance is classified into one of 9 DA classes with the confidence score being the output of the dense layer. We act according to the policy described in Figure 1 if the confidence level of the DA classifier is higher than a certain threshold. We ignore the detected DA class if confidence level is below the decided threshold and we act only according to the intent detected in the utterance.

6 CONCLUSION

In this paper, we introduced a dialogue policy along with a DA tag-set for task-oriented Turkish dialogue agents. We first analyzed real-life conversations collected from a banking domain chatbot and developed 9 DA classes based on interlocutor behavior observed. We presented a dialogue policy based on this tag-set that improves the user experience. We designed it for human-machine hybrid agents in a way that it can handle split utterances, better understand the context, and accept feedback from the user. We trained a DA classifier using the annotated corpus collected according to the proposed DA tag-set. Through a series of experiments, we showed the self-attentive bi-LSTM model performs the best on sentence-level classification metrics. We have currently deployed the dialogue policy presented in this paper with the trained DA classifier in chatbots of Yapi Kredi that are helping the users with their questions, needs, applications, and financial calculations.

ACKNOWLEDGEMENT

Authors would like to thank Secil Arslan and Atilberk Celebi for their fruitful discussions and insights they provided about the paper.

REFERENCES

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Austin, J. L. (1975). *How to do things with words*, volume 88. Oxford university press.

Barahona, L. M. R., Lorenzo, A., and Gardent, C. (2012). Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents.

Bunt, H. (2009). The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. R. (2012). Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Bunt, H., Petukhova, V., and Fang, A. C. (2017). Revisiting the iso standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Bunt, H., Petukhova, V., Malchanau, A., Wijnhoven, K., and Fang, A. (2016). The dialogbank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3151–3158.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Cerisara, C., Kral, P., and Lenc, L. (2018a). On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, 47:175–193.

Cerisara, C., Kral, P., and Lenc, L. (2018b). On the effects of using word2vec representations in neural networks for dialogue act recognition. *Computer Speech & Language*, 47:175–193.

Choi, W. S., Cho, J.-M., and Seo, J. (1999). Analysis system of speech acts and discourse structures using maximum entropy model. In *Proceedings of the 37th Annual Meeting of the Association for computational Linguistics*, pages 230–237.

Chowdhury, S. A., Stepanov, E., and Riccardi, G. (2016). Transfer of corpus-specific dialogue act annotation to iso standard: Is it worth it? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 132–135.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Ji, G. and Bilmes, J. (2005). Dialog act tagging using graphical models. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–33. IEEE.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing. Vol. 3*. Pearson London London.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard swbd-damsl labeling project coder's manual. *Draft 13. Technical Report 97-02*.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

Kay, M., Norvig, P., and Gawron, M. (1992). *Verbmobil: A translation system for face-to-face dialog*. University of Chicago Press.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Paul, S., Goel, R., and Hakkani-Tür, D. (2019). Towards universal dialogue act tagging for task-oriented dialogues. *arXiv preprint arXiv:1907.03020*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Purver, M., Howes, C., Healey, P. G., and Gregoromichelaki, E. (2009). Split utterances in dialogue: a corpus study. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 262–271. Association for Computational Linguistics.

Quarteroni, S. and Riccardi, G. (2010). Classifying dialog acts in human-human and human-machine spoken conversations. In *Eleventh Annual Conference of the International Speech Communication Association*.

Ramadan, O., Budzianowski, P., and Gašić, M. (2018). Large-scale multi-domain belief tracking with knowledge sharing. *arXiv preprint arXiv:1807.06517*.

Sak, H., Güngör, T., and Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *International Conference on Natural Language Processing*, pages 417–427. Springer.

Shah, P., Hakkani-Tur, D., Liu, B., and Tur, G. (2018). Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Young, S. (2007). Cued standard dialogue acts. *Report, Cambridge University Engineering Department, 14th October*, 2007.