Experiments with Domain Dependent Dialogue Act Classification using Open-Domain Dialogue Corpora

Swapnil Hingmire **Apoorv Shrivastava** **Girish K. Palshikar** **Saurabh Srivastava**
{swapnil.hingmire, apoorv.shrivastava}@tcs.com
{gk.palshikar,sriv.saurabh}@tcs.com
TCS Research, Pune, India

Abstract

Dialogue Act (DA) classification plays a major role in the interpretation of an utterance in a dialogue and hence in the development of a dialogue agent. Learning a DA classifier requires large corpora of annotated dialogues which require extensive human efforts and cost. Additionally, nature of dialogue varies based on domain (e.g. tourism, healthcare, finance) as well as the nature of dialogues (e.g. dialogues that involve only queries and responses or dialogues that involve planning or recommendation). Hence, DA classifier trained on a particular corpus may not perform as per the expectations on another domain-*dependent* or task-*dependent* dialogues. In this paper, we propose Conditional Random Field (CRF) based DA classifier, which we train on an open-domain corpus and extend it for a domain-*dependent* corpus by enabling a domain expert to incorporate her domain knowledge in the form of simple rules. Hence, our approach does not need domain-*dependent* labeled corpora. We show the effectiveness of our proposed approach on two real-world datasets.

1 Introduction

Dialogues are an integral part of human interactions, and of arts such as literature, theatre, and films. The presence of discernible common structures in dialogues, despite endless manifestations, has fascinated linguists and artists for ages. With the advent of virtual assistants (chatbots or dialogue agents), there is keen interest in building systems that are capable of having natural and meaningful dialogues with a human user. Dialogues also occur in other major applications, including emails (Cohen et al., 2004), chats (Carpenter and Fujioka, 2011), and web forums (like Wikipedia discussions (Ferschke et al., 2012), students discussion forums (Kim et al., 2010a), comments sections in newspapers, and in community QA systems (Bhatia et al., 2014) like StackOverflow.com.

The theory of dialogue acts provides an important building block in efforts to understand and model structure, function, and flow in dialogues. A *dialogue act* (DA) represents an abstract category of the essential meaning of an utterance (often a sentence or a fragment) in the context of an ongoing dialogue. The *meaning* here usually refers to the agent's intention, the role and relationship of the utterance to the overall dialogue, etc. The *context* of an utterance includes the dialogue state, the mental state, beliefs, and agenda of the human user, and in general, any information contained in previous utterances in the dialogue.

There is a well-accepted set of 43 DAs for English (Stolcke et al., 2000), which have been used to annotate several dialogue corpora; e.g., the human-human telephone English speech Switchboard corpus, Berkeley ICSI Meeting Recorder Digits corpus, etc. Labeling of the DAs to various utterances in a dialogue bring to the fore various relationships among the utterances. For instance, if an utterance is tagged with the DA YES-NO-QUESTION then it is likely that the next utterance will have the DA of either YES-ANSWER, NO-ANSWER, NON-UNDERSTANDING or perhaps OTHER-ANSWER like *"I don't know"*. These annotated dialogue corpora have been used to build classifiers that automatically identify the DA for any given utterance as part of a given dialogue. DA classification is useful in applications where a computer system is one of the participants in the dialogues; examples: customer help-desk (Bangalore et al., 2006),

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 305–311

tutoring systems (Litman and Silliman, 2004), speech recognition, etc. But it is also used in other applications such as machine translation.

Various machine learning techniques have been used to build classifiers for DA, including HMM (Stolcke et al., 2000), Bayesian Networks (Keizer, 2001), logistic regression (Boyer et al., 2011), language models (Reithinger and Klesen, 1997), multi-layer perceptrons (Wright, 1998), Conditional Random Field (CRF) (Kim et al., 2010b) etc. Recently, several authors have explored deep learning based methods for DA classification (e.g. (Li and Wu, 2016; Khanpour et al., 2016)). An important limitation of these classification approaches is they need a large annotated corpus. More often, they are evaluated on the same domain on which they are trained.

Recently there is increasing trend of building domain specific chat-bots (e.g. domains like Insurance, Finance, Healthcare, IT services helpdesks, Tourism, etc.). It is important to note that conversations in different domains have different characteristics. For example, conversations recorded in IT services help-desks are frequently occurring queries (questions) and their responses, while conversations recorded in Tourism help-desk may involve planning of a tour, purchase of insurance in insurance domain, or recommendation of a product are likely to involve long, and detailed conversations. Additionally, some words have a domain-*dependent* sense, for example, the word "escalation" is used as a synonym to "complaint" in IT services help-desks.

Hence, we hypothesize that a DA classifier trained on an open-domain corpus may not capture characteristics of conversations for different domains and hence, its performance may not be optimal. One way to overcome this problem is to build domain-*dependent* DA classifiers. However, the creation of such classifiers requires huge cost and human efforts. Hence it is important from the practical point of view to build a DA classifier that requires minimum cost and human efforts, at the same time it can be used across multiple domains.

In this paper, we propose a CRF based DA classifier that uses a richer set of features which incorporate lexical, syntactic and semantic information as well as dialogue history. Initially, we learn a DA classifier on an open-domain corpus and then allow a domain expert to incorporate her domain knowledge in the form of simple rules. In our approach, we combine both statistical learning and domain knowledge to build a domain-*dependent* DA classifier.

The paper is organized as follows: In Section 2, we propose our CRF and cue based approach for DA classification. Section 3 discusses evaluation of our proposed DA classifier with respect to a Deep Recurrent Neural Network (RNN) based DA classifier proposed by (Khanpour et al., 2016). In Section 4 we conclude and discuss future prospects of our work.

2 Our Approach

We use cue-based approach for DA classification. Cue phrases are single words, or combinations of words in phrases, that can serve as reliable indicators of some discourse function. A cue-based model uses different sources of knowledge (cues) for detecting a DA such as lexical, collocational, syntactic, prosodic, or conversational-structure cues. This knowledge can then be fed to a machine learning system for training a DA classifier. There is a wide range of features used in DA classification, including the words in each utterance, syntactic information such as Part of Speech (PoS) tags, pragmatic information, including the discourse context as captured by the DAs of preceding utterances, whether there has been a change of speaker, and prosodic information from the acoustic signal if the audio data is available.

Conditional Random Field (CRF) are often applied in machine learning for structured predictions and can be thought of as the sequential version of logistic regression, where logistic regression is a log-linear model for classification, CRF is a log-linear model for sequential labeling. Whereas an ordinary classifier predicts a label for a single sample without regard to neighboring samples, a CRF can take context into account, which is the best match for a problem like conversation analysis as in any conversation most of the utterances are contextually dependent. For example, a lot of information has already been discussed in the conversation till the current utterance, and any new utterance will most likely to keep the already discussed information in mind instead of repeating the information.

We use CRF for training a model with features that provide enough cues for classification of dialogue acts, whether clearly distinguishing DAs like THANKING and APOLOGY, or closely related DAs which are hard to distinguish, e.g. all question-related dialogue acts.

2.1 Modeling Steps

The steps for our model creation starts with text cleaning from correction of spelling mistakes and normalization of repeated symbols. In the next step, we change each word to its lemma form, and the corresponding PoS tags are obtained for each of them. In the third step, word bi-grams and PoS bi-grams are also added as features. After adding these features (words, word bi-grams, PoS, PoS bigrams), we introduce a few cue based features for accuracy improvement. We observed that most of the `QUESTION` classes have at least one of the cues for a question, like any one word from WH-Words (what, why, who, where, how) or a question mark "?". Hence, we add a feature to indicate an utterance starting with a WH-Word is likely to be a `QUESTION`. To discriminate `QUESTION` classes further, we add features like presence of WH-Words or collocations based question phrase like *"can I", "are you"*, etc.

We also add separate features for DAs where cues for expressing gratitude, apology or back-channel acknowledgment (like *"Yeah", "okay", "uh-huh"*, etc) are present in an utterance. Additionally, we add a feature for `CONVENTIONAL-OPENING` as the opening utterances of conversations contain words and phrases along with expression of greeting like *"Hello", "Welcome"*, etc. and making it prone to be tagged as `STATEMENT`.

In the end, we created following set of semantic and syntactic features for training of model:

1. Lemmas of words
2. PoS tags
3. PoS tag and word lemma bigrams
4. presence of words that express apology
5. presence of Wh-word
6. presence of words that express gratitude
7. presence of words that indicate start of a conversation
8. presence of a question phrase
9. presence of a question phrase at the beginning of an utterance
10. presence of words that express agreement with the last utterance

3 Experimental Evaluation

We evaluate the performance of our algorithm with Recurrent Neural Network based dialogue act classifier proposed in (Khanpour et al., 2016).

3.1 Datasets

Training datasets:

Since our study focuses on classifying DAs in open-domain conversations, we chose to evaluate our model on Switchboard (SwDA) (Jurafsky et al., 1997) and Dialog State Tracking Challenge 2 (DSTC2[1]) datasets:

- SwDA: The Switchboard corpus (Godfrey et al., 1992) contains 1,155 five-minute, spontaneous, open-domain dialogues. (Jurafsky et al., 1997) revised and collapsed the original DA tags into 43 DAs, which we use to evaluate our model. SwDA has 19 conversations in its test set.

- DSTC2: The Dialog State Tracking Challenge-2 dataset is a conversational dataset of an automated restaurant assistance system and its users, having a total of 2118 different conversations and a total of 19 different user goals which are mapped to 19 different dialogue acts based on similarity of meaning.

Test datasets:

- DSTC2: we used DSTC2 for both training and testing as it is a domain specific dataset.

- Mutual Funds: This dataset contains conversations between customers of an online money management platform and customer service associate through online chat. This dataset is about queries regarding mutual funds transactions through the platform. It contains 26 conversations with total 572 conversational utterances. An example conversation between a customer and a help-desk assistant with manually tagged DAs is given in Table 1

[1] `http://camdial.org/~mh521/dstc/`

Speaker	Utterance	Dialogue Act
Customer	Please assist me in payment of MF	ACTION-DIRECTIVE
Assistant	Hi !	CONVENTIONAL-OPENING
Assistant	This is Jim from ZZZ Mutual Funds Online Assistance.	STATEMENT
Assistant	How may I assist you ?	WH-QUESTION
Customer	I have started mf last month onlly	STATEMENT
Customer	please assist me how can I transfer amount for this month	ACTION-DIRECTIVE
Customer	Hello	CONVENTIONAL-OPENING
Customer	anyone is there ?	STATEMENT
Assistant	Surely I will assist you with the same.	STATEMENT
Assistant	Could you please help me with your registered Email ID and contact number for verification purpose ?	YES-NO-QUESTION
Customer	fname.lname@xyz.com	ABANDONED/UNINTERPRETABLE
Customer	99XX99XX99	ABANDONED/UNINTERPRETABLE
Assistant	Thank you for the details provided.	THANKING
Assistant	Have you schedule any SIP from your mutual fund account	WH-QUESTION
Customer	I dont know much about this	STATEMENT
Assistant	Please provide your PAN No , Date of Birth and Ending 4 Digits of your bank account linked with Myuniverse Investment account	ACTION-DIRECTIVE
Customer	ABCDE0000G	ABANDONED/UNINTERPRETABLE
Customer	DD / MM / YYYY	ABANDONED/UNINTERPRETABLE
Customer	9999	ABANDONED/UNINTERPRETABLE
Assistant	Thank you for the details provided.	THANKING
Assistant	Please be online , I shall check this for you.	STATEMENT
Assistant	Hello sir	CONVENTIONAL-OPENING
Assistant	As checked , you have schedule SIP from your Account	STATEMENT
Customer	Okay	AGREEMENT/ACCEPT
Customer	can you call on my number please	YES-NO-QUESTION
Assistant	Yes sir	YES-ANSWERS
Assistant	Thank you for contacting us.	THANKING
Assistant	Have a nice day.	CONVENTIONAL-CLOSING
Customer	thank you	THANKING

Table 1: An Example Conversation from Mutual Funds Domain

3.2 Experimental Settings

RNN based approach (RNN_{DA})

We used the SwDA and DSTC2 dataset to train RNN_{DA} based model with LSTM layers as described by (Khanpour et al., 2016). All conversations in the training set were preprocessed, and a randomized selection of one-third of them was utilized as a development set to allow the LSTM parameters to be trained over a reasonable number of epochs. We used pre-trained Glove (Pennington et al., 2014) word embeddings of 300 dimension vectors[2]. We used the NN packages provided by (Lei et al., 2015a) and (Lei et al., 2015b). We trained the model with following parameters kept constant (dropout = 0, decayrate = 0.7, dimension of hidden layer = 100, number of layers = 10 and

[2] http://nlp.stanford.edu/data/glove.6B.zip

learning rate = 0.01)

CRF based approach (CRF_{DA})

As both SwDA and DSTC2 datasets are conversational datasets we trained CRF_{DA} model for sequence labeling of dialogue acts. We used CRF implementation from MALLET[3] for training the model.

3.3 Enhancing performance of CRF_{DA}

In the CRF_{DA} classifier for a given sentence output is given as probability distribution across all DAs. We analyzed these output distribution and found that sometimes the correct DA is having slightly less probability than the highest probability DA, so to improve the prediction accuracy we used priority rules for DAs.

Priority Rules:

If the probability difference of top two DAs is within specified threshold and lower probability DA is defined as the high priority then we override the algorithm predicted DA to the high priority DA. For instance, suppose we have defined the threshold as 0.2 probability difference and we have a priority rule defined as: `DA1`→`DA2`, then `DA2` is having higher priority than `DA1` and when in CRF_{DA} output, `DA1` is having higher probability than `DA2` and their probability difference is less than or equal to our threshold 0.2 than the `DA2` (second highest probability dialogue act) is given as prediction in place of `DA1` (highest probability dialogue act).

For example, an utterance with text *"But how come we weren't doing this, say, twenty years ago"* which got tagged with `STATEMENT` and `WH-QUESTION` as top two suggestions with a probability difference of around 0.15 and we can clearly say that the utterance is more of a question than a statement. To handle such cases we defined a priority rule like `STATEMENT`→`WH-QUESTION` with a acceptable probability difference threshold of 0.3 Using this rule whenever a sentence gets `STATEMENT` and `WH-QUESTION` as top two predictions and have a probability difference less than or equal to 0.3 than we change the algorithm prediction from `STATEMENT` to `WH-QUESTION`. We defined few more priority rules based on similar observations.

3.4 Analysis of Results

Table 2 show results of our experiments. We can observe the impact of the domain on the performance of both CRF_{DA} and RNN_{DA} classifiers. When we trained RNN_{DA} classifiers using SwDA- an open-domain corpus as a training dataset and evaluated on the domain-*dependent* datasets, the performance was poor. We can also observe that when we trained RNN_{DA} classifiers using DSTC2- a domain-*dependent* corpus and evaluated on the test dataset of DSTC2, the performance is significantly higher when the classifiers are evaluated on SwDA or Mutual Funds dataset. In summary, a RNN_{DA} classifier trained on one corpus of one domain performs poor on dialogues in another domain.

In Table 2, we can observe that CRF_{DA} outperforms RNN_{DA} on both DSTC2 and Mutual Funds dataset when SwDA corpus is used for training. The performance CRF_{DA} is comparable to RNN_{DA} when the dialogues from the same domain are used for both for training and testing. Hence, we can say that performance of both RNN_{DA} and CRF_{DA} is sensitive to the domain of dialogues.

We can also observe in Table 2 that addition of a few manually defined rules to CRF_{DA} classifier (CRF_{DA} + Rules) significantly improves its performance.

4 Conclusions and Future Work

DA classification is an important task in building Dialogue Agents. However, the creation of a sufficiently large tagged dataset for a domain is a highly challenging task as it exerts a high cognitive load on the domain experts (which are likely to be expensive). One approach is to use an open-domain tagged dataset and use it across different domains. In this paper, we proposed a CRF based approach for learning a DA classifier on an open-domain dataset and evaluated it on two different domain-*dependent* datasets. In our approach, we did feature engineering for linguistically motivated features so that the features will capture how *in-general* a dialogue takes place. However, for each domain and further for each domain-specific task, dialogues have different characteristics. To handle such a domain-*dependent* dialogues, we extended our approach through the incorporation of a few easy to define rules which improved the performance of DA classification on

[3] `http://mallet.cs.umass.edu`

Training Corpus	**Test Corpus**	RNN_{DA}	CRF_{DA}	CRF_{DA} + Rules
SwDA	SwDA	68.9	66.9	67.1
	DSTC2	21.9	**46.4**	**61.7**
	Mutual Funds	14.5	**58.0**	**63.2**
DSTC2	SwDA	11.1	**21.5**	**21.7**
	DSTC2	94.1	89.8	90.1
	Mutual Funds	33.2	32.9	**43.3**

Table 2: Comparison of DA Classification Accuracy for Different Datasets

domain-*dependent* datasets. In summary, towards the goal of reducing knowledge acquisition overhead in creating domain-*dependent* tagged corpora for different domains, our approach uses existing open-domain corpus to learn a DA classifier and enhances it using a set of manually defined rules.

In future, we would like to do experiments with a few more open-domain and domain-*dependent* dialogues. We would also like to explore transfer learning techniques for DA classification.

References

S. Bangalore, G. Di Fabbrizio, and A. Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proc. 21st COLING, and 44th ACL*, pages 201–208.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, October.

Kristy Boyer, Joseph Grafsgaard, Eun Young Ha, Robert Phillips, and James Lester. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1190–1199, June.

Tamitha Carpenter and Emi Fujioka. 2011. The role and identification of dialog acts in online chat. In *Proc. Workshop on Analyzing Microtext at the 25th AAAI, Conference on Artificial Intelligence*.

W.W. Cohen, V.R. Carvalho, and T.M. Mitchell. 2004. Learning to classify email into ”speech acts”. In *Proc. Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 309–316.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, April.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP’92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95, December.

Simon Keizer. 2001. A bayesian approach to dialogue act classification. In *Proc. BI-DIALOG*.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jihie Kim, Jia Li, and Taehwan Kim. 2010a. Towards identifying unresolved discussions in student online forums. In *Proc. NAACL HLT, 2010 Fifth Workshop on Innovative Use of NLP, for Building Educational Applications*, pages 84–91.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010b. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’10, pages 862–871, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015a. Molding cnns for text: non-linear, non-consecutive convolutions. *arXiv preprint arXiv:1508.04112*.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez i Villodre. 2015b. Denoising bodies to titles: Retrieving similar questions with recurrent convolutional models. *CoRR*, abs/1512.05726.

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1970–1979, Osaka, Japan, December. The COLING 2016 Organizing Committee.

D. J. Litman and S. Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Proc. HLT/NAACL.*

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Norbert Reithinger and Martin Klesen. 1997. Dialog act classification using language models. In *Proc. EuroSpeech-97*, pages 2235–2238.

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.

Helen Wright. 1998. Automatic utterance type detection using suprasegmental features. In *Proceedings of the International Conference on Spoken Language Processing 1998*, page 1403.