Modeling Temporality of Human Intentions by Domain Adaptation

Xiaolei Huang[1], Lixing Liu[2], Kate B. Carey[3], Joshua Woolley[4], Stefan Scherer[2], Brian Borsari[4]
[1]Dept. of Information Science, University of Colorado Boulder
[2]Inst. for Creative Technologies, University of Southern California
[3]Dept. of Behavioral and Social Sciences, Brown University
[4] Dept. of Psychiatry, University of California, San Francisco
xiaolei.huang@colorado.edu, {lxliu, scherer}@ict.usc.edu,
kate_carey@brown.edu, {josh.woolley, brian.borsari}@ucsf.edu

Abstract

Categorizing a patient's intentions during clinical interactions in general and within motivational interviewing specifically may improve decision making in clinical treatments. Within this paper, we propose a method that models the temporal flow of a conversation and the transition between topics by using domain adaptation on a clinical dialogue corpus comprising Motivational Interviewing (MI) sessions. We deploy Bi-LSTM and topic models jointly to learn theme shifts across different time stages within these hour-long MI sessions to assess the patient's intent to change their habits or to sustain them respectively. Our experiments show promising results and improvements after considering temporality in the classification task over our baseline. This result confirms and extends related literature that has manually identified that certain phases within MI sessions are more predictive of patient outcomes than others.

1 Introduction

Motivational Interviewing (MI) (Miller and Rollnick, 2012) is a collaborative communication style used to address a variety of health problems such as alcohol and drug use. Accurately understanding the patient's intentions to change from his/her speech during the session could greatly enhance the efficacy of MI. Motivational Interviewing Skill Code (MISC) is a coding system that captures client language, specifically change talk (CT) and sustain talk (ST) (Miller et al., 2008). However, reliable MISC coding is labor-intensive and requires domain expertise. Recent computational annotation methods have been proposed to automatically classify patients' behaviors within MI (Xiao et al., 2016; Pérez-Rosas et al., 2017; Gibson et al., 2017). To this end, Recurrent Neural Networks (RNN) that capture sequential information are applied for the classification of patient behavior.

Recent research shows that themes and words within a conversation change across time (Dufour et al., 2016). Similarly within MI, topics and the patient's attitude towards their willingness to change might shift. Within this work, we investigate how shifts in themes across time affects performance of the intention classifications for the dialogue.

Specifically, we focus on the patient intent classification task and propose a method that adopts the temporal factor by domain adaptation to improve performance of the classifiers. We evaluate our approach on a dataset of college alcoholism (Carey et al., 2009; Borsari et al., 2012), containing transcripts of MI conducted with U.S. college students. Specifically, we first explore the theme shift and give a brief analysis by topic modeling (Blei et al., 2003). We then utilize Bi-directional Long Short-Term Memory (Bi-LSTM) (Graves and Schmidhuber, 2005) to encode utterances from both word and topic embeddings. Next, we concatenate both contextual information with the encoded utterance representations. Finally, we jointly train a unified representation of utterance by domain adversarial training and patient intent classification. We show that this approach can lead to improvements in classification performance.

2 Dataset

We conduct our experiments on a clinical dataset of college student alcoholism (Carey et al., 2009; Borsari et al., 2012), where we obtain 193 MI transcripts with a total of 83677 utterances. Each of the MI session ranged between 60 and 90 minutes. Each client utterance was coded using the MISC. In this paper, we focus on classifying patient behavior on the utterance-level. Specifically, we classify patient behavior on collapsed MISC

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 696–701
Brussels, Belgium, October 31 - November 4, 2018.

Table 1: Examples of utterances in Alcoholism Treatment. "I" stands for interventionist and "P" for patient. MISC codes are provided in the third column. P and I codes are coded following MISC. Change talk (CT), sustain talk (ST), and follow neutral (FN) codes are also provided.

Role	Conversation	MISC
I	Maybe you could tell me a little bit about what you do on the weekends, what your weekends have been like.	quo
P	Well we go out, but before we go out we just drink in the dorm room.	FN
I	Has this sort of changed your thinking, are things different than they were when you came in?	quc
P	I mean, I feel guilty about drinking,	CT (o+3)
I	Yeah. So it feels like, or it sounds like social social drinking is a big part of how you meet other people.	res
P	It's just, like ...and I don't mean to sound mean, but about the kids who don't drink, and people think that, "Oh, the kids who dont drink are losers".	ST (o-3)

annotation codes into with three categories: "CT": *Change talk* indicates utterances that reflect motivating factors related to change; "ST": *Sustain talk* indicate the patient has no intentions to change; "FN": *Follow neutral* means there is no indication of patient inclination. An example conversation snippet, highlighting all three sources of information is provided in Table 1. The intention labels (o+3, o-3) are only available for patients, whose '+' and '-' refer to change vs sustain talk (CT vs ST) and the number measures the "strength of client language," which represents a subjective assessment by human annotators, and the 'quo' and 'quc' refer to "open question" and "closed questions", which are only for interventionist (see (Borsari et al., 2015) for details regarding the coding strategy). While the MISC codes of client utterances within MISC are more complex and comprise other types of annotations, we focus on human intention modeling (i.e., CT vs. ST vs. FN) only.

How the theme of dialogue shift overtime?

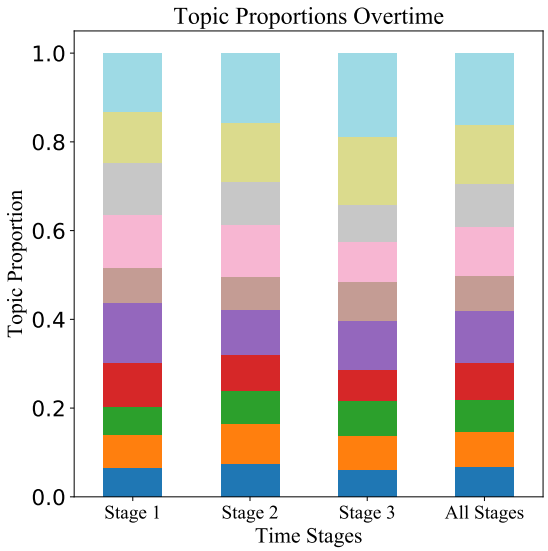

Figure 1: 10 topic proportions of patient's utterance across time stages.

We qualitatively examined how the distribution of content changes across different time stages. To measure the distribution of content, we trained a topic model with 10 topics using Gensim (Řehůřek and Sojka, 2010) with default parameters. The data doesn't have associated timestamps, thus we empirically split each MI transcript by the number of patient utterances equally into three time stages, stage 1, stage 2 and stage 3. We calculated the proportion of each topic within the same time period by take the average of all transcripts. We then normalized the topic distributions and finally visualize the extent to which distributions of the 10 topics varies by time.

We can observe the varied topic distributions across different stages of conversations, where the topic distributions are plotted from the bottom to the top. There are some topics have more variations, such as topic 4, and some topics are very stable such as topic 1 [1]. Recent research shows the performance of classification tasks might be impacted by the temporal character of language (Huang and Paul, 2018). Thus, it might be desirable to model the temporality in the computational classifiers.

[1]The 5 top words of topic 4 and 1: yeah, go, friends, know, people; beer, alcohol, games, meeting, playing.

3 Model

The architecture of the proposed model is shown in Figure 2. We feed four types of information to the model: topic- and word-level data of the utterance (*content*), preceding interventionist verbal behavior (*context*) and prior MISC annotations of utterances (*MISC*). Particularly, we empirically extracted previous 5 utterances as context and 10 previous codes as MISC[2], where we set "unk" as the default.

Embeddings. We built two types of embeddings, word embedding and topic embedding. We created word embeddings from Googles pre-trained Word2Vec (Mikolov et al., 2013) and created topic embeddings from a trained LDA (Blei et al., 2003) specific to the corpus. We treated each MISC as one document and trained an embedding model.

Unified Representation. We apply Bi-directional LSTM (Bi-LSTM) (Graves and Schmidhuber, 2005) on the inputs. Dropouts (Srivastava et al., 2014) are applied on the outputs of Bi-LSTM. We merge the outputs by concatenation and feed the outputs to the dense layer to learn a unified representation of the utterance.

Joint Learning. We apply domain adversarial training (Ganin et al., 2016) only on the topic inputs from learned topic representations. Our intuition is that the topic distributions across different stages of the MI session could track the variations of patients' intents. We empirically split the conversation into three time stages: Stage 1-3 (i.e, beginning, middle, and end). The goal of domain adversarial training is converted to a time stage prediction task, which aims to differentiate topic themes both locally and globally. We used one-hot encoding to represent labels of the prediction tasks. We deploy softmax functions for both time stage and intention predictions. We use categorical cross entropy to jointly optimize the training process of the two classification tasks: domain classification and patient intent classification.

4 Experiments

Each utterance is lowercased and tokenized by NLTK (Bird et al., 2009). We filter out the utterances that are shorter than 5 tokens and then remove punctuations. Finally, we obtain 22432 patient utterances. The dataset is stratified and split into training set (80%), validation set (10%) and testing set (10%), as shown in Table 2. We train our models on the training set and run grid search to find the optimal parameters on the validation set by the weighted F1 score.

Table 2: Statistics of the processed dataset.

Datasets	Train	Valid	Test	Total
CT	6246	779	768	7794
ST	3099	406	401	3906
FN	8600	1058	1074	10732
All				22432

The details of optimized parameters are listed as follows. The models were trained for 15 epochs with a batch size of 64. Each utterance and its context are padded to 50 words. The utterance's previous MISC codes are padded to 10. We pad the sequences with an "unknown"-token. The size of LSTMs was tuned in the range of $[100, 150, 200]$ and the size of dense layer tuned within $[100, 150, 200]$. We select the activation function of the Dense layer within {relu (Hahnloser et al., 2000), tanh, softplus} (Hahnloser et al., 2000). We tried different flip gradient value within $[0.05, 0.01, 0.005]$ for the domain adversarial training. We tuned the dropout rate between $[0.1, 0.2]$. The optimizer was selected either RMSprop (Hinton et al., 2012) or Adam (Kingma and Ba, 2014) with a fixed learning rate of 0.001. Finally, we empirically set the loss weight of the domain adversarial training to 0.05.

We trained the topic model on the MI corpus using Gensim (Řehůřek and Sojka, 2010). The number of topics was selected by coherence scores among 5, 10, 20 topics. We used Google pre-trained word embedding with 300 dimensions (Mikolov et al., 2013). We obtained 50-dimension code embedding by Word2vec (Mikolov et al., 2013) for the MISC codes, where each sequence of MISC were treated as a document.

We select three different approaches as our baselines with the inputs: content, context, MISC, and topic.

- (Pérez-Rosas et al., 2017) with rich linguistic features (denote as Perez2017_lin): We reproduced their method. We used scikit-learn (Buitinck et al., 2013) to ex-

[2]We encode 10 MISC codes prior to the current one as a sequence of 10 "words", then we treat the sequence as an additional input document.

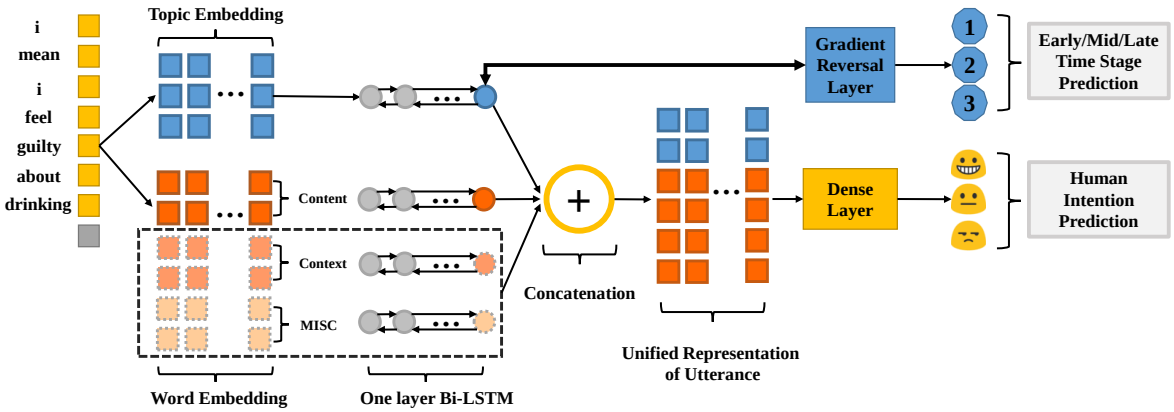

Figure 2: The proposed model learns from tow different channels: word and topic channels. The topic channel captures the theme shifts across stages while the word channel aims to capture the behavior indicators within each utterance.

tract n-gram features, and applied Stanford Parser (Manning et al., 2014) to extract Part-of-Speech (POS). We replaced LIWC (Pennebaker et al., 2001) by a free and open source lexicon (Mohammad and Turney, 2013) to extract semantic features. We use Ngram+ to denote the rich linguistic features.

- (Pérez-Rosas et al., 2017) with embeddings (denote as Perez2017_vec): Word embedding shows superior results over the n-gram features in the classification tasks (Mikolov et al., 2013) in the recent past. We experiment feeding the classifier with word vectors while we keep the same parameter settings as the Perez2017_lin baseline. We deploy the strategy of concatenating word embeddings to build representations of utterances, which is denoted as "Vec-con".

- (Xiao et al., 2016) (denoted as Xiao2016): Their approach applies Bi-directional RNN to encode each utterance by both the utterance itself and its preceding one. There are two major differences between their method and ours: first, they did not consider temporality in their model, second, they did not use the previous MISC sequences as inputs. They used Gated Recurrent Unit (GRU) (Chung et al., 2014) as the RNN cell.

We use the "Co", "Ct", "MISC" to denote the utterance (*content*), preceding interventionist verbal behavior (*context*) and prior MISC annotations of utterances (*MISC*) respectively. And we use "All" to denote all of the inputs[3]. We use the "T" to denote temporal shifts proposed in our paper. We balance training weights for the classification labels. We use metrics from scikit-learn (Buitinck et al., 2013) to evaluate the classification performance by precision, recall and weighted F1 on the intention labels.

4.1 Results and Discussion

Table 3: Results of classification evaluations.

Features	Method	Precision	Recall	F1
Perez2017_lin				
Ngram+	Co	0.62	0.62	0.62
	Co+Ct	0.60	0.61	0.61
	All	0.61	0.61	0.62
Perez2017_vec				
Vec-con	Co	0.60	0.58	0.59
	Co+Ct	0.61	0.59	0.60
	All	0.61	0.57	0.58
Xiao2016				
Vec	Co	0.65	0.63	0.64
	Co+Ct	0.68	0.64	0.65
	All	0.67	0.67	0.67
Proposed model				
Vec	Co+T	0.65	0.64	0.65
	Co+Ct+T	0.70	0.66	0.68
	All+T	0.74	0.67	0.70

The results of our experiments are summarized in the Table 3. Findings indicate that our proposed approach leads to a small performance boost after using the topic embeddings. Thus, our sim-

[3]The baselines did not use one or more inputs (the context and MISC) in the original publications. We used different combinations for fair comparison.

ple feature augmentation approach has the potential to make classifiers more robust. In addition, the contextual information ("Ct") is quite useful to identify the patients' current intentions, and the sequential information through time stages has strong indications of human intentions.

Significance Analysis

We conducted significance analysis to compare Xiao2016 and our proposed method. Because Xiao2016 only used content and context inputs, in this analysis, we train our method with the same inputs (Co+Ct). We followed the method of bootstrap samples (Berg-Kirkpatrick et al., 2012) to create 50 pairs of training and test datasets with replacement, where we keep the sizes the same in the Table 2. We keep the same experimental steps and use the parameters that achieved the best performances in the Table 3 to train the models.

To compare the two approaches, we conduct a paired t-test comparing the achieved F1 scores of both models. We used a two-tail test instead of one tail test used in the paper due to its increased rigor and lack of prior assumptions (Dror et al., 2018). The test reveals a significant result with $t(85) = 3.084$ and $p = 0.00275$. The result shows that we can reject the null hypothesis that our proposed method is not better than Xiao2016.

5 Conclusion

In this paper, we focus on the temporal characteristics of the MI corpus and propose a simple method that models the temporal factor within a single MI session. We jointly learn the utterance representation via time stage and intention predictions and the proposed model improves the performance of the classification task. The identified intent of clients could help therapists adjust their treatment strategy. In future work, we will investigate other external sources of knowledge, such as acoustic cues and videos to further improve the performance of the model.

6 Acknowledgements

We thank the anonymous reviewers for their constructive comments. Partial work done while the first author was an summer intern at ICT, USC. The idea of modeling temporal factor was inspired by the paper (Huang and Paul, 2018), co-authored with Michael J. Paul. This work was supported by National Institute on Alcohol Abuse and Alcoholism grants R01 AA015518 and R01 AA017427 to B. Borsari, and R01 AA012518 to K. Carey. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health, or the Department of Veterans Affairs or the United States Government. The authors would like to thank the students and therapists who allowed their audiotapes to be utilized for this study.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Brian Borsari, Timothy R Apodaca, Kristina M Jackson, Nadine R Mastroleo, Molly Magill, Nancy P Barnett, and Kate B Carey. 2015. In-session processes of brief motivational interventions in two trials with mandated college students. *Journal of consulting and clinical psychology*, 83(1):56.

Brian Borsari, John TP Hustad, Nadine R Mastroleo, Tracy O'Leary Tevyaw, Nancy P Barnett, Christopher W Kahler, Erica Eaton Short, and Peter M Monti. 2012. Addressing alcohol use and problems in mandated college students: a randomized clinical trial using stepped care. *Journal of consulting and clinical psychology*, 80(6):1062.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Kate B Carey, James M Henson, Michael P Carey, and Stephen A Maisto. 2009. Computer versus in-person intervention for students violating campus alcohol policy. *Journal of consulting and clinical psychology*, 77(1):74.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1383–1392.

Richard Dufour, Mohamed Morchid, and Titouan Parcollet. 2016. Tracking dialog states using an author-topic based representation. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 544–551. IEEE.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

James Gibson, Dogan Can, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. In *Proc. Interspeech*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2047–2052. IEEE.

Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Lecture 6a overview of mini-batch gradient descent.

Xiaolei Huang and Michael J Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 694–699.

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

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The {Stanford} {CoreNLP} Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2008. Manual for the motivational interviewing skill code (misc 2.1). Unpublished manuscript. Retrieved from: https://casaa.unm.edu/download/misc.pdf.

William R Miller and Stephen Rollnick. 2012. Motivational interviewing: Helping people change (applications of motivational interviewing).

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence Ann, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1128–1137.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Bo Xiao, Dogan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Interspeech*, pages 908–912.