Machine Learning Methods for Discourse Segmentation of Communications in E-Mail Based Behavioral Interventions

Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], Sylvie Naar, PhD[2], Gwen L. Alexander, PhD[3], April Idalski Carcone, PhD[4]
[1]Department of Computer Science, Wayne State University, Detroit, Michigan
[2]Center for Translational Behavioral Research, Department of Behavioral Sciences and Social Medicine, Florida State University, Tallahassee, Florida
[3]Department of Public Health Sciences, Henry Ford Health System, Detroit, Michigan
[4]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *Communication science approaches to developing effective behavior interventions, such as motivational interviewing (MI), are limited by the traditionally manual qualitative coding of communication exchanges, which is a very resource-intensive and time-consuming process. This study focuses on the analysis of e-Coaching sessions, behavior interventions that are delivered via email and grounded in the principles of MI. A critical step towards automated annotation of e-Coaching communication exchanges is segmentation of emails into textual fragments that correspond to MI behaviors. In this work, we formulate this task as a classification problem and propose lexical, punctuation and part-of-speech features to address it. We experimented both with traditional machine learning method, conditional random fields (CRF) and deep learning methods, such as multilayer perceptrons (MLP), bidirectional recurrent neural networks (BRNN) and convolutional recurrent neural networks (CRNN). Results indicate that CRNN outperformed CRF, MLP and BRNN achieving 0.989 macro F1-score overall and 0.825 macro F1-score for detecting new segment.*

Introduction

The emergence of e-Health technologies opened up new ways to deliver a variety of behavioral interventions to any demographic group of patients in any geographical location. Motivational interviewing (MI), an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change[1–3], is one type of these interventions. MI sessions are generally aimed at eliciting "change talk", or statements of intrinsic motivation about patients' own desire, ability, reasons, need for and commitment to behavior change, which have been established by previous research[4] as a reliable mediator of health behavior change. However, communication science approaches to understanding the efficacy of MI are inherently limited by traditional qualitative coding methods.

Qualitative coding of motivational interviews with pre-defined codes has been traditionally performed manually by trained annotators, which is a tedious and resource-intensive process that involves several iterations of reading, comprehension and interpretation of interview transcripts. Rapidly developing computational technologies, specifically, machine learning methods, offer a unique opportunity to accelerate this process. In particular, machine learning methods have been successfully applied to a variety of analytical tasks involving textual data, such as classification[5] and sentiment analysis[6]. In our previous work, we examined the utility of machine learning methods for automated annotation[7,8] and analysis[9] of in-person MI sessions. Specifically, we demonstrated that machine learning methods can be utilized for annotation of MI transcripts according to a simple communication code scheme with the accuracy comparable to human coders[7]. Experimental data utilized in these studies, however, were prepared by transcribing audio conversations, which were clearly segmented into utterances by a counselor, a patient and a caregiver.

In this study, we focus on the analysis of e-Coaching sessions, behavior interventions that are delivered via email and grounded in the principles of motivational interviewing. Specifically, the e-Coaches involved in this study used emails to communicate motivation-enhancing messages that encourage healthy eating among GenY adolescents. e-Coaching data is comprised of email responses, which are free-text documents, unlike more traditional dyadic clinical interviews that are naturally segmented into utterances due to their conversational nature.

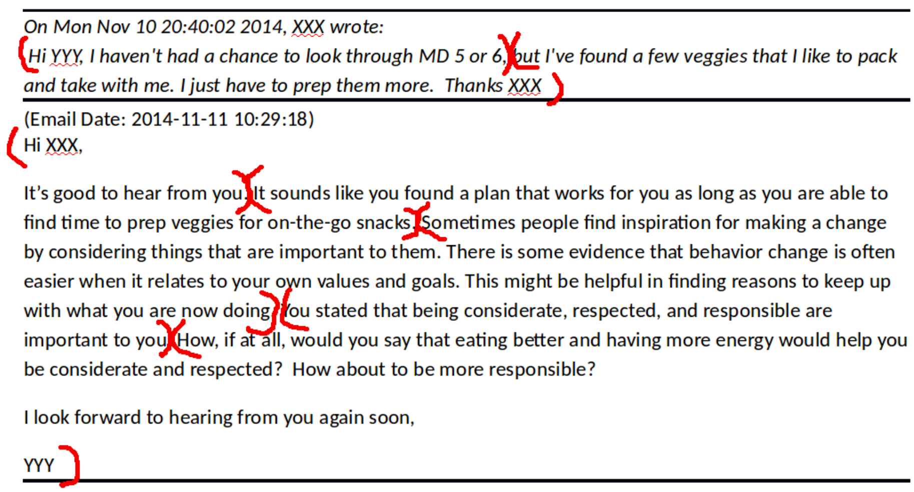
The unstructured nature of e-Coaching exchanges poses a unique set of challenges for their qualitative analysis. A

[*]Authors provided an equal contribution.

significant barrier to fully automating the behavior coding process of e-Coaching emails is their segmentation into textual fragments that correspond to distinct communication behaviors. Automating this task is a unique and challenging problem due to the following major reasons:

1. Emails are unstructured text that contains informal information exchange in a non-traditional format.
2. Discourse segments in e-Coaching emails do not necessarily correspond to sentences or collection of sentences. One sentence can be segmented into multiple MI behavior fragments. On the other hand, an MI behavior may comprise several sentences.

Figure 1 illustrates a segmentation of an e-Coaching email exchange, in which the first sentence is segmented into 2 MI behavior fragments, while the fourth and fifth segments correspond to one and three sentences, respectively. Segmentation of e-Coaching emails corresponds to a special type of discourse analysis[10] aimed at better understanding the effective e-Coaching communication strategies and revealing the unique socio-psychological characteristics of a patient.



On Mon Nov 10 20:40:02 2014, XXX wrote:
Hi YYY, I haven't had a chance to look through MD 5 or 6, but I've found a few veggies that I like to pack and take with me. I just have to prep them more. Thanks XXX

(Email Date: 2014-11-11 10:29:18)
Hi XXX,

It's good to hear from you. It sounds like you found a plan that works for you as long as you are able to find time to prep veggies for on-the-go snacks. Sometimes people find inspiration for making a change by considering things that are important to them. There is some evidence that behavior change is often easier when it relates to your own values and goals. This might be helpful in finding reasons to keep up with what you are now doing. You stated that being considerate, respected, and responsible are important to you. How, if at all, would you say that eating better and having more energy would help you be considerate and respected? How about to be more responsible?

I look forward to hearing from you again soon,

YYY

Figure 1: Example of e-Coaching emails segmented into fragments that correspond to MI behaviors of an e-Coach and a patient.

The goal of this research study is to assess the applicability of machine learning methods for automated segmentation of e-Coaching emails into textual fragments corresponding to individual behaviors, which is the first step of the coding process of e-Coaching communications. In particular, we introduced lexical, punctuation and part-of-speech (POS) features and experimented with both traditional supervised machine learning method, linear-chain conditional random fields (CRF) and deep learning methods, such as multilayer perceptrons (MLP), bidirectional recurrent neural networks (BRNN) and convolutional recurrent neural networks (CRNN), to find the best performing method and feature combination.

Relevant previous work in the biomedical domain primarily focused on segmentation of text into sections and headers[11–14] or sentence boundary detection[15–18]. Apostolova et al.[11] applied SVM along with word-vector cosine similarity metric combined with several heuristics to segment clinical reports into sections, such as demographics, history, procedure, finding and impression. After identification of each line in the document, Tepper et al.[13] trained Maximum Entropy models for section classification. Denny et al.[12] proposed a SecTag algorithm, which combined natural language processing techniques, terminology-based rules and Naive Bayes classifier to identify the sections and headers that achieved 99% recall with 95.6% precision. On the other hand, SVM based on prosodic and part of speech features[16] and recurrent convolutional neural networks using word embeddings[15] were utilized for detecting sentence boundaries. Liu et al. have shown that a linear chain CRF performed better than the hidden markov model and Maxent on the NIST sentence boundary detection[19]. Recently, Treviso et al. proposed a recurrent convolutional neural network for automatic identification of sentence boundary in a neuropsychological text in Portuguese language[17]. Their proposed model combined prosodic and part of speech features with pre-trained GloVe word vectors[20], which achieved 0.74 F1-score in sentence boundary detection in the Brazilian constitution dataset.

In this study, we adopt the model proposed by Treviso et al.[17] and apply it differently from the aforementioned studies. The main difference is that our study focuses on segmentation of e-Coaching emails as distinct from determining the sentence boundary. Another important difference is that punctuation features were utilized in combination with word embeddings and part-of-speech features. Furthermore, we train our model with pre-trained word2vec word vectors[21], instead of using GloVe word vectors. Segmentation of e-Coaching emails is also different from a traditional shallow discourse analysis of conversations[22] in that the focus is on segmentation, rather than on determining the types of transitions between the utterances or assigning utterances to speakers.

Recently, an online clinical intervention called MENU GenY[23] (Making Effective Nutrition Choices for Generation Y) was proposed and evaluated. MENU GenY is a technology-based public health intervention that relies on personalized e-coaching to encourage increased fruit and vegetable intake among young adults, aged 21-30. The goal of MENU GenY was to develop a better coding dictionary among GenY to improve eating habits. However, segmentation of clinical conversation in the context of electronically delivered interventions, in particular, segmentation of clinical interaction text into groups of MI behaviors, is still performed manually, which slows down qualitative analysis of these interventions. This study proposes several machine learning methods to address this problem and the authors are unaware of any other work that focused on the same problem.

Methods

Data collection

The experimental dataset for this work was constructed from 49 e-coaching sessions, which include a total of 3,138 segmented and annotated MI behaviors. Each session represents an MI intervention delivered via email. We formulate the segmentation task as a binary classification problem where each word or punctuation marks annotated with a label 1 or 0, to indicate whether it precedes a new segment or not. In total, we obtained 95,777 words and 7,140 punctuation marks, which include 3,138 "new segment" and 99,779 "same segment" instances. In this study, we experimented with traditional machine learning method, CRF and deep learning methods, such as MLP, BRNN and CRNN. For MLP models, samples were created with $2l$ words/punctuation marks at each step in a word sequence, where a sample contains next l words/punctuation marks and prior l words/punctuation marks including current word/punctuation mark. Each sample is classified into either "new segment" or "same segment" based on whether the current word/punctuation mark precedes a new segment or not. For CRF, BRNN and CRNN models, an email was taken as input sequence, such that POS tags and word embedding of each word or punctuation marks were used as input and binary labels (1 or 0) corresponding to "new segment" and "same segment" classification decision were considered as the output of the model at each step. In the gold standard, words within the same segment were assigned the label of 0 and the last word or punctuation mark of a segment were assigned the label of 1.

Features

We utilized three types of features in conjunction with CRF, MLP, BRNN and CRNN methods: word embeddings, punctuation and POS features. Since POS tags have been shown to be effective semantic abstractions of individual words, we used topic features for our experiment[17,19]. To extract our POS features, we tag the e-Coaching emails using the NLTK POS tagger[24]. Punctuation marks, which correspond to one of the symbols {'.', ',', '!', '?', ':', ';'} between a pair of words, are also employed as a feature since punctuation marks designate the boundary of a sentence, clause and phrase and often also correspond to a segment boundary. For natural language processing (NLP) tasks, inputs are received as a text where words can be considered as the basic processing unit. Therefore, it is important to represent a word in such a way that it carries all relevant information. Word embeddings are one of this representation where each word represented as a real-valued vector in a high dimensional vector space. The distributed representation of word allows capturing semantic, syntactic and morphological information from large unannotated corpora[20,21]. Since pre-trained word2vec word vectors provide the best results, in our experiment, we utilized pre-trained word2vec word vectors for words. Trained word vectors were also utilized when punctuation marks are not available in the pre-trained word2vec word vectors. Figure 2 illustrates the performance of CRNN model on e-Coaching email segmentation by varying the dimension of pre-trained and trained word embeddings with GloVe and word2vec models.

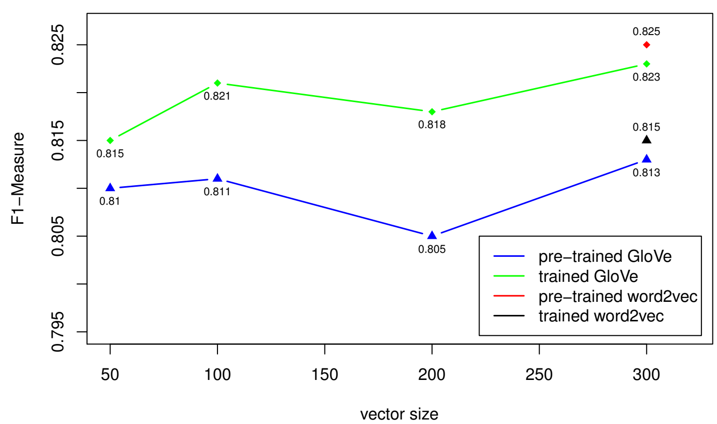


Figure 2: Performance of CRNN model on e-Coaching email segmentation by varying the dimension of pre-trained and trained word embeddings with GloVe and word2vec models.

Classifiers

In the present study, we experimented with four different classifiers, including one traditional machine learning method, CRF and three neural networks models. Since neural networks provide a flexible architecture for constructing and synthesizing complex models, we take the advantage of this flexibility to test MLP, BRNN and CRNN models for the task of segmentation of e-Coaching emails.

Conditional Random Fields: CRF have been widely used in various NLP tasks including part-of-speech tagging and segmentation tasks[25,26]. Unlike a maximum entropy markov model which uses per-state exponential models for the conditional probabilities of next states given the current state, a CRF is a discriminative model of the conditional distribution of the entire sequence of labels given the observation sequence. A traditional linear-chain CRF model is defined as a conditional probability distribution $p(y|x)$ for output and input sequences, y and x, by the Eq. 1

$$p(y|x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{1}$$

Here Z_x is a normalization factor over all possible labelings of x, and $f_k(y_{t-1}, y_t, x, t)$ is a feature function, and λ_k is a learned weight associated with feature f_k. The optimal output sequence y^* for an input sequence x, $y^* = arg\, max_y p(y|x)$, is obtained efficiently by the Viterbi algorithm. In our experiments, the following features were utilized with CRF models: i) current word/punctuation ii) next and previous 3 words/punctuations iii) whether the word/punctuation is special characters (';', '?', '.', ',', '!', ':', etc) or not iv) whether the word is title or not? (e.g. "The" is a title word but "the" is not) v) POS tags

Multilayer Perceptrons: MLP is a feed-forward artificial neural network model which maps set of input onto a set of appropriate output. This study utilized MLP with a nonlinear activation function (relu) and one hidden layer consisting of 128 hidden units. In order to prevent over-fitting, we implement dropout on fully connected layers, which randomly hide some neurons during the training phase[27]. Dropout is also applied in the fully connected layers in BRNN and CRNN models. For a given sample with $2l$ words/punctuations, the first layer input of MLP was created by the summation of first l word vectors and last l word vectors. Figure 3 demostrates the performance of MLP model on e-Coaching email segmentation by varying value of l.

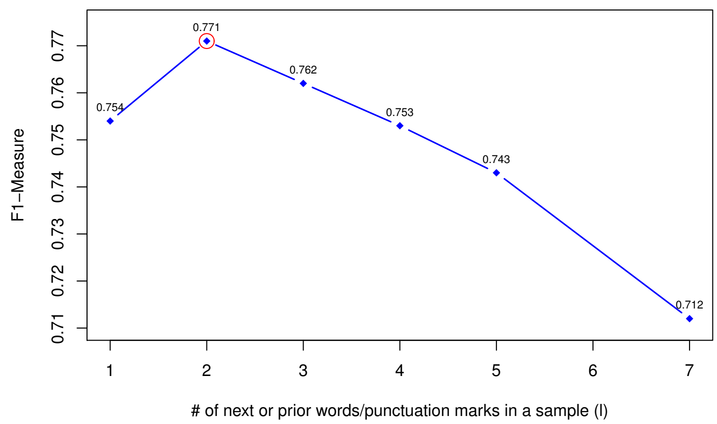


Figure 3: Performance of MLP model on e-Coaching email segmentation by varying the length of sample size.

Bidirectional Recurrent Neural Networks: BRNN is a neural network designed to capture sequential patterns by considering both past and future inputs as well as complex relationships between input feature and output labels. The output of BRNN layer is computed as the summation of the forward RNN output with backward RNN output. Gated Recurrent Units (GRU)[29] are a special type of RNN that are capable of handling variable size input sequence and have an internal memory that can be reset. Figure 4 will represent an architecture of BRNN if convolution layer is removed.

Convolutional Recurrent Neural Networks: CRNN is a deep neural networks architecture, shown in Figure 4, consisting of 5 layers: 1) input layer 2) embedding layer 3) convolution layer with max pooling 4) BRNN layer 5) fully connected layer with dropout and sigmoid output.

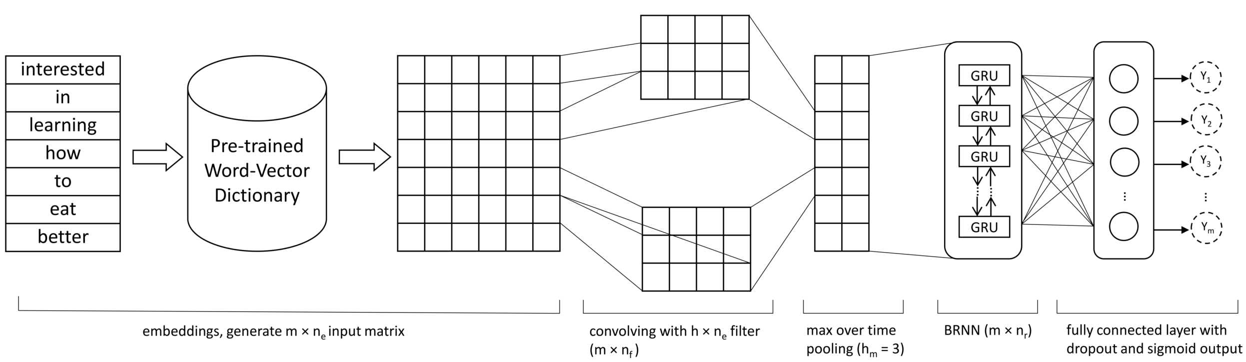


Figure 4: Architecture of convolutional recurrent neural networks for automatic segmentation of e-Coaching emails.

A sequence of m words is fed into the input layer which provides a $m \times n_e$ matrix after fetching the pre-trained word vectors. When POS tags are used in combination with word embedding, 10-dimensional POS vectors are concatenated with 300-dimensional word vectors to obtain new vectors $n_e = n_w + n_p$ of size 310. The 1D convolutional layer generates $m - h_c + 1$ new features with $h_{c.n_e}$, a filter of h_c words and zero-padding on both sides of the text. A pooling operation was performed over time to find the most significant features in a region. After the convolutional layer, $m \times n_f$ are fed into bidirectional recurrent neural networks with 200 hidden units. Dropout is used to randomly drop 50% of the neurons and passes them through a fully connected layer, where a sigmoid operation is calculated, giving us the probability of whether the word precedes a new segment or not. We experimentally determined that the best performance is achieved when filter length in convolution layer is 7, pool size is 3, batch size is 32, adam[30] is used for optimization and the early stopping strategy is used with dropout to prevent over-fitting. Source codes of all experiments will be available at https://github.com/teanalab/xxxxx upon acceptance of this paper for publication.

Evaluation metrics

We report standard metrics for experiments (precision, recall and F1-measure) to evaluate the performance of binary classifiers[31]. However, accuracy is not reported as a performance metric because accuracy is highly sensitive to the prior class probabilities and does not fully describe the actual difficulty of the decision problem for an unbalanced dataset. The results are reported based on 5 folds cross-validation and weighted macro-averaging over the folds. We also estimate the area under the receiving operating characteristics (ROC) curve[32] (AUC) metric due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[33].

Results

Our experimental results have two important directions. First, results are reported with respect to "new segment" and "same segment" classes as well as their weighted average in Table 1. Second, classification performance of different machine learning methods are outlined in Tables **??**, **??** and 2 when topic features are extracted with different probabilistic generative latent variable models.

Table 1: Performance of CRF, MLP, BRNN and CRNN methods for identification of "new segment" class as well as weighted average over "new segment" and "same segment" classes when word embeddings or lexical features are used. The highest value for each performance metric is highlighted in bold.

Method	**new segment**			**weighted macro average**		
	Precision	**Recall**	**F1-Measure**	**Precision**	**Recall**	**F1-Measure**
CRF	0.782	0.691	0.733	0.983	0.984	0.984
MLP	**0.836**	0.593	0.694	0.982	0.983	0.982
BRNN	0.606	0.680	0.641	0.977	0.976	0.976
CRNN	0.775	**0.797**	**0.785**	**0.986**	**0.986**	**0.986**

As follows from Table 1, KNN outperforms all other models in terms of precision and F1-measure achieving 0.808 precision with 0.728 F1-measure for new segment detection. KNN also shows superior performance in all performance metrics for "same segment" class and weighted average over "new segment" and "same segment" classes. NB demonstrates the lowest performance among all models in terms of all performance metrics. On the other hand, SVM has the highest recall of 0.705 when only lexical features are used to identify "new segment". Results indicate that performance of classifiers is remarkably higher for "same segment" class compared to "new segment" class, which is expected since 96.71% instances belong to "same segment" categories. For example, KNN achieves 22.40%, 50.07% and 36.26% higher precision, recall and F1-measure, respectively, in "same segment" identification compared to "new segment" detection. When lexical features are used in combination with punctuation mark, LSTM demonstrates the best performance in terms of precision and F1-measure for "new segment" identification while GRU shows 2.64% higher recall than LSTM. GRU and LSTM show exactly same performance in terms of all performance metrics for "same segement" and overall classifications.

Table 2 summarizes the weighted average results of the models for segmentation of e-Coaching emails. Overall, SVM has the best performance across all topic model-based features in terms of all performance metrics, achieves 0.990 precision, recall and F1-measure when DL-LDA-based topic features are used. Similar to results in Tables 1, **??** and **??**, NB demonstrates the lowest performance among all methods while best performance achieves with LCA-based topic features compared to other model-based topic features. Influence of the additional features is also consistent with the results in Tables **??** and **??**. Precision increases by 0.9% and 0.2%; recall increases by 0.9% and 0.2%; and F1-measure increases by 0.9% and 0.3% for SVM and KNN methods, respectively, when lexical feature is used in

Table 2: Performance of CRF, MLP, BRNN and CRNN methods for identification of “new segment” class as well as weighted average over “new segment” and “same segment” classes when all features are used together. The highest value for each performance metric is highlighted in bold.

Method	new segment			weighted macro average		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
CRF	0.813	0.772	0.792	0.988	0.988	0.988
MLP	**0.814**	0.732	0.771	0.987	0.987	0.987
BRNN	0.683	0.820	0.745	0.985	0.983	0.984
CRNN	0.789	**0.864**	**0.825**	**0.990**	**0.989**	**0.989**

combination with punctuation and best model-based topic features.

Discussion

This study is the first effort to evaluate the automatic segmentation of e-Coaching emails. Experimental results indicate that SVM is the best model among all machine learning methods considered for this study. SVM achieved 0.990 F1-measure in overall, 0.845 and 0.995 F1-measures for detecting “new segment” and “same segment”, respectively. The robust performance of SVM provides the evidence that machine learning models are capable to learn conceptual information from clinical exchange. It also indicates that topic features are more important than lexical features, even when deep learning methods are employed. Although the domain of this study was intentionally quite small, we believe that our study is not limited to the e-Coaching domain and it can be successfully applied to other domains, in which discourse segmentation is a preliminary step for annotation.

Punctuation mark and topic model-based features made a significant improvement in performance of all machine learning methods. Among all topic model-based feature, the DL-LDA-based topic feature provides the highest performance in “new segment” and “same segment” classification as well as weighted average over “new segment” and “same segment” classification results. In nearly all cases, ML methods perform better, when lexical features are used in combination with punctuation and topic model-based features. This results also indicate that segmentation performance might be improved by adding additional relevant features.

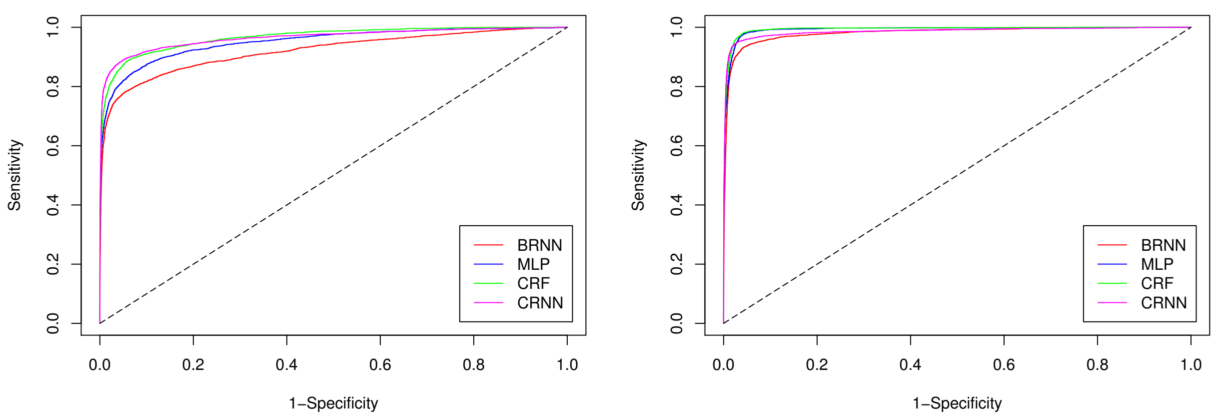


Figure 5: Receiver operating characteristic curves showing the performance of binary classifiers for the segmentation of e-Coaching text when only word embeddings/lexical (left) and combination of word embeddings/lexical, POS and punctuation features (right) are used.

In this paper, results are reported by each class to avoid confusion regarding overall model performance due to severe class imbalance. In addition, standard metrics: precision, recall and F1-measure were used to eliminate doubt about the model performance since accuracy is misleading for imbalanced datasets. AUC is also utilized due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[33], which are demonstrated by the ROC curves in Figure 5. Although NB provides lowest classification results, it shows moderate AUC values, achieving 0.978 AUC when only lexical features are used and 0.977 AUC when a combination of lexical, punctuation and topic features are used. On the other hand, KNN and SVM exhibit 0.972 and 0.986 AUCs when only lexical features are used; and 0.966 and 0.980 AUCs when a combination of lexical, punctuation and topic model-based features are used. Finally, LSTM demonstrates the highest AUC value among all machine learning models, achieving 0.996 AUC, while GRU achieves only 0.2% lower AUC than LSTM.

Table 3: AUC values of all classifiers when only word embeddings/lexical and combination of word embeddings/lexical, POS and punctuation features are used. Highest value for each experiment across all models is highlighted in boldface.

Features	**Area Under the ROC Curve (AUC)**			
	CRF	**MLP**	**BRNN**	**CRNN**
word embeddings or lexical features	**0.966**	0.951	0.920	0.965
all features	**0.994**	0.992	0.981	0.986

We observed moderate performance of RNN, in particular, LSTM and GRU for the text segmentation. We believe that RNN will perform better if a larger dataset is utilized. In this study, we employed 3,138 examples of boundary case which limit to achieve the best performance. We also observed that GRU performs better than LSTM, which was observed in previous studies[29].

Although punctuation mark plays an important role in segmentation boundary detection and large numbers of errors were encountered by the false positive of boundary identification. For example, a text segment from an e-Coaching email *"A typical day in regards to fruit and vegetable has me eating about a serving at breakfast (our caf has cut up fruit) and then maybe a piece of fruit later in the day or as a snack. Vegetable tends to be a side serving at lunch and dinner and I get celery or carrot cuts with dressing for a snack a lot of times. I could probably add some sort of vegetable into my breakfast (like spinach in an omelet) and snack on another piece of fruit when I quotemark m hungry rather than the junk food I tend to eat."* incorrectly segmented after the first sentence where a punctuation mark was encountered. Similarly, additional information is a common cause for misclassification of an email segment into multiple segments. For instance, the first sentence of the above email segment represents a positive commitment to adolescent's behavior change then next two sentences provide an additional information to support their commitment.

Our proposed approach is novel for the segmentation of e-Coaching emails since previous studies mainly focused on the segmentation of text into sections, headers and sentences. However, this study segmented clinical exchange into groups of MI behaviors which will significantly reduce the amount of resource and time required to segment clinical exchange manually. Furthermore, these segmentation models could be integrated with auto coding classifiers to build a software pipeline for automated annotation of exchanges.

The limitation of this study is that e-Coaching data is collected from a single medical institute; formatting, style and email segment can be different in other settings. Therefore, there is a need to replicate the experiments with different data sets. As our future work, we plan to evaluate our approach to other datasets for discourse analysis.

Conclusion

Segmentation of e-Coaching emails is an integral part of developing e-Coaching interventions. Although several studies have focused on clinical interventions, they are limited by the qualitative coding of clinical interactions. In addition, previous studies in the medical domain mainly segmented clinical text into sections and sentences, none of them considered segmentation of text into groups of MI behaviors in the setting of discourse analysis with emails. In this paper, we compared the performance of machine learning models for the task of segmentation of e-Coaching text. We found out that CRNN provides the best performance for the segmentation of text in terms of all performance metrics. Manual segmentation of e-Coaching data is very resource-intensive and time-consuming task, which can significantly decrease the time and effort required to develop an effective behavioral intervention. Our proposed methods can help to identify individual text segments, which can be annotated directly with a classification model. This approach will also help for developing fully automated e-Coaching and accelerate the pace of identifying effective communication strategies.

Acknowledgments

This study was supported by a grant from the National Institutes of Health, NIDDK R21DK108071, Carcone and Kotov, MPIs. We would like to thank research staff and student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help in preparing the training dataset.

References

[1] Miller WR, Rollnick S. Motivational interviewing: Helping people change. Guilford press; 2012.

[2] Miller WR, Rollnick S. Ten things that motivational interviewing is not. Behavioural and cognitive psychotherapy. 2009;37(2):129–140.

[3] Miller WR, Rose GS. Toward a theory of motivational interviewing. American psychologist. 2009;64(6):527.

[4] Apodaca TR, Longabaugh R. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. Addiction. 2009;104(5):705–715.

[5] Nigam K, McCallum AK, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. Machine learning. 2000;39(2-3):103–134.

[6] Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics; 2012. p. 90–94.

[7] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics. 2016;62:21–31.

[8] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, Hartlieb KB. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 785.

[9] Hasan M, Kotov A, Carcone AI, Dong M, Naar-King S. Predicting the outcome of patient-provider communication sequences using recurrent neural networks and probabilistic models. In: Proceedings of the 2018 AMIA Informatics Summit. American Medical Informatics Association; 2018. .

[10] Webber B, Egg M, Kordoni V. Discourse structure and language technology. Natural Language Engineering. 2012;18(4):437–490.

[11] Apostolova E, Channin DS, Demner-Fushman D, Furst J, Lytinen S, Raicu D. Automatic segmentation of clinical texts. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. IEEE; 2009. p. 5905–5908.

[12] Denny JC, Spickard III A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association. 2009;16(6):806–815.

[13] Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. In: LREC; 2012. p. 2001–2008.

[14] Cho PS, Taira RK, Kangarloo H. Text boundary detection of medical reports. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2002. p. 998.

[15] Griffis D, Shivade C, Fosler-Lussier E, Lai AM. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. AMIA Summits on Translational Science Proceedings. 2016;2016:88.

[16] Kreuzthaler M, Schulz S. Detection of sentence boundaries and abbreviations in clinical narratives. In: BMC medical informatics and decision making. vol. 15. BioMed Central; 2015. p. S4.

[17] Treviso M, Shulby C, Aluísio S. Sentence Segmentation in Narrative Transcripts from Neuropsychological Tests using Recurrent Convolutional Neural Networks. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. vol. 1; 2017. p. 315–325.

[18] Fraser KC, Ben-David N, Hirst G, Graham N, Rochon E. Sentence segmentation of aphasic speech. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2015. p. 862–871.

[19] Liu Y, Stolcke A, Shriberg E, Harper M. Using conditional random fields for sentence boundary detection in speech. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05); 2005. p. 451–458.

[20] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

[21] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

[22] Galley M, McKeown KR, Fosler-Lussier E, Jing H. Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics; 2003. .

[23] Alexander GL, Lindberg N, Firemark AL, Rukstalis MR, McMullen C. Motivations of Young Adults for Improving Dietary Choices: Focus Group Findings Prior to the MENU GenY Dietary Change Trial. Health Education & Behavior. 2017;p. 1090198117736347.

[24] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009.

[25] Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning; 2001. p. 282–289.

[26] Hirohata K, Okazaki N, Ananiadou S, Ishizuka M. Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I; 2008. .

[27] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014;15(1):1929–1958.

[28] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–1780.

[29] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014;.

[30] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

[31] Aas K, Eikvil L. Text categorisation: A survey. Technical report, Norwegian computing center; 1999.

[32] Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian pediatrics. 2011;48(4):277–287.

[33] Hu J, Yang H, King I, Lyu MR, So AMC. Kernelized Online Imbalanced Learning with Fixed Budgets. In: AAAI; 2015. p. 2666–2672.