Deep Neural Architectures for Discourse Segmentation in E-Mail Based Behavioral Interventions

Mehedi Hasan, BS[1a], Alexander Kotov, PhD[1a], 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 develop effective behavior interventions, such as motivational interviewing (MI), are limited by traditional qualitative coding of communication exchanges, a very resource-intensive and time-consuming process. This study focuses on the analysis of e-Coaching sessions, behavior interventions delivered via email and grounded in the principles of MI. A critical step towards automated qualitative coding of e-Coaching sessions is segmentation of emails into fragments that correspond to MI behaviors. This study frames email segmentation task as a classification problem and utilizes word and punctuation mark embeddings in conjunction with part-of-speech features to address it. We evaluated the performance of conditional random fields (CRF) as well as multi-layer perceptron (MLP), bi-directional recurrent neural network (BRNN) and convolutional recurrent neural network (CRNN) for the task of email segmentation. Our results indicate that CRNN outperforms CRF, MLP and BRNN achieving 0.989 weighted macro-averaged F1-measure and 0.825 F1-measure for new segment detection.*

Introduction

The emergence of e-Health technologies has greatly expanded the reach of behavioral interventions. One such intervention is email-delivered Motivational Interviewing (MI). MI is an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change[1,2]. MI is linked to behavior change through the elicitation of patient "change talk", or statements of intrinsic motivation expressing patients own desires, abilities, reasons, need for, and commitment to behavior change[3]. However, communication science approaches to understanding the efficacy of MI are inherently limited by traditional qualitative coding methods. The specific provider behaviors responsible for the elicitation of change talk, however, are less clear and may vary by treatment context. Thus, current MI research is focused on understanding which provider behaviors and in which contexts lead to patient change talk. To date, no study has examined this process in email-delivered MI.

The study of patient-provider communication in MI relies on the behavioral coding of intervention transcripts, an iterative, resource-intensive, and cognitively demanding process of qualitative analysis involving several iterations of reading, comprehension and interpretation of interview transcripts. Rapidly developing computational technologies, specifically, machine learning methods, offer an opportunity to dramatically accelerate this process. In particular, machine learning methods have been successfully applied to a variety of analytical tasks involving textual data, such as classification[4] and sentiment analysis.[5] In our previous work, we examined the utility of machine learning methods for automated annotation[6,7] and predicting the outcome[8] of in-person MI sessions. Experimental data utilized in these studies were transcribed auto recordings of in-person MI sessions with a counselor, which were segmented into counselor and client utterances during the transcription process.

In this study, we focus on the analysis of email-delivered MI, or e-Coaching, to promote healthy eating among young adults. The e-Coaching dataset is composed of email correspondence between an MI counselor and the young adult patient. Unlike transcribed in-person exchanges, email correspondence is not clearly segmented into codable speech acts (i.e., utterances). Thus, the unstructured nature of e-Coaching exchanges poses a unique set of analytic challenges. Segmentation of e-Coaching exchanges into textual fragments that correspond to distinct e-Coach and patient communication behaviors is a significant barrier to qualitative analysis of this type of clinical conversation. Automating this task is a unique and challenging problem due to the following reasons:

---

[a]These authors provided an equal contribution.

1. Emails are unstructured text containing informal information exchange in a non-traditional format. For example, an e-Coach usually responds to several previous patient statements in one email. In contrast, in a traditional, in-person MI session, each utterance is assumed to be a response to an immediately preceding utterance.

2. Discourse segments in e-Coaching do not have a clear breakpoint, such as the end of a sentence or a paragraph. One sentence may be divided into fragments corresponding to multiple MI behaviors. On the other hand, an MI behavior may comprise several sentences.

Figure 1: Example of an e-Coaching exchange segmented into fragments corresponding to MI behaviors of an e-Coach and a patient

Figure 1 illustrates a segmentation of an e-Coaching exchange, in which the first sentence is segmented into 2 MI behavior fragments, while the fourth and fifth MI behavior fragments comprise one and three sentences, respectively. Segmentation of e-Coaching exchanges constitutes a special case of clinical discourse analysis[10] aimed at better understanding the effective communication strategies specific to this type of behavioral interventions.

The goal of this study is to assess the effectiveness of deep learning methods for the task of automated segmentation of e-Coaching emails into textual fragments corresponding to individual patient and provider behaviors. For this study, we utilized the data from MENU GenY (Making Effective Nutrition Choices for Generation Y)[11], a web-delivered public health intervention with email-based coaching to encourage increased fruit and vegetable intake among young adults, aged 2130. A secondary goal of the MENU GenY project was to identify the specific communication strategies used by e-Coaches to elicit change talk for healthier eating among young adult patients. Segmentation of clinical conversation in the context of electronically delivered interventions into groups of MI behaviors is traditionally performed manually by MI researchers, which significantly slows down its qualitative analysis. *This paper is the first work to evaluate the empirical effectiveness of deep learning architectures in addressing the problem of discourse segmentation in the context of email-based behavioral interventions.*

Specifically, we evaluate the effectiveness of distributed representations (i.e. embeddings) of words and punctuation marks as well as part-of-speech (POS) features in conjunction with both traditional supervised machine learning methods, such as linear-chain Conditional Random Fields (CRF)[12] and deep learning methods, such as Multi-Layer Perceptron (MLP),[13] Bidirectional Recurrent Neural Network (BRNN)[14] and Convolutional Recurrent Neural Network (CRNN),[15] to determine the best performing method and feature combination for the task of segmentation of e-Coaching emails into MI behaviors.

Related work

Prior work on textual segmentation in the biomedical domain primarily focused on sentence boundary detection[15,20,21] and segmentation of clinical documents in patients' electronic health records (EHR) into sections and headers.[16–19] In particular, maximum entropy models[18] and Support Vector Machine (SVM) along with word vector similarity metrics

and several heuristics[16] have been applied to identify specific sections in EHR, such as general patient information, medical history, procedures, findings, etc. Denny et al.[17] proposed SecTag algorithm, which combined natural language processing techniques, terminology-based rules and a Naïve Bayes classifier to identify sections and headers in EHR. Segmentation of e-Coaching emails, however, is different from segmentation of other clinical documents, since the focus is on dialog acts in clinical conversation.

SVM in conjunction with prosodic and part-of-speech features[21] and recurrent convolutional neural networks[20] have also been utilized for *sentence boundary detection* in general text. Liu et al.[22] demonstrated that a linear-chain CRF outperforms Hidden Markov and maximum entropy models for this task.

Segmentation of e-Coaching emails is also different from traditional shallow discourse analysis[23], which besides identification of speech acts, also aims to determine the types of transitions between speech acts and label speech acts with the speakers who performed them in a multi-speaker conversation. The proposed methods will automate the process of segmenting clinical exchanges into MI behaviors, which will significantly reduce the time and resources required to perform such segmentation manually. Furthermore, these methods can be integrated with the automated MI behavior coding methods[6,7] to create a software pipeline for fully automated analysis of email-delivered behavioral interventions.

Materials and methods

Dataset

The experimental dataset for this work was constructed from 49 e-Coaching sessions, which include 330 and 281 emails by e-Coaches and patients, respectively. Various statistics of the experimental dataset are provided in Table 1. Each e-Coaching session represents an MI intervention delivered via email. Emails were segmented into 3,138 text fragments and annotated with the Minority Youth-Sequential Coding of Process Exchanges (MY-SCOPE),[9] a qualitative coding scheme to characterize patient-provider communication during MI sessions with 115 distinct behavior codes. Email segmentation can be considered as sequence tagging, which can be framed as a binary classification problem, in which each word or punctuation mark is annotated with one of the two class labels ("new segment" or "same segment") to indicate whether it is a beginning of a new MI behavior segment or not. In total, the dataset consists of 95,777 words and 7,140 punctuation marks and includes 3,138 "new segment" and 99,779 "same segment" instances. In this study, we experimented with traditional machine learning methods, such as Conditional Random Fields (CRF)[12] and deep learning methods, such as Multi-Layer Perceptron (MLP),[13] Bi-directional Recurrent Neural Network (BRNN)[14] and Convolutional Recurrent Neural Network (CRNN).[15] In the case of MLP, training and testing samples were created based on a sliding window of $2n$ words or punctuation marks over each position (which could be a word or a punctuation mark) in a given input sequence, such that each sample consists of the n words or punctuation marks after the current position and n words or punctuation marks prior to the current position, including the position itself. In the case of CRF, BRNN and CRNN models, an e-Coaching email was taken as an input sequence, POS tags and embeddings of each word or punctuation mark were used as input and binary labels corresponding to "new segment" and "same segment" classification decisions were considered as the model output. In the gold standard, words or punctuations 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 (an example of an input with the assigned labels is provided in the 2nd row of Table 1).

Table 1: Summary of statistics of the experimental dataset and example of a segmented sequence

Sessions	Instances	Class labels		Tokens		Emails		Annotation	
		new	same	words	punctuations	patient	provider	method	codes
49	102,917	3,138	99,779	95,777	7,140	281	330	MYSCOPE	115
segmentation example: Hi[0] XXX[0] ,[0] its[0] good[0] to[0] hear[0] from[0] you[0] .[1] it[0] sounds[0] like[0]...									

Features

We utilized three types of features in conjunction with CRF, MLP, BRNN and CRNN: word embeddings as lexical features, punctuation and POS features. Syntactic abstractions of individual words, such as POS tags, have been

previously shown to be effective features for similar natural language processing tasks.[15,22] To extract POS features, we pre-processed e-Coaching emails using the NLTK POS tagger[b]. Punctuation marks, which correspond to one of the symbols {'.', ',', '!', '?', ':', ';'} between a pair of words, were also used as a feature, since punctuation marks designate the boundary of a sentence, clause or a phrase and often also correspond to a segment boundary.[19] For natural language processing (NLP) tasks, inputs are received as textual fragments, in which individual words are as the basic lexico-semantic units. Therefore, it is important to represent a word in such a way that preserves all relevant lexical and semantic information. Embedding is a form of distributed representation, when each word is associated with a dense real-valued vector in low-dimensional space. Embeddings have been previously shown to effectively capture semantic, syntactic and morphological properties of words.[24,25] For experiments reported in this paper, we utilized word embeddings pre-trained on Google News corpus consisting of 1.6 billion words using word2vec software package.[c] For words or punctuation marks, which do not have pre-trained embeddings, we utilized the embeddings of the same dimensionality trained on the experimental dataset. CRF utilized lexical features, POS tags and the preceding label.

Methods

We experimented with 4 different classifiers, including one traditional machine learning model (CRF) and three deep learning methods (MLP, BRNN and CRNN). Since deep learning architectures provide a flexible mechanism for constructing complex models, we take advantage of this flexibility to test different variations of MLP, BRNN and CRNN models for the task of segmentation of e-Coaching emails.

Conditional Random Fields (CRF): CRF has been widely used in various NLP tasks that involve sequence annotation, such as part-of-speech tagging.[12,26] Unlike the maximum-entropy Markov model, which uses per-state exponential models for conditional probability of the next state given a current state, CRF model directly estimates a distribution of the entire output sequence conditioned on the observation sequence. A linear-chain CRF model is defined as a conditional probability distribution $p(y|x)$ of output sequence y, given input sequence x:

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

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

Multi-Layer Perceptron (MLP): MLP is a neural network, which consists of multiple fully connected layers that map an input to one or several outputs.[13] Figure 2 illustrates a multi-layer perceptron with a single hidden layer. MLPs have no cycles or loops. Information in them flows only forward, from the input layer through the hidden layer(s) to the output layer. The MLP in this study utilizes one hidden layer consisting of 128 neurons and rectified linear unit (ReLU) as a nonlinear activation function. In order to prevent over-fitting, we applied dropout (random masking of neurons)[27] to fully connected layers during training. Dropout was also applied to a fully connected layer in CRNN.

Bi-directional Recurrent Neural Network (BRNN): BRNN is a neural network designed to capture sequential patterns by considering both past and future inputs as well as complex relationships between input features and output labels.[14] The hidden state of BRNN is an aggregation of the hidden states of a forward and backward recurrent neural networks (RNNs). Gated Recurrent Units (GRU)[28] capable of handling variable size input sequence and having internal memory, which can be reset, were utilized as an RNN in this work.

Convolutional Recurrent Neural Network (CRNN): CRNN[15] shown in Figure 3 is a deep neural architecture, which

---

[b]https://www.nltk.org/
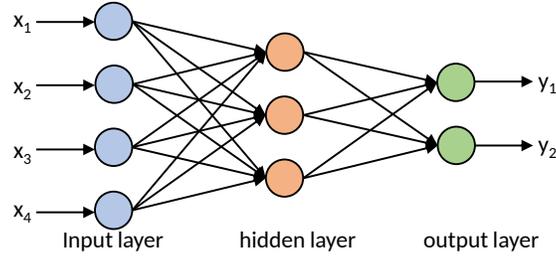[c]https://code.google.com/p/word2vec/

Figure 2: Multi-layer perceptron with a single hidden layer

combines convolutional and recurrent layers. Our implementation of CRNN consists 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. E-coaching email exchanges are represented as a sequence of m words and punctuations, which are fed into the input and embedding layers to produce a $m \times n_e$ matrix after fetching the embeddings for words and punctuations in the input sequence. This matrix is a distributed representation of an input email exchange, which contains rich morpho-syntactic information that can be utilized for its segmentation. When POS tags are utilized along with word embeddings, they are represented with a 10-dimensional vector, which is concatenated with 300-dimensional word embeddings to obtain new embedding vectors $n_e = [n_w; n_p]$ of size 310. The primary purpose of a convolution layer is to extract new features for each word or punctuation mark based on the neighboring words or punctuation marks. A one-dimensional (1D) convolution operation is utilized in this layer in our implementation of BRNN. In 1D convolution, one filter is responsible for the extraction of one feature. After applying n_f different filters with zero-padding on both sides of the input text, n_f features are produced by the convolution layer for each word. A max pooling over time operation is then applied to find the most significant features in a textual fragment. The bi-directional recurrent layer receives new features extracted from the convolution layer. Unidirectional RNNs are typically utilized to capture long-range dependencies in a sequence of observations. Bi-directional RNNs, on the other hand, are capable of capturing both past and future contexts through forward and backward traversals of a sequence. The purpose of the fully connected layer in CRNN is to use the output of the bidirectional RNN layer for classifying each word or punctuation into "new segment" or "same segment" classes. Since a fully connected layer has a larger number of parameters, they are more likely to excessively co-adapt to other parameters in the network and result in over-fitting. To prevent this, we utilized dropout by randomly ignoring 50% of the connections in the fully connected layer of CRNN. Finally, logistic sigmoid outputs the probability of classifying or labeling each word or punctuation mark with "same segment" class. We experimentally determined the optimal parameters using 5-fold cross-validation and found out that the best performance is achieved when filter length in the convolution layer is 7, number of filters is 100, max pooling size is 3, ReLU is used as an activation function in the convolution layer, hyperbolic tangent is used as an activation function in the bi-directional RNN layer and the number of dimensions in the hidden state of RNNs is 200. Adam[29] with 50 epochs, the batch size of 32 and learning rate of 0.001 was used for optimization and the early stopping strategy was applied.[d]

Evaluation metrics

We report standard metrics of precision, recall and F1-measure to evaluate the performance of the classifiers.[30] Accuracy is not reported as a performance metric, since it is highly sensitive to the distribution of prior class probabilities, which is skewed when datasets with unbalanced classes are involved. The results are reported based on 5-fold cross-validation (one fold was used as a test set and the remaining 4 folds were used as a training set) and weighted macro-averaging over the folds. We also report the area under the precision-recall curve (AUPR), due to its effectiveness in measuring the performance of binary classifiers in the case of the datasets with imbalanced class distribution.[31]

---

[d]source code of all methods is available at `https://github.com/teanalab/eCoaching-Text-Segmentation`
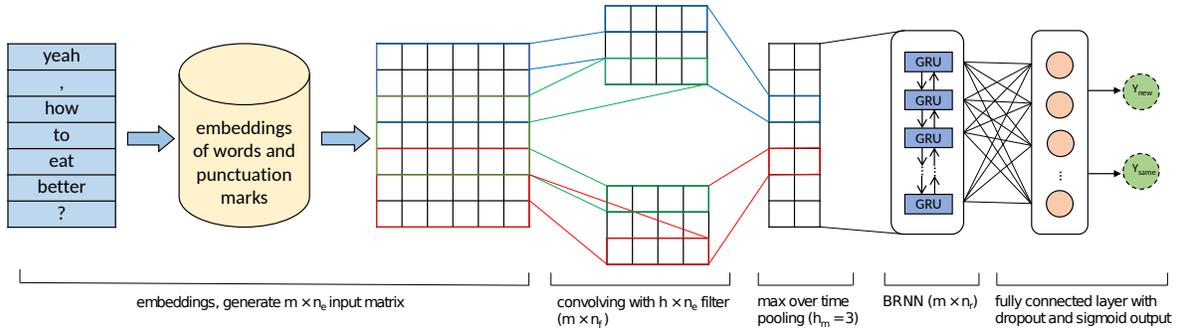
Figure 3: Architecture of a convolutional recurrent neural network for automated segmentation of e-Coaching emails into fragments corresponding to MI behaviors

Results

Our experiments spanned three dimensions. First, we determined the optimal sizes of word embedding vectors and the sliding window of MLP. Second, we evaluated the performance of different methods with respect to detecting "new segment" as well as the weighted average over "new segment" and "same segment" classes. Third, we assessed the impact of different types of features as well as their combination on the performance of different machine learning methods on the e-Coaching email segmentation task.
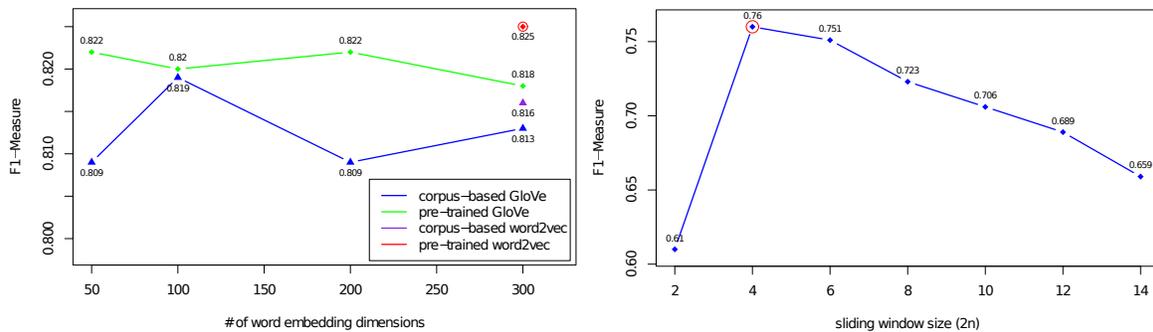

Figure 4: F1-measure of CRNN on the task of e-Coaching email segmentation by varying the number of dimensions in pre-trained and corpus-based GloVe and word2vec embeddings (left). F1-measure of MLP on the task of e-Coaching email segmentation by varying the size of the sliding window (right).

Figure 4 (left) illustrates the performance of CRNN on the task of e-Coaching email segmentation by varying the number of dimensions in pre-trained and corpus-based GloVe [e] and word2vec embeddings. We observed that the best performance is achieved with pre-trained 300-dimensional word2vec word vectors, when three types of features are used together. Therefore, we report the results for other deep learning models used in this study when 300-dimensional word2vec embedding vectors are utilized. The input layer of MLP consists of a sum of embeddings of n words or punctuation marks before and the sum of embeddings of n words or punctuation marks after the word or punctuation mark, which is the center of a sliding window of $2n$ words or punctuation marks. Figure 4 (right) demonstrates the performance of MLP on e-Coaching email segmentation by varying the size of the sliding window. It can be observed that the best performance of MLP is achieved when the size of the sliding window is 4 (or $n = 2$). Therefore, MLP results in the remaining experiments are reported when n is set to 2.

As follows from Table 2, CRNN outperforms all other methods in terms of recall and F1-measure achieving 0.797 recall and 0.785 F1-measure for new segment detection. CRNN also shows superior performance according to all performance metrics calculated as a weighted average over "new segment" and "same segment" classes. BRNN had

[e] https://nlp.stanford.edu/projects/glove/

Table 2: Performance of CRF, MLP, BRNN and CRNN on "new segment" detection as well as the weighted average over "new segment" and "same segment" classes when only lexical features are used. The highest value for each performance metric is highlighted in boldface.

Method	New Segment			Overall			AUPR
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	
CRF	0.782	0.691	0.733	0.983	0.984	0.984	0.780
MLP	**0.836**	0.593	0.694	0.982	0.983	0.982	0.736
BRNN	0.606	0.680	0.641	0.977	0.976	0.976	0.655
CRNN	0.775	**0.797**	**0.785**	**0.986**	**0.986**	**0.986**	**0.818**

the lowest performance among all models in terms of precision and F1-measure. On the other hand, MLP had the highest precision of 0.836 when lexical features are used to identify "new segment". CRF achieves 0.733 F1-measure, the second highest in identifying "new segment". CRF also demonstrated the second best performance among all models according to all metrics calculated as a weighted average over both classes. Experimental results indicate that the performance of all classifiers according to all metrics calculated as a weighted average over both classes is significantly higher than their performance on "new segment" detection, which is expected since 96.95% of instances belong to the "same segment" class and 99.3% of them are correctly classified. For example, CRNN achieves 27.23%, 23.71% and 25.61% higher precision, recall and F1-measure calculated as a weighted average over "new segment" and "same segment" classes, compared to the "new segment" detection.

Table 3: Performance of CRF, MLP, BRNN and CRNN on "new segment" detection as well as the weighted average over "new segment" and "same segment" classes when all types of features are used together. The highest value for each performance metric is highlighted in boldface.

Method	New Segment			Overall			AUPR
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	
CRF	0.813	0.772	0.792	0.988	0.988	0.988	**0.877**
MLP	**0.817**	0.710	0.760	0.986	0.987	0.986	0.842
BRNN	0.683	0.820	0.745	0.985	0.983	0.984	0.770
CRNN	0.789	**0.864**	**0.825**	**0.990**	**0.989**	**0.989**	0.867

Table 3 summarizes the results of all models on the task of segmentation of e-Coaching emails when word embeddings or lexical features are used in combination with punctuation and POS features. Similar to results in Tables 2, CRNN demonstrates the best performance among all methods achieving 0.864 recall with 0.825 F1-measure for "new segment" detection and 0.990 precision with 0.989 recall and F1-measure overall. BRNN and CRF demonstrated the lowest and second highest performance on the task of email segmentation among all methods, respectively. We observed that classification performance significantly improved for "new segment" detection when lexical features are used in combination with punctuation and POS features. Specifically, precision increases by 3.96%, -2.27%, 12.71% and 1.81%; recall increases by 11.72%, 19.73%, 20.59% and 8.41%; and F1-measure increases by 8.05%, 9.51%, 16.22% and 5.1% for CRF, MLP, BRNN and CRNN methods, respectively, on new segment detection when all types of features are utilized together. Similarly, precision increases by 0.51%, 0.41%, 0.82% and 0.41%; recall increases by 0.41%, 0.41%, 0.72% and 0.3%; and F1-measure increases by 0.41%, 0.41%, 0.82% and 0.3% for CRF, MLP, BRNN and CRNN methods, respectively, as a weighted average over "new segment" and "same segment" classes when lexical features are used in combination with punctuation and POS features.

Table 4 illustrates the impact of different types of features as well as their combination on e-Coaching email segmentation performance. Punctuation and POS features have similar effect measured by the AUPR, which increases by 12.44%, 14.4%, 17.56% and 6% for CRF, MLP, BRNN and CRNN, respectively, when all features are used together. Individually, although punctuation features improve the performance of all classifiers, POS features improve the performance of only CRF and MLP. CRF achieved the highest AUPR when all types of features are used together. On the other hand, POS features degraded the AUPR of BRNN and CRNN.

Table 4: Area under the precision-recall curve (AUPR) values of all classifiers demonstrating the impact of different types of features on e-Coaching email segmentation performance. Highest AUPR value for each feature set across all models is highlighted in boldface.

Features	CRF	MLP	BRNN	CRNN
word embeddings only	0.780	0.736	0.655	**0.818**
word embeddings + POS	0.797 (+2.18%)	0.746 (+1.36%)	0.647 (-1.22%)	**0.798 (-2.44%)**
word embeddings + punctuation	**0.876 (+12.31%)**	0.835 (+13.45%)	0.774 (+18.17%)	0.874 (+6.85%)
all features	**0.877 (+12.44%)**	0.842 (+14.4%)	0.770 (+17.56%)	0.867 (+6%)

Discussion

This study is the first effort to design and evaluate machine learning methods for automated segmentation of e-Coaching sessions. Experimental results indicate that CRNN is the best model among all machine learning methods considered for this study. CRNN achieved 0.989 F1-measure overall and 0.825 F1-measure for detecting "new segment". The robust performance of CRNN provides an evidence that deep learning models are capable of detecting the boundaries of patient and provider behaviors in email delivered behavioral interventions. Our experiments also highlight the importance of punctuation and POS features along with word embeddings for all machine learning methods employed this study. Although the domain of this study was intentionally focused, we believe that the proposed methods are not limited to e-Coaching sessions and our conclusions can be generalized to other domains, which require discourse segmentation.

Punctuation marks and POS features resulted in significant improvement in the performance of traditional machine learning and deep learning methods. Punctuation features had a stronger individual impact on model performance than POS features. In all cases, CRF and MLP performed better, when word embeddings were used in conjunction with punctuations and POS features. Considering punctuations improved the performance of BRNN and CRNN measured by precision, recall and F1-measure, while POS features lowered their AUPR.

The convolution layer made a significant difference between the performance of CRNN and BRNN in MI session discourse segmentation. CRNN had 22.46% and 10.74% higher F1-measure in "new segment" detection and 1.02% and 0.51% higher F1-measure overall compared to BRNN, when word embeddings and all other features were used, respectively. In CRNN, a convolution layer performs a series of convolution and pooling operations, which produce a number of important high-level features from input embeddings. These high-level features are then utilized by the bidirectional RNN layer in CRNN, which translates to a significant increase in performance. In contrast, BRNN utilizes the input embeddings directly as features.

Although punctuation marks play an important role in segmentation boundary detection, a few errors were triggered by the presence of punctuation marks. 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 cafe 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 am hungry rather than the junk food I tend to eat."* was incorrectly segmented after the first sentence, when period was encountered. Similarly, additional information is a common cause for misclassification of an email segment into multiple segments. For instance, although the first sentence in the above email segment represents a positive commitment to behavior change, the next two sentences provide additional information to support the patient's commitment.

Conclusion

Segmentation is the first step of qualitative analysis of unstructured clinical communications, such as e-Coaching. Although several studies have focused on the segmentation problem in biomedical context, they are limited to segmenting clinical text in EHR into sections and sentences. No previous studies considered the task of automated segmentation of clinical communications into groups of MI behaviors in the context of unstructured MI sessions. By comparing the performance of machine learning methods for the task of segmentation of e-Coaching emails, we found out that con-

volutional recurrent neural networks demonstrate the best performance in terms of most performance metrics. Manual segmentation of e-Coaching sessions is a very resource-intensive and time-consuming task, which can significantly decrease the time and effort required to develop effective behavioral interventions. Our proposed methods can help to identify textual segments corresponding to MI behaviors in unstructured clinical dialog, which can then be annotated with MI behavior annotation methods in a pipeline setting. Automated segmentation and annotation of e-Coaching emails can significantly decrease the time to identify effective communication strategies in email-based MI. As future work, we are planning to evaluate the methods presented in this work on the datasets from other types of behavioral interventions.

Acknowledgments

References

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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. .

[9] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. Journal of developmental and behavioral pediatrics: JDBP. 2013;34(8):599.

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

[11] 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.

[12] 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.

[13] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986;323(6088):533.

[14] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 1997;45(11):2673–2681.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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. .

[24] 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.

[25] 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.

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

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

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

[31] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. ACM; 2006. p. 233–240.