Machine Learning Methods for Discourse Segmentation of E-Mail Based Clinical Communication

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 to understand clinical process like Motivational Interviews (MI) is limited by traditional qualitative coding methods. Qualitative coding of communication data is very resource-intensive and time-consuming process, which requires auto coding. This study utilized e-coaching data where email is used to deliver motivation-enhancing coaching to encourage healthy eating. A critical step toward automatic annotation of communication coding process is the segmentation of text data. In this study, we transformed segmentation task into a classification task and developed several state-of-the-art machine learning models including Support Vector Machine, Naive Bayes, K-Nearest Neighbor (KNN), Recurrent Neural Networks by utilizing contextual, topic and punctuation mark features. Results indicate that KNN is the best model and achieved 0.986 F1-measure in overall, 0.779 and 0.993 F1-measures for detecting "boundary" and "not boundary" classes, respectively. This study has a great implication to save money and accelerate the pace of identifying effective communication strategies.*

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, in some cases, a caregiver.

In this study, we focus on the analysis of e-Coaching sessions, MI 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 significant barrier to fully automating the behavior coding process of e-Coaching emails is their segmentation into tex-

*Authors provided equal contribution.

tual 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.

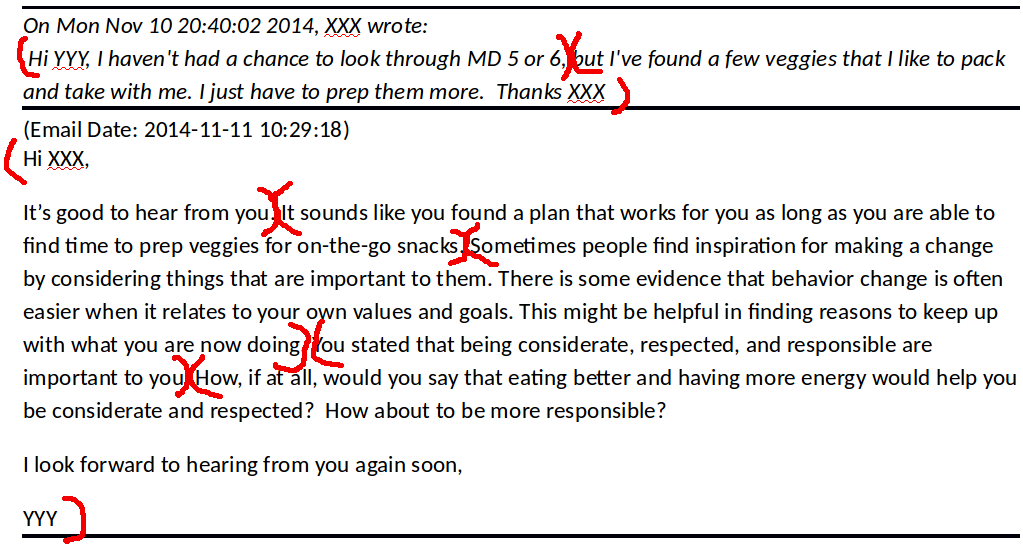

Figure 1: Example of e-Coaching emails segmented into fragments that correspond to MI behaviors of an e-Coach and a patient.

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 contextual, topic and punctuation features and experimented with both traditional supervised machine learning methods, such as Support Vector Machine (SVM), Naive Bayes (NB) and k-Nearest Neighbor (KNN) classifiers, and deep learning methods, such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), 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[15–17]. 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. Segmentation of e-Coaching emails is different from traditional shallow discourse analysis of conversa-

tions[18] 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[19] (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 introduces a novel computational approach 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. To filter out noise from the dataset, non-ascii characters are removed and then applied stemming to obtain a general form of word from different word representations, such as "eating", "eats", and "eat". We formulate the text segmentation task into a binary classification, as shown in Figure 2. For NB, SVM and KNN models, clinical exchange is given as the input then it is partitioned as adjacent word pairs by sliding them. Each pair is classified into either "boundary" or "not boundary" class. The original text is segmented at the position, where an adjacent word pair classified into "boundary" class. If all adjacent pairs of a block of text classified into "not boundary", the whole block of text is then treated as one segment about a single MI behavior. Totally, we obtained 95,421 word pairs, which include 3,138 "boundary" and 92,283 "not boundary" instances. For RNN, a chunk of text was taken as input sequence which gets one hot encoding of word vector for the model input. A binary label was assigned as output sequence where "boundary" and 'not boundary" words were associated with label 1 and 0, respectively.

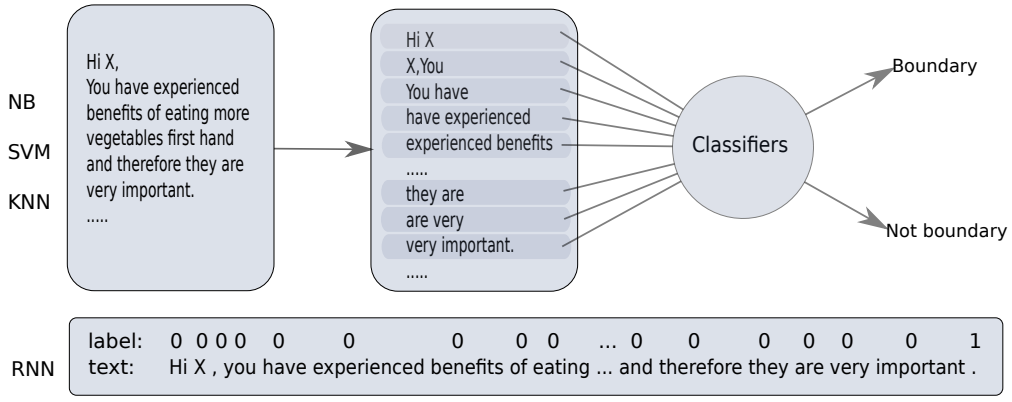

Figure 2: Transformation of text segmentation task into text classification task.

For the experiment, we utilized three type of features including word (contextual feature), punctuation mark, and topic model-based features. In contextual features, each word represented in a binary format, where 1 indicates the appearance of the word and 0 for absence. Since topic models are very effective[8,20,21] to represent text documents, we derived features from a topic model named Labeled LDA[22]. To obtain feature values, we utilize the experimental data which is segmented as MI code or class. First, we derived a class-specific multinomial $p(w|c)$ per each class c from class-specific topics: $p(w|c,z)$ by marginalizing over topic z, where the number of topics is experimentally determined by the model performance[8].

$$p(w|c) = \sum_{z=1}^{K} p(w|z,c)$$

Here, K is the number of topics per class c. After that, word-specific class distributions $p(c|w)$ were estimated with class-specific multinomials $p(w|c)$ by using the following formula:

$$p(c|w) = \frac{p(w|c)p(c)}{p(w)}$$

where $p(c)$ is the prior distribution of class c in the training set and $p(w)$ is a probability of word w in the collection language model estimated using maximum likelihood. In our experiment, we represent $p(c|w)$ as the feature showing that how indicative each word w for a particular class c. Punctuation mark containing one of the symbols {'.', ',', '!', '?', ':', ';', '-'} is also employed as feature. This is one of the most important features as they indicate the boundary of a sentence, clause, and phrase.

Segmentation classifiers

Several state-of-the-art classifiers, including Naive Bayes (NB)[23], Support Vector Machine (SVM)[24], K-Nearest Neighbor (KNN)[23], two variant of Recurrent Neural Networks (RNN)[25]: Long Short Term Memory (LSTM)[26] and Gated Recurrent Unit (GRU)[27], are employed to estimate the classification performance.

Naive Bayes: this model is constructed by using the training data and estimate the prior probability of classes, and each feature has given the class. Then, the posterior probability is computed to predict the class label by applying the Bayes theorem with the assumption that features are conditionally independent. This study utilized a specialized version of Naive Bayes called Multinomial Naive Bayes, which is best suitable for discrete features such as word.

Support Vector Machine: we used this model as one of the state-of-the-art classification technique proven to perform well in text categorisation[28] for its ability to cope with very high dimensional input feature space. SVM finds the best hyperplane in the feature space that maximizes the separation between the closest "boundary" and "not boundary" training examples. In this experiment, the polynomial kernel is employed to train the SVM model for the segmentation of e-coaching text.

K-Nearest Neighbor: By this model, each training sample represented as a point in the input feature space. For a new test sample, Euclidean distance is calculated to find the k-nearest neighbors. Finally, the test sample is classified into majority class of the k-nearest neighbors. We experimentally determined that best performance was achieved with k = 3 for the classification of word pairs.

Recurrent Neural Networks: RNN is a neural network architecture designed to capture sequential patterns present in temporal sequence such as text data. When we predict the "boundary" point, adjacent word pair will help to understand the pattern of the sequence. Long Short Term Memory networks usually reffered as LSTMs[26], are a special type of RNN capable of handling variable size input sequence, contains internal memory. GRU[27] is a variant of LSTM mathematically represented by the following formula:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{3}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{4}$$

In Eq. 1-4, σ corresponds to sigmoid function and $\odot$ designates an element-wise product. The update gate z_t and reset gate r_t at time step t are computed by the Eq. (1) and (2), where W_z, W_r, W_h, U_z, U_r, U_h are the weight matrices and b_z, b_h and b_r are bias vectors. The activation h_t of the GRU at time t is a linear combination of previous activation

h_{t-1} and the candidate activation $\tilde{h}_t$, which is represented by Eq. (4) and (3). We build our RNN model with one hidden layer, output layer, and input layer. We reset our model state after feeding each input sequence where input was given as one-hot encoding of word vector. Since one-hot vector is given in the input layer, results are reported with contextual and punctuation mark features only. We experimentally determined that the best performance is achieved when the number of hidden units = 25, batch size = 1, and optimizer = adam.

Evaluation metrics

In this experiment, standard metrics: precision, recall, and F1-measure, are applied to evaluate the performance of binary classifiers[29]. 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. We conduct the experiment with 5 folds cross-validation and weighted macro-averaging of these metrics over the folds. All models have trained on 80% of the data and remaining 20% of the data is used as a test set for reporting the performance of the model. We also estimate the area under the receiving operating characteristics (ROC) curve[30] (AUC) metric due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[31].

Results

Experimental results are evaluated with "boundary" and "not boundary" classes as well as their weighted average, which are shown in Table 1, 2, and 3, respectively.

Table 1: Performance of NB, SVM, KNN, and RNN methods for detecting segmentation boundary in e-coaching text. The highest value for each performance metric is highlighted in bold.

Method	contextual features only			contextual + punctuation marks (+topic-based except RNN)		
	Precision	**Recall**	**F1-measure**	**Precision**	**Recall**	**F1-measure**
NB	0.594	0.662	0.626	0.590	0.666	0.626
SVM	0.742	**0.679**	0.709	0.774	0.696	0.733
KNN	**0.808**	0.663	**0.728**	**0.820**	**0.742**	**0.779**
LSTM	–	–	–	0.800	0.646	0.714
GRU	–	–	–	0.800	0.715	0.741

As follows from Table 1, KNN performs best among all machine learning models in terms of precision and F1-measure, achieved 0.808 precision with 0.728 F1-measure when contextual features are used; and 0.820 precision with 0.779 F1-measure when a combination of contextual, punctuation mark and topic model-based features are used. However, NB demonstrates the lowest performance among all models in terms of all performance metrics. In this study, GRU appears as the second highest model, obtains 10.68% higher recall with 3.78% higher F1-measure than LSTM. On the other hand, SVM exhibits highest recall 0.679 when only contextual features are used. When contextual features are used in combination with punctuation mark and topic model-based features, recall increases by 0.6%, 2.5%, and 11.92%; and F1-measure increases by 0%, 3.39%, and 7% for NB, SVM, and KNN models, respectively. Nevertheless, precision increases by 4.31% and 1.49% for SVM and KNN methods while decreases by 0.7% in NB.

Table 2: Performance of NB, SVM, KNN, and RNN methods for the identification of "not boundary" class. The highest value for each performance metric is highlighted in bold.

Method	contextual features only			contextual + punctuation marks (+ topic-based except RNN)		
	Precision	**Recall**	**F1-measure**	**Precision**	**Recall**	**F1-measure**
NB	0.988	0.985	0.987	0.989	0.984	0.986
SVM	**0.989**	0.992	0.991	0.990	0.993	0.991
KNN	**0.989**	**0.995**	**0.992**	0.991	**0.994**	0.993
LSTM	–	–	–	0.993	**0.994**	**0.994**
GRU	–	–	–	**0.994**	**0.994**	**0.994**

Table 2 summarizes the performance of NB, SVM, KNN, and RNN models for detecting "not boundary" class in e-coaching text. We observed that performance is remarkably high in "not boundary" class compared to boundary detection which is expected because 96.71% instances are from "not boundary" class. KNN achieves 22.40%, 50.07%, and 36.26% higher precision, recall, and F1-measure, respectively, for contextual features; and 20.85%, 33.96%, and 27.47% higher precision, recall, and F1-measure, respectively, for combined features compared to boundary class. In contrast to boundary detection, RNN demonstrates the highest performance among all models. LSTM obtains 0.993 precision with 0.994 recall and F1-measure while GRU exhibits 0.994 for all performance metrics. Impact of punctuation mark and topic model-based features is also consistent with "not boundary" classification. Results show that F1-measure increases by 0%, and 0.1% for SVM and KNN models although decreases by 0.1% for NB.

Table 3: Weighted average performance of NB, SVM, KNN, and RNN methods for the segmentation of e-coaching text in detecting both "boundary" and "not boundary" classes. The highest value for each performance metric is highlighted in bold.

Method	contextual features only			contextual + punctuation marks (+ topic-based except RNN)		
	Precision	**Recall**	**F1-measure**	**Precision**	**Recall**	**F1-measure**
NB	0.975	0.974	0.975	0.976	0.974	0.975
SVM	0.981	0.982	0.981	0.983	0.983	0.983
KNN	**0.983**	**0.984**	**0.983**	**0.986**	**0.986**	**0.986**
LSTM	–	–	–	**0.986**	0.983	0.984
GRU	–	–	–	**0.986**	0.985	**0.986**

Table 3 outlines the weighted average results of the experiment on the models for the segmentation of e-coaching text by classifying them into "boundary" and "not boundary" classes. Overall, KNN obtains the best performance with all metrics and NB denotes the lowest performance among all methods. GRU demonstrates the same result as KNN for precision and F1-measure except for recall. SVM shows moderate performance, obtains 0.981, 0.982, and 0.983 for precision, recall, and F1-measure, respectively, when contextual features are used. Influence of the additional features is also consistent as mentioned in Table 1 and 2. Precision increases by 0.1%, 0.2%, and 0.3%; recall increases by 0%, 0.1%, and 0.2%; and F1-measure increases by 0%, 0.2%, and 0.3% for NB, SVM, and KNN methods, respectively, when combined features are used.

Discussion

This study is the first efforts to evaluate the automatic segmentation of e-coaching text. Experimental results indicate that KNN is the best model among all machine learning methods considered for this study. KNN achieved 0.986 F1-measure in overall, 0.779 and 0.993 F1-measures for detecting "boundary" and "not boundary", respectively. The robust performance of KNN provides the evidence that machine learning models are capable to learn information from clinical exchange. 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 domain as well.

Punctuation mark and topic model-based features made a significant improvement in performance of all machine learning methods. Nearly all cases, the model performs better when contextual features are used in combination with punctuation mark and topic model-based features. This results also mean that segmentation performance might be improved by adding more relevant features including human insight into the problem.

In this paper, results are reported by each class to avoid confusion about the overall model performance. In addition, standard metrics: precision, recall, and F1-measure were used to eliminate doubt about the model performance because accuracy is misleading for imbalance dataset. AUC values are also outlined due to its effectiveness in measuring the quality of binary classifiers for imbalanced datasets[31], which are demonstrated by the ROC curves in Figure 3. NB shows the highest AUC values, achieved 0.978 for both cases while provides lowest classification results. On the other hand, KNN and SVM exhibit 0.972 and 0.835 AUCs when only contextual features are used; and 0.959 and 0.844 AUCs when a combination of contextual, punctuation mark and topic model-based features are used. Finally, LSTM

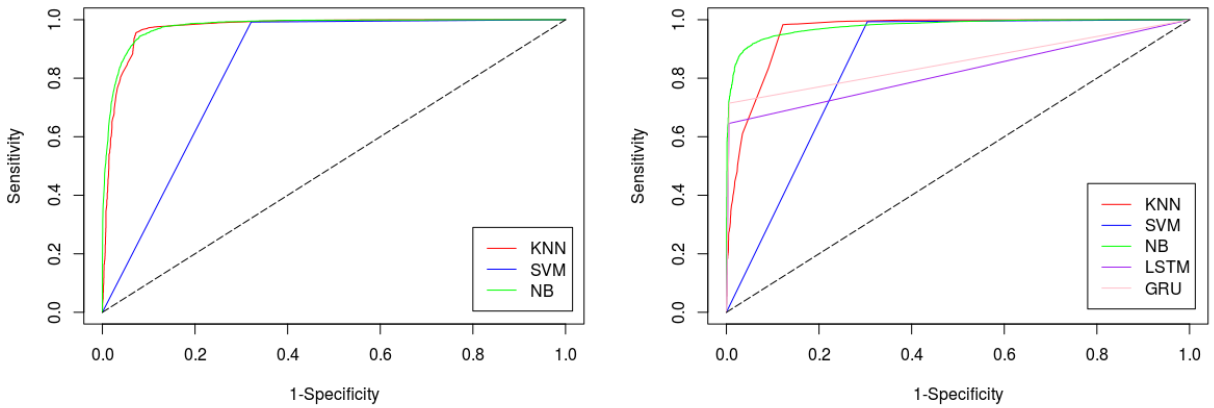

Figure 3: Receiver operating characteristic curves showing the performance of binary classifiers for the segmentation of e-coaching text when contextual features (left) and combination of contextual and other features (right) are used.

demonstrates lowest AUC values among all machine learning models, achieved 0.82 AUC while GRU achieves 4.15% higher AUC than LSTM. The conclusion drawn from the ROC curves also confirmed the robustness and superiority of KNN model for the segmentation of clinical exchange.

We observed the second highest performance of RNN, in particular, LSTM and GRU for the text segmentation. We believe that RNN will perform better if a large set of data 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 other previous study[32].

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 block "A1 A2 A3. B1 B2 B3. C1 C2 C3 C4." can be incorrectly segmented at position A3, B3, and C4 where a punctuation mark was encountered. Similarly, additional information is the common reason for the classified original segment into multiple segments. For instance, the above text block can also be incorrectly segmented at position B3 and C4 because third sentence (C) only supports the MI code confirmed by first two sentences (A and B).

Our proposed approach is novel for the segmentation of e-coaching text because previous studies mainly focus on the segmentation of text into sections, headers, and sentences in other medical domain. 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 package of automatic coding procedure of clinical exchange.

The limitation of this study is that e-coaching text 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 on other datasets involves in discourse analysis. We also plan to use more relevant features to improve model performance. For example, part-of-speech tagging[33] and distance from the beginning and end of the current sentence might significantly enhance the classification performance.

Conclusion

Segmentation of e-coaching text is an integral part of developing an automated e-coaching intervention. Although several studies have done in 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 are used for the segmentation of text into groups of MI behaviors in the setting of discourse analysis with email

under the principle of motivational interviews. In this paper, we compared the performance of machine learning models for the task of segmentation of e-coaching text. We found out that k-nearest neighbor 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 MV, Shulby C, Aluísio SM. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. arXiv preprint arXiv:161000211. 2016;.

[18] 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. .

[19] 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.

[20] Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. Journal of biomedical informatics. 2016;62:59–65.

[21] Lu HM, Wei CP, Hsiao FY. Modeling healthcare data using multiple-channel latent Dirichlet allocation. Journal of biomedical informatics. 2016;60:210–223.

[22] Ramage D, Hall D, Nallapati R, Manning CD. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics; 2009. p. 248–256.

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825–2830.

[24] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011;2(3):27.

[25] Bengio Y, Frasconi P, Simard P. The problem of learning long-term dependencies in recurrent networks. In: Neural Networks, 1993., IEEE International Conference on. IEEE; 1993. p. 1183–1188.

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

[27] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259. 2014;.

[28] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Springer; 1998. p. 137–142.

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

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

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

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

[33] Hasan M, Kotov A, Mohan A, Lu S, Stieg PM. Feedback or Research: Separating Pre-purchase from Post-purchase Consumer Reviews. In: European Conference on Information Retrieval. Springer; 2016. p. 682–688.