WIP: Automating anonymous Processing of peer evaluation comments

Mr. Siqing Wei, Purdue University-Main Campus, West Lafayette (College of Engineering)

Siqing Weir received both bachelor's and master's degrees in electrical and Computer Engineering from Purdue University. He is currently pursuing a Ph.D. degree in Engineering Education at Purdue University. After years of experience of serving a peer teacher and a graduate teaching assistant in first-year-engineering courses, he is a research assistant at CATME research group studying the existence, causes and interventions on international engineering teamwork behaviors, the integration and implementation of team-based assignments and projects into STEM course designs and using mixed-method, especially natural language processing to student written research data, such as peer-to-peer comments. Siqing also works as the technical support manager at CATME research group.

Mr. Rui Wang, Purdue University

Rui Wang is an undergraduate student at School of Electrical and Computer Engineering at Purdue University. His research interests include interpretable machine learning, robust computer vision and natural language processing.

Dr. Matthew W. Ohland, Purdue University at West Lafayette

Matthew W. Ohland is Associate Head and Professor of Engineering Education at Purdue University. He has degrees from Swarthmore College, Rensselaer Polytechnic Institute, and the University of Florida. His research on the longitudinal study of engineering students, team assignment, peer evaluation, and active and collaborative teaching methods has been supported by the National Science Foundation and the Sloan Foundation and his team received for the best paper published in the Journal of Engineering Education in 2008, 2011, and 2019 and from the IEEE Transactions on Education in 2011 and 2015. Dr. Ohland is an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS.

Dr. Gaurav Nanda, Purdue University at West Lafayette

Dr. Gaurav Nanda is an Assistant Professor of Practice in the School of Engineering Technology at Purdue University. He completed his Ph.D. in Industrial Engineering from Purdue University and Masters and Bachelors from Indian Institute of Technology Kharagpur, India.

His research interests include application of text mining and machine learning methods to analyze real-world data. Currently, he is studying learner experiences in online courses by applying text mining approaches on user generated data such as discussion forums and open-ended feedback.

WIP: Automating anonymous processing of peer evaluation comments

Abstract
De-identifying qualitative datasets is time-consuming and expensive but is a critical step in protecting the confidentiality of study participants. Peer-to-peer comments are an important supplement to peer evaluation ratings in team-based learning courses. Those comments comprise valuable research data for educational study to investigate but they usually contain identifiable information, such as names. In this work in progress, we study and propose a pipeline tool to identify all names appearing in CATME team peer evaluation comments and replacing those names with pseudonyms such as Rater 1 and Rater 2. We explored several natural language processing techniques empowered by machine learning methods and then optimized to the final algorithm. At its core, the algorithm combines the long short-term memory (LSTM) and conditional random field (CRF) approaches most common in the field of named entity recognition. The current algorithm performs well, with a high recall of 0.8 with reasonable precision scores resulting in 76 of F_1 score as we want to put an emphasis on recalls. We also propose this as a tool to make a large amount of data available for research that would otherwise be sensitive due to personally identifiable information.

Introduction

Peer evaluations are used as the assessment and educational tool in various settings including academia [1]. CATME SMARTER teamwork system enables students to rate peers according to behaviorally anchored scale [2] and also provide textual feedback [1], [3]. When used as a tool for collecting research data, CATME researchers currently can only collect deidentified peer numerical rating results based on the consent of instructor users. The peer-to-peer comments, another vital research asset, are left out due to legal regulation on protecting educational records. Educational records, including names of students, are protected by the Family Educational Rights and Privacy Act (FERPA) and other international regulation [4], [5]. To extend the research capability of the system, there is a need to remove identifiable information from peer-to-peer comments, so we propose an automatic pipeline to achieve this goal.

Automating deidentification has been extensively explored in health disciplines. Many researchers have already proposed algorithms, pipelines and tools to resolve the issues based on the U.S. Health Insurance Portability and Accountability Act (HIPAA)'s requirement on protecting protected health information [6]–[8]. However, HIPAA requires protection on lots of unexpected information in the academic setting, such as locations, dates, telephone numbers, fax numbers, social security numbers, etc. [9]. In the education context, Rudniy reported an automating deidentification project using peer feedback textual data for online writing projects via MyR [10]. However, our peer to peer comment data is structured in groups to facilitate teamwork learning so that it is highly possible that the commenter mentions more than one group member, which might result that the traditional treatment of removing the identifiable entity causes confusion about the relative reference. Therefore, we need additional functionality to replace appeared names with pseudonyms such as Rater 1 and Rater 2 to make the context clear about those pronouns. However, as this work is marked as working in progress, we have not completed the research work for this functionality. Therefore, we are not going to discuss replacing names with pseudonyms.

Literature Review

In this section, we provide the background knowledge of the natural language processing technique we propose later to help readers to get familiar with this field. We first introduce machine learning and deep learning and then discuss technical domains related to our research as token-level classification and named entity recognition. Finally, we describe LSTM-CRF algorithm chosen to implement for this study.

Machine learning (ML) is considered a subfield of artificial intelligence, where practitioners and researchers do not hard-code rules into machines to perform certain intelligent actions or inference but instead make the model to learn from data [11], which could be understood that as its simplest form as performing a regression that forces the universal function approximator to approximate the true function $p(y|x)$ where x is the input and y is the output. Deep learning refers to the type of machine learning that is associated with the use of the deep artificial neural network. Since the introduction of AlexNet [12], deep learning has shown tremendous potential serving as a class of universal function approximators. And the algorithm we choose to implement is classified as a deep learning algorithm. Usually, a deep learning model has millions or more parameters whose values are all determined during training. However, the performance of such an approximator depends largely on the dataset chosen, the compatibility of the model architecture (the prior imposed to the final model), the dataset and the given tasks. The detailed elaboration of deep learning is beyond the scope of this work, we refer the readers to this review [8].

Token-level classification tasks or linguistic sequence labeling have long been a major topic in the computational linguistic community [13]. Such tasks are an essential component in larger natural language processing systems for information retrieval, relation extraction, and question answering. Specific tasks under token-level classification tasks include part-of-speech tagging, named entity recognition, tokenization. Tokens in natural languages are discrete units, to perform numerical computation, we create a look-up table for the tokens, that is, one entry in the table per token. Each entry of the table is a continuous space vector with a pre-determined size. In our setting, we aim to perform token-level classification, i.e. a name or not a name, which resembles named entity recognition (NER).

Standard datasets for NER tasks have been created and CoNLL 2003 is one of the most widely used dataset and we use this dataset as our measure for literature review [14]. The solutions to NER has seen tremendous progress since the renaissance of the artificial neural network since the highly impactful AlexNet [12]. And the progress has been made along the lines of LSTM & CRF and now to large, learned language models. LSTM-CRF is proposed as a neural-network based sequence labeling model as a combination of Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) [15]. Long Short-Term Memory is proposed as an effort to address the problems of long-term dependency for a long sequence [16] and CRF is employed in the model to learn a dependency between the output labels. LSTM-CRF achieved 90.94 in F_1 score at CoNLL 2003 [17]. LM-LSTM-CRF introduced character-level knowledge into the model to leverage such information with a F_1 score at 91.24 [18]. BERT is short for Bidirectional Encoder Representations from Transformers [19]. Bidirectional means this model extract information from one sentence not only in left-to-right direction but also the other way around, mimicking what we often find ourselves doing when reading a sentence. Encoder refers

that BERT is encoding each input sentence into numerical space with continuous vector representations. The transformer is an architecture devised, in the first place, for neural machine translation, in which field despite its smaller model size and faster training time surpassed then state-of-the-art results [20]. Researchers since then have been trying to incorporate the Transformer architecture into other tasks in addition to neural machine translation. Because of its ability to model long term dependencies [16], a problem faced by NLP community from the start, and to parallelize computation for faster runtime with multiple computation cores [20], Transformers have gained tremendous focus and proved the architecture's performance is better than those of the previous models. At the time of its release, BERT Large achieved 92.8 in F_1 score.

Method

Our dataset contains CATME peer evaluation results in first-year engineering courses for five semesters conducted at a large midwestern research university, which contains 74061 pieces of comments in 601 teams from 21 sections. Students were instructed to write constructive self-reflection and feedback to other team members based on a validated teamwork behavior model that was introduced and assessed via CATME [21].

In our formulation, our problem is to determine whether a word in a sentence is the name of a person. We can perform such analysis from two perspectives. One is a cloze-like task where given the context of the word of interest, we can make predictions on the grammatical and semantical representations, and the other is to perform classification on words with character-level information. To extract as much information as possible, we need to leverage both word-level and token-level information. We propose to use a machine learning model, an ensemble of 5 models trained on different parts of the training dataset. The model is based on word-level information of the sentence.

For the decision on whether to code (whether there is relevant information), we use F_1 score as our metric [22],

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} * 100,$$

where,

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive},$$

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}.$$

For a higher F_1 score, a binary classification model has to be able to produce both good precision and good recall metrics.

To generate labels, we run a word match and consider every pronoun a true label, leveraging the free labeling in addition to the actual names.

Following the common practice of supervised learning fashion, we split the dataset into a training set and a testing set by random selection, leaving 25% of the original data for testing and the rest for training. During training, we optimize the parameters of the model based on a certain

objective function that serves as a surrogate objective because the direct objectives (F1 in our case) are usually non-differentiable, which is a prerequisite for first-order optimization algorithms. Usually, such objective functions are based on the idea to improve alignment between the model prediction of certain inputs and their corresponding labels, as is our setting. The optimization, however because of the complexity of the geometry of the surrogate function, mostly ends up in local minima, which have been suggested to be good enough in most situations [23]. Training is performed on the dataset. The optimization algorithm employed usually is stochastic gradient descend (SGD) or its variants such as adaptive momentum [24], which is one of the most prevalent optimizers and the one we employ in this work. At its core, SGD is about to perform the following update to parameter w,

$$w = w - \alpha \times w',$$

where w is the original parameter value, α is called learning rate and w' is the derivative of w. We use the chain rule of differentiation to propagate loss signal Variants of SGD usually perform some augmentation with momentum or utilize the approximation of the second-order derivative since the exact computation is computationally inefficient. The differentiation is performed with respect to the loss function, which is usually computed by a mini-batch of the entire dataset for its feasible memory requirements, and smoothed loss function geometry [25].

Training is performed as long as the model is not overfitting, which is determined by the trend of the validation loss, that is, the loss function value computed using the validation dataset. During training, the validation set has never been seen by the model, which guarantees the determination process of the overfitting criterion is adequate. Before the overfitting, both the training loss, computed by training dataset and the validation loss should share the same trend of decreasing. For overfitting, however, one can observe that with the decrease of training loss, which is the result of the optimization algorithm, the same no longer holds for validation loss.

When training ends, the model performs inference to either carry out its design purpose. We then use the testing set to report the performance of our model. For our work, we use the embeddings trained from GloVe [26] and paragram [27]. We combine the two embeddings by taking the element-wise arithmetic mean of each entry. Our word-level model consists of a single LSTM layer with an intentionally mismatch between the forward pass and backward pass of LSTM. For an illustration of LSTM, we refer to the explanation in [28].

To extract information from both sides, we use bidirectional LSTM and creates a mismatch as in Figure 1. To get rid of the information from the word itself, we pad the forward output from the beginning and the backward output from the end, move in the way as in Figure 1, and we get rid of the information from the current word. We then concatenate the output mismatched vectors and feed the result into the fully connected layer for classification on each token.

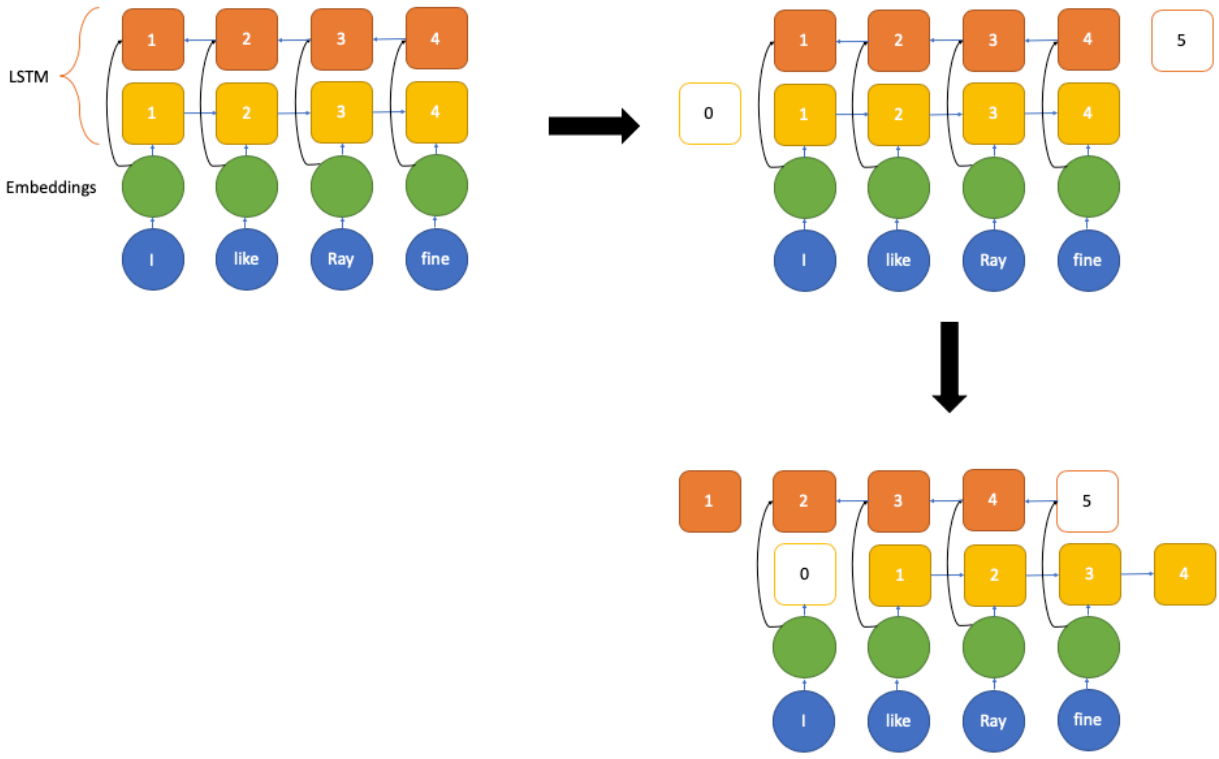

Figure 1. the operation of mismatch that masks the current word and only extracts the context information.

In-progress results and discussion

We report our F_1 score as 76 on the testing dataset with the threshold of determination determined on the validation set. For each sentence, this task contains 5 binary classification tasks for F_1 score to be computed on, we deem this result very positive. Though our model achieves good results for its design purpose, it is very reasonable to assume that with more data we are expecting substantial performance boost on the next iteration of the model. Moreover, we are excited to be able to bring this feature online for instructors to use. We will train an ensemble of 5 models in total to perform name removal for the CATME platform. And with a threshold of 0.3761, we have the end F_1 score of 76% on solely CATME data.

This work in progress paper reveals the possibility and feasibility of using natural language processing techniques to automatically remove students' names. This work builds up the pipeline which could be quick to implement in other similar educational settings, which could significantly speed up the de-identification process for textual data and benefit researchers who need to use this type of data to conduct studies.

Limitations

The way we create ensemble creation is crude, in the sense that it does not yet include the interaction between the character-level information and word-level information. The training data size is a bit constrained in the sense that we have not yet used any large available dataset for pre-training that would allow our model to better generalize.

The current method to calculate precision and recall is not ideal enough because students might refer to each other with nicknames or abbreviated names, which causes the inaccuracy. In our future work, we will manually check for random samples to report precision, recall, and F1 score more precisely. As the hottest research area, the machine learning community creates and proposes new approaches to solve standard problems speedily. To expand this work, we will try several different well-accepted algorithms for our proposed problems to select the best approach to implement. We will also further conduct research to realize replacing those named entities with pseudonyms.

References

[1] C. Brawner, O. Murch, D. Ferguson, and M. Ohland, "Comparing Peer-to-Peer Written Comments and Teamwork Peer Evaluations.," in *Association for Engineering Education - Engineering Library Division Papers*.

[2] M. W. Ohland *et al.*, "The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation," *Acad. Manag. Learn. Educ.*, vol. 11, no. 4, pp. 609–630, 2012, doi: 10.5465/amle.2010.0177.

[3] M. L. Loughry, M. W. Ohland, and D. J. Woehr, "Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools," *J. Mark. Educ.*, vol. 36, no. 1, pp. 5–19, 2014, doi: 10.1177/0273475313499023.

[4] "Family Educational Rights and Privacy Act of 1974 (FERPA)." [Online]. Available: https://epic.org/privacy/student/ferpa/. [Accessed: 02-Jan-2020].

[5] "Invasion of pravacy: Penalties and remedies: Review of the law of privacy: Stage 3. " New Zealand Law Commission, ISBN 978-1-877316-67-8, 2009.

[6] I. Neamatullah *et al.*, "Automated de-identification of free-text medical records," *BMC Med. Inform. Decis. Mak.*, vol. 8, no. 32, pp. 1–17, 2008, doi: 10.1186/1472-6947-8-32.

[7] O. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the State-of-the-Art in Automatic De-identification," *J. Am. Med. Informatics Assoc.*, vol. 14, no. 5, pp. 550–563, Sep. 2007, doi: 10.1197/jamia.M2444.

[8] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *J. Am. Med. Informatics Assoc.*, p. ocw156, Dec. 2016, doi: 10.1093/jamia/ocw156.

[9] "Standards for privacy of individually identifiable health information. Final rule," 2002.

[10] A. Rudniy, "De-Identification of Laboratory Reports in STEM | Journal of Writing Analytics," *J. Writ. Anal.*, vol. 2, pp. 176–202, 2018.

[11] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, 27-May-2015, doi: 10.1038/nature14539.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012.

[13] V. Yadav and S. Bethard, "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models," in *the 27th International Conference on Computational Linguistics*, 2018.

[14] E. F. T. K. Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *CoNLL*, 2003.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural

Architectures for Named Entity Recognition," in *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 260–270.
[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[17] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.
[18] L. Liu *et al.*, "Empower sequence labeling with task-aware neural language model," in *32nd AAAI Conference on Artificial Intelligence*, 2018, pp. 5253–5260.
[19] J. Devlin, M.-W. Chang, L. Kenton, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for Language understanding," in *arXiv preprint arXiv:1810.04805*, 2018.
[20] A. Vaswani *et al.*, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, 2017.
[21] A. C. Loignon, D. J. Woehr, J. S. Thomas, M. L. Loughry, M. W. Ohland, and D. M. Ferguson, "Facilitating peer evaluation in team contexts: The impact of frame-of-reference rater training," in *Academy of Management Learning and Education*, 2017, vol. 16, no. 4, pp. 562–578, doi: 10.5465/amle.2016.0163.
[22] C. J. Van Rijsbergen, *Information retrieval*, 2d ed. London ; Boston: Butterworths, 1979.
[23] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in Neural Information Processing Systems*, 2014, vol. 2, pp. 2933–2941.
[24] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
[25] R. Kleinberg, Y. Li, and Y. Yuan, "An Alternative View: When Does SGD Escape Local Minima," in *the 35th International Conference on Machine Learning*, 2018.
[26] P. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing*, 2014.
[27] J. Wieting, M. Bansal, G. Kevin, and K. Livescu, "Towards Universal Paraphrastic Sentence Embeddings," in *ICLR*, 2016.
[28] I. Goodfellow, Y. Bengio, and A. Courville, "Sequence modeling: Recurrent and recursive nets," in *Deep learning*, The MIT Press, 2016, pp. 367–415.