Improving Performance of PRIDE Algorithm through Advanced Methods

Korea University COSE461 Final Project

Younsuk Yeom
Department of Biomedical Engineering
Team 1
2018250028

Chaehyeon Kim
Department of Biomedical Engineering
Team 1
2020250028

Donghyeon Ki
Department of Biomedical Engineering
Team 1
2018250034

Abstract

PRIDE is an algorithm to solve relationship classification tasks by capturing word and utterance representation in conversations. In PRIDE, utterance chunks were made to overcome input limits of BERT. But in aspect of capturing the full context of conversations, it has poor expressiveness. To address the limitation, in this project, we propose 2 methods: **1)** shuffling the order of utterances and **2)** creating overlapped chunk to add additional context of conversation at word level. Using a FiRe—a Film Relationship dataset, we compare performances of our input formation methods to prior state-of-the-art algorithms. Through our experiments, we show the effectiveness of a hierarchical representation of conversations.

1 Introduction

Relationship Extracting Tasks Relationship Extraction is the task of identifying and classifying relationships between entities in text. It involves extracting relationships from sentences by recognizing the entities involved and understanding the interactions between them. This field utilizes natural language processing techniques to analyze text and infer meaningful relationships. The extracted relationships can have various applications, such as social network analysis or product recommendation systems. We focus on the Relationship Classification task within Relationship Extraction for our project.

Prior Work and its Limitation PRIDE proposed by Tigunova et al. [2021] is a neural multi-label classifier for predicting relationships in dialogue. It makes inference among 12 fine-grained directed speaker's relationships from dialogue, by hierarchically creating utterance representations and combining them with external knowledge about speaker features and conversational style. It uses BERT (Devlin et al. [2018]) to create contextual word embeddings for each utterance, and Transformer (Vaswani et al. [2017]) encoders to build conversation representations that preserve information about the sequence and speakers of utterances. In this way, automatically extracted interpersonal relationships of conversation interlocutors can enrich personal knowledge bases.
Due to the input length limitation of BERT, PRIDE splits the input sequence of utterances into chunks, and runs BERT for each chunk. It also creates enriched utterance representations for conversational context. At this point, we found some possible problems. First, since this is the way to obtain context

between generated utterance representations by putting them into transformer encoder as input, it is difficult to say it as "conversational context" that takes into account the level of word where the actual conversational element exists. Second, while saying that "We chose BERT to create word representations, because this model efficiently captures contextual information.", they actually did not make good use of BERT in identifying relationship between chunks, which is important for grasping the entire context. Lastly, they assume that each chunk in the split has the maximal possible length that fits into one run without breaking individual utterances. However, in reality, an utterance itself may be longer than 512 tokens, which can lead to somewhat unreasonable assumption. Therefore, we suggest some corresponding solutions for improving context representations.

Contribution To enhance the expressive power of words in the conversation, novel two approaches are attempted to form inputs. First, by randomly shuffling the order of utterances, we induce BERT to make word embeddings with non-sequential utterances. Second, by overlapping utterance chunks, we can add additional context of conversation. We compare the performances of our new approaches with original PRIDE. Additionally, we also present the performance of combining the two methods. Through experiments, the usefulness of PRIDE's hierarchical representation in conversations is examined, and we observe that important factors such as the order of utterance or specific words may differ when classifying each relationship between speakers.

2 Related Work

BERT BERT, short for Bidirectional Encoder Representations from Transformers, proposed by Devlin et al. [2018] has been developed to generate comprehensive representations of text by considering both the left and right context in all layers during its pretraining phase. This allows the pretrained BERT model to be easily fine-tuned by adding a single output layer. In addition, BERT has a maximum input length restriction of 512 tokens.

PRIDE: Predicting Relationships in Conversations PRIDE hierarchically creates word and utterance representations, which are then combined with representations of personal attributes and interpersonal dimensions to create a representation of the full conversation history. Given this representation of the conversation, a multi-label classification layer predicts one or more of the twelve relationship labels. First, PRIDE creates contextual word representations with BERT, which efficiently captures contextual information. The input for a pair of speakers (sp_A, sp_B) in PRIDE is shown in Figure 1. It has a total of N utterances $u_1, ..., u_N$, where i-th utterance consists of words $w_i^1, ..., w_i^{n_i}$. The word representations r_i^j are created with a function (BERT) $f^{word}(w_1^1, .., w_1^{n_1}, ..., w_N^{n_N}) = r_i^j$, which takes as input the concatenation of all utterances and produces the representations for each word. Because of the input length limitation of BERT, which is 512 tokens, the input sequence of utterances is split into chunks and then run BERT several times. Next, PRIDE creates utterance representations. The aggregation function aggregates word representations r_i^j within each utterance into utterance representations r_i, $a^{word}(r_i^1, ..., r_i^{n_i}) = r_i$. The aggregation is performed on the utterances from all runs of BERT and outputs $r_1, ..., r_N$ as the representations of utterances. The best word aggregation strategy was attention-weighted average.

3 Approach

3.1 Random shuffling utterances in utterance chunks

We notice that utterance chunks, which was used to overcome the BERT input length limit within PRIDE, has limitations in capturing the entire context of the conversation script. Despite being formed as a "fragment" of the conversation to run BERT, it is challenging to capture the context of the entire content at the word representation level. To tackle the problems, we shuffle the utterances randomly. This method preserves the word order within each utterance while randomly mixing of the order between utterances. By randomly arranging the combinations of utterances within the utterance chunks, which serves as the input for a single BERT run, it enables information exchange between words appearing earlier and words within subsequent utterances at the word-level embedding stage. The process of Bert input after Randomly utterance chunks formation can be seen in Figure2. Subsequently, by comparing the results, it would be possible to determine whether the expressiveness

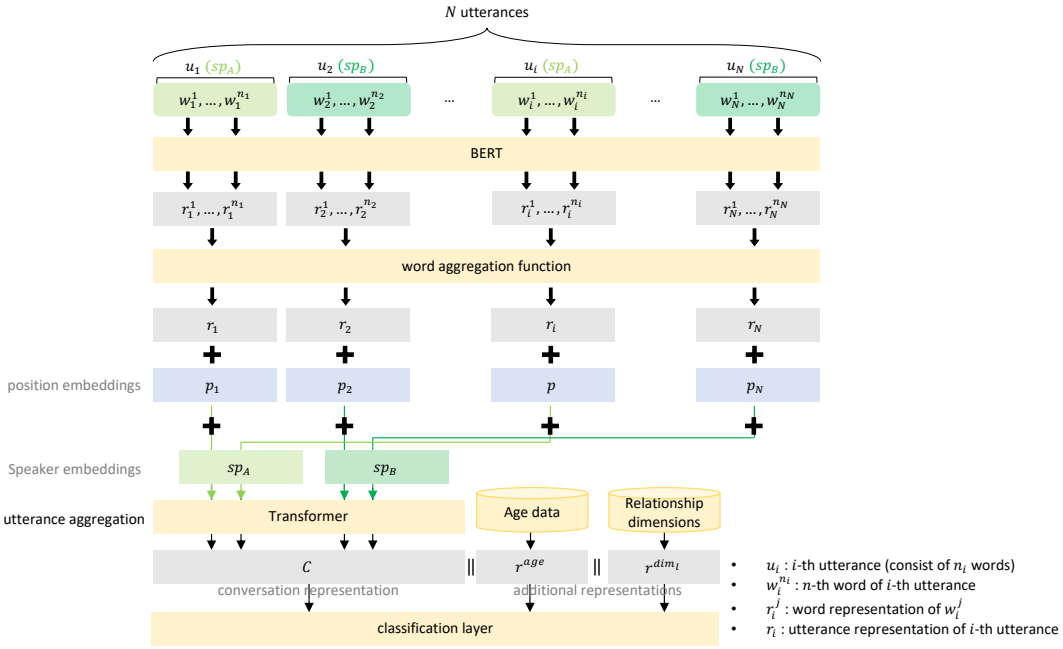

Figure 1: PRIDE model

of utterances or the expressiveness of words is more important in the actual relationship classification through dialogue. Creating utterance chunks through shuffling the utterances randomly means removing the order between utterances in the overall classification task. Therefore, if we observe that the results improve through this approach, it would serve as the evidence that the selected words in the conversation were more crucial in classifying the relationships.

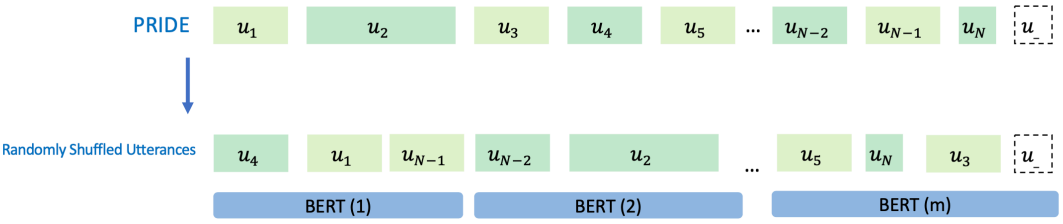

Figure 2: Randomly shuffled utterances before constructing utterance chunks.

3.2 Enhancing conversational context of PRIDE by Overlapping chunks

In order to overcome the limitations of PRIDE mentioned above and enhance the conversational context, we create overlapped chunks. PRIDE creates chunks and runs BERT several times, but there isn't any connection between chunks at word level. Our approach takes full advantage of BERT to capture contextual information efficiently when using BERT at word level, and literally seeks to understand "smooth" contextual information between chunks. The comparison between PRIDE and our method is shown in Figure 3. Original PRIDE simply forms a chunk using each utterance only once to fit the BERT max limitation. In contrast, our method uses rear utterances of previous chunk once more as the front utterances of next overlap chunk, allowing the connectivity between the utterances that makeup the entire conversation to be preserved and inputted at the word level. The following is related to implementation. If the entire conversation is fit with max limitation, that is, there is one chunk, meaning the entire context can be grasped with single BERT run, we do not overlap any chunk. Also, overlap after the last chunk is not carried out, since only the connection with the previous chunk is needed. In other cases, overlap is conducted. How much to overlap is set as a hyperparameter, which is set to 1/3 of total number of utterances constituting the previous chunk. Therefore, the overlap chunk is created by including the utterances corresponding to one-third of the

total number of utterances of the previous chunk and the utterances of the following chunk until they fit the limitation. An overlap chunk is added between existing chunks, maintaining order and context of the conversation. The way of generating the overlapped chunks is same as in PRIDE, so that there is no difference between the existing chunks and the overlapping chunks.

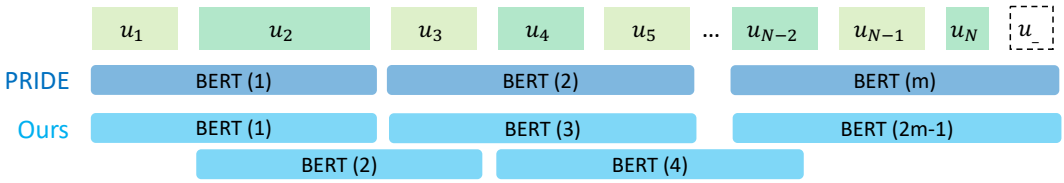

Figure 3: Comparison between original PRIDE and our method over running BERT on each chunk.

4 Experiments

4.1 Data

In this work, we use FiRe—a Film Relationship dataset presented by Tigunova et al. [2021], which contains labeled relationships of fictional characters from well-known movies, which was collected through crowdsourcing. FiRe is based on movie scripts, offering a representative approximation of real-world conversations. To make the dataset, the authors of PRIDE use the `Jinni Movie Dataset` collected in Gorinski and Lapata [2018], which provides speaker labels for each utterance as well as the film genre metadata. The movie selection was guided by two primary considerations: **1)** the feasibility of automated linkage to their corresponding Wikipedia page for annotation purposes, and **2)** the inclusion of genres that reflect actual life situations, such as drama or family, thereby better approximating real-life conversations. Summary statistics of FiRe are given in Table 1.

	avg	max
words per utterance	13	602
utterances per pair	99	597
words per pair	1,087	3,977

Table 1: Statistics for FiRe

4.2 Evaluation method

Following the evaluation procedure as specified in Tigunova et al. [2021], we performed five-fold cross-validation, in which the arrangement of folds ensures that the movies are not overlapping. The models are trained on three folds and we select the best hyperparameters based on the performance on a 1-fold validation set. We show the results on the remaining 1-fold test set in Table 3. In our multi-label classification task, we predict all labels that surpass a particular score threshold, regarded as a hyperparameter. We compute macro-averaged multilabel precision, recall and F1 scores as our evaluation metrics.

4.3 Experimental details

During training processes, we use Binary Cross Entropy loss function. To handle the issue of class imbalance, we apply an oversampling strategy, effectively augmenting the less-represented labels. We follow the hyperparameter settings specified in Tigunova et al. [2021], which is presented in Table 2.

4.4 Results

The main quantitative results are presented in Tabel 3. We show that, excluding the high recall values from PRIDE-shuffle, the proposed algorithm has shown less than satisfactory performance, contrary to our expectations. Additionally, contrary to our hypothesis that combining two methods would enhance performance, the PRIDE-shuffle-overlap shows poorer performance than when not combined.

The PRIDE-shuffle algorithm, despite displaying low precision, achieved high recall rates. However, its F1 score was marginally weaker, suggesting a possible imbalance between recall and precision, which led to a drop in the F1 score. We suspect that the disruption of the utterance sequence might be

hyperparameters	value
BERT learning rate	3e-6
Learning rate for the rest of the model	0.01
Transformer hidden layer size	2048
Age embedding size *m*	64
Training epoch	100
Word aggregation	attention-weighted average
Utterance aggregation	max

Table 2: Hyperparameter settings of our models.

the primary reason for this outcome. On the other hand, PRIDE-overlap has a lower F1 score and precision than PRIDE, with consistent recall. The inclusion of additional context information does not appear to contribute positively, but rather seems to obstruct the main objective of relationship reasoning.

model	cross-val on FiRe		
	F1	**P**	**R**
RNN	0.11	0.11	0.15
BERT$_{ddrel}$	0.2	0.25	0.2
HAM	0.23	0.25	0.22
BERT$_{conv}$	0.27	0.25	0.33
PRIDE (ours)	**0.36**	**0.37**	0.38
PRIDE-shuffle	0.35	0.34	**0.39**
PRIDE-overlap	0.34	0.35	0.38
PRIDE-shuffle-overlap	0.32	0.35	0.34

Table 3: Performance of our algorithms compared to previous algorithms.

In Table 4, we show per class F1 scores for our algorithms. PRIDE-shuffle generally exhibits a higher recall (sensitivity) compared to PRIDE. However, neither PRIDE-overlap nor PRIDE-shuffle-overlap presents a significant performance improvement over PRIDE. Losing balance in terms of accuracy is reflected in the changes observed in the F1 scores for each label class. Shuffling the utterances results in an increase in the F1 scores for some relationships (friend, colleague, sibling, employee, boss, spouse, and commercial). We discuss more analytic details of experiments in Section 5.

class	**PRIDE** (ours)	**PRIDE** -shuffle	**PRIDE** -overlap	**PRIDE** -shuffle-overlap
friend	0.466	**0.517**	0.505	0.505
lover	**0.622**	0.611	0.621	0.588
spouse	0.354	0.391	0.**403**	0.316
colleague	0.222	**0.263**	0.142	0.125
child	**0.606**	0.5	0.517	0.529
parent	0.621	0.566	0.588	**0.622**
sibling	0.3	**0.419**	0.346	0.333
employee	0.267	**0.296**	0.262	0.185
boss	0.136	0.179	**0.2**	0.172
enemy	0.122	0.086	0.034	**0.167**
medical	0.467	0.276	**0.483**	0.25
commercial	0.077	**0.08**	0.065	0.071

Table 4: Class F1 scores of our algorithms compare to PRIDE.

5 Analysis

In PRIDE, they said that strategy that splitting into chunk that utilizes 512 tokens at most and running BERT is effective. Since we propose a method that take advantage of BERT at most for identifying conversation context at word level, we expected that the overall performance of inferring relationships would be further enhanced by strengthening the overall contextualization of the conversation. Through PRIDE-shuffle or PRIDE-overlap algorithms, we thought that understanding the conversational context once more when creating contextual word embedding would help model to understand the context of the entire conversation.

However, for PRIDE-shuffle, the observed decrease in precision, despite the increase in F1 score, indicates that the shuffling method disrupts the balance of model accuracy. This suggests that the algorithm has developed a tendency to predict everything as positive, affecting its overall performance. Interestingly, for specific relationships, the F1 scores are higher than the original PRIDE. This can be interpreted as the shuffling of utterances and word embedding aiding in the classification of those specific labels. Ultimately, it implies that word choice is more important than the order of utterances for those labels.

And for PRIDE-overlap, there has been no improvement in performance. It can be interpreted that the process of identifying the context of the conversation has been repeated twice, providing too much information about the entire context above the appropriate level, which is not helpful of hindered in inferring the relationship. Conversational representations that preserve information about the sequence and speakers of utterances, build by Transformer encoders, are enough as conversational context for predicting relationships of speakers in dialogue.

Lastly, for PRIDE-shuffle-overlap, after shuffling, providing excessive interval information through overlap deteriorates the model's performance. It is speculated that this leads to the loss of meaningful word embedding representations due to the disruption of context.

Taking into consideration that it is a multiple binary classification problem, let's interpret the performance results for each relationship label. First, let's briefly explain the effects of each model. PRIDE, while preserving the order of words and utterances, hierarchically understands the input. PRIDE-shuffle, on the other hand, focuses more on understanding the words within the entire script while breaking the order of utterances. PRIDE-overlap reintegrates the partial discontinuity between utterances while emphasizing the order of utterances even more.

- **Friend, colleague, sibling, commercial and employee**
  These labels show significant improvement with PRIDE-shuffle. This suggests that the selected words from the overall conversation are more crucial in classifying the relationships than the order of utterances.

- **Lover, spouse, boss, and medical**
  These labels show similar or higher improvement with PRIDE-overlap compared to PRIDE. In these relationships, it is important to understand the words without breaking the order of utterances. In other words, the sequence of utterances plays a more significant role in classifying the relationships in these cases.

- **Child and Parent**
  In the case of these labels, PRIDE performs better or similarly to other models. It seems that hierarchical understanding by preserving the order of words and utterances, from word to utterance and from utterance to the overall conversation, is more important in grasping the relationships.

- **Enemy**
  Interestingly, for the label "enemy", PRIDE-shuffle-overlap shows the highest performance compared to other models, particularly surpassing PRIDE. Considering that the previous two models(shuffling and overlapping) have lower performance, this can be interpreted as a mere coincidence. PRIDE-shuffle-overlap, as a combination of shuffling and overlapping, may have unintentionally led to improved performance in this case.

6 Conclusion

In this work, we proposed 2 new methods to improve performance of PRIDE, an algorithm for predicting fine-grained relationships from conversations. we randomly shuffle the order of utterances, assuming that certain words are more important than context when predicting relationships. In addition, we creates overlapped chunks to add additional context of conversation at word level, by making the most of BERT.

Contrary to the expectation that overall performance will be improved, two methods showed slightly lower performance than PRIDE, and the performance decreased even more when two methods were used together. In the case of PRIDE-shuffle, we can find that observing only certain words by mixing the order of utterances did not help much in understanding the context. Anyway, this method improved or did not improve performance by class. It can be interpreted that the method interfered if the order of utterances is important and that the method helped if a specific word is important when identifying relationships. In the case of PRIDE-overlap, we can find that providing too much context information hindered the relationship prediction. The transformer encoder in PRIDE is sufficient to grasp the overall conversational context.

In conclusion, rather than observing only certain words by mixing the order of utterances or over-understanding contextual information by overlapping utterances, it is appropriate to understand the context hierarchically as in the original PRIDE. Although we have not achieved performance improvement, the contributions of our work are as follows. We tried two new methods to improve PRIDE performance: shuffling and overlapping. They were compared to the prior methods in terms of performance. Simultaneously, the performance for each relationship label was compared to observe which approach helped classify specific relationships. Through experiments, we considered the effectiveness of methods in PRIDE. We presented points that need to be improved in relationship-extracting tasks by identifying limitations in the PRIDE model.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Philip Gorinski and Mirella Lapata. What's this movie about? a joint neural network architecture for movie content analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1770–1781, 2018.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. Pride: Predicting relationships in conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4636–4650, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Appendix: Team contributions

- **Younsuk Yeom(2018250028)**
  Code writing for making inputs of PRIDE-shuffle. Writing a paper.
- **Donghyeon Ki(2018250034)**
  Implement experiments of PRIDE, PRIDE-shuffle, PRIDE-overlap and PRIDE-shuffle-overlap. Writing a paper.
- **Chaehyeon Kim(2020250028)**
  Code writing for making inputs of PRIDE-overlap. Writing a paper.