Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes

Hong Guan, MS and Murthy Devarakonda, PhD

Biomedical Informatics, College of Health Solutions
Arizona State University, Tempe, AZ

Abstract

Relation extraction from biomedical text is important for clinical decision support applications. In post-marketing pharmacovigilance, for example, Adverse Drug Events (ADE) relate medical problems to the drugs that caused them and were the focus of two recent shared challenges. While good results were reported, there was a room for improvement. Here, we studied two new improved methods for relation extraction: (1) State-of-the-art deep learning contextual representation model called BERT, Bidirectional Encoder Representations from Transformers; (2) Selection of negative training samples based on the "near-miss" hypothesis (the Edge sampling). We used the datasets from MADE and N2C2 Task-2 for performance evaluation. BERT and Edge together improved performance of ADE and Reason (indication) relations extraction by 6.4-6.7 absolute percentage (and error rate reduction of 24%-28%). ADE and Reason relations contained longer text between the entities, which BERT and Edge were able to leverage to achieve the performance improvement. While the performance improvement for medication attribute relations was smaller in absolute percentages, error rate reduction was still considerable.

Background and Significance

Relation extraction is an important problem in biomedical NLP because it helps to identify critical elements of patient care, such as the medications causing adverse drug events, reasons for treatments, and bases for diagnosis, outcomes, and treatment changes, from clinical text. Even in the general domain NLP, relation extraction is an active and ongoing research area. We studied two new strategies for exploiting contextual information to improve relation extraction performance.

One strategy was to use the state-of-the-art neural network model called BERT,[1] Bidirectional Encoder Representations from Transformers, which was shown to achieve significant performance improvement in general domain NLP tasks over task-specific neural architectures that used word2vec,[2] GloVe,[3] or ELMo[4] word embeddings as features. BERT achieves performance improvement in two ways. First, by using multiple layers of bidirectional Transformer encoder blocks which are based on the self-attention model,[5,6] and second by using a learning model that predicts randomly masked tokens in a sequence, called Masked Learning Model (MLM) which is based on the Cloze procedure in Journalism.[7] The self-attention model was shown to better leverage context compared to the neural networks that use LSTMs (Long Short-Term Memory neural networks) and CNNs (Convolutional Neural Networks).[5] Furthermore, BERT pre-trained model only required fine-tuning with task-specific training data and a simple feed forward layer with softmax to predict relation labels.

The second strategy was to use the "near-miss" hypothesis[8] in selecting negative training samples. Typically, negative samples far exceed positive samples in a training set, and the standard approach is to down sample using random selection. But, the strategies used in active learning suggest potential benefits of using near miss samples.[9] A near miss is a negative sample that differs from the learned concept in only a small number of significant points. In psychology of game playing[10] and in AI learning and reasoning,[11] it was observed that the near misses have a significant effect on the outcome. Here, we propose the Edge sampling, which selects negative samples such that relation entities in it are at the "edge" of (or close to) the corresponding positive sample entities in the text. Thus, an Edge negative sample shares a significant text with the corresponding positive sample and yet differs from it by some text (i.e. follows the near-miss hypothesis).

Two recent biomedical NLP challenges focused on Adverse Drug Event (ADE) relation extraction from clinical notes, Medication and Adverse Drug Events from Electronic Health Records 1.0 (MADE)[12] and National NLP Clinical Challenges Task 2 (N2C2).[13] Two obfuscated examples from MADE are shown in Figure 1. While the top performing systems achieved high accuracy (mid to high 0.9 F measures) in extracting medication attribute relations, such as drug-dosage, drug-route, and drug-strength, they substantially under performed on all important ADE and Reason

(indication) relations (mid to high 0.7 F measures). The ADE and Reason relations typically occur over long distances (e.g. up to 501 words apart in MADE) (see Figure 1) and hence there is a challenge and an opportunity to leverage contextual information in these long-distance relations. Here we, therefore, studied the impact of BERT and the Edge sampling on extracting long distance relations using the two datasets.

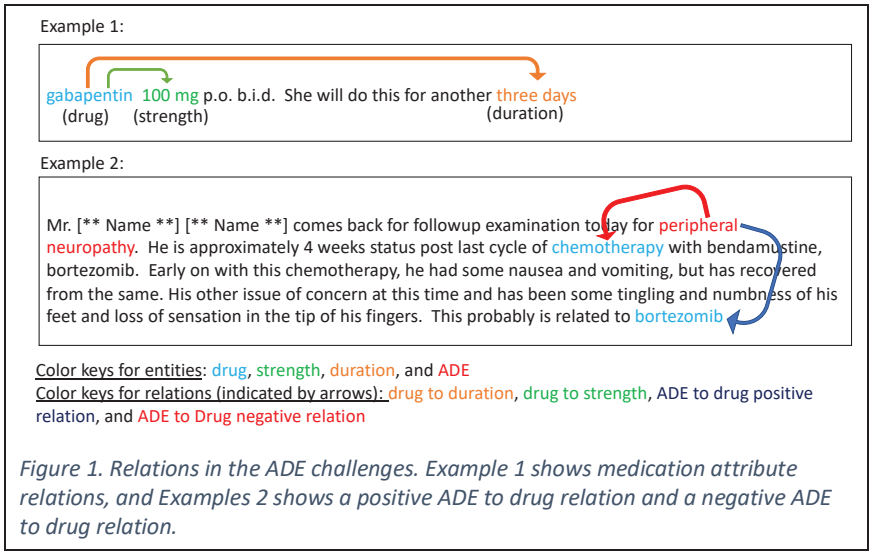


Figure 1. Relations in the ADE challenges. Example 1 shows medication attribute relations, and Examples 2 shows a positive ADE to drug relation and a negative ADE to drug relation.

Our results showed that BERT alone achieved substantial improvement over the top performing systems for the ADE/Reason relations and the Edge sampling provided additional improvement over BERT with random sampling. Further analysis showed distinct differences in the relation distances between ADE/Reason and medication attributes relations. Thus, this study established new state of the art performance for such long-distance relations in clinical NLP.

Methods

Datasets

The MADE dataset contained 1092 de-identified clinical notes of 21 cancer patients. Each note was annotated with medication information (i.e. drug name, dosage, route, frequency, and duration), ADEs, indications (reasons), other signs and symptoms (SSLIFs), and relations among those entities. The data was split into a training set of 900 notes and a test set of 180 notes.

The N2C2 dataset consisted of 505 discharge summaries from the MIMIC-II clinical care database. Each note was annotated with drug names, dosages, durations, and other entities, and relations of drugs with adverse drug events and other entities. The dataset was split into a training set of 303 notes and a test set of 202 notes.

Table 1. The MADE dataset statistics

Relation Type	Training Dataset			Test Dataset		
	Positive Samples	*Possible Negative Samples*	*Ratio*	*Positive Samples*	*Possible Negative Samples*	*Ratio*
ADE-Drug	2,057	27,288	**13.3**	512	7,485	14.6
Reason-Drug	4,530	54,522	**12.0**	871	9,821	11.3
Duration-Drug	901	9,998	**11.1**	146	1,219	8.4
Severity-SSLIF	3,459	25,630	**7.4**	553	3,450	6.2
Dosage-Drug	5,150	45,578	**8.9**	863	6,140	7.1
Frequency-Drug	4,407	39,923	**9.1**	728	4,991	6.9
Route-Drug	2,544	20,776	**8.2**	454	3,188	7.0
Form-Drug	N/A					
Strength-Drug	N/A					
Overall	23,048	223,715	**9.7**	4,127	36,294	8.8

Table 2.The N2C2 dataset statistics

Relation Type	Training Dataset			Test Dataset		
	Positive Samples	*Possible Negative Samples*	*Ratio*	*Positive Samples*	*Possible Negative Samples*	*Ratio*
ADE-Drug	1,061	4,430	**4.2**	724	2,912	4.0
Reason-Drug	4,991	29,751	**6.0**	3,392	20,029	5.9
Duration-Drug	642	3,305	**5.1**	425	2,022	4.8
Severity-SSLIF	N/A					
Dosage-Drug	4,206	20,591	**4.9**	2,690	12,801	4.8
Frequency-Drug	6,303	48,993	**7.8**	4,031	31,472	7.8
Route-Drug	4,995	34,370	**6.9**	3,544	19,724	5.6
Form-Drug	6,647	34,011	**5.1**	4,371	21,383	4.9
Strength-Drug	6,703	45,766	**6.8**	4,242	29,109	6.9
Overall	35,548	221,217	**6.2**	23,419	139,452	6.0

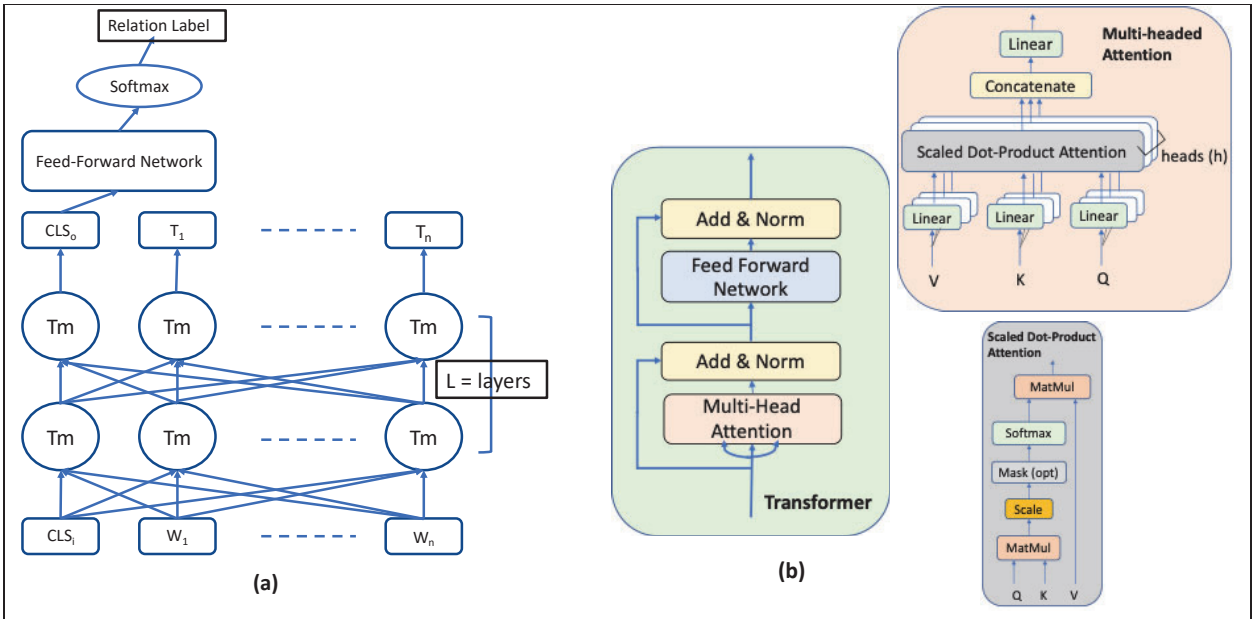
Detailed characteristics of the two datasets are shown in Tables 1 and 2, which include the number of positive relations and potential negative relations in the training and test sets broken down by relation types. The ratios of potential negative relations to positive relations are also shown in the tables. It can be seen that the ratios are high in the training data of both sets, i.e. substantially more potential negative relations than the positive relations, and the ratios are even

higher in the MADE training dataset. Specifically, the ratio is 13.3 for the ADE-Drug relation, 12.0 for Reason-Drug, and 9.7 overall in the MADE training dataset. These high ratios indicate an opportunity to strategically select negative samples to optimize training.

BERT (Bidirectional Encoder Representations from Transformers)

BERT is a new exciting development in neural network models research, demonstrating significantly improved state-of-the-art performance on various general domain NLP tasks,[1] including sentence classification which is relevant to us here. BERT is a pre-trained model that produces sequence (e.g. sentence) and word level representations, which can be fine-tuned for task-specific outcomes such as relation classification and concept extraction. Only a simple feed-forward network with a softmax layer is needed to process the BERT output for task-specific objectives.

In the last few years, word2vec[2] has become the de facto standard for producing feature representations of words in biomedical NLP, which are then processed by task-specific neural architectures such as RNNs, LSTMs, CNNs, and heavily-engineered combinations thereof. On occasion, GloVe[3] or ELMo[4] were used instead to generate word representations, but still required complicated task-specific neural network engineering. BERT approach is to do away with the task-specific architectures and provide a broadly applicable pre-trained model which only need to be fine-tuned for the task, using task-specific training data. Other such approaches were also proposed, for example, OpenAI Generative Pre-trained Transformer (GPT),[14] but BERT was shown to outperform them with a multi-layer bidirectional architecture.[1]



(c) <source> peripheral neuropathy </source> . He is approximately 4 weeks status post last cycle of <target> chemotherapy </target>

Figure 2. (a) BERT model, Tm is the transformer encoder and w_1 to w_n are input sequence tokens; (b) Transformer encoder details and the multi-headed attention component, reproduced from Vaswani et al[5]; (c) an example input sequence to BERT without the CLS_0 toke.

As shown in Figure 2a, BERT uses layers of neural network components known as Transformer encoders, shown as *Tm* in the figure, to generate representations of input sequences in the output. Each BERT layer processes its input sequence in the forward and backward directions *simultaneously*, using a novel pre-training objective known as the masked learning model (explained later). The BERT Transformer encoder contains two sublayers (see Figure 2b), the first sublayer is a multi-head self-attention[5] mechanism that allows modeling of the context for each word position and the second is a feed-forward network that provides non-linear activation. The encoder architecture also included a residual connection around each sublayer which was shown to simplify optimization[15] and a sublayer of normalization[16] which reduced computational requirements. Fundamentally, the attention function provides a mapping of a query and a key-value pair to an output. Intuitively and as applied here, an attention layer produces output (say, a word representation) that is based on any arbitrary word positions (of the input sequence) by comparing each

sequence member with each other sequence member (self-attention) and producing a series of probability distributions to assign importance. Multi-headed attention can simultaneously optimize for different input combinations.

As mentioned earlier, in pre-training, BERT uses a novel pre-training objective known as the masked learning model (MLM),[7] where some random words in the input are masked, and the pre-training objective is to predict the original word based on the context. In other approaches, typically next word prediction was used, which limited a multi-layered model to process the input either in the forward direction only or process in the forward and backward directions separately and then aggregate the representations[14] – both these approaches fail to leverage the forward and backward contexts at the same time. The use of MLM was a key invention that enabled simultaneous bidirectional input processing in a multi-layer model, without allowing words-to-be-predicted appearing in the input of an upper layer.

The BERT model we used here came pre-trained with BookCorpus[17] and English Wikipedia (general domain corpus). Corpus words were tokenized using the WordPiece dictionary[18] of 30,000 words and as needed words were split into pieces using ## (two hash marks). Word piece representations using biomedical corpus were not readily available at the time of our study, however, recently a BERT model was pre-trained on a biomedical corpus[19] and it was shown to improve entity extraction.[20] In the future, we plan to study performance of this BERT model, that was pre-trained on biomedical corpus, in relation extraction.

We employed BERT in its base configuration of 12 layers (=24 sublayers), 768 hidden size, 12 self-attention heads. In our configuration, BERT produced representations for each word (token) in the input as well as a single sequence representation (shown as CLS_o). This sequence classification representation from the top layer of BERT (see Figure 2a) was used as the input to a feed-forward layer containing one hidden layer. A softmax layer provided the final relation classification label for the pair of entities in the input sequence.

In our method, as shown in Figure 2c, the input to BERT was a sequence of words that started with the first entity of a relation and ended with the second entity of the relation. As previous studies demonstrated,[21] this is a convenient choice since sentence segmentation of clinical notes text is error prone due to embedded lists and tables which are not well handled by the standard NLP code such as NLTK.[22] We were also constrained by our BERT model input limit which was 512 word-pieces. Entity spans were further marked using the entity tags.

In the training (fine-tuning) stage, the input sequences included all positive samples in the gold standard and an equal number of negative samples, that were either randomly down sampled from all possible negative samples or Edge sampled as described in the next section. Only the sequences that were less than or equal to a heuristically determined (in the validation stage) maximum sequence length were used in training.

The Edge Sampling

In the MADE training dataset, there were 13.3 times as many potential negative samples as the positive samples for the ADE relation. In the N2C2 training dataset the ratio was 4.2. Ideally, the negative and positive samples should be equal for optimum model training. The new sampling method we propose here, called the Edge sampling, tries to select negative samples that share the largest common text (and hence context) with the corresponding positive samples, i.e. near-miss samples.

The Edge sampling works as follows. For each entity of a gold standard relation, the immediately preceding and following entities of the same type as the gold standard entity are identified (if they exist) in the text. From these four neighboring entities and from the original two entities from the gold standard, valid relations are formed by having one (and exactly one) of the gold standard entities as a part of the relations. Valid means the relation meets its entity type requirements and the entities occur in a sequence of words in the document. If entities in a gold standard relation are of different types (as in the case of relations in MADE and N2C2), at most four such new relations can be formed. This is illustrated in Figure 3. Note that, depending on the location and entities in a clinical note, we may find four or less, including zero, such new relations. If any of these relations were already in the gold standard or were picked already as a negative sample, they were removed from the list. Among the remaining relations, the Edge sampling randomly picks one.

If an Edge sample cannot be found for a positive sample, a random negative sample (that was not already picked) was selected from the rest of possible negative samples. Negative example selection takes place only in the training phase. During the evaluation and validation phases, all potential relations were assessed by the model and outcome measured.

Due to practical limitations, our model enforces a maximum sequence length, which is heuristically determined during the validation phase. In training, positive samples longer than the max length or ignored, and similarly in the Edge

Entity types of a relation type R are:	T_1 and T_2
Entity instances of type T_1 in the text (p and f indicate immediately preceding and following i)	T_{1p}, T_{1i}, T_{1f}
Entity instances of type T_2 in the text (p and f indicate immediately preceding and following i)	T_{2p}, T_{2i}, T_{2f}
Gold standard positive relation of type R, R_i	[T_{1i}, T_{2i}]
For R_i, Edge relation sequences are: R_1, R_2, R_3, R_4	[T_{1i}, T_{2p}], [T_{1i}, T_{2f}], [T_{1p}, T_{2i}], [T_{1f}, T_{2i}]

Figure 3. The Edge sampling illustrated.

sampling, potential samples longer than the max length are not considered. In the evaluation phase also, the longer sequences were ignored and the system was penalized for it in performance calculations.

Experiments and Metrics

We fine-tuned our BERT-based neural network model separately for MADE and N2C2, each with and without the Edge sampling. We therefore trained (fine-tuned) the model four different times and tested each one. We mostly adopted the default settings of BERT hyperparameters – i.e., training batch size of 16, 10 epochs, and a learning rate of 2e-5. However, we experimentally determined the optimum sequence length using validation sets - 20% of clinical documents from the N2C2 training set and 10% of clinical documents from the MADE training set. The fine-tuned models were tested on the full MADE and N2C2 test datasets.

For each of the four experiments, we calculated standard Recall, Precision, and F measures individually for all relations as well as for all the relations combined. We compared our results with the published results of the systems that performed best in the task of relation extraction given gold entity labels from the two challenges: the University of Utah system[23] for MADE (denoted as the MADE_Best), and the UTHealth developed system[24,25] for N2C2 (denoted as the N2C2_Best). Since the published results did not aggregate performance for the ADE and Reason relations, we obtained their weighted average, weighted by the number of relations evaluated in the test phase, from individual relation results. We used two metrics for comparison: (1) Absolute F measure difference; and (2) Error rate reduction in the F measure achieved by model Y compared to model X, which is calculated as:

$$error\ rate\ reduction = \frac{F_y - F_x}{1 - F_x}$$

where F_y and F_x are the F measures of models Y and X respectively. While the absolute F measure difference shows the net improvement in the measure, the error rate reduction is a sound relative measure that shows reduction in the remaining performance gap of the previous model. It is usually expressed as a percentage. Recent studies in the general domain NLP have adapted this metric for effective comparison.[1,2]

We suspected that for certain relation types, such as the medication attributes, the distance between the entities may be short and in which case the potential to leverage contextual information is rather limited. The two datasets might also be different in terms of the distances between entities, since N2C2 contains discharge summaries whereas MADE contains clinical notes. In order to quantify such differences, we studied the sequence length distributions of positive samples, Edge samples, and all negative samples for each dataset and plotted their cumulative distribution frequency.

Results

Table 3 shows performance evaluation on the MADE dataset. Precision, recall, and F measures were shown for BERT (using random negative instances sampling) and BERT with the Edge sampling (shown as BERT+Edge) for each relation type, for the aggregate of ADE and Reason relations, and for all relations together. The table also shows

absolute F measure differences between MADE_Best and BERT, BERT and BERT+Edge, and between MADE_Best and BERT+Edge. Percentage error rate reduction was shown between MADE_Best and BERT+Edge.

In terms of absolute F measure, BERT achieves 1.1% improvement overall compared to MADE_Best, but 4.1% improvement for the ADE and Reason relations. Edge improved the overall F measure by additional 1.6% compared to MADE_Best and by additional 1.9% for the ADE and Reason relations. BERT improved the Reason relation extraction F measure significantly (by 6.6%), while Edge improved the ADE relation F measure significantly (by 3.3%). Performance improvement from BERT+Edge over MADE_Best is substantial: absolute F measure improved by 2.7% overall and by 6.4% for ADE and Reason. Notice that the error rate was reduced by 22.7% overall and 23.6% for ADE and Reason. The error rate reduction was substantial for most relations.

Table 3. Performance results for the MADE dataset

Relation Type	MADE_Best (P/R/F)	BERT (P/R/F)	BERT over MADE_Best: Δ F measure	BERT+Edge (P/R/F)	BERT+Edge over BERT: Δ F measure	BERT+Edge over MADE_Best	
						Δ F measure	err. reduced%
ADE-Drug	0.787/0.683/0.731	0.652/0.848/0.737	0.006	0.730/0.814/0.770	0.033	0.039	14.5%
Reason-Drug	0.780/0.739/0.758	0.728/0.948/0.824	0.066	0.772/0.904/0.833	0.009	0.075	31.0%
Duration-Drug	0.937/0.912/0.924	0.759/0.973/0.853	(0.071)	0.946/0.952/0.949	0.096	0.025	32.9%
Severity-SSLIF	0.911/0.962/0.936	0.959/0.971/0.965	0.029	0.966/0.977/0.971	0.006	0.035	54.7%
Dosage-Drug	0.957/0.962/0.960	0.924/0.983/0.952	(0.008)	0.939/0.970/0.954	0.002	(0.006)	(15.0%)
Frequency-Drug	0.971/0.923/0.947	0.935/0.966/0.950	0.003	0.942/0.967/0.955	0.005	0.008	15.1%
Route-Drug	0.961/0.921/0.941	0.941/0.976/0.958	0.017	0.957/0.980/0.968	0.01	0.027	45.8%
Form-Drug	N/A						
Strength-Drug	N/A						
Overall	0.903/0.859/0.881	0.839/0.953/0.892	0.011	0.880/0.938/0.908	0.016	0.027	22.7%
ADE+Reason	0.783/0.712/0.746	0.700/0.911/0.791	0.041	0.757/0.871/0.810	0.019	0.064	23.6%

Table 4 shows performance evaluation on the N2C2 dataset in the same way as in Table 3. The general trend of the results is similar to that of the MADE results but notably the absolute F measure improvement for the aggregate of all relations is somewhat smaller. BERT achieves a modest 0.7% F measure improvement overall and Edge managed to add additional 0.4% F measure to a total of 1.1% improvement. On the other hand, BERT improved the F measure of the Reason relation significantly, by 6.8%. The highest Edge F measure improvement was for the ADE relation, adding an additional 1.3% to BERT. While the F measure improvement from BERT+Edge over N2C2_Best for all relations combined is relatively small, 1.1%, but the error rate reduction was a substantial 18.3% indicating that the

Table 4. Performance results for the N2C2 dataset

Relation Type	N2C2_Best F measure	BERT (P/R/F)	BERT over N2C2_Best Δ F measure	BERT+Edge (P/R/F)	BERT+Edge over BERT Δ F measure	BERT+Edge over N2C2_Best	
						Δ F measure	err. reduced%
ADE-Drug	0.795	0.795/0.825/0.810	0.015	0.818/0.828/0.823	0.013	0.028	13.7%
Reason-Drug	0.758	0.784/0.872/0.826	0.068	0.807/0.858/0.832	0.006	0.074	30.6%
Duration-Drug	0.883	0.915/0.890/0.902	0.019	0.915/0.913/0.914	0.012	0.031	26.5%
Severity-SSLIF	N/A						
Dosage-Drug	0.972	0.970/0.977/0.973	0.001	0.974/0.977/0.976	0.003	0.004	14.3%
Frequency-Drug	0.969	0.969/0.981/0.975	0.006	0.980/0.977/0.979	0.004	0.010	32.3%
Route-Drug	0.974	0.967/0.980/0.974	0.000	0.973/0.973/0.973	(0.001)	(0.001)	(3.8%)
Form-Drug	0.977	0.980/0.987/0.983	0.006	0.991/0.984/0.987	0.004	0.010	43.5%
Strength-Drug	0.987	0.968/0.987/0.977	(0.010)	0.980/0.983/0.982	0.005	(0.005)	(38.5%)
Overall	0.94	0.935/0.960/0.947	0.007	0.947/0.955/0.951	0.004	0.011	18.3%
ADE+Reason	0.763	0.786/0.864/0.823	0.060	0.809/0.853/0.830	0.007	0.067	28.3%

improvement in relative terms is still significant. The F measure improvement for the aggregate of ADE and Reason relations is 6.7%, and the corresponding error rate reduction was 28.3%, both of them indicate significant improvement. As in the case of MADE results, the error rate reduction was substantial for most relations. Precision and recall details were publicly unavailable for the N2C2_Best at the time of writing this paper.

Statistical significance test: Previous studies[26] have used the McNamara test (and is generally accepted as a good test) for determining the statistical significance of F measure improvement of an NLP task. The test requires the contingency (confusion) table from the performance study. Using the data in our study, we determined that the F measure improvement with the Edge sampling was statistically significant at $p<0.001$ for the MADE dataset for all relations combined and for ADE+Reason relations. For the N2C2 dataset, the improvement was also significant at $p<0.001$ for all relations combined but was only significant at $p<0.03$ for ADE+Reason relations. We could not determine statistical significance of performance improvements relative to the MADE_Best and N2C2_Best models because the contingency tables for them are not publicly available at the time of this article.

Another important observation from Tables 3 and 4 is that the Edge sampling consistently improved precision, while often losing ground on recall. Edge improved precision for the overall and ADE+Reason by 4.1% and 5.7% for the MADE dataset, and by 1.2% and 2.3% for the N2C2 dataset respectively. Recall reduced by small percentages across the broad. These results indicate an important characteristic of our Edge sampling approach. We also note that both BERT and BERT+Edge consistently improved recall over MADE_Best.

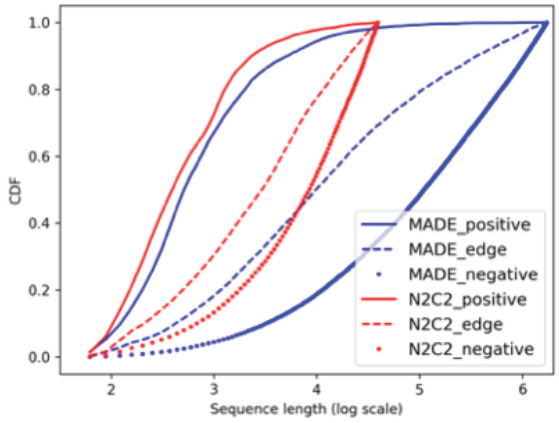


Figure 4. Cumulative Distribution Frequencies (CDF) of relation sequence lengths of the ADE and Reason relations.

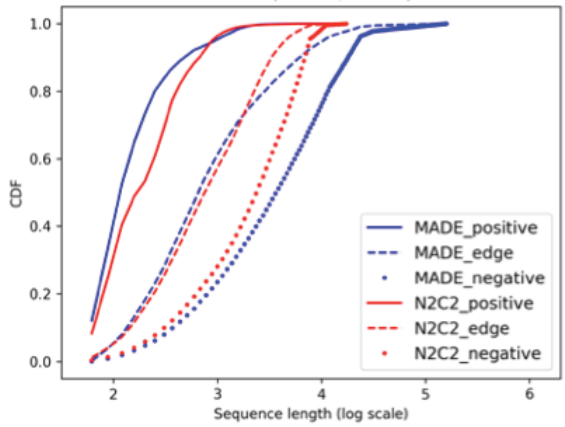


Figure 5. Cumulative Distribution Frequencies (CDF) of relation sequence lengths of the rest of the relations.

The sequence length distributions of the positive, all negative, and Edge-sampled relations in the training datasets of MADE and N2C2 are shown in Figures 4 and 5 respectively. We showed the distributions for the ADE and Reason relations and for all the rest of relations. We showed detailed statistics of the distributions in Table 4. Two important observations can be made from the Figure and the Table.

First, the ADE and Reason relation lengths are significantly different from the lengths of the rest of the relations, especially those of the potential negative samples. In the MADE dataset, the median length of the ADE and Reason negative relations is 155 words, which is 4.3 times the median length of the rest of the relations (see Table 5). In N2C2, the ratio is more modest 1.7 times but it is still significant.

Second, the negative sample relation lengths in MADE are significantly longer than in N2C2 for the ADE and Reason relations. The median length of the relations in MADE is 3.04 times the median length of N2C2 negative relations, i.e. 155 versus 51. The median lengths of the rest of the relations are similar in both datasets.

Table 5. Sequence length statistics for positive, all potential negative samples, and Edge samples

Samples	Statistic	MADE		N2C2	
		ADE and Reason	The Rest of Relations	ADE and Reason	The Rest of Relations
Positive Samples	Mean	22.7	9.8	17.6	11.0
	Std. Dev.	26.7	4.5	12.1	4.6
	Median	**16**	**8**	**14**	**10**
	Max	501	59	99	68
All negative samples	Mean	190.5	40.0	52.0	29.8
	Std. Dev.	139.9	24.3	25.6	12.8
	Median	**155**	**36**	**51**	**30**
	Max	511	181	99	69
Edge samples	Mean	98.0	22.8	37.5	20.4
	Std. Dev.	107.8	17.0	23.8	10.0
	Median	**55**	**17**	**32**	**18**
	Max	511	181	99	69

While some differences exist between the two datasets, the ADE and Reason relations are consistently longer than the rest, and therefore, these sequences are likely to contain more context, which can be leveraged by BERT and the Edge sampling. The Figure and the Table also show that the Edge sampling, reduces the negative sample lengths. For example, the median lengths for ADE/Reason are 55 and 32 for the two datasets, which are substantially smaller than the median lengths of all negative samples (i.e. 155 and 51).

Discussion

The methods we used here, BERT and the Edge sampling, better leverage contextual information to improve relation extraction compared to the top performing systems from the shared challenges, MADE and N2C2. BERT makes better use of context in two ways:

1. Multiple layers of Transformer encoders based on the Attention model, rather than the RNN, LSTM, CNN, or combination models thereof, was shown to better leverage context;
2. Using the Masked Learning Model that randomly masks a word in a sequence for conditioning word representations to simultaneously analyze the input bidirectionally, rather than predicting merely an association of words (as in word2vec), or the next words (as in OpenAI GPT), or even a concatenation of representations predicting next and preceding words (as in ELMo).

In addition, BERT only required fine-tuning of a pre-trained model rather than complex task-specific neural networks that use word representations as features thus achieving transfer learning from a large text corpus. For these reasons, BERT in our study performed better than the top performing systems from the challenges. It is interesting to note that the MADE_Best employed a carefully feature-engineered Random Forest and N2C2_Best used complex composition of neural networks consisting of LSTMs and CNNs, and BERT improved upon both approaches.

Selection of negative samples in model training is known to be a challenge,[27–29] especially when relation entities are multiple sentences apart which gives rise to a very large number of potential negative samples. Most previous studies simply down sampled the larger population, but one study[21] considered a feature-engineered Alternating Decision Tree machine learning model for selecting candidate samples, both for training and testing. The Edge sampling, by preferring negative training samples that share significant context with positive samples, provides a simpler and in combination with BERT higher accuracy compared to the previous candidate selection approach on the MADE dataset.

In can be observed from the results that the performance improvement achievable from contextual information depends on the length of context between relation entities. Quantitatively, the relation distance differences between ADE/Reason and medication attribute relations can be seen in Figures 4 and 5, and in Table 5. Qualitatively, as can be seen from the examples in Figure 1, medication attributes tend to appear in short segments and often have a simple and easy to recognize patterns. These simple patterns can be easily recognized more easily and indeed, the previous studies have achieved very high F measures, often above 0.95, by recognizing such patterns. Whereas, the entities of ADE and Reason appear anywhere and in complex discourse in a clinical document. These typically longer relations offer an opportunity for the sophisticated neural architecture of BERT to create better representations, and hence increase F measure accuracy substantially. It should however be noted that the error rate reduction is significant across most relations with BERT plus Edge.

The Figures 4 and 5, and the Table 5 also show that the Edge sampling, not only takes advantage of the contextual information (by definition) but also reduce the negative sample lengths and make their lengths similar to the positive samples. As we noted earlier, the Edge sampling disproportionately improves precision over recall, which we plan to study further to understand the reasons for it and how it can be leveraged further.

Conclusion

This study makes the following important contributions:

1. We showed that the BERT model, which leverages bidirectional contextual information with multi-layer Transformers, requiring only fine-tuning can provide excellent performance in biomedical relation extraction without complicated, task-specific neural network designs containing RNSs, LSTMs, and CNNs.
2. We showed that the "near-miss" based the Edge sampling of negative instances, rather than random selection, can improve training and therefore model performance especially when there is a large population of potential negative samples to choose from and when relation sequences are long enough to form "near-misses" from positive samples.

3. Both BERT and Edge use contextual information in long distance relations to achieve significant performance improvement. The performance improvement, compared to the previous methods, was substantial for such relations (i.e. ADE and Reason relation): 6.4% absolute F measure improvement (23.6% error reduction) for the MADE dataset and 6.7% absolute F measure improvement (28.3% error reduction) for the N2C2 dataset.

Conflict of Interest

The authors do not have any conflicts of interest.

References

1. Devlin J, Chang. M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT*. Minneapolis, MN; 2019.
2. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Lake Tahoe, California; 2013:1-9. doi:10.1162/jmlr.2003.3.4-5.951
3. 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. doi:10.3115/v1/D14-1162
4. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: *Proceedings of NAACL-HLT 2018*. New Orleans, Louisiana; 2018.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA; 2017. doi:10.1017/S0952523813000308
6. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: *Proceedings of ICLR 2015*. ; 2015. doi:10.1146/annurev.neuro.26.041002.131047
7. Taylor WL. Cloze procedure: A new tool for measuring readability. *Journal Bull*. 1953;30(4):415-433.
8. Reid RL. The Psychology of the Near Miss. *J Gambl Behav*. 1986;2(1):32-39.
9. Gurevich N, Markovitch S, Rivlin E. Active Learning with Near Misses. In: *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*. Boston, Massachusetts; 2016:362-367.
10. Habib R, Dixon MR. Neurobehavioral evidence for the "Near-Miss" effect in pathological gamblers. *J Exp Anal Behav*. 2010;93(3):313-328. doi:10.1901/jeab.2010.93-313
11. Winston PH. Learning structural descriptions from examples. In: Winston PH, ed. *The Psychology of Computer Vision*. New York: McGraw-Hill Book Company; 1975.
12. UMass BioNLP. NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE1.0). https://bio-nlp.org/index.php/projects/39-nlp-challenges. Accessed February 5, 2018.
13. National NLP Clinical Challenges Task 2: Adverse Drug Events and Medication Extraction in EHRs. https://n2c2.dbmi.hms.harvard.edu/track2. Published 2018. Accessed May 8, 2018.
14. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understandingby Generative Pre-Training. OpenAI. https://blog.openai.com/language-unsupervised/. Published 2018.
15. He K. Deep Residual Learning for Image Recognition. *arXiv Prepr arXiv151203385v1*. 2015.
16. Ba JL, Kiros JR, Hinton GE. Layer Normalization. *arXiv Prepr arXiv160706450v1*. 2016.
17. Zhu Y, Kiros R, Zemel R, et al. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. ; 2015:19-27. doi:10.1109/ICCV.2015.11
18. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv Prepr arXiv160908144v2*. 2016.
19. Lee J, Yoon W, Kim S, et al. BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *arXiv Prepr arXiv190108746*. 2019.
20. Si Y, Wang J, Xu H, Roberts K. Enhancing Clinical Concept Extraction with Contextual Embeddings. *arXiv Prepr arXiv190208691v3*. 2019.
21. Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations using Neural Networks. *Drug Saf*. 2019:0-2. doi:10.1007/s40264-018-0764-x

22. Natural Language Toolkit (NLTK). https://www.nltk.org/. Accessed February 1, 2019.
23. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson O V. Detecting adverse drug events with rapidly trained classification models. *Drug Saf*. 2018. doi:10.1007/s40264-018-0763-y
24. Stubbs A, Buchan K, Filannino M, Uzuner O. National NLP Clinical Challenges Task 2 Results. https://portal.dbmi.hms.harvard.edu/projects/n2c2-t2/. Published 2018.
25. Wei Q. UTH: Identifying Medications and Corresponding Attributes in Electronic Health Records (Slides Presented at AMIA N2C2 Workshop https://n2c2.dbmi.hms.harvard.edu/). 2018. https://n2c2.dbmi.hms.harvard.edu/.
26. Dror R, Reichart R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In: *Proceedings of the 56th Annual Meeting Ofthe Association for Computational Linguistics*. Melbourne, Australia; 2018:1383-1392.
27. Swampillai K, Stevenson M. Extracting relations within and across sentences. In: *Proceedings of Recent Advances in Natural Language Processing*. Hissar, Bulgaria; 2011:25-32.
28. Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain; 2017:1171-1182.
29. Peng N, Poon H, Quirk C, Toutanova K, Yih W. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Trans Assoc Comput Linguist*. 2017;5:101-115.