Argument Structure Mining Based on Effective Usage of Contextual Information

Menglong Xu[1], Yanliang Zhang[1*], and Yapu Yang[2]
[1] School of Physics & Electronic Information Engineering, Henan Polytechnic University,
Jiaozuo, 454000, China.
[e-mail: xml@home.hpu.edu.cn, ylzhang@hpu.edu.cn]
[2] Xu Ji ELECTRIC CO., LTD.,
Xuchang, 461000, China.
[e-mail: yangyapu@163.com]
[*]Corresponding author: Yanliang Zhang

*Received December 4, 2023; revised November 19, 2024; accepted January 10, 2025;
published January 31, 2025*

Abstract

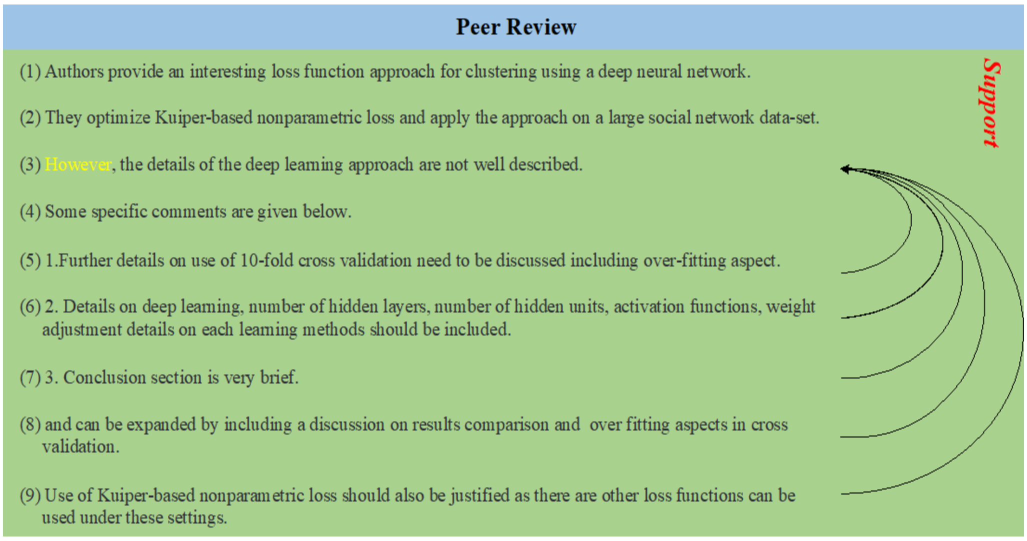
Argument Structure Extraction (ASE) is increasingly prominent for its role in identifying discourse structure within documents. Many pioneering works have demonstrated that the contextual information in the document is vital for the final performance of ASE. Traditional context-aware methods relying on concatenation of contextual sentences have proven insufficient, introducing noise, inefficiency, and bias due to the reliance on discourse markers. To overcome these issues, we introduce Efficient Argument Structure Extraction (E-ASE), which eschews sentence concatenation in favor of encoding sentences separately and applying sentence-level attention to integrate context. To mitigate discourse marker bias, E-ASE employs a novel data augmentation technique, substituting discourse markers with a [MASK] token and leveraging Masked Language Modeling (MLM) loss. Our empirical research, conducted across five diverse datasets, demonstrates E-ASE's state-of-the-art (SOTA) performance, save for on the ECHR dataset, marking a significant advancement in the field of ASE by optimizing contextual information usage and enhancing both the training and inference processes.

Keywords: Argument structure extraction; Contextual information; Discourse markers; Masked Language Modeling; Data augmentation.

http://doi.org/10.3837/tiis.2025.01.001 **ISSN : 1976-7277**

1. Introduction

Argument Structure Extraction (ASE), which aims to identify the discourse structure of arguments in documents, has been an important sub-task of argument mining [1-7]. ASE is usually formatted as the problem of automatic argumentative relation prediction: given any proposition in a document, predicting the existence and polarity (support or attack) of relation from any other proposition within the full document or a context window [2] [5] [6] [8]. Being an important role in discovering the central theses and reasoning process, ASE has aroused more and more attention from academic and industrial communities in a wide spectrum of domains, such as legal documents [9-12], scientific articles [13-16], online posts [17-20], and biomedical literature [21] [22]. **Fig. 1** provides an example for ASE, where the fifth through ninth sentences describe the shortcomings of the review paper, thereby supporting the conclusion sentence(the third sentence).



Peer Review

(1) Authors provide an interesting loss function approach for clustering using a deep neural network.

(2) They optimize Kuiper-based nonparametric loss and apply the approach on a large social network data-set.

(3) However, the details of the deep learning approach are not well described.

(4) Some specific comments are given below.

(5) 1.Further details on use of 10-fold cross validation need to be discussed including over-fitting aspect.

(6) 2. Details on deep learning, number of hidden layers, number of hidden units, activation functions, weight adjustment details on each learning methods should be included.

(7) 3. Conclusion section is very brief.

(8) and can be expanded by including a discussion on results comparison and over fitting aspects in cross validation.

(9) Use of Kuiper-based nonparametric loss should also be justified as there are other loss functions can be used under these settings.

Fig. 1. An example of Argument structure extraction in peer reviews. The fifth through ninth sentences support the third sentence. The argument discourse markers are marked in yellow.

Traditional methods for ASE usually rely on high-quality labeled data from domain experts and manually designed customized features for addressing long dependencies and encoding task-specific language [6] [23] [24] [25]. It is very time-consuming for these methods to be utilized in application scenarios, where the queries have very long contexts and range from different domains [23] [26] [27]. To mitigate this problem, [26] firstly propose the context-aware ASE, which can be directly fine-tuned from pre-trained Transformers [28] [29]. In contrast to the prior works which only encode pairwise propositions while ignoring the contexts [25], the context-aware ASE model takes a different approach. It constructs contextual information for a given proposition by concatenating it with its neighboring sentences within a constant context window. Subsequently, this contextual information is encoded utilizing a pre-trained encoder. Concurrently to our work, [30] also proposed an efficient method for leveraging context information. Specially, [30] followed the strategy of constructing context information by concatenating neighboring sentences within a context window, as introduced by [26]. In contrast, our method involves feeding individual propositions within a document into Roberta separately and in parallel. Extensive experimentation has demonstrated the critical importance of contextual information for the final performance of ASE.

Table 1. Statistics of five datasets. We report the discourse markers rate of propositions pairs in test sets and the previous experimental results on two test sets with and without discourse markers. 'w' refers to with, 'w/o' refers to without.

	discourse markers rate	**test sets**	
		***w/* discourse markers**	***w/*o discourse markers**
AMPERE	17.72%	77.40	71.20
Essays	13.17%	72.81	65.46
AbstRCT	12.16%	75.13	69.87
ECHR	16.97%	69.15	60.49
CDCP	14.07%	68.35	62.91

While the context-aware ASE has achieved some improvements, our motivated experiments find that their way of utilizing the contextual information, i.e., simply concatenating the sentences in the context window, cannot make full use of the contextual information. The main drawback of their method can be three-fold. Firstly, concatenating all sentences in the contextual window may introduce much noise, misleading the model to pay attention to less informative contextual sentences and thus degrading the final performance. Additionally, concatenating all sentences in the contextual window results in an excessive length input fed into the encoder. To the best of our knowledge, the performance of the Transformer degrades with the increase of the input length [31]. Secondly, since the reconstruction of contextual information for each proposition during inference, the efficiency of both training and inference is notably diminished. At each step of training or inference progress, the model encounters substantial redundancy in encoding such lengthy contextual inputs. Furthermore, as the size of the context window is set as a constant during the modeling process, the propositions lacking enough contextual sentences are essentially disregarded, which causes low efficiency for data utilization. Thirdly, the ubiquitous discourse markers in the corpus, such as `but`, `hence`, and et., behave as significant signals for ASE, which may introduce much bias for the model's prediction and enable the model to give the right predictions but neglect the detailed context. **Table 1** illustrates the statistics of the discourse markers rate of proposition pairs within the test sets, along with our previous experimental results on two distinct test sets: one inclusive of discourse markers and another where discourse markers are absent. From **Table 1**, we can find that the performance on the test set with discourse markers is much higher than those without discourse markers. These results show that the discourse markers have a significant effect on the ultimate performance.

To better leverage contextual information and address the issues outlined above, we introduce E-ASE, a model aimed at enhancing ASE performance through effective utilization of contextual information. Specifically, to remove the contextual noise and enhance the training and inference efficiency, the proposed E-ASE encodes each sentence separately without concatenating contextual sentences as a whole. Therefore, the model only needs to encode the current proposition rather than a very long context. Moreover, it enables the parallel computation of all propositions within a document without the necessity of re-constructing and re-encoding. On top of the encoder, E-ASE applies sentence-level attention for context aggregation. This mechanism assigns higher weights to informative contextual sentences and lower weights to the less informative ones. To mitigate the influence of the discourse markers, we propose a data augmentation method where the training corpus is augmented by randomly replacing the discourse markers with the token [MASK]. Subsequently, we employ the masked language modeling loss to improve the model's capacity to gather contextual information. We conduct extensive experiments on five publicly available data sets, namely AMPERE [26], Essays [6], AbstRCT [25], ECHR [9], and CDCP [32], originating from different domains. The experimental results show that the proposed E-ASE achieves substantial improvements on all datasets consistently, compared to the context-aware ASE.

In summary, the main contributions of this paper can be concluded as follows:

1. We propose the E-ASE, which encodes each sentence separately without concatenating the contextual sentences and incorporating the contextual information with sentence-level attention. The novel architecture is very effective in removing contextual noise and enhancing training and inference efficiency.

2. We propose a simple and effective data augmentation method for relieving the effects of the discourse markers. Working together with the MLM loss, the data augmentation method enhances the model's ability to comprehend contextual information substantially.

3. We conducted extensive experiments on five datasets from different domains. The results show that the proposed method can improve the ASE performance consistently.

2. Related work

ASE Argument structure extraction has garnered increasing attention in recent years. This task involves identifying and extracting the relationship among the different argument propositions within documents. Conceptually, the argument structure extraction can be separated into two subtasks: premise detection, which aims to identify the targeted propositions(head), and relation classification, which involves classifying the relations of other propositions(tail) to the head. Early research drew inspiration from discourse parsing [4] [33] and some methods employed statistics and manually crafted features for classification [5] [24]. With the increasing adoption of deep learning in recent years, ASE models based on pre-trained language models have been proposed and have achieved remarkable performance [26] [30] [34] [35] [36]. In comparison to the traditional methods in this context, ASE models based on pre-trained language models consistently achieve superior performance. For example, [36] employs Bert [37] as the backbone network and introduces probing for the purpose of extracting additional semantic information from the language model. [26] propose a context-aware model based on Roberta [30], encoding the head propositions in a context window for ASE tasks. Our model builds upon the framework presented in [26], and further explores for the efficient utilization of contextual information and elimination of contextual noise.

Data augmentation Data augmentation is a technique aimed at enhancing the diversity of examples trained without the need for collecting the new data. It has yielded significant results across various deep learning and machine learning tasks [38]. During recent years, there has been a growing focus on data augmentation within the field of Natural Language Processing (NLP). This surge is particularly notable due to the widespread availability of large pre-trained language models, which has led to the exploration of numerous tasks and domains. Many of these are resource-constrained, characterized by limited training samples, thus underscoring the pivotal role played by data augmentation [39]. Building on the foundation laid by [38] and [39], more recent works take advantage of cutting-edge architectures like BERT [37] to achieve better results. For instance, [40] utilize BERT (Bidirectional Encoder Representations from Transformers) to address bottlenecks in the LSTM-based language model proposed by [41]. They introduce a conditional BERT model, which once well-trained, serves as a tool for augmenting sentences, in this approach, random words are masked within a labeled sentence, and conditional BERT is employed to predict new words consistent with the label of the given sentence. This method is compared with the identical methods studied in [40], using the same datasets, and consistently demonstrates superior performance in all evaluated scenarios. Additionally, [42] employs a pair of corruption and reconstruction functions to move randomly on a data manifold using BERT. The experiment consistently outperforms existing data augmentation methods and baseline models. In a similar vein, [43] introduce Augmented ABERT which employs a cross-encoder to label new input pairs, augmenting the training data. This augmentation strategy improves its performance in pairwise sentence scoring tasks.

MLM In general, the primary objectives of pretraining models can be categorized mainly into auto-regressive (AR) language modeling and auto-encoding (AE), with Masked language model (MLM) falling under the latter category. An illustrative example is provided by [37], who introduced a masked language model as the BERT architecture. This model adopts a random masking strategy to select the mask tokens, with the goal of pretraining deep bidirectional representations by reconstructing the original token from unmasked data. Roberta [29], on the other hand, represents a replication study of Bert's pretraining approach. It incorporates a comprehensive evaluation of the effects of hyperparameter tuning and training set size to determine the most influential factors. The experiments demonstrate that performance can be substantially promoted by removing the next sentence prediction objective, training the model on a larger corpus with a bigger batch size, and dynamically choosing tokens for masking in the input sequence. In recent years, several variants of Masked Language Models (MLMs) based on BERT, employing different masking strategies, have been proposed. Examples include XLNet [44] and MASS [45]. Furthermore, some research endeavors have introduced knowledge-enabled masking strategies, aiming to incorporate domain-specific knowledge into language models. For instance, [46-48] choose to mask named entities and [49-51] propose masking units such as spans during pre-training.

3. Method

In this section, we will introduce the proposed approach for enhancing the utilization of contextual information in ASE. We begin by providing an in-depth description of the task ASE, followed by detailed explanation of the model architecture of the proposed E-ASE. Subsequently, we introduce our data augmentation methods for mitigating the impact of discourse markers. Finally, the training loss will be presented.

3.1 Task of ASE

Argumentation Structure Extraction (ASE) is commonly formulated as a classification task. In formal terms, we define a dataset $D = \{d^i\}_{i=1}^{N}$ consisting of N documents, where each document comprises multiple propositions$\{s_k\}^i$. The objective of our task is to predict the presence of specific relationships, such as '***attack***', '***support***', or '***no - relation***', between s_k and s_j, where the target propositions s_j is ***head*** and s_k refers to ***tail***. Our end-to-end model considers all propositions pairs.

For each proposition $\boldsymbol{S_i}$ within documents, we fed them into Roberta separately in parallel. $\boldsymbol{H_i^s}$, the sentence representation of a given proposition $\boldsymbol{S_i}$. $\boldsymbol{C_i}$, the context-aware sentence representation for a given proposition $\boldsymbol{S_i}$.

3.2 E-ASE

Following the approach outlined in [26], our proposed E-ASE is constructed based on pre-trained Roberta model, which has achieved great success in a wide spectrum of natural language understanding tasks and has served as foundation model in numerous classification tasks. As Roberta has been widely investigated in many previous works, this paper will only focus on our novel modifications while omitting detailed discussions of Roberta's internal architecture. Interested readers are referred to [29] for comprehensive information regarding Roberta.

Given a document $S = (s_1, \dots, s_i, \dots, s_N)$, where s_i is the i-th proposition and N is the amount of sentence in the document. For each proposition, the [CLS] token is added at the head position. Diverging from the approach employed in [26], which concatenates context sentences together, we feed N propositions into Roberta separately in parallel. For a given proposition s_i, its sentence representation H_i^s is calculated as:

$$H_i = \text{Roberta}(s_i) \tag{1}$$

$$H_i^s = H_i[\text{index}_{\text{cls}}] \tag{2}$$

where $Roberta(x)$ means to get the hidden state by feeding x into Roberta, and $index_{cls}$ represents the index of [CLS] token. As the sentence representation h_i does not contain the context information, the sentence-level self-attention is applied to incorporate the contextual information. The context-aware sentence representation $C_i$for proposition i is calculated as:

$$C_i = \text{Attention}([H_1^s, \dots, H_i^s, \dots, H_N^s])[i] \tag{3}$$

In the self-attention, Q, K and V are identical, and they are all set as the sentence-level representation, i.e., $[h_1, \dots, h_i, \dots, h_N]$. For a comprehensive understanding of self-attention calculation, we refer the readers to [28]. Based on the calculated context-aware sentence representations, we calculate the relation for proposition i and j as:

$$p(y \mid (s_i, s_j)) = \text{Softmax}(\tanh(W_1 * C_i + W_2 * C_j)) \tag{4}$$

where C_i and C_j are the context-aware sentence representations for propositions i and j separately, W_1 and W_2 are two trained weight matrices. The whole architecture of the proposed E-ASE is illustrated as **Fig. 2**.

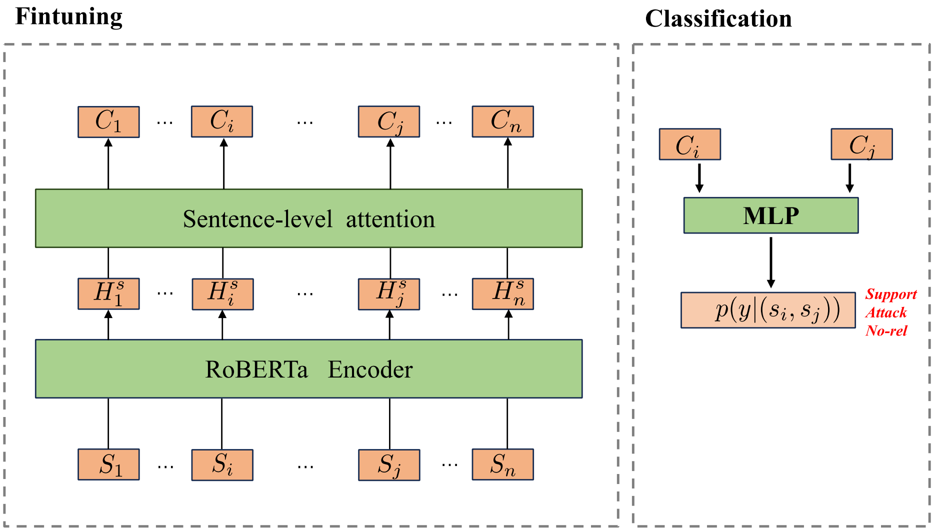

Fintuning
Classification
C_1 ··· C_i ··· C_j ··· C_n
Sentence-level attention
H_1^s ··· H_i^s ··· H_j^s ··· H_n^s
RoBERTa Encoder
S_1 ··· S_i ··· S_j ··· S_n
C_i
C_j
MLP
$p(y|(s_i, s_j))$
Support
Attack
No-rel

Fig. 2. The architectural design of our proposed model for argument structure extraction.

3.3 Data Augmentation

Since discourse markers exert considerable influence on the model's predictions, they can lead the model to appear `arrogant` and `lazy`, as it tends to make accurate predictions based solely on these discourse markers without comprehending the whole context. However, in general cases, the ability to grasp the context is crucial for the model to provide correct predictions. Therefore, to mitigate the adverse impact of discourse markers, we propose data augmentation by randomly noising the discourse markers.

The data augmentation method can be divided into two distinct steps. Firstly, we augment the training corpus by noising discourse markers. Specifically, for training examples with discourse markers, we augment them by either removing the discourse markers or replacing the discourse markers with any other word. With the data augmentation, the ratio of the discourse markers in the training corpus has declined by a large margin. Secondly, following the idea of masked language modeling, we employ self-supervised training to enhance the model's capacity to comprehend the context information. In particular, we randomly replace some words with the [MASK] token and then use the MLM loss to predict the original words. Contrary to the traditional works which assign equal importance to each word, we treat the discourse markers differently from the normal words, where the replacing probability for discourse markers is two times higher than normal words. This is mainly because we aim to encourage the model to gather more richer context information to predict the discourse makers.

3.4 Training

The training of the proposed E-ASE can be divided into two separate processes: pre-training and finetuning. In the pre-training process, the MLM loss is applied to pre-train the encoder, which enhances the encoder's ability to comprehend the context information; After pre-training, we finetune the whole model with the classification loss.

4. Experiment and Results

Experiments are conducted on publicly available corpora spanning various domains. This section begins by outlining our experimental setup, which includes the corpora and baselines employed in this paper. Subsequently, implementation details and the evaluation will be presented. Finally, we will report our main results.

4.1 Experiment Setups

This subsection describes the experimental setups in detail for easy reproduction. Specifically, we will present the corpora used in this paper and introduce our baseline works.

4.1.1 Corpus

In this paper, five openly accessible corpora, namely AMPERE, Essays, AbstRCT, ECHR, and CDCP, are utilized. These corpora originate from diverse domains and have been widely employed in prior research.

AMPERE The dataset AMPERE consists of 400 reviews from ICLR 2018, which are collected from OpenReview[1] [26]. Each example within this dataset represents a paper review, enriched with annotated segmented propositions and the corresponding types, such as **evaluation**, **request**, **fact**, and more. Annotators are required to further annotate the relations among these segmented propositions, namely **support** and **attack**. In accordance with prior studies, we use 300, 20, and 80 samples for training, validation, and testing respectively.

Essays The dataset Essays contains 402 essays, as assembled by [5], and obtained from[2] . The propositions within these essays are meticulously annotated at the sub-sentence level, categorized into types such as **premise, claim**, or **major claim.** Additionally, the relations among these propositions (support or attack) are annotated from a **premise** to a **claim** or to another **premise**. Following previous works, we allocate 228 samples for training, 40 for validation, and 80 for testing.

AbstRCT The dataset **Biomedical Paper Abstract** contains 700 paper abstracts from [25]. Note that the dataset contains fewer propositions compared to the previous mentioned ones, with only 70 attack links presents. Following previous works, we regard **attack** as **no-rel**, and only make classification on **support** and **no-rel**, given the significantly low occurrence of **attack** instances. Concretely, we use 350, 50, and 300 for training, validation, and testing.

ECHR The dataset ECHR, as described in [9], comprises 42 documents from the European Court of Human Rights. Within this dataset, the links are annotated from premises to conclusions.

CDCP The dataset CDCP, as detailed in [32], contains annotated comments related to Consumer Debt Collection Practices. This dataset includes information on supporting relations.

[1] https://openreview.net
[2] essaysforum.com

4.1.2 Baselines

We compare our model with the following baselines. SVM-linear and SVM-RBF [6] are the feature-based models relying on support vector machines to extract key features, aiding in identifying argumentative relations between propositions. In contrast, the other models are neural network based. SEQPAIR encodes the head and tail separately, while SEQCON encodes them within each other's context. CASE and ECASE are built on the RoBERTa model, with different strategies for leveraging context. CASE simply concatenates contextual propositions, whereas ECASE enhances this approach with a sentence-level attention mechanism to improve the utilization of contextual information.

SVM-linear The traditional lexical ASE model, SVM-linear, was implemented using features adapted from Table 10 of [6], except for features specific to the essays domain. [26] construct experiments with linear kernel and adjust regularization coefficients during validation.

SVM-RBF The model SVM-RBF introduces a radial-basis function(RBF) kernel in the experiments, and the regularization factors are tuned on validation.

SEQPAIR The model SEQPAIR uses BERT to encode the head and trail propositions in an individually sentences separately and concatenates their features to predict the argumentative relation.

SEQCON SEQCON, a model extended to a context-aware version from SEQPAIR, which encodes the head and tail in the context of each other and concatenates [CLS] representations of the head and tail for classification.

CASE CASE (context-aware ASE model) as introduced in [26], concatenates the head sentence with forward context propositions and backward propositions in a variable contextual window with the goal of predicting the argumentative relations.

ECASE ECASE (efficient context-aware ASE model) as presented in [30], concatenates the head sentence with forward context propositions and backward propositions in a variable context window and employ sentence-level attention for further extracting the relations between sentences.

4.2 Implementation and Evaluation

Consistent with previous works, we employ the pre-trained Roberta as our encoder, which is the official release version of Huggingface[3]. Our implementation is based on the code base released by the baseline work of CASE [26]. The proposed model and baseline models are trained on one V100 GPU with Adam optimizer. As for the hyper-parameters, the learning rate is set as 1e-5, and its scheduler is set as constant. For all corpus, we train our model with 10 epochs. For the data augmentation, the probability of mask replacement is 0.45, and the probability of replacing words with others is 0.2. For evaluation, we use the official toolkit for calculating the macro-F1 score. For each experiment, we report the averaged result with five different random seeds.

[3] https://huggingface.co/

4.3 Main Results

In this section, we present the primary results obtained from the five datasets, as shown in **Table 2**. In the end-to-end setting, E-ASE demonstrates superior performance, achieving Macro F1 scores of 74.15%, 70.07%, 74.29%, 69.21%, and 72.71% across the five datasets, outperforming the baselines. On average, E-ASE surpasses CASE-20 by 2.49% and ECASE by 1.29% in Macro F1 scores. Notably, E-ASE achieves significant improvements on the AMPERE, Abstract, and CDCP datasets, highlighting its ability to effectively leverage contextual information for identifying argumentative relations. On the ECHR dataset, E-ASE attains a competitive Macro F1 score of 69.21%, slightly lower than ECASE-20 (69.49%), likely due to the unique complexities of legal texts. Furthermore, models incorporating backward and forward contextual inputs, such as CASE and ECASE, consistently outperform SEQPAIR and SEQCON-10, emphasizing the importance of input structure and context-awareness. Overall, these results underscore that efficiently utilizing contextual information is critical for improving the recognition of argumentative structures in discourse.

5. Analysis

5.1 Analysis of Discourse Markers

In our previous experiments, we found that discourse markers have much effect on the final performance, which motivates us to relieve our model's dependence on the discourse markers with data augmentation and MLM loss. Therefore, it is a natural question of whether our model can relieve the effects of discourse markers. To answer this question, we construct two different test sets, which are named Test Normal and Test Discourse Marker respectively. Test Normal only includes examples without discourse markers and Discourse Marker includes examples with discourse markers. Both test sets have the same number of examples (i.e., the number of examples with discourse markers in the original test set), which are extracted from the original test set. We report the performance of our model and the baseline model, i.e., CASE, on the two test sets. The results are reported in **Table 3**. It's evident that when transitioning the Test Discourse Marker to Normal, the performance of CASE-20 experiences a significant decrease of 4.7 percent from the original level. In contrast, our model exhibits a notably smaller decrease of only 0.5 percent. This observation serves as a demonstration of our model's ability to mitigate the effects of discourse markers effectively.

Table 2. Results of our model on the tested five datasets. For results of the baseline models, we directly copy their results from the corresponding original papers. \`-\` represents that the original paper did not report the results. \`-10' and \`-20' refer to the context length used in the baseline models. \`Supp.', \`Atk.' and 'Macro' represent the F1 scores of \`Support', \`Attack', and final macro respectively.

	AMPERE			**Essays**			**AbstRCT**		**ECHR**		**CDCP**	
	Supp	Atk	Macro	Supp	Atk	Macro	Supp	Macro	Supp	Macro	Supp	Macro
SVM-Linear	-	-	24.82	-	-	28.69	-	-	-	21.18	-	29.01
SVM-RBF	-	-	26.38	-	-	31.68	-		-	21.36	-	30.34
SEQPAIR	17.34	7.40	26.42	31.42	24.32	30.37	17.32	32.34	23.11	33.23	14.16	28.44
Head Given												
SEQCON-20	43.04	49.67	63.34	47.71	30.14	57.40	60.60	69.20	41.81	65.48	38.34	63.10
CASE-20	64.45	68.56	77.41	72.14	40.14	69.13	63.35	70.94	35.18	69.35	45.63	69.56

End-to-End												
CASE-20	60.33	60.55	73.23	70.73	36.31	68.58	64.85	72.79	31.77	65.66	39.56	66.22
ECASE-20	60.82	57.56	72.82	73.74	37.54	69.51	63.58	73.59	**39.63**	**69.49**	**46.87**	70.16
E-ASE(Ours)	**63.04**	**59.12**	**74.15**	**73.90**	**38.21**	**70.07**	**64.91**	**74.29**	37.21	69.21	45.25	**72.71**

Table 3. Results for analyzing on discourse markers. '-20' refer to the context length used in the baseline models.

	Test Normal	Test Discourse Marker
CASE-20	68.4	71.8
E-ASE(Ours)	71.9	72.3

Table 4. Ablation study of *E-ASE* with context length 20. '$\boldsymbol{w/o}$' refers to without.

	Macro Average
E-ASE	**72.1**
w/o MLM	71.3
w/o Sentence-level Attention	70.1
w/o Data Augmentation	70.8

Table 5. Results of inference efficiency between *CASE* and *E-ASE*.

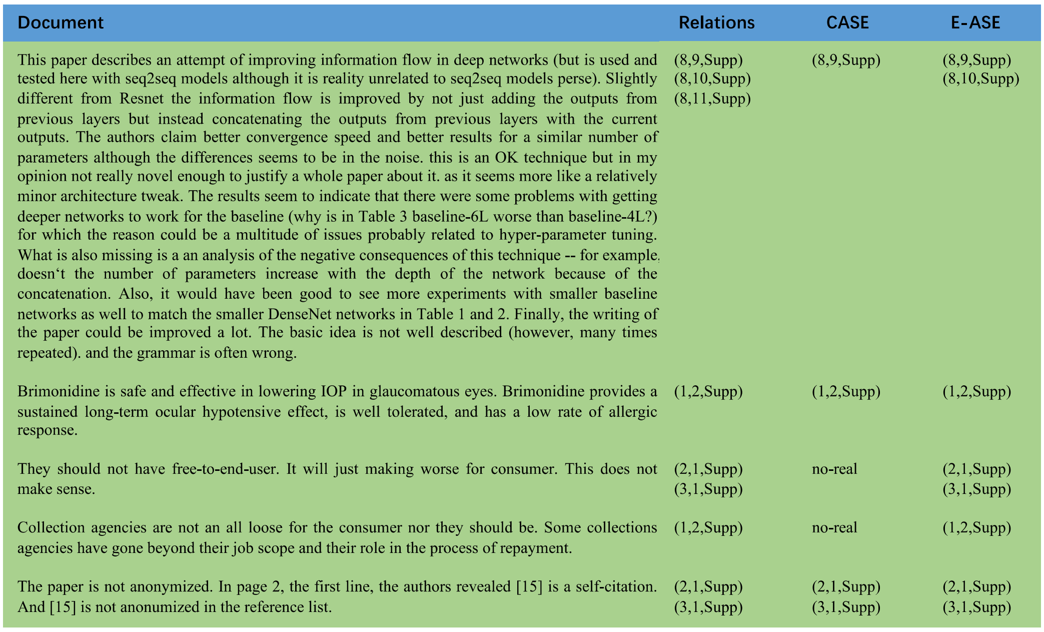
	Average Inference Time (s)
CASE-20	51
E-ASE(Ours)	37

5.2 Ablation Study

We conduct an ablation study to assess the effectiveness of each component in our model and methodology, namely data augmentation, sentence-level attention and MLM loss. The experimental results are presented in **Table 4**, where each row corresponds to a different configuration tested during the ablation study. As demonstrated in the ablation tests, removing the MLM loss component resulted in a 1.1 percent decrease in the macro average, removing the sentence-level attention led to a 2.7 percent decreasing and removing data augmentation led to a 1.8 percent decrease. Our ablation study underscores the critical importance of sentence-level attention in achieving the state-of-the-art results.

5.3 Inference Efficiency

We compare the inference efficiency between our model and the baseline model, CASE. To calculate the inference efficiency, we run the inference process of each model on the same machine and GPU device, i.e., one V100 GPU device. The average inference time of each model on the five test sets is reported in **Table 5**. The results indicate that the average inference time for E-ASE is 27.5 percent faster than that for CASE. These results also serve as a compelling demonstration of our efficient strategy, which involved feeding individual propositions into the model in parallel.



Document	Relations	CASE	E-ASE
This paper describes an attempt of improving information flow in deep networks (but is used and tested here with seq2seq models although it is reality unrelated to seq2seq models perse). Slightly different from Resnet the information flow is improved by not just adding the outputs from previous layers but instead concatenating the outputs from previous layers with the current outputs. The authors claim better convergence speed and better results for a similar number of parameters although the differences seems to be in the noise. this is an OK technique but in my opinion not really novel enough to justify a whole paper about it. as it seems more like a relatively minor architecture tweak. The results seem to indicate that there were some problems with getting deeper networks to work for the baseline (why is in Table 3 baseline-6L worse than baseline-4L?) for which the reason could be a multitude of issues probably related to hyper-parameter tuning. What is also missing is a an analysis of the negative consequences of this technique -- for example. doesn‘t the number of parameters increase with the depth of the network because of the concatenation. Also, it would have been good to see more experiments with smaller baseline networks as well to match the smaller DenseNet networks in Table 1 and 2. Finally, the writing of the paper could be improved a lot. The basic idea is not well described (however, many times repeated). and the grammar is often wrong.	(8,9,Supp) (8,10,Supp) (8,11,Supp)	(8,9,Supp)	(8,9,Supp) (8,10,Supp)
Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response.	(1,2,Supp)	(1,2,Supp)	(1,2,Supp)
They should not have free-to-end-user. It will just making worse for consumer. This does not make sense.	(2,1,Supp) (3,1,Supp)	no-real	(2,1,Supp) (3,1,Supp)
Collection agencies are not an all loose for the consumer nor they should be. Some collections agencies have gone beyond their job scope and their role in the process of repayment.	(1,2,Supp)	no-real	(1,2,Supp)
The paper is not anonymized. In page 2, the first line, the authors revealed [15] is a self-citation. And [15] is not anonumized in the reference list.	(2,1,Supp) (3,1,Supp)	(2,1,Supp) (3,1,Supp)	(2,1,Supp) (3,1,Supp)

Fig. 3. Case study of ASE. In the first column of Relations: the term in order is head, tail, and argumentative relations from tail to head. The case is tested in end-to-end setting.

6. Case Study and limitation

Fig. 3 shows some cases in the test set. It is evident that E-ASE outperforms CASE in scenarios involving lengthy context documents. This observation underscores the capacity of our model to leverage contextual information more effectively. Additionally, we conduct tests on various samples within ECHR and CDCP. Notably, our model consistently outperforms CASE in the identification of argumentative relations across these datasets.

Based on the discussion in section 5.1, our analysis of discourse markers has demonstrated that our model is effective in mitigating the influence of discourse markers. This success is primarily attributed to the utilization of data augmentation and MLM loss. Indeed, the inclusion of MLM loss during data augmentation can potentially lead to an increase in the model's training time. Therefore, in future work, we plan to explore more efficient strategies for optimizing training time during data augmentation.

7. Conclusion and Future Research

In this work, we have proposed an efficient context-aware Argument Structure Extraction (ASE) model by augmenting the training corpus and employing sentence-level attention, significantly improving the utilization of contextual information. Experimental results demonstrated that our approach consistently outperformed strong baseline models across five diverse datasets. Ablation studies further confirmed the importance of each module in our framework. This study underscores the potential of efficiently leveraging contextual information and highlights the critical role of sentence-level attention in reducing redundancy and enhancing performance.

However, the use of MLM loss during data augmentation increases training time, presenting a trade-off between performance and efficiency. In future work, we plan to explore more efficient strategies to optimize training time during data augmentation. Additionally, we aim to incorporate large language models to improve generalization and accuracy. Techniques such as in-context learning, and instruction fine-tuning will be investigated to further enhance the model's adaptability and performance across a broader range of ASE tasks.

Acknowledgement

This research is supported by Henan Academy of Science and Technology Information through Grant 2121022110504.

References

[1] E. Cabrio and S. Villata, "Five Years of Argument Mining: a Data-driven Analysis," in *Proc. of the 27th International Joint Conference on Artificial Intelligence*, pp.5427-5433, Stockholm, Sweden, 2018. Article(CrossRefLink)

[2] J. Lawrence and C. Reed, "Argument Mining: A Survey," *Computational Linguistics*, vol.45, no.4, pp.765-818, 2020. Article(CrossRefLink)

[3] M. Lippi and P. Torroni, "Argumentation Mining: State of the Art and Emerging Trends," *ACM Transactions on Internet Technology (TOIT)*, vol.16, no.2, pp.1-25, 2016. Article(CrossRefLink)

[4] A. Peldszus and M. Stede, "From Argument Diagrams to Argumentation Mining in Texts: A Survey," *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, vol.7, no.1, pp.1-31, 2013. Article(CrossRefLink)

[5] C. Stab and I. Gurevych, "Identifying Argumentative Discourse Structures in Persuasive Essays," in *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp.46-56, Doha, Qatar, 2014. Article(CrossRefLink)

[6] C. Stab and I. Gurevych, "Parsing Argumentation Structures in Persuasive Essays," *Computational Linguistics*, vol.43, no.3, pp.619-659, 2017. Article(CrossRefLink)

[7] E. M. Vecchi, N. Falk, I. Jundi, and G. Lapesa, "Towards Argument Mining for Social Good: A Survey," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol.1: Long Papers, pp.1338-1352, Online, 2021. Article(CrossRefLink)

[8] C. Lyu and W. Feng, "Analyzing Chinese text with clause relevance structure," *Neurocomputing*, vol.519, pp.82-93, 2023. Article(CrossRefLink)

[9] P. Poudyal, et al., "ECHR: Legal Corpus for Argument Mining," in *Proc. of the 7th Workshop on Argument Mining*, pp.67-75, Online, 2020. Article(CrossRefLink)

[10] H. Xu, et al., "Using Argument Mining for Legal Text Summarization," *IOS Press*, vol.334: Legal Knowledge and Information Systems, pp.184-193, 2020. Article(CrossRefLink)

[11] M. Elaraby and D. Litman, "ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining," in *Proc. of the 29th International Conference on Computational Linguistics*, pp.6187-6194, Gyeongju, Republic of Korea, 2022. Article(CrossRefLink)

[12] M. Elaraby, Y. Zhong, and D. Litman, "Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking," in *Proc. of the Findings of the Association for Computational Linguistics: ACL 2023*, pp.7601-7612, Toronto, Canada, 2023. Article(CrossRefLink)

[13] P. Accuosto and H. Saggion, "Transferring Knowledge from Discourse to Arguments: A Case Study with Scientific Abstracts," in *Proc. of the 6th Workshop on Argument Mining*, pp.41-51, Florence, Italy, 2019. Article(CrossRefLink)

[14] P. Accuosto and H. Saggion, “Mining arguments in scientific abstracts with discourse-level embeddings,” *Data & Knowledge Engineering*, vol.129, 2020. Article(CrossRefLink)

[15] K. Al Khatib, et al., “Argument Mining for Scholarly Document Processing: Taking Stock and Looking Ahead,” in *Proc. of the Second Workshop on Scholarly Document Processing*, pp.56-65, Online, 2021. Article(CrossRefLink)

[16]A. Binder, L. Hennig, and B. Verma, “Full-Text Argumentation Mining on Scientific Publications,” in *Proc. of the first Workshop on Information Extraction from Scientific Publications*, pp.54-66, Online, 2022. Article(CrossRefLink)

[17] J. Park and C. Cardie, “Identifying Appropriate Support for Propositions in Online User Comments,” in *Proc. of the First Workshop on Argumentation Mining*, pp.29-38, Baltimore, Maryland, 2014. Article(CrossRefLink)

[18] F. Boltužić and J. Šnajder, “Back up your Stance: Recognizing Arguments in Online Discussions,” in *Proc. of the First Workshop on Argumentation Mining*, pp.49-58, Baltimore, Maryland, 2014. Article(CrossRefLink)

[19] J. Park, A. Katiyar, and B. Yang, “Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments,” in *Proc. of the 2nd Workshop on Argumentation Mining*, pp.39-44, Denver, CO, 2015. Article(CrossRefLink)

[20] M. Hansen and D. Hershcovich, “A Dataset of Sustainable Diet Arguments on Twitter,” in *Proc. of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pp.40-58, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Article(CrossRefLink)

[21] B. Liu, V. Schlegel, R. Batista-Navarro, and S. Ananiadou, “Entity Coreference and Co-occurrence Aware Argument Mining from Biomedical Literature,” in *Proc. of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pp.54-60, Toronto, Canada, 2023. Article(CrossRefLink)

[22] J. Si, et al., “Biomedical Argument Mining Based on Sequential Multi-Task Learning,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.20, no.2, pp.864-874, 2023. Article(CrossRefLink)

[23] H. Nguyen and D. Litman, “Context-aware Argumentative Relation Mining,” in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.1127-1137, Berlin, Germany, 2016. Article(CrossRefLink)

[24] V. Niculae, J. Park, and C. Cardie, “Argument Mining with Structured SVMs and RNNs,” in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.985-995, Vancouver, Canada, 2017. Article(CrossRefLink)

[25] T. Mayer, E. Cabrio, and S. Villata, “Transformer-based Argument Mining for Healthcare Applications,” in *Proc. of the 24th European Conference on Artificial Intelligence - ECAI 2020*, pp.2108-2115, IOS Press, 2020. Article(CrossRefLink)

[26] X. Hua and L. Wang, “Efficient Argument Structure Extraction with Transfer Learning and Active Learning,” in *Proc. of the Findings of the Association for Computational Linguistics: ACL 2022*, pp.423-437, Dublin, Ireland, 2022. Article(CrossRefLink)

[27] J. Opitz and A. Frank, “Dissecting Content and Context in Argumentative Relation Analysis,” in *Proc. of the 6th Workshop on Argument Mining*, pp.25-34, Florence, Italy, 2019. Article(CrossRefLink)

[28] A. Vaswani, et al., “Attention Is All You Need,” in *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp.6000-6010, California, USA, 2017. Article(CrossRefLink)

[29] Y. Liu, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” in *Proc. of the ICLR 2020 Conference*, Addis Ababa, Ethiopia, 2020. Article(CrossRefLink)

[30] Y. Luo, et al., “Enhancing Argument Structure Extraction with Efficient Leverage of Contextual Information,” in *Proc. of the Findings of the Association for Computational Linguistics: EMNLP 2023*, pp.7563-7571, Singapore, 2023. Article(CrossRefLink)

[31] Z. Dai, et al., “Transformer-XL: Attentive Language Models beyond a Fixed-Length Context,” in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.2978-2988, Florence, Italy, 2019. Article(CrossRefLink)

[32] J. Park and C. Cardie, "A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments," in *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. Article(CrossRefLink)

[33] A. Peldszus and M. Stede, "Joint prediction in MST-style discourse parsing for argumentation mining," in *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp.938-948, Lisbon, Portugal, 2015. Article(CrossRefLink)

[34] K. Mitsuda, R. Higashinaka, and K. Saito, "Combining Argumentation Structure and Language Model for Generating Natural Argumentative Dialogue," in *Proc.of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, vol.2: Short Papers, pp.65-71, Online, 2022. Article(CrossRefLink)

[35] J. Bao, et al., "A Generative Model for End-to-End Argument Mining with Reconstructed Positional Encoding and Constrained Pointer Mechanism," in *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp.10437-10449, Abu Dhabi, United Arab Emirates, 2022. Article(CrossRefLink)

[36] Y. Sun, et al., "Probing Structural Knowledge from Pre-trained Language Model for Argumentation Relation Classification," in *Proc.of the Findings of the Association for Computational Linguistics: EMNLP 2022*, pp.3605-3615, Abu Dhabi, United Arab Emirates, 2022. Article(CrossRefLink)

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1 (Long and Short Papers), pp.4171-4186, Minneapolis, Minnesota, 2019. Article(CrossRefLink)

[38] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Applied Soft Computing*, vol.132, 2023. Article(CrossRefLink)

[39] S. Y. Feng, et al., "A Survey of Data Augmentation Approaches for NLP," in *Proc. of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp.968-988, Online, 2021. Article(CrossRefLink)

[40] X. Wu, et al., "Conditional BERT Contextual Augmentation," in *Proc. of the 19th International Conference on Computational Science – ICCS 2019*, LNTCS, vol.11539, pp.84-95, Portugal, Springer, Cham, 2019. Article(CrossRefLink)

[41] S. Kobayashi, "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," in *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.2 (Short Papers), pp.452-457, New Orleans, Louisiana, 2018. Article(CrossRefLink)

[42] N. Ng, K. Cho, and M. Ghassemi, "SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1268-1283, Online, 2020. Article(CrossRefLink)

[43] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks," in *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.296-310, Online, 2021. Article(CrossRefLink)

[44] Z. Yang, et al., "XLNet: generalized autoregressive pretraining for language understanding," in *Proc. of the 33rd International Conference on Neural Information Processing Systems*, pp.5753-5763, Vancouver, Canada, 2019. Article(CrossRefLink)

[45] K. Song, et al., "MASS: Masked Sequence to Sequence Pre-training for Language Generation," in *Proc. of the 36th International Conference on Machine Learning*, pp.5926-5936, Long Beach, California, USA, 2019. Article(CrossRefLink)

[46] Y. Sun, et al., "ERNIE: Enhanced Representation through Knowledge Integration," *arXiv preprint arXiv:1904.09223*, 2019. Article(CrossRefLink)

[47] Z. Zhang, et al., “ERNIE: Enhanced Language Representation with Informative Entities,” in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.1441-1451, Florence, Italy, 2019. Article(CrossRefLink)

[48] C. Lin, et al., “EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain,” in *Proc. of the 20th Workshop on Biomedical Language Processing*, pp.191-201, Online, 2021. Article(CrossRefLink)

[49] M. Joshi, et al., “SpanBERT: Improving Pre-training by Representing and Predicting Spans,” *Transactions of the association for computational linguistics*, vol.8, pp.64-77, 2020. Article(CrossRefLink)

[50] Y. Li and H. Zhao, “Pre-training Universal Language Representation,” in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol.1: Long Papers, pp.5122-5133, Online, 2021. Article(CrossRefLink)

[51] Y. Levine, et al., “PMI-Masking: Principled masking of correlated spans,” *arXiv preprint arXiv:2010.01825*, 2020. Article(CrossRefLink)

MENGLONG XU received his bachelor's degree in electrical engineering and automation from Nanchang Institute of Technology in 2017 and is currently studying for his master's degree in the the School of Physics and Electronic Information Engineering at Henan Polytechnic University. His current research interests include natural language processing and machine learning.

YANLIANG ZHANG received the B.Sc. degree from Henan University, in 2001, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, in 2011. He is currently a Professor with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, China. His research interests include machine vision and affective computing.

YAPU YANG received the B.Sc. degree from Hunan University of Technology, in 2002, and the M.Sc. degree from the School of Electrical Engineering, ChongQing University, in 2012. He is currently working in Xu Ji ELECTRIC CO., LTD., Xu Chang, China. His research interests include machine learning.