symmetry

MDPI

Article

MRE: A Military Relation Extraction Model Based on BiGRU and Multi-Head Attention

Yiwei Lu [1], Ruopeng Yang [1,*], Xuping Jiang [1], Dan Zhou [1], Changsheng Yin [1] and Zizhuo Li [2]

[1] College of Information and Communication, National University of Defense Technology, Wuhan 430019, China; yiweilu89@nudt.edu.cn (Y.L.); jiangxuping17@nudt.edu.cn (X.J.); zhoudan17@nudt.edu.cn (D.Z.); yinchangshen@nudt.edu.cn (C.Y.)

[2] Electronic Information School, Wuhan University, Wuhan 430000, China; zizhuo_li@whu.edu.cn

* Correspondence: yangruopeng17@nudt.edu.cn; Tel.: +86-186-2771-1887

Abstract: A great deal of operational information exists in the form of text. Therefore, extracting operational information from unstructured military text is of great significance for assisting command decision making and operations. Military relation extraction is one of the main tasks of military information extraction, which aims at identifying the relation between two named entities from unstructured military texts. However, the traditional methods of extracting military relations cannot easily resolve problems such as inadequate manual features and inaccurate Chinese word segmentation in military fields, failing to make full use of symmetrical entity relations in military texts. With our approach, based on the pre-trained language model, we present a Chinese military relation extraction method, which combines the bi-directional gate recurrent unit (BiGRU) and multi-head attention mechanism (MHATT). More specifically, the conceptual foundation of our method lies in constructing an embedding layer and combining word embedding with position embedding, based on the pre-trained language model; the output vectors of BiGRU neural networks are symmetrically spliced to learn the semantic features of context, and they fuse the multi-head attention mechanism to improve the ability of expressing semantic information. On the military text corpus that we have built, we conduct extensive experiments. We demonstrate the superiority of our method over the traditional non-attention model, attention model, and improved attention model, and the comprehensive evaluation value F1-score of the model is improved by about 4%.

Keywords: military relation extraction; bi-directional encoder representations from transformers (BERT); BiGRU; multi-head attention

check for updates

Citation: Lu, Y.; Yang, R.; Jiang, X.; Zhou, D.; Yin, C.; Li, Z. MRE: A Military Relation Extraction Model Based on BiGRU and Multi-Head Attention. *Symmetry* **2021**, *13*, 1742. https://doi.org/10.3390/sym13091742

Academic Editor: José Carlos R. Alcantud

Received: 11 August 2021
Accepted: 7 September 2021
Published: 19 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the progress in science and technology, and the evolution of war patterns, operational information and intelligence data have exponentially increased. This massive amount of information forms a "war fog", which directly interferes with the commander's command decision making. Military data abound in the form of unstructured text. Therefore, understanding how to extract valuable operational information from unstructured text, and how to build a military knowledge base for command and decision support, has become a topic of intense research in the field of military information extraction. As one of the basic tasks in military information extraction technology, military relation extraction is a key approach to creating military knowledge bases and a military knowledge graph [1]. This approach also facilitates improvements in the quality of operational information services, assisting commanders in decision making.

The main purpose of military relation extraction is to identify the semantic relations between symmetric entity pairs from unstructured military texts, and to express them according to the structured form of a triplet (entity e1, entity e2, relation r) [2]. For example, a sentence marked with military named entities are as follows:

<e1>第 1步兵师 </e1>命令<e2>第 16步兵团 </e2>在 <e3>奥马哈海滩 </e3>登陆 .

(<e1> 1st Infantry Division </e1> orders < e2> 16th Infantry Regiment </e2> to land at <e3> Omaha Beach </e3>).

As shown in Figure 1, symmetric entity pairs (第 1步兵师 , 第 16步兵团) and (第 16步兵团 , 奥马哈海滩) in this sentence have military relations "Command", "Deploy", obtaining triples (第 1步兵师 , 第 16步兵团 , Command) and (第 16步兵团 , 奥马哈海滩 , Deploy).

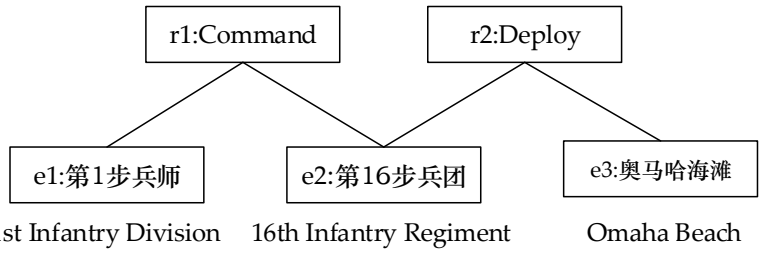


Figure 1. The symmetric relation between military entities.

Relation extraction will automatically identify the relations between the symmetric entity pairs "1st Infantry Division" and "16th Infantry Regiment" as "command relation", and generate relation triple (1st Infantry Division, 16th Infantry Regiment, Command). Consequently, the extraction of symmetric entity relations has a wide spectrum of applications in the fields of combat data processing, military knowledge map construction, commander's critical information requirements (CCIRs), and question–answer on military knowledge [3].

At present, the most prevalent method of entity relation extraction in domain-specific fields is supervised learning. In particular, the effect of relation extraction is significantly improved with in-depth application of the deep neural network model. Nevertheless, this method requires considerable time and effort to construct a large number of artificial features and to label numerous document requirements, which directly affect relation extraction. Compared with other fields, the artificial construction features of military text are not obvious, Chinese word segmentation is not so accurate, and, sometimes, and between input and output, the correlation is poor. Common relation extraction takes a single sentence as its processing unit, without taking the semantic association between sentences into account.

To address the above issues, we design a feature representation method that combines word embedding with position embedding, based on the pre-trained model, using BiGRU networks and the multi-head attention mechanism to capture the semantic features of military text, and achieve the effective extraction of military relations. We offer the following three contributions:

(1) We encode the input military text using the pre-trained language model. The word features and position features of military text are combined to generate the vector feature of military text, and then the semantic features of military text can be expressed more effectively.
(2) We apply a multi-head attention mechanism combined with BERT into military relation extraction. As a variant of self-attention, the core idea of this approach is to calculate self-attention from multiple dimensional spaces, so that, based on effective expression of semantic features in military texts from BERT, the model can learn more semantic features in military texts from different subspaces, and thus capture more contextual information.
(3) We establish the types and tagging methods of military relations, and construct a certain scale corpus of military relations via analyzing the semantic features of military texts.

2. Related Works

Relation extraction is one of the important and critical tasks in natural language processing (NLP). At present, relation extraction methods mainly include the following three types: feature-based methods [4,5], kernel-based methods [6,7], and deep learning-based methods [8–11].

Feature-based methods are based on feature vectors. First, different feature sets are constructed manually; next, they are transformed into feature vectors, then input into appropriate classifiers to realize relation extraction. For example, Kambhatla et al. [4] combined word features, syntactic features, and semantic features, and designed a classifier based on a maximum entropy model. The F1-score reached 52.8% with ACE RDC2003 evaluation datasets. Che et al. [12] used the entity type, order of occurrence of the two entities, and the number of words around the entity as the features, and de-signed a classifier based on support vector machines (SVM). In the evaluation of the dataset of ACE RDC2004, the F1-score reached 73.27%. However, since the extraction efficiency relies heavily on artificially constructed features, it is difficult to improve the performance of this method.

The kernel-based method was first introduced by Zelenko et al. [13]. This method does not require construction of feature vectors. Instead, it mainly calculates the similarity of two nonlinear structures by analyzing the structural information of corpus and by adopting appropriate kernel functions, so as to realize the relation extraction. Extensive experiments have indicated that this method can achieve useful results. Plank et al. [14] proposed the introduction of structural information and semantic information into kernel functions simultaneously, to cope with the problem of relation extraction. However, since all data must be fully summarized by the kernel function, the validity of kernel functions is the key to the extraction efficiency of kernel-based methods.

In recent years, deep learning methods have been enthusiastically applied to various fields of NLP, due to their superiority of learning and expression of deep features, making good progress with entity relation extraction task. For instance, Liu et al. [15] first applied the convolutional neural network theory to tackle relation extraction problems. Specifically, they built an end-to-end network based on convolutional neural networks, and coded sentences according to synonymous vectors and lexical features. The optimality of the model of the ACE 2005 dataset was 9% higher than that of the most advanced kernel-based model at that time. Zeng et al. [16] proposed a novel, piecewise convolutional neural networks (PCNN) model based on multi-instance learning, which can not only automatically extract the internal features of sentences, but also effectively reduce the impact of noise. Although convolutional neural networks (CNN) can effectively improve the efficiency of relation extraction, CNN is not suitable for learning long-distance-dependent information [10]. Although recurrent neural networks (RNN) can effectively learn long-distance-dependent information, there is a gradient disappearance problem in the trained process, which limits the processing of context [17]. To address the above problems, Hochreiter and Schmiduber [18] designed a long- and short-term memory network (LSTM), which effectively alleviates the gradient disappearance problem of RNN by introducing a gating unit. Zhou et al. [19] applied a BiLSTM neural network to learn sentence features, and a self-attention mechanism to capture more semantic information in sentences. Experiments with the SemEval2010 dataset show that attention-based mechanisms can effectively boost the efficiency of relation extraction. As a simplification of the LSTM model, the GRU neural networks model was first used in machine translation tasks. This model has the advantages of simple calculation and high execution efficiency [20], and has been used recently in relation extraction tasks. Luo et al. [21] achieved good results by combining BiGRU neural networks and attention mechanisms, to build a geographic data analysis model, from which entity relations are extracted. Zhang et al. [22] proposed a model combining a dual-layer attention mechanism and BiGRU neural network to realize the extraction of character relations. The experimental results showed a significant improvement in extraction efficiency. Zhou et al. [23] proposed a neural network-based attention

model (NAM) for the extraction of chemical–disease relation (CDR). Li et al. [24] proposed a relation extraction model based on dual attention-guided graph convolutional network, and they used the dual attention mechanism to capture rich context dependencies and achieved better performances. Thakur et al. [25] proposed a model of entity and relation extraction in IoT. Liu et al. [26] proposed a relation extraction method based on CRF and the syntactic analysis tree, and created a military knowledge graph.

Compared with the relation extraction in open domains and other specific domains, military texts contain numerous abbreviations, combinations, nesting, and other complicated grammatical forms, and, as shown in Figure 2, military texts are usually very concise, but contain many kinds of relations in a long sentence. The semantic relationship of Chinese noun phrases is more complex than that of English [27]. Therefore, it is difficult to obtain effective entity and relation features. While, the existing word segmentation tools are mainly applicable to the general domain, and it is difficult to achieve good results in the military domain. Thus far, there is no public corpus in the military domain, which makes the extraction of military relation more difficult.

第16步兵团（欠第2骑兵营）当面之敌为德军352师，接第1步兵师命令，于6月6日6:30之前在奥马哈海滩登陆，并立即向燧发枪营发起攻击。

The enemy of the 16th Inf Regiment(Remove 2th Cavalry Battalion) was the German 352nd Inf Div. and 1st Inf Div. ordered that it should land on Omaha Beach before 6:30 on June 6, and immediately attacked the Fusiliers Battalion.

（第16步兵团，Remove，第2骑兵营）（16th Inf Regiment，Remove，2th Cavalry Battalion）
（第16步兵团，Enemy，德军352师）（16th Inf Regiment，Enemy，German 352nd Inf Div.）
（第16步兵团，Command，第1步兵师）（16th Inf Regiment，Command，1st Inf Div.）
（第16步兵团，Deploy，奥马哈海滩）（16th Inf Regiment，Deploy，Omaha Beach）
（第16步兵团，Target，燧发枪营）（16th Inf Regiment，Target，Fusiliers Battalion）

Figure 2. Example of relations in a long sentence in military texts.

Military texts are usually long and there are many long-dependent sentences. As to BiGRU structure can gain rich contextual features in military text, relation extraction is more effective.

3. Military Relation Extraction Model

Based on the pre-trained language model, this paper presents an explicitly military relation extraction model that combines BiGRU and multi-head attention mechanisms. The structure of this model is shown in Figure 3. First, all the characters in the input sentence are vectorized by using the pre-trained language model [28], and the relative position vectors of each character are calculated. The word embedding and the position embedding are then joined to generate the sentence eigenvectors, which are input into the bi-directional gate recurrent unit (BiGRU), which can capture the high-dimensional semantic features of sentences; more context information of sentences can be captured by establishing a multi-head attention mechanism; and finally, the conditional probability of each relation type can be calculated through a SoftMax classifier, which outputs the classification results.

3.1. Embedding Layer

Before being inputted into the neural network model, the sentences in the natural language text form must first be represented by vectors. To achieve the embedding of sentences, we combine word embedding with position embedding.

3.1.1. Word Embedding

There are many ways to achieve word embedding. The main models of word embedding are Word2Vec [29] and GloVe [30]. These models are usually static and fixed, and cannot change with context, so that they cannot effectively express the word features

in the context of military texts. Pre-trained language models can express rich sentence syntax and grammar information, and can model the ambiguity of words. These models are widely used in natural language processing, such as information extraction and text classification [31]. Bi-directional encoder representations from transformers (BERT) are one such pre-trained language model proposed by Google in 2018 [28].

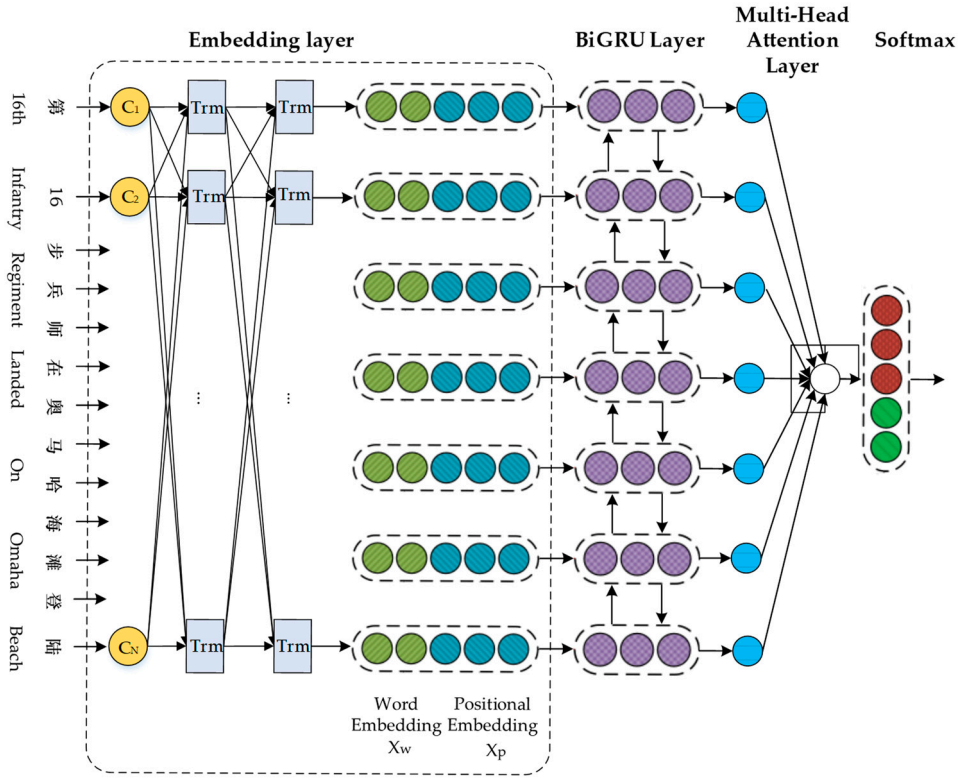


Figure 3. Framework of military relation extraction model based on BERT-BiGRU-MHATT.

We use BERT to achieve the word embedding, as shown in Figure 4. Given a sentence in Chinese military text containing N characters, it can be represented as $S = (s_1, s_2, \ldots, s_N)$. Each character contains the following three types of features: character features, sentence features, and position features. We represent the character features of X as $\left(e_1^t, e_2^t, \ldots, e_N^t\right)$, the sentence features as $\left(e_1^s, e_2^s, \ldots, e_N^s\right)$, and the position features as $\left(e_1^p, e_2^p, \ldots, e_N^p\right)$. The input of the word vector representation layer of BERT is the sum of character features, sentence features, and position features, as follows: $C_i = e_i^t + e_i^s + e_i^p$, $C = (C_1, C_2, \ldots, C_N)$.

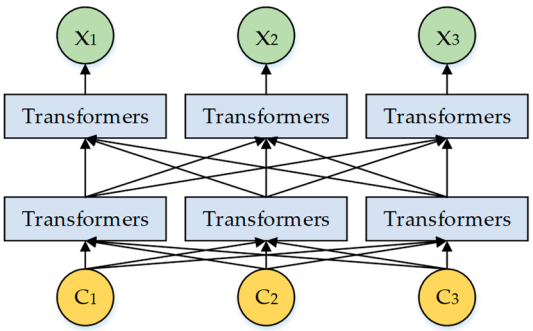


Figure 4. Language model based on BERT.

After inputting C into the multi-layer transformers, we can obtain the following final word embedding: $X_w = (x_1, x_2, \ldots, x_N) = \mathrm{BERT}(C_1, C_2, \ldots, C_N)$, and the dimension of x_i is d_w.

3.1.2. Position Embedding

Although word embedding can effectively capture the word information in a sentence, it is difficult to obtain the structural information of the sentence. The distance relation between the word and the entity directly affects the determination of the entity relation. Therefore, the position embedding X_p is used in this paper to denote the relative distance between the current word and two entities. As shown in Figure 5, the relative positions of the current word "在 "(on) with the military named entity "第六步兵团 "(16th Infantry Regiment) and "奥马哈海滩 " (Omaha Beach) are 6 and –1, and both relative positions correspond to the dimension d_p position embedding X_{p1}, X_{p2}.

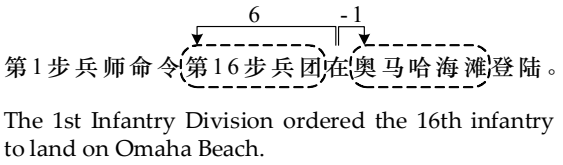


Figure 5. Relative position of the current word and the military named entities.

Finally, word embedding is cascaded with the position embedding to generate the following complete feature representation vector: $X_i = X_w^i + X_p^i$; the dimension of X_i is $d = d_w + 2d_p$.

3.2. BiGRU Layer

GRU neural networks are essentially a variant of recurrent neural networks (RNN). In order to solve the problem that traditional RNN rewrites its own memory in unit steps, and has the problem of gradient dispersion, based on RNN, Hochreiter et al. [18] proposed a neural network named long short-term memory (LSTM). The LSTM neural networks mainly include input gates, forget gates, and output gates. As shown in Figure 6, GRU is a simplified LSTM neural network, which can be calculated more easily, while maintaining the effect of LSTM neural networks.

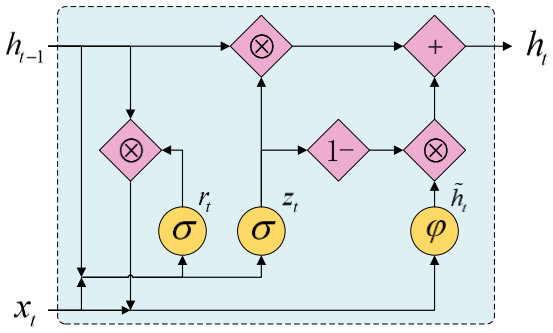


Figure 6. Structure of GRU.

As show in Figure 6, x_t is the input vector, h_{t-1} is hidden state at time t−1, and h_t is the output vector of current GRU. At the time t, x_t and h_{t-1} are the input into the GRU networks, and we can obtain the output h_t. Further, h_t is expressed as Formulas (1)–(4), as follows:

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

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (2)$$

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

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

where σ is the symbol of Sigmoid function, which can help the GRU neural networks to retain or forget information, and ⊗ is an elementwise production, z_t is the update gate, and

r_t is the reset gate. Further, $\tilde{h}_t$ is the candidate implied state at the time t. W_z, W_r, W_h are the input weights for the current time, and U_z, U_r, U_h are the weights for the cyclic input. Additionally, b_z, b_r, b_h are the corresponding offset vectors for W_z, W_r, W_h, U_z, U_r, U_h.

In order to make full use of the contextual information in military texts, we chose the BiGRU structure, which includes a forward hidden layer and a backward hidden layer. As shown in Figure 7, each input sequence is input into forward GRU networks and backward GRU networks, and two symmetrical hidden layer state vectors are obtained. These two state vectors are symmetrically merged, and then we can obtain the final coded representation of the input sentence, as follows:

$$H_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \tag{5}$$

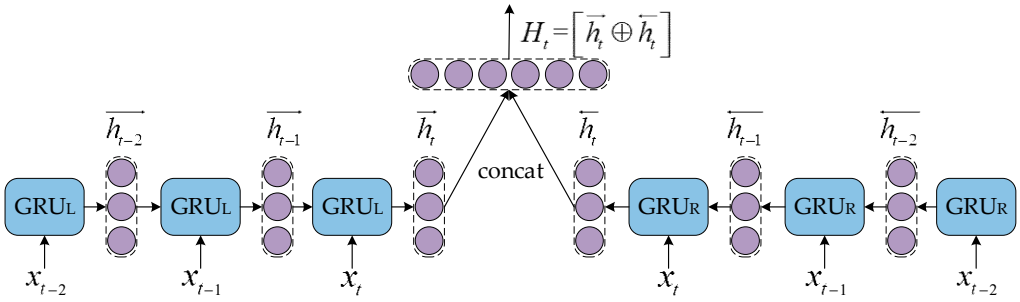


Figure 7. Symmetric CONCAT of output from BiGRU.

3.3. Multi-Head Attention Layer

The multi-head attention mechanism can be used to represent the correlation between the input and the output during text processing tasks. In military text, there are usually a large number of technical terms and abbreviations, and the referential relationships are complex and the sentence structures are diverse. Using the multi-head attention mechanism, the entity and relation information can be effectively analyzed and extracted.

After the sentence $X = (x_1, x_2, \ldots, x_T)$ is computed from the BiGRU layer, we can obtain the vector $H = (H_1, H_2, \ldots, H_T)$, where T is the length of X, and γ is the weighted average of H. We can construct the general attention model as follows:

$$\alpha = \mathrm{SoftMax}\left(w^T \tanh(H)\right) \tag{6}$$

$$\gamma = H\alpha^T \tag{7}$$

where $H \in R^{d^w \times T}$, in which d^w is the dimension of the embedding layer, w is a parameter vector in training, and w^T is its transpose. After the single-head attention calculation, we can obtain the output eigenvalue as follows:

$$h^* = \tan h(\gamma) \tag{8}$$

As shown in Figure 8, the multi-head attention mechanism [32] can help our model to derive more features from different representation subspaces and capture more contextual information from military texts. In a single self-attention calculation, after H is transformed linearly [28], we can obtain $W_i^h H$, and $W_i^h \in R^{d^h/k \times d^h}, i \in \{1, 2, \ldots, k\}$. Additionally, by using the mechanism of multiplicative attention, we can achieve a highly optimized matrix multiplication. The Formulas (6)–(8) are used for k times of calculation. After splicing and linearly mapping the results of the calculation, we can obtain the final result, as follows:

$$h_s = w_s \otimes \mathrm{concat}(h_1^*, h_2^*, \ldots, h_k^*) \tag{9}$$

where the dimension of w_s is $k \times d^h$, and $\otimes$ means point-by-element multiplication.

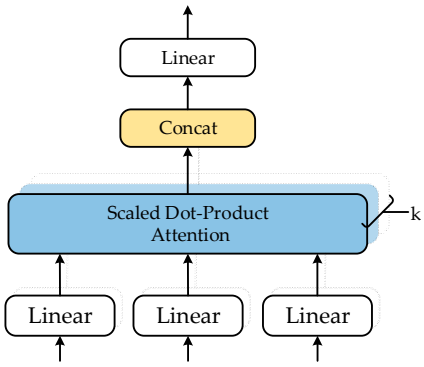


Figure 8. The structure of the multi-head attention mechanism.

3.4. Ouput Layer

Military relation extraction is also essentially a multi-classification problem. There are a few common classifiers, such as k-NN, random forest, SoftMax, etc. [33]. SoftMax is specific, with a simpler calculation and more-remarkable results than others, so, in the output layer, we chose SoftMax to calculate the conditional probability of each relation type, and it chose the relation category corresponding to the maximum conditional probability as the output of the prediction result. As the relation type of the entity pair in sentence S, y' is predefined. SoftMax calculates with h_s as the input. The formulas of the predicted relation type y are as follow:

$$p(y'|S) = \text{SoftMax}(W_o h_s + b_o) \tag{10}$$

$$y = \underset{y'}{\text{argmax}} P(y'|S) \tag{11}$$

where $W_o \in R^{c \times kd_w}$, in which c is the number of relation types in the dataset. We chose the cross-entropy loss function with L2 penalty as the objective function, as follows:

$$J(\theta) = \sum_{i=1}^{m} \log(y'_i) + \lambda ||\theta||_2^2 \tag{12}$$

where m is the number of relations in sentence S, y'_i is the probability of each relation type obtained through SoftMax, and λ is the L2 regularization factor.

4. Experiments and Results

4.1. Dataset

At present, the research on entity relation extraction tasks in the open domain, and in medical and judicial fields, is mature, and there are many open datasets, such as ACE2003-2004 [34], SemEval2010 Task8 [35], and FewRel [36]. Research on relation extraction in Chinese is also developing gradually, but relation extraction in military fields is basically in its infancy, with no public datasets. Military scenarios, an important form of military text, contain a large amount of military information, such as subordinate relation, location relation, attacking relation, etc. Therefore, we chose military scenarios as the research object; through the analysis of a large number of military scenarios, we have organized experts in the military field to conduct the research and discussion, and define the relations in the military field in combination with the specifications of military documents. Table 1 lists six coarse-grained categories and 12 fine-grained categories of military relationships, and Table 2 shows an example of a military relation labeling corpus.

Table 1. Types and description of military relations.

Coarse-Grained	Fine-Grained	Type Description
Command Relation	Command (Com)	Superior commands subordinate.
	Affiliation (Aff)	A subordinate is subordinate to a superior.
	Equivalent (Eq)	No superior-subordinate relation.
	Coreference (CoF)	Two entities represent the same organization.
	Remove (Rem)	Remove superior-subordinate command relation.
Position Relation	Enemy (En)	The two organizations are hostile.
	Alliance (Alli)	The two organizations are alliance.
Location Relation	Deploy (Dep)	Entity is in a specific location.
	Route	Entity is in a position.
Equipment Relation	Own	Organization configures some equipment.
	Link	Organization links some equipment. Equipment links some equipment.
Target Relation	Target	Organization attacks some organization. Organization attacks some location.

Table 2. Military named entity relation.

Entity 1	Entity 2	Relation	Military Sentence
第 1步兵师 1st Inf Div.	第 16步兵团 16th Inf Regiment	Command	第 1步兵师命令第 16步兵团进攻 1st Inf Div. ordered 16th Inf Regiment to attack
第 1步兵师 1st Inf Div.	第 4骑兵团 4th Cavalry Regiment	Remove	第 1步兵师 (欠第 4骑兵团) 1st Inf Div. (Remove 4th Cavalry Regiment)
第 16步兵团 16th Inf Regiment	奥马哈海滩 Omaha Beach	Deploy	第 16步兵团登陆奥马哈海滩 16th Inf Regiment landed on Omaha Beach
第 16步兵团 16th Inf Regiment	德军 352师 German 352nd Inf Div.	Enemy	第 16步兵团当面之敌为德军 352师 The enemy of 16th Inf Regiment was the German 352nd Inf Div.
第 16步兵团 16th Inf Regiment	燧发枪营 Fusiliers Battalion	Target	第 16步兵团向燧发枪营发起攻击 16th Inf Regiment attacked Fusiliers Battalion

4.2. Evaluation Criterion

We selected military scenario texts randomly as analysis objects, annotated 50 texts (about 320,000 words) manually, and have obtained 6105 military text corpus' as the datasets of the experiments. The distribution of relations in the military text corpus is shown in Table 3.

Table 3. Statistics of the military text corpus.

Coarse-Grained	Fine-Grained	Number of Samples
Command Relation	Command (Com)	1354
	Affiliation (Aff)	1110
	Equivalent (Eq)	320
	Coreference (CoF)	120
	Remove (Rem)	106
Position Relation	Enemy (En)	876
	Alliance (Alli)	332
Location Relation	Deploy (Dep)	102
	Route	1380
Equipment Relation	Own	149
	Link	158
Target Relation	Target	98

In the experiments, TP represents the number of correctly classified relations, FP represents the number of incorrectly classified relations, and FN represents the number of classified relations that should be correctly classified, but have not been classified. Precision, recall, and F1-score are chosen as the evaluation criterions, and can be calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}} \tag{15}$$

4.3. Parameters Setting

In the experiment, we used our previously developed cross-validation methods to optimize the parameters of our model, and the data are verified in the literature [9]. The specific parameters are shown in Tables 4–6.

Table 4. Main parameters setting in BERT.

Hyperparameters	Property Value
Hidden Layers	12
Hidden Size	768
Hidden Dropout Prob	0.1
Attention Heads	12
Position Embeddings	512

Table 5. Main parameters setting in BiGRU.

Hyperparameters	Property Value
Word Embedding	200
Position Embedding	50
Hidden Layer Node Number	240
Batch Size	64
Learning Rate	0.002
Epoch	30

Table 6. Main parameters setting in MHATT.

Hyperparameters	Property Value
Word Embedding	200
Learning Rate	0.002

It should be noted that if the head number k of self-attention in the multi-head layer is too large or too small, we should first determine the value of the parameter k before starting the comparative experiment. Referring to the experiment of Vaswani et al. [32], we take k = {1,2,4,6,10,15,30} as the candidate value (k should be divisible by d^h); the results are shown in Table 7. With the increase in k, when k = 6, the comprehensive evaluation index of the model reaches the highest value. Therefore, the value of parameter k in this experiment is six.

Table 7. Results of different k values.

k	F1-Score (%)
1	88.2
2	88.7
4	89.5
6	**90.2**
10	89.3
15	89.2
30	88.1

4.4. Results and Analysis

The military text corpus is selected as the training corpus and test corpus, and the experiment is conducted according to the set experimental parameters. To verify the validity of our model, we have designed a number of comparative experiments.

4.4.1. Comparison of Result on Different Embedding Methods

In this section, we employ the commonly used tool Word2Vec (dimension of Word2Vec is set to 100) [29] for the comparative experiments. We compared three feature representation methods to verify the effectiveness of the embedding-based pre-trained model (BERT), combined with word embedding and position embedding. These methods include the following:

Feature representation of Word2Vec + word;
Feature representation of Word2Vec + word + position;
Feature vector representation of BERT + word;
Feature vector representation of BERT + word + position.

As shown in Table 8, embedding methods based on the pre-trained model are superior to those based on Word2Vec. The F1-score of the BERT + word is 7.6% higher than that of the Word2Vec + word method, and that of the BERT + word + position method is 7.9% higher than that of the Word2Vec + word + position method. At the same time, the input embedding method based on word embedding and position embedding is better than that based on word (among which the F1-score of the Word2Vec + word + position method is 4.1% higher than that of the Word2Vec + word method, and the BERT + word + position method is 4.4% higher than that of the BERT + word method). This analysis indicates that the feature vector combined with word embedding and position embedding can better express the semantic features in military text.

Table 8. Comparison of different embedding methods.

Model	Precision (%)	Recall (%)	F1-Score (%)
Word2Vec + word	77.9	78.6	78.2
Word2Vec + word + position	81.9	82.7	82.3
BERT + word	85.6	86.1	85.8
BERT + word + position	90.8	89.6	**90.2**

4.4.2. Comparison of Result on Different Feature Extraction Models

To verify the advantages of the BiGRU-MHATT model, several classical relation extraction models are set up in this paper, as follows:

Traditional non-attention models: BiLSTM, BiGRU;
Based on the traditional attention models: BiLSTM-ATT model, BiGRU-ATT;
Based on the improved attention models: BiLSTM-2ATT, BiGRU-2ATT.

(1) The structure of BiGRU. As shown in Table 9, the model of the BiGRU networks can extract military relations more effectively. The F1-score with the BiGRU structure is

1.8–2.5% higher than those with the BiLSTM structure. Thus, the GRU network, as a variant of LSTM networks, can not only acquire memory sequence characteristics effectively, but can also learn long-distance dependency information. Military texts are usually long and there are many long-dependent sentences. As the BiGRU structure can acquire rich contextual features in military text, relation extraction is more effective.

Table 9. Comparison of different feature extraction methods.

Model	Precision (%)	Recall (%)	F1-Score (%)
BiLSTM	74.6	78.1	76.3
BiGRU	75.3	81.1	78.1
BiLSTM-ATT	79.1	82.4	80.7
BiGRU-ATT	81.2	85.3	83.2
BiLSTM-2ATT	82.5	85.9	84.2
BiGRU-2ATT	87.2	85.1	86.1
BiGRU-MHATT (**ours**)	90.8	89.6	**90.2**

(2) The influence of the attention mechanism. From Table 8, we can observe that the attention mechanism is better than the non-attention mechanism. The F1-score of the BiLSTM-ATT model is 4.4% higher than those of the BiLSTM model, and the F1-score of the BiGRU-ATT model is 5.1% higher than those of the BiGRU model, indicating that the attention mechanism effectively improves the accuracy in extracting military relations. At the same time, the F1-score of our proposed model, which combines BiGRU with a multi-head attention mechanism, is at least 4.1% higher than that of other attention mechanism models, indicating that our model can learn more sentence characteristics from military text, thus improving the extraction efficiency of military relations. Therefore, we believe that the model proposed in this paper has several advantages in extracting military relations.

4.4.3. Comparison of Result on Different Training Data Sizes

To test the training efficiency of the model, we designed six training corpus' of different sizes, between 1000 and 5000 words, and evaluated the performance of the BiLSTM-2ATT, BiGRU-2ATT, and BiGRU-MHATT models. As shown in Figure 9, the performance gap between the five models becomes more significant as the size of the training set increases. When the dataset reaches 4000, the F1-score of BiGRU-MHATT approaches the maximum value of BiGRU-2ATT. This indicates that the BiGRU-MHATT model proposed in this paper can make full use of the training document.

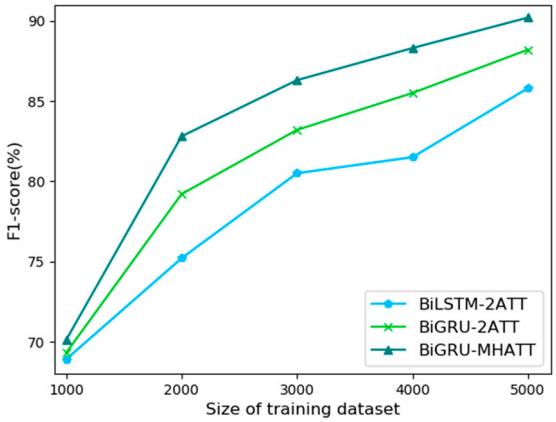


Figure 9. Comparison of three models under different sizes of training dataset.

4.4.4. Comparison of Result on Different Sentence Length

To test the sensitivity of the model to sentence length, we classified the test corpus by sentence length, including (<20, [20, 30], [30, 40], [40, 50], >50), and evaluated the performance of BiLSTM-2ATT, BiGRU-2ATT, and BiGRU-MHATT. As shown in Figure 10, BiGRU-MHATT is superior to the BiLSTM-2ATT and BiGRU-2ATT models, in terms of sentence length. With the increase in sentence length, the information acquisition performance of the three models showed a downward trend, and the BiGRU-MHATT model was slower than the other two models. The results show that BiGRU-MHATT can acquire the semantic features in long text more effectively.

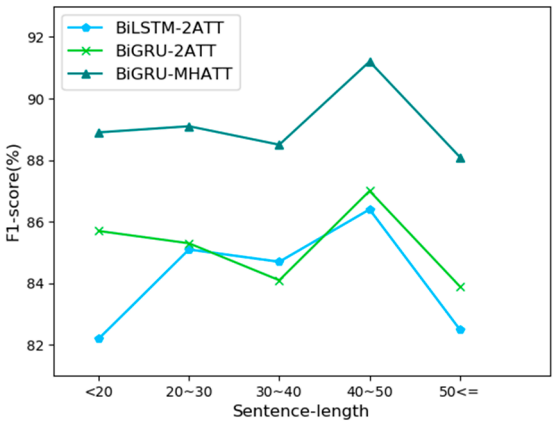


Figure 10. Comparison of three models under different sentence lengths.

4.4.5. Comparison of Result on Dataset SemEval-2010 Task 8

To verify the generalization ability of our model, we conducted experiments on public corpus: SemEval2010-Task8, which contains 10,717 sentences, including 8000 training and 2717 testing instances.

As shown in Table 10, the extraction effect of our model on the English dataset is not the best, which indicates that the generalization ability of our model needs to be improved. It can also prove that our model can learn more features from military texts based on the pre-trained language model and multi-head attention mechanism.

Table 10. Comparison of different feature extraction methods.

Model	SemEval-2010 Task 8	Military Corpus
BiLSTM-ATT	84.0	80.7
BiGRU-ATT	**85.2**	83.2
BERT-BiGRU-MHATT	84.2	**90.2**

5. Conclusions and Future Work

In this paper, we construct a military relation extraction model based on the characteristics of Chinese military text. Through the constructed military text corpus, the experimental results show that our model can achieve better performance than traditional non-attention models, traditional attention models, and improved attention models. In further experiments, although in a different language dataset, the generalization ability of our model needs to be further improved, but it is verified that our model has stronger robustness and generalization ability in different training data sizes and sentence lengths.

In the future, we plan to expand the military text corpus, distinguish fine-grained semantic information to achieve fine-grained military relation extraction, and we will try to extract military entity and relation jointly.

Author Contributions: Conceptualization, Y.L. and R.Y.; methodology, Y.L. and X.J.; software, C.Y. and D.Z.; validation, C.Y. and Z.L.; formal analysis, Y.L. and R.Y.; investigation, Y.L. and C.Y.; resources, Y.L. and C.Y.; data curation, Y.L. and C.Y.; writing—original draft preparation, Y.L. and X.J.; writing—review and editing, Z.L.; visualization, D.Z. and Z.L.; supervision, X.J. and D.Z.; project administration, R.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Foundation of China (No. 2019-SKJJ-C-083).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank all of the reviewers for their comments on this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ding, K.; Liu, S.; Zhang, Y.; Zhang, H.; Zhou, X. A knowledge-enriched and span-based network for joint entity and relation extraction. *Comput. Mater. Contin.* **2021**, *680*, 377–389. [CrossRef]
2. Zhao, K.; Xu, H.; Cheng, Y.; Li, X.; Gao, K. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowl.-Based Syst.* **2021**, *219*, 106888. [CrossRef]
3. Li, Y.; Huang, W. Weak supervision recognition of military relations. *Electron. Des. Eng.* **2018**, *1*, 77–78.
4. Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2014 on Interactive Poster Demonstration Sessions, Barcelona, Spain, 21–26 July 2004.
5. Suchanek, F.M.; Ifrim, G.; Weikum, G. Combining linguistic and statistical analysis to extract relations from Web documents. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 712–717.
6. Bunescu, R.C.; Mooney, R.J. A shortest path dependency kernel for relation extraction. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, BC, Canada, 6–8 October 2005; pp. 724–731.
7. Wang, M. A re-examination of dependency path kernels for relation extraction. In Proceedings of the IJCNLP, Hyderabad, India, 7–12 January 2008; pp. 841–846. Available online: https://aclanthology.org/I08-2119.pdf (accessed on 8 September 2021).
8. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation classification via convolutional deep neural network. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 2335–2344.
9. Zhang, S.; Zheng, D.; Hu, X.; Yang, M. Bidirectional long short-term memory networks for relation classification. In Proceedings of the PACLIC; 2015; pp. 73–78. Available online: https://aclanthology.org/Y15-1009.pdf (accessed on 8 September 2021).
10. Dos Santos, C.; Xiang, B.; Zhou, B. Classifying relations by ranking with convolutional neural networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; pp. 626–634.
11. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying relations via long short-term memory networks along shortest dependency paths. *arXiv* **2015**, arXiv:1508.03720.
12. Che, W.X.; Liu, T.; Li, S. Automatic Entity Relation Extraction. *J. Chin. Inf. Process.* **2005**, *19*, 1–6.
13. Zelenko, D.; Aone, C.; Richardella, A. Kernel methods for relation extraction. *J. Mach. Learn. Res.* **2003**, *3*, 1083–1106.
14. Plank, B.; Moschitti, A. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, 4–9 August 2013; Volume 1, pp. 1498–1507.
15. Liu, C.Y.; Sun, W.B.; Chao, W.H.; Che, W.X. Convolution neural network for relation extraction. In Proceedings of the 9th International Conference on Advanced Data Mining and Applications, Hangzhou, China, 14–16 December 2013; pp. 231–242.
16. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015.
17. Zhang, D.; Wang, D. Relation classification via recurrent neural network. *arXiv* **2015**, arXiv:1508.01006.
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
19. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.-W.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, 7–12 August 2016; pp. 207–212.

20. Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 14–16 October 2014; pp. 1724–1734.
21. Luo, X.; Zhou, W.; Wang, W.; Zhu, Y.; Deng, J. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. *IEEE Access* **2018**, *6*, 5705–5715. [CrossRef]
22. Zhang, L.X.; Hu, X.W. Research on character relation extraction from Chinese text based on Bidirectional GRU neural network and two-level attention mechanism. *Comput. Appl. Softw.* **2018**, *35*, 136–141.
23. Zhou, H.; Yang, Y.; Ning, S.; Liu, Z.; Lang, C.; Lin, Y. Combining Context and Knowledge Representations for Chemical-Disease Relation Extraction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16*, 1879–1889. [CrossRef] [PubMed]
24. Li, Z.; Sun, Y.; Zhu, J.; Tang, S.; Ma, H. Improve relation extraction with dual attention-guided graph convolutional networks. *Neural Comput. Appl.* **2021**, *33*, 1773–1784. [CrossRef]
25. Thakur, N.; Han, C.Y. An Ambient Intelligence-Based Human Behavior Monitoring Framework for Ubiquitous Environments. *Information* **2021**, *12*, 81. [CrossRef]
26. Liu, C.; Yu, Y.; Li, X.; Wang, P. Application of Entity Relation Extraction Method under CRF and Syntax Analysis Tree in the Construction of Military Equipment Knowledge Graph. *IEEE Access* **2020**, *8*, 200581–200588. [CrossRef]
27. Wang, C.; He, X.; Zhou, A. Open Relation Extraction for Chinese Noun Phrases. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 2693–2708. [CrossRef]
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-trained of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
29. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv* **2013**, arXiv:1310.4546.
30. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
31. Kwon, H. Friend-Guard Textfooler Attack on Text Classification System. *IEEE Access* **2021**, *99*. [CrossRef]
32. Vaswani, A.; Shazeer, N.; Parmar, N. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
33. Thakur, N.; Han, C.Y. Multimodal Approaches for Indoor Localization for Ambient Assisted Living in Smart Homes. *Information* **2021**, *12*, 114. [CrossRef]
34. Doddington, G.; Mitchell, A.; Przybocki, M.; Ramshaw, L.; Strassel, S.; Weischedel, R. The automatic content extraction (ACE)program-tasks, data, and evaluation. In Proceedings of the COLING, Stroudsburg, PA, USA, 23–27 August 2004; pp. 2014–2329. Available online: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf (accessed on 8 September 2021).
35. Girju, R.; Nakov, P.; Nastase, V.; Szpakowicz, S.; Turney, P.; Yuret, D. Classification of semantic relations between nominals. In Proceedings of the SemEval, Prague, Czech Republic, 23–24 June 2007; pp. 13–18. Available online: https://aclanthology.org/S07-1.pdf (accessed on 8 September 2021).
36. Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; Sun, M. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv* **2018**, arXiv:1810.10147.