applied sciences

MDPI

Article

Few-Shot Relation Extraction on Ancient Chinese Documents

Bo Li, Jiyu Wei, Yang Liu, Yuze Chen, Xi Fang and Bin Jiang *

School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai 264209, China; libo18@mail.sdu.edu.cn (B.L.); weijiyu@mail.sdu.edu.cn (J.W.); liuyangproven@mail.sdu.edu.cn (Y.L.); cyz122699@mail.sdu.edu.cn (Y.C.); fangxi@mail.sdu.edu.cn (X.F.)
* Correspondence: jiangbin@sdu.edu.cn

Featured Application: Our work can be applied to ancient Chinese documents or other digital humanity research fields with limited data.

Abstract: Traditional humanity scholars' inefficient method of utilizing numerous unstructured data has hampered studies on ancient Chinese writings for several years. In this work, we aim to develop a relation extractor for ancient Chinese documents to automatically extract the relations by using unstructured data. To achieve this goal, we proposed a tiny ancient Chinese document relation classification (TinyACD-RC) dataset annotated by historians and contains 32 types of general relations in ShihChi (a famous Chinese history book). We also explored several methods and proposed a novel model that works well on sufficient and insufficient data scenarios, the proposed sentence encoder can simultaneously capture local and global features for a certain period. The paired attention network enhances and extracts relations between support and query instances. Experimental results show that our model achieved promising performance with scarce corpus. We also examined our model on the FewRel dataset and found that outperformed the state-of-the-art no pretraining-based models by 2.27%.

Keywords: ancient Chinese document; relation extraction; few-shot learning; sentence encoder; digital humanity

Citation: Li, B.; Wei, J.; Liu, Y.; Chen, Y.; Fang, X.; Jiang, B. Few-Shot Relation Extraction on Ancient Chinese Documents. *Appl. Sci.* **2021**, *11*, 12060. https://doi.org/10.3390/app112412060

Academic Editor: Valentino Santucci

Received: 5 July 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

According to reference [1], traditional humanity studies are currently experiencing a crisis in which humanity and humanistic meaning must compete with the social, economic, popular culture, and other forms of prevailing entertainment. This crisis also has an influence on society's pedagogy and people's job-seeking tendency. Humanity research also suffers from various problems, such as information overload, lack of specific recommendations, and limited dissemination in the mass media. In contrast with traditional humanity research, digital humanity uses computer technology to process the large amounts of disorganized data, thereby can effectively reduce the difficulty of massive data organization and usage [2], the scarcity of diversity in information dissemination and influence, and the limitations of humanity research applications in the society. Therefore, digital methods, such as Natural Language Processing (NLP), Computer Vision (CV), and Data Science (DS), must be applied to support humanity research.

This work mainly focuses on automatically extracting relations in ancient Chinese documents under limited data (our proposed dataset and model is accessible at https://github.com/boss66757979/MASCOT-PA, accessed on 16 December 2021), which has similar application scenarios to the real-world tasks. Nonetheless, training and evaluation samples are difficult to acquire in these tasks. For example, the research on the ShihChi corpus necessitates the creation of a dataset by historian annotators, which would take a significant amount of time and effort and would be expensive and unattainable. Their progress has also been hampered due to their failure to use digital research in these fields.

Despite the difficulties, research in these fields still has great significance. The most difficult task for humanity researchers is to find specific pieces of information in demand from a mess of documents, which require several years of study to learn and understand the inner relations in this information, while it would only consume a few minutes handled digitally. This work might also be useful in other fields, including biological science, health care [3], law and jurisdiction [4], and chemistry [5], if similar training and evaluation data criteria are met. These situations, which play significant roles in scientific study, are also important to society.

According to the above reasons, research and analysis in ancient Chinese documents with limited data are significant. A feasible measure for extracting and analyzing these limited data must be established because a standard training dataset requires numerous resources (time, human, and finance) to create.

The common deep learning solutions for information extractions are based on standard or distantly supervised learning methods, which extract information, such as entities and relations. However, these methods are limited in real-world scenarios. A training corpus for standard learning methods is difficult to obtain, and distantly supervised learning methods inevitably experience the long-tail problem [6] in datasets. For example, NYT-10 [7] comprises news data that have an imbalanced distribution of topics when used in other areas.

To deal with the problems under a limited data scenario, the few-shot learning method, which only requires five or fewer support instances to predict the query instance label, was proposed to extract the relations. The few-shot learning task for relation classification has been proposed using the FewRel [6] dataset. The pretraining-based method has achieved the best result using this dataset [8], surpassing human performance on relation classification tasks. However, pretraining-based models require a huge amount of corpus for model pretraining, which consumes great time and resources. The pretraining models of BERT-large [9] with 340 M parameters are trained using 64 TPU (Tensor Processing Unit) chips for 4 days. The training requires at least 12–16 GB GPUs to address memory constraints while training with only one batch (https://github.com/google-research/bert, accessed on 16 December 2021). The need for huge computation resources makes rapid feedback or individual experiments impractical.

The need for a large corpus cannot be met in research fields with limited text. BERT used 16 GB uncompressed plain text from BookCopus [10] and English WikiPedia. RoBERTa [11] (158 GB) and XLNET [12] (161 GB) used more corpus for models pretraining. However, even the complete collection SiKuQuanShu (http://skqs.guoxuedashi.net, accessed on 16 December 2021) has merely a total of 1.5 GB plain text data for the ancient Chinese documents. Pretraining-based models are impractical due to a lack of training corpus. Moreover, the corpus cannot be collected from contemporary Chinese documents because of their different sentence structures, idioms, and words. Pretraining is difficult because of these factors.

Accordingly, word representation-based models, such as Meta Networks (CNN), Prototypical Networks (CNN) [6], and MLMAN [13], were suggested. These models achieved good performance when GloVe [14] was used as the static word embedding vector. However, a large gap is still observed between humans and the models. The best result of these models lags 10% behind human performance on the FewRel test dataset. Moreover, these models require pretrained word representation (e.g., a standard GloVe model generally requires 42 billion tokens to generate a global word representation matrix). Although static word embedding models may be faster than dynamic models, such as BERT, they still require an external corpus.

We examined several classical few-shot learning methods that do not have a satisfying performance to address the above problems. Then, we developed a multi-head self-attention with convolution encoder (MASCOT) and paired attention (PA) relation extractor to fill the gap. The MASCOT encoder can capture more local and global features. Our experiment has shown that this encoder has a better sentence encoding ability

than convolution or transformer layer. This encoder achieves better performance than the standard GloVe-based convolution encoder, which means that its structure relieves the resource dependence to a large extent even without an external corpus for word embedding pretraining. Additionally, the paired attention layer is used to enhance instance's feature information and expressive capability. Our experimental results show that the paired attention achieved a 3.68% improvement over the base model.

For the ancient Chinese documents, we tested our proposed model on the ShihChi documents and analyzed the prediction results in Section 5.5. The result shows that the relation extraction on the ancient Chinese document is useful, and researchers can analyze or plot the structured predicted data to support their research or arguments. We performed a demo analysis for ShihChi documents to demonstrate a usage for our model, which might inspire researchers using manners of this model.

Finally, our contribution may be summarized as follows: First, we examined the prevailing few-shot learning methods and adopted them in the few-shot relation extraction task to find a better solution. Second, we used a local GloVe word representation and MASCOT-PA to avoid problems with the high pretraining corpus requirement. Finally, we applied our model to an ancient Chinese document relation extraction task and achieved promising results.

2. Related Work

2.1. Prevailing Few-Shot Learning Methods

Few-shot learning has been proposed and largely implemented in computer vision research, such as GNN [15], SNAIL [16], meta network [17], prototypical networks [18], feature pre-processing and rectify based methods [19–21], metric-based methods [22], end-to-end few-shot learning framework similar to relation network [23], and memory-augmented neural networks (MANN) [24]. The aforementioned methods are of great benefit to the performance of models in computer vision tasks, such as image classification.

GNN regards the instance as a node in the graph using label embedding as the node representation and propagates or receives it to predict a node label. SNAIL uses temporal convolutional neural networks and attention for the support and query instance pairs. Meta Network and Prototypical Networks use a distance or metric to evaluate the relation between support and query instances, where Meta Network measures all support instances distance from query instance in one class, and Prototypical Networks generate a prototype for this class to measure the distance.

In the feature adjustment, such as power transforms in [19], which modifies the instance feature vector using Equation (1), according to their experiment, 0.5 would be a suitable γ value, and ϵ is usually a small number 1×10^{-6}. This operation increases model performance by 5.88% on the mini-ImageNet dataset with WRN [25] backbone.

$$\bar{x} = \frac{(x+\epsilon)^{\gamma}}{||(x+\epsilon)^{\gamma}||_2} \tag{1}$$

Similar to power transforms, prototype rectification [20,21] adopts a pseudo-labeling strategy to augment the support set and diminish the prototype bias. Prototype rectification sums pseudo-labeled instances in $\{S, Q\}$ with their distance or similarity from the original prototypes. The rectified instance is generated through Equation (2) by using this weighted combine features of support and query instances to correct the data distribution

$$\bar{m}_c = \frac{1}{|S_c| + |Q_c|} \sum_{I \in \{S_c, Q_c\}} \frac{exp(dist(I, m_c))}{\Sigma_{c=1}^{C} exp(dist(I, m_c))} \cdot I, \tag{2}$$

where $|S_c|$ and $|Q_c|$ represents the support and query set sizes of class c respectively; $dist(\cdot, \cdot)$ is a euclidean distance function; and m_c is the mean of all class c support instances. Accordingly, the model performance increased by 5.91% on the mini-ImageNet from this rectification.

The IADM is proposed by [22], which is set up to improve the transductive learning performance on all instances and alleviate the improper confidence weight when calculating prototypes. The original procedure is shown in Equation (3). This metric significantly outperforms the original ProtoNets by 3.72%.

$$\bar{D} = \left\| \frac{q/||q||_2}{g_\phi(q)} - \frac{c_n/||c_n||_2}{g_\phi(c_n)} \right\|_2^2 \quad (3)$$

Relation network [23] is a general framework for few-shot learning, which concatenates all support instances with a query instance and produces their relation score or similarity by a relation model g_ϕ. The g_ϕ uses convolutional block as a few-shot relation predictor, (Deep Neural Network) DNN as zero-shot relation predictor. This framework can be flexible for other task transplantation and powerful because of the end-to-end learning for embedding and nonlinear metrics, which achieved state-of-the-art performance on the Omniglot dataset.

Using external memory to enhance neural network performance is a general idea that shares a similar mechanism as the human brain learning procedure. The MANN or Neural Turing Machines (NTMs) [26,27] use this strategy to reinforce the models' ability to store variables and data structures when training at long timescales, which means that it can relieve the forgetting problem of detail information on each training step. In few-shot learning, the meta-learning with memory-augmented neural networks [24] adopt an external memory dump to store and retrieve previously input memory. The external memory matrix uses a Least Recently Used Access method to refresh memories and cosine similarity to retrieve memories. This method has a much faster converge speed than no external memory methods and also achieved a better performance than humans on the Omniglot dataset.

For the above models, some of them are good at classifying semantic and sentences, and some of them perform poorly at this classification task. We developed these models and analyzed their performance in Section 5.3, the results show that the prototypical network has a strong ability handling the few-shot learning tasks, therefore, we consider it as our baseline model. Other models such as relation networks or memory aggregated neural networks perform poorly, and we analyzed its reason in Section 5.3.

2.2. Few-Shot Relation Classification

Relation classification or extraction has been proposed for several years [28,29], and much progress has been achieved on this task [30–33]. By contrast, few-shot relation classification has only been recently proposed, and few research results have been accumulated. FewRel [6] proposed few-shot relation classification tasks, which use the FewRel dataset to evaluate model performance. In addition, FewRel examined some of the most competitive few-shot learning methods, and the examination result shows that these methods cannot achieve acceptable performance on the FewRel dataset. Therefore, a few novel models were proposed to deal with this problem.

Proto-HATT [34] focuses on noisy data influence over the vanilla prototypical networks and data sparsity problem caused by the feature extraction of the support instance set. To address these problems, researchers proposed a hybrid attention mechanism to alleviate the deviation of prototypes and measure the sparse space distance. Proto-HATT has a 1.07% improvement over the vanilla prototypical networks on the FewRel dataset, especially in the presence of noisy data, where it achieves an enhancement of 3.66%. TPN [35] uses a transformer to enhance a prototypical network, and it outperforms Proto-HATT in FewRel dataset by 1.76% because of its superior sentence and relation representation ability. These two models are based on a prototypical network, which is a baseline for our proposed task. Meanwhile, MLMAN [13] used another strategy to generate each class prototype.

The MLMAN attempts to combine many types of instance aggregate methods. First, a CNN context encoder is adopted to derive the sentence representation. Then, MLMAN

uses local matching and an aggregation layer to improve different and similar semantics of query instance and support instance representations. Next, MLMAN generates all classes' prototypes by using a softmax operation with a matching degree from a matching function. Finally, MLMAN determined all query instance classes using a class-level matching function. This method achieved 5% improvement over the baseline and surpassed the above prototypical network-based models. However, we adopt a superior sentence encoder MASCOT to enhance their representation and feature modeling capability in comparison with this model. We utilize a pair function, which is also used in BERT-PAIR [36], to increase the intensity of related instance correlation and predict instance relation after we produce the prototype for each instance. Our experiment shows that this pair function surpasses prototypical network-based metric measurement and the matching function from MLMAN.

Pretraining-based models have better performance than original models. BERT-PAIR is the first model to adopt BERT and pair mechanism on this task due to its excellent language modeling performance, it has a great improvement over the baseline. Matching the Blanks [8] proved that a model could even surpass human performance by learning text representation. Based on BERT, their model modifies an output and input data form. The result shows that the model outperforms the baseline even without training on the FewRel dataset. Although these two models have superior performance, they cannot be transferred to limited data fields, which is our model most excels at.

MICK [37] proposed a framework that works well when only little training data are provided. Furthermore, MICK proposed the task enrichment that introduces other domain datasets to enhance the training data and randomly extracts instances from in-domain or cross-domain and training models to distinguish them. This strategy is highly effective when there is an extreme lack of training data.

However, according to their experiments, this framework only works well for a small number of instances (under 700) and very simple models. This framework can hardly have an obvious improvement in complex models, such as Proto-HATT, MLMAN, and BERT-PAIR (about 1% accuracy). This is because MICK aims to help models learn better and doesn't change the core part of the models. For strong baselines, the given data is sufficient to train a satisfied model. In addition, this framework inherits the original models' shortcomings. These models require enough corpus to conduct a word pretraining. Meanwhile, our model fixed these two problems through a local GloVe word embedding and an excellent sentence encoder.

2.3. Sentence Encoding

Sentence encoding helps a machine encode a sentence from the original word embedding. The four types of encoder are as follows: a recurrent neural network (RNN)-based encoder [38], a CNN-based encoder [39], a transformer encoder [40], and their combinations. Encoders, such as LSTM with attention [41] or ConvS2S [42], can easily surpass the basic phrase-based encoder [43] with larger N-gram language models [44]. Meanwhile, ConvS2S is slightly better than the LSTM models. The transformer encoder [45] has better performance on language modeling but worse performance on a long-range dependency problem. These models can capture only temporal or spatial features of a sentence, which forces researchers to adopt a composite model.

Composite models can capture more information. The CNN-RNN structured networks outperform conventional CNNs and BLSTM in terms of text semantic modeling. For example, the LSTM-CNN network [46] used a convolution network to form character-level representation. Then, the character-level representation was concatenated with word embedding to construct an LSTM input. The output result of the LSTM input was used as the final sequence feature, and it was sent to the CRF layer to accomplish a NER task. The performance of the LSTM-CNN network broke the record on the Penn Treebank WSJ and CoNLL-2003 datasets.

However, the traditional RNN-based model can suffer from its sequential structure when dealing with long-term dependencies [47]. Accordingly, the Simultaneously Self-Attending model [48] adopts a novel encoder called BRAN, which is similar to the Conformer [49], to fix the above problems by replacing RNN with a self-attention mechanism. This encoder uses a transformer encoder structure with a convolution module and multi-head attention as each block component. An instance sentence passes through the novel encoder and is sent to two MLPs to produce head or tail position-specified sentence representation. The relation between the head and the tail is calculated by performing a bi-affine operation of two sentence representations. This method achieved state-of-the-art performance on the biocreative V chemical disease relation dataset.

Although BRAN eliminates the problem that RNN brings in, this model still has problems. First, the original BRAN model uses a sequential combination of multi-head attention and convolution modules, which may lead to the loss of information during data transfer between two different modules. In our experimentation, the original model showed drawbacks on few-shot learning tasks. Our proposed MASCOT encoder relieved this problem through a parallel convolution and multi-head attention structure. In comparison with BRAN, MASCOT has better performance under the deeper layer, faster training speed through parallel structure, and stronger capability on capturing and storing features.

3. Background

The general few-shot relation classification problem is considered a classification task. Given the support set in Table 1 with N classes, each class has K instances, a query set consists of Q instances, and all Q instances belong to one of the N classes. Then, a few-shot relation classification task is expected to predict the Q instance classes by analyzing the given support set. In this section, we use a simple baseline model to demonstrate the overall procedure of the few-shot relation classification task. This baseline model consists of two components: (1) an instance encoder that converts text data into instance representation and (2) a class predictor that distinguishes the correct class from those given instances.

Table 1. Example of the support and query sets, where cyan represents head entities, and magenta denotes tail entities. In the query instance, the quark is part of a hadron; hence, it should be the part of relations.

Support Set	
Relation Classes	**Instances**
(A): crosses	Instance 1: Wilton Bridge was a major crossing of the River Wye and was protected by Wilton Castle . Instance 2: Albion Riverside, in London, is a high-end residential development located between Albert Bridge and Battersea Bridge on the River Thames.
(B): part of	Instance 1: Lava Tree State Monument is a public park located southeast of Pahoa in the Puna District on the island of Hawaii. Instance 2: "Peace Train" is the title of a 1971 hit song by Cat Stevens, taken from his album "Teaser and the Firecat".
Query Instance	
Class (A) or (B)	It takes place when a quark of one hadron and an antiquark of another hadron annihilate , creating a virtual photon or Z boson which then decays into a pair of oppositely-charged lepton.

3.1. Instance Encoder

In the few-shot relation classification task, the initial text of Q instances should be converted to a representation vector. Some well-known pretraining-based language models, such as BERT, would be the best choice. However, most of these models are cumbersome and intricate, which makes them unsuitable to be used as lightweight and fast models. To

meet these two requirements, some static word embedding models, such as CBOW [50], Skip-gram [51], and GloVe [14] can be used. GloVe performs fairly better than CBOW and other models under the same feature dimension and corpus size in terms of word analogies and word similarity tasks. We chose GloVe as the word embedding model because of its simplicity and intelligibility.

The GloVe model transforms words into an embedding matrix using a dictionary trained by corpus. The pretrained dictionary mapped words to a d_w-dimensional vector, with T words in the instance sentence. An instance could be represented as an $\mathbb{R}^{T \times d_w}$ matrix. After the word embedding for each instance in the support set and construction of the query set will be inputted into a simple CNN encoder, which consisted of a single convolution and pooling layer to learn the inner semantics, the matrix will be reduced from $\mathbb{R}^{T \times d_w}$ to $\mathbb{R}^H$, where H is the embedding hidden size. The CNN encoder can build up an instance semantic representation, which can also be trained by using an RNN or other related models. Most prevailing models choose the CNN encoder for better parallel computing acceleration.

3.2. Class Predictor

The instance encoder builds up instance representations S_n and Q for all instances in the support and query sets, respectively. Then, we introduce layer normalization for the input matrix to normalize the distribution of input representations. After the data have been normalized, similar to the prototypical network, the model will generate each class prototype and find the nearest query instance class prototype.

In class n, the prototype c_n for each class will be represented as a mean of all K instance embedding vectors S_n^k:

$$c_n = \frac{1}{|S_n|} \Sigma_{k=1}^{K} S_n^k. \tag{4}$$

Thus, all support class prototypes will be represented as an $\mathbb{R}^{N \times H}$ matrix.

With the produced prototype of each class, the distance between prototype c_n and query instance q will be regarded as a metric. A query instance class could be predicted from a softmax operation of all class distances, which maps the distance of all instances to be a probability:

$$P(q = k|S) = \frac{exp\{f_d(q, c_n)\}}{\Sigma_{n'=1}^{N} exp\{f_d(q, c_{n'})\}}, \tag{5}$$

where f_d represents the distance function. Prototypical networks use the regular Bregman divergence [52] as a distance function. Hence, squared Euclidean distance $f_d(x, y) = ||x - y||^2$ and Mahalanobis distance [53] would be adopted as a metric function.

4. Methodology

4.1. Sentence Encoder

In text sources, each instance sentence should be transformed into a word embedding matrix using the local GloVe. We first examined a standard transformer encoder to capture the time series structure and features of sentences. However, the transformer has drawbacks in capturing local features. Then, we evaluated the BRAN encoder to enhance different feature types. Nonetheless, results showed that BRAN has low performance on this task. We proposed the MASCOT encoder to promote their ability on capturing local and global semantic features in a sentence and alleviate the convolutional influence on original sentence feature (the overall structure is presented in Figure 1). Based on BRAN and Conformer, the MASCOT, which shares a similar structure with transformer encoder, shows a better performance than BRAN and transformer in our experimental results.

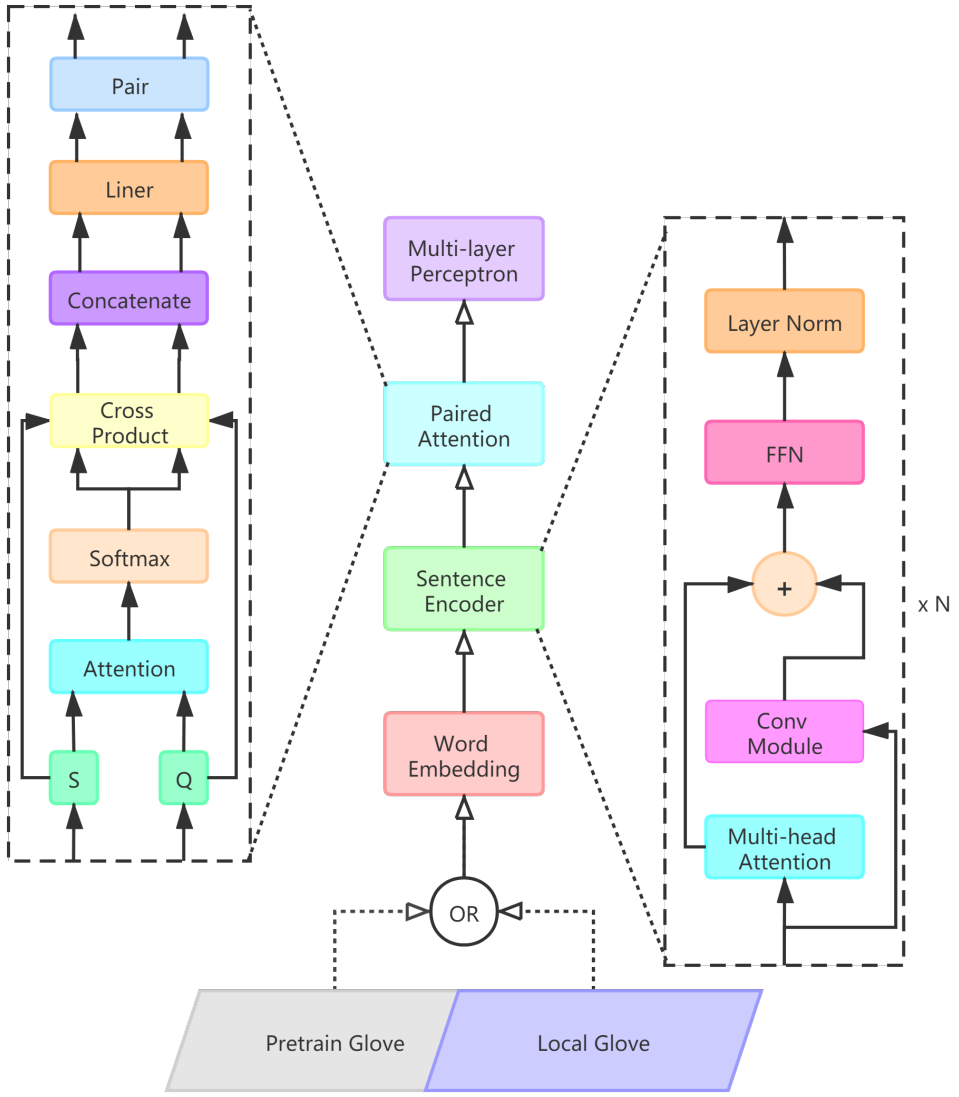

Figure 1. Overall structure of the proposed model.

In MASCOT, we introduced a three-fold convolution module along with multi-head self-attention into the proposed model to simultaneously process input source data and calculate them using a feed-forward network (FFN). FFN is a multi-layer perceptron that uses the activation function to solve the linearly inseparable problems. We compared our proposed model with Conformer and BRAN encoder. The comparison results show that our model is more suitable for the few-shot relation classification task. Moreover, the sequential structure modules in the BRAN encoder largely slow down the converging speed, resulting in the loss of information in input sources, which is the main drawback of the BRAN encoder.

The three-ford convolution module structure is shown in Figure 2. There are two convolution layers with the Mish activation function [54]. At the end of the final convolution layer, we used a residual learning method to connect the input of this module. Finally, we included a layer normalization to realize a better output distribution.

After passing through the two encoding layers, the matrix will go through a pooling layer to be reshaped. Finally, the instance matrix with the shape $\mathbb{R}^{N \times D}$ will be sent to the next module, where N is the instance number, and D is the instance embedding representation.

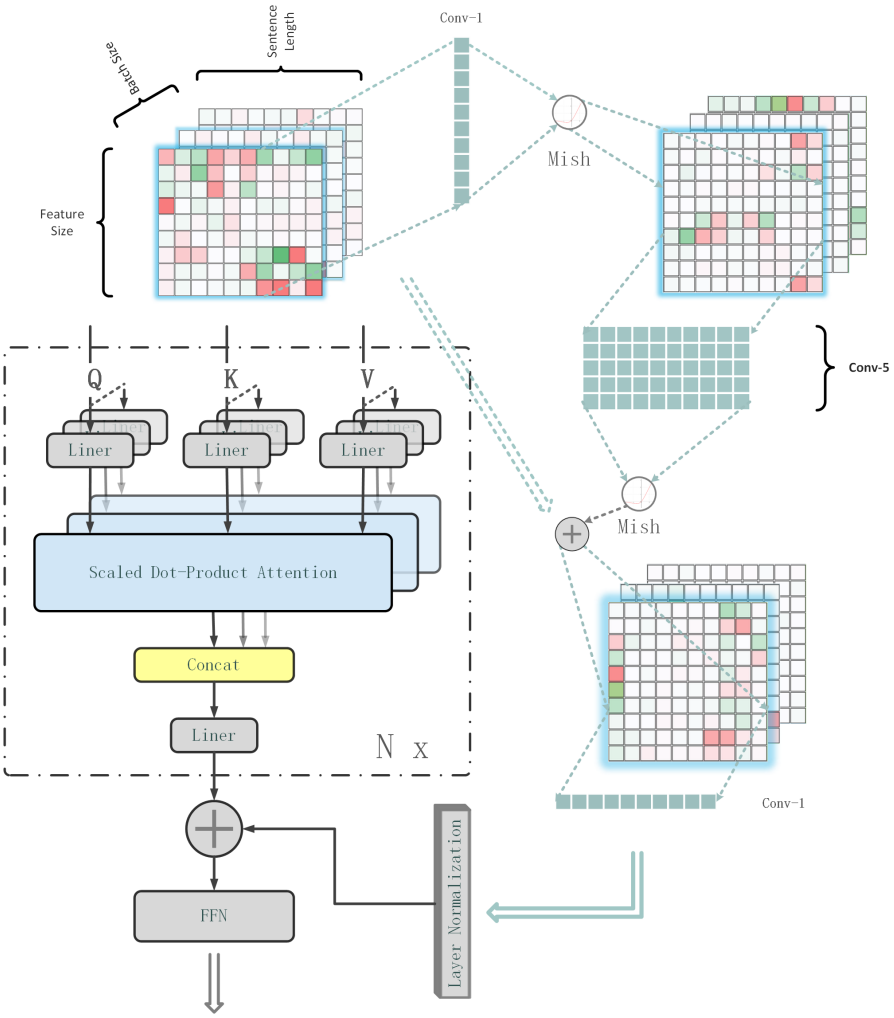

Figure 2. MASCOT encoder structure of our model.

4.2. Paired Attention

On the basis of the soft-align elements [55] for input representation and the enhancement of local inference information [35], we used a paired instance attention mechanism in this study, which consists of three parts. The first part is an input sentence representation attention, S and Q, which are generated by the sentence encoder, for the support and query instance representations. Similar to MLMAN, we first reshape S from $\mathbb{R}^{N\times K\times H}$ to $\mathbb{R}^{T_s\times H}$. Similar to S, query instance Q also has the shape $\mathbb{R}^{T_q\times H}$, where T_q represents the total number of query instances. Then, attention weights between the support and query instances can be calculated by

$$e_{ij} = Q_i^{\top} \cdot S_j, \tag{6}$$

where Q_i and S_j represent the i-th query instances and the j-th support instances, respectively; and e_{ij} is the attention weight matrix.

With the attention weight matrix, the attention-weighted support and query instance representations can be calculated as follows:

$$\widetilde{S}_i = \Sigma_{j=1}^{T_s} \frac{exp(e_{ij})}{\Sigma_{j'=1}^{T_s} exp\left(e_{ij'}\right)} \cdot S_j, \tag{7}$$

$$\widetilde{Q}_j = \Sigma_{i=1}^{T_q} \frac{exp(e_{ij})}{\Sigma_{i'=1}^{T_q} exp\left(e_{i'j}\right)} \cdot Q_i, \tag{8}$$

where $\widetilde{S}_i$ collects the attention weights of each query instance and attempts to identify related and irrelevant parts; these relations will be magnified by a product operation. The same process is performed with $\widetilde{Q}_j$.

The second part is an aggregate of attention-weighted and original instance representations, and it is designed to enhance instance information. In support instance embedding S, the attention-weighted embedding $\widetilde{Q}_j$ has attention information about S and Q, and the instance discrepancy metric between S and Q can be measured using the difference $|S - \widetilde{Q}|$ and element-wise product $S \odot \widetilde{Q}$. Thus, the aggregate operation for support instance embedding will be expressed as follows:

$$\bar{S} = f_\phi([S; \widetilde{Q}; |S - \widetilde{Q}|; S \odot \widetilde{Q}]), \tag{9}$$

and it is the same for query instance embedding:

$$\bar{Q} = f_\phi([Q; \widetilde{S}; |Q - \widetilde{S}|; Q \odot \widetilde{S}]), \tag{10}$$

where the function $f_\phi : \mathbb{R}^{T_s \times 4H} \rightarrow \mathbb{R}^{T_s \times D}$ compressed aggregated instance embedding into a denser space. Then, the support instance embedding will be split into N classes. Next, an instance-encoded sample, which is a conjunction of the support and query instance embedding, will be sent to instance pair unit.

$$Paired = MASCOT(Concat([\bar{S}; \bar{Q}])) \tag{11}$$

The third part is a paired instance relation extractor, which is designed to discover the inner relations between two paired instances. The extractor uses two layers of MASCOT to learn the relation connection of each word in the support and query instances, which is the same as Equation (11). Then, the paired instance embedding will be compressed using the max and average pooling operations in Equation (12).

$$Out = Concat([MaxPooling(Paired); AvgPooling(Paired)]) \tag{12}$$

Finally, the adjusted instance vector will be normalized by layer normalization and sent to a class predictor to find their corresponding labels.

5. Experiments

In this section, we demonstrate our experiment results. The following subsections introduce the dataset and evaluation method we used, then the training parameter configuration details and the overall performance comparison, ablation study and analysis for each component in our proposed model, and the models' appliance on the Tiny Ancient Chinese Document Relation Classification (TinyACD-RC) dataset.

5.1. Dataset and Evaluation

For models implemented in this work, we used the FewRel dataset as the training and evaluation datasets, because this dataset proposed the few-shot relation classification task and was widely adopted by models focusing on this task. As shown in Table 2, the FewRel dataset contains 70,000 sentences with 100 relations using a distantly supervised method extracted relations from Wikipedia; then, raw data were filtered and annotated by annotators to generate the standard dataset.

Table 2. Detailed structures of the used datasets, where Classes is total relation types in a dataset, and Training/Test is the ratio of Training/Test dataset.

	Classes	**Total Instances**	**Training/Test**
FewRel 1.0	100	70,000	64/16
TinyACD-RC	32	1600	24/8

The standard FewRel 1.0 task is a general few-shot relation classification task, the four types of tasks are as follows: 5-way-1-shot, 5-way-5-shot, 10-way-1-shot, and 10-way-5-shot. In an N way K shot task, each training step inputs a data matrix, which consists of support and query sets. The support set consists of N random classes extracted from the FewRel dataset, and K samples are extracted by each class as an instance. Similar to the support set, the query set will also extract $Q \times N$ samples from N classes. The standard few-shot relation classification task is used to predict all query instance classes by analyzing other support set instances.

The TinyACD-RC dataset is proposed to simulate the digital humanity analysis tasks. This dataset is generated from the ShihChi corpus (http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm, accessed on 16 December 2021) and manually annotated by historians, a sample of this dataset can be seen in Figure 3. As discussed in [56], people could have different interpretations for the same topic or sentence; therefore, these two datasets are annotated twice. First, the data will be individually labeled by two annotators and re-evaluated by the third annotator. Given that this dataset mainly focuses on an ancient Chinese document, the instance sentence consists of traditional Chinese characters. This dataset has 11 entity pair types with 32 relation classes, and the detailed compositions for this dataset are displayed in Table 3.

Support Set	
Relation classes	Instances
(A) : 生育/give birth to	黃帝者，少典之子 Huangdi, son of Shaodian 帝顓頊生子曰窮蟬 Emperor Zhuanxu gave birth to a son called Qiongchan
(B) : 婚娶/marry	姜原為帝嚳元妃 Jiang yuan was the First concubine of Emperor Ku 莊公取齊女為夫人曰哀姜 Duke Zhuang took Qi woman as his wife, Aijiang
(C) : 留居/stay at	黃帝居軒轅之丘 Huangdi lives on the hill of Xuanyuan 祝茲侯軍棘門 Marquis Zhuzi stationed at Thorny Gate
Query Instance	
Class (A), (B) or (C)	季歷娶太任 Ji Li married Tai Ren

Figure 3. Samples for TinyACD-RC dataset, which has same structure with Table 1. For query instance, where Ji Li married Tai Ren, and because the word married, Ji Li should have a marital relation with Tai Ren.

To make the dataset more accurate and fine-grained, we refer to the entity pair types in SemEval 2010 task 8 [29], while the original dataset only contains 9 types of entity pairs, and some of their types are unsuitable for our dataset, we modify and append some pair types. In our dataset, the entity subject focuses on person or object, then as the original entity pairs cause-effect and message-topic have no connection with person or object, we delete these two entity pairs. For the types we add, the executant-recipient contains relations that

represent an action the executant performs to the recipient. The entity pair giver-receiver contains the giving relation between entities, the giving thing can be an object or message. And the rest of two pair types represent the relations between the giving thing and the giver or receiver, where the giving thing can be tangible or intangible, such as a letter or a few words.

The TinyACD-RC dataset is smaller compared with FewRel. However, the instance sentence in TinyACD-RC contains an extremely complex structure and frequent ellipsis for sentence components, which makes their analysis more difficult than the normal sentence in FewRel. The traditional Chinese words in the TinyACD-RC dataset come from the hieroglyph, which means the word mainly focuses on describing the event or object. In comparison with other languages, such as Latin, hieroglyphs are less concerned about grammar, which is a complication of the TinyACD-RC dataset. Almost all ancient documents, such as ShihChi, are written with limited materials. The author has to notch words on bamboo slips without ink and pen, which means that they will omit a considerable number of words that humans can infer and understand after learning ancient Chinese for several years. For example, in ShihChi the subject for an action or event can stay far from the object, usually across a few paragraphs or the whole chapter, because author believes that the reader can infer the real subject by reading the previous paragraphs. Moreover, the document used a mass of pronouns and implications to reduce the word usage, which increases the difficulty in understanding and analyzing them. Even for humans, these papers are difficult to analyze because of their intricacy.

Table 3. All relation classes for TinyACD-RC dataset, where entity type is the category for entities in sentence and relation classes is their corresponding relation name.

Entity Pair	Relation Classes
Operator-Object (OO)	operate
Producer-Object (PO)	give birth to, create, think of
Container-Content (CC)	contain
Entity-Origin (EO)	tribute from, come from
Entity-Destination (ED)	leave, attend, enfeoff
Whole-Component (WC)	attribute, alias
Member-Collection (MC)	family member of, take office, inside, buried in, stay at
Executant-Recipient (ER)	attack, defy, pardon, recommend, manage, meet, appraise, marry, appoint, request
Giver-Receiver (GR)	give to, talk to, worship
Receiver-Giving (RG)	accept
Giver-Giving (GG)	speak

5.2. Training Configurations

Common parameter settings: The model experimental parameters are illustrated in Table 4. As mentioned in Section 3.1, we initiated word embedding using two types of GloVe matrices: the local GloVe generated by a training corpus and the pretrained GloVe matrix adopted from FewRel2.0. The local GloVe generated the 17 k frequently used word embedding vectors through an unsupervised method using FewRel texts. By contrast, pretrained GloVe consists of the 400 k frequently used word embedding vectors.

As previously mentioned, the experimentation has four types of standard FewRel 1.0 tasks. To compare the performance of the different models, we regard the model mentioned in Section 3 as the base model. We illustrated the effect of changing other components of the model by comparing their performance. The result is shown in Section 5.3.

MANN parameter setting: Similar to [57], we used 64 and 128 memory locations, Each location contains a 256-dimensional vector. In addition, the forget rate λ is set to 0.5; there is a dropout layer behind the memory-aggregate module, and the dropout rate is set up to 0.1.

Table 4. Model parameter settings.

Parameters	Value
Batch Size	1
Learning Rate	0.05/0.1
Optimization Strategy	SGD
Train Iteration	50,000
Test Iteration	10,000
Weight Decay	1×10^{-5}
Word Embedding Dimension	50
Position Embedding Dimension	5
Hidden Size	60
Sentence Max Length	40
Training Class N	20/10
Evaluate Class N	10/5
Support Instance K	5/1
Query Instance Q	5/1

5.3. Overall Results

To find a proper few-shot learning method for relation classification, we introduce and modify a few prevailing methods to few-shot relation classification tasks, while GNN, Meta Network, SNAIL, and Prototypical Networks have been implemented and clearly introduced in FewRel, We examined the rest of the models, and the comparison of all the results is illustrated in Table 5.

Table 5. Class predictor comparison, Model is the abbreviation for proposed model name and No. is their identification number (the same below). PT represents power transforms, PR represents prototype rectification, Relation represents the relation network for few-shot learning, MANN represents the memory aggregated neural network, and IADM represents the input-adaptive distance metric from the MCT model.

Model	No.	5-w-1-s
Base model	1	73.64
+PT	2	74.50
+PR	3	74.40
+Relation	4	56.49
+MANN	5	73.52
+IADM	6	75.93

Power transform has more improvement compared with the base model, partly because it maps original spread data to a Gaussian distribution, which revises the skew distribution of the data. Prototype rectification also performs pseudo-labeling that augments support samples, which rectifies the prototype and distributes it closer to the ideal one. These adjustments will prevent the deformative distribution of variables and degrade the effect of extreme values, which may predominate other values.

For IADM, We eliminated the CNN layer and retained FFN as a scaling function $g_\phi(x)$ according to the differences between the two environments. IADM has an improvement of 2.29% over the original model. In comparison with power transform, IADM also normalizes the data distribution. Moreover, IADM can better adjust the extreme value across the overall distribution with a scaling function, and this scaling can capture global features and map sparse data into a limited space. Accordingly, the model will be less influenced by the sparse distribution caused by a large-scale transformation operation.

By contrast, the relation network did not achieve a noteworthy result. This model, which is proposed for image classification, has low performance on few-shot relation

classification tasks. This defect stems from the disparate structure of data, and the model cannot use transfer learning on these two tasks.

MANN [26] can easily achieve great performance on the training dataset, but it seldom improved the performance on the development dataset.

This phenomenon resulted from the MANN memory dump's overfitting. As shown in Figure 4, after 90,000 steps, the development accuracy remains unchanged, and the model cannot learn more patterns. With the refreshing memory dump saving all advantageous and suitable information, the model for each training step only overfits the current training query and support instances.

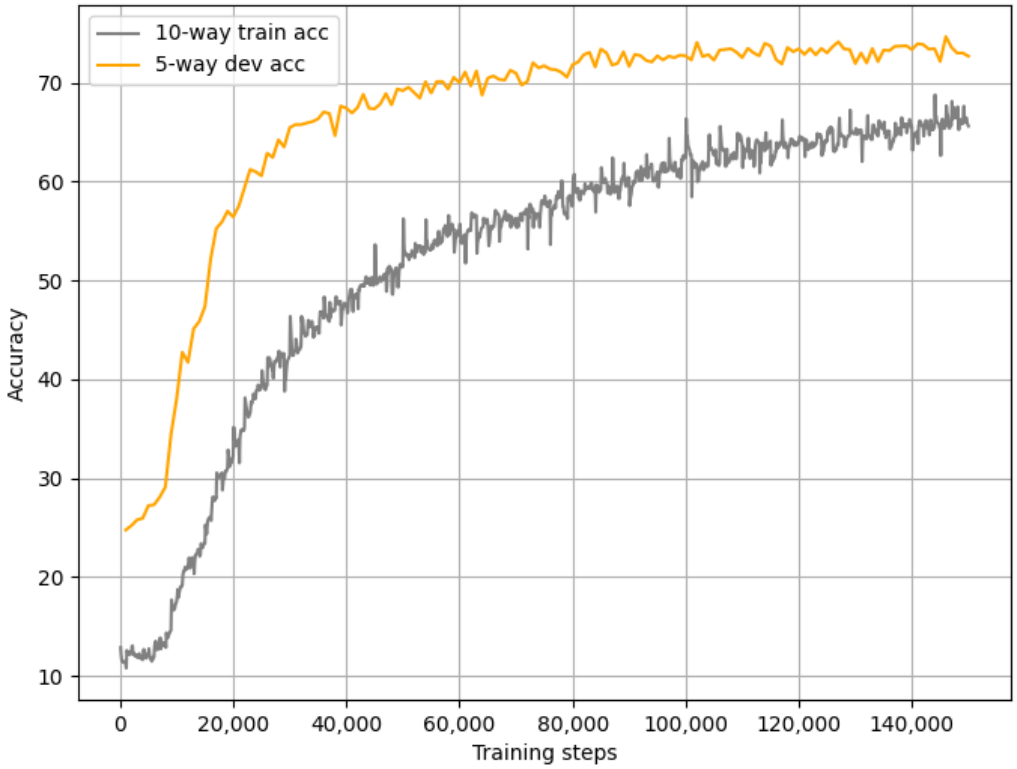

Figure 4. Comparison of training steps for training and development set performance for MANN.

In Table 6, we compared our model with other extant models on the FewRel 1.0 task. The result shows that MASCOT-PA has advantages over other no pretraining-based models, even outperforms the BERT-based TPN model and prototypical network in the 5-way-1-shot task, which proves that MASCOT-PA is excel at handling few-shot relation classification tasks. While it also exposes a drawback for MASCOT-PA. When the model adopts local GloVe as word embedding, it only performs well on *N*-way-1-shot tasks, the test result on *N*-way-5-shot even weaker than the baseline. This is because MASCOT-PA focuses on expressing more information in a sentence, trying to excavate more interconnection and relations in it, which also can be done by using more training data. This is why improving the number of the shot data has a lesser promotion for the MASCOT-PA than baseline.

Our proposed model is essential when facing the dilemma that no pretraining or external corpus is available for the pretraining procedure. Moreover, the expense of huge model training and inference makes BERT not an ideal choice for edge computing, portable computer, and devices. Therefore, we overcome the prevent models' problem with limited corpus.

Table 6. Overall result for the FewRel 1.0 task, where *N*-w-*K*-s represent the *N*-way-*K*-shot task performance.

	5-w-1-s	**5-w-5-s**	**10-w-1-s**	**10-w-5-s**
Proto (CNN) [6]	69.20	84.79	56.44	75.55
Proto-HATT [58]	75.01	87.09	62.48	77.50
MLMAN [13]	79.01	88.86	67.37	80.07
TPN (BERT) [59]	80.14	93.60	72.67	89.83
Proto (BERT) [36]	80.68	89.60	71.48	82.89
ConceptFERE(s) [60]	84.28	90.34	74.00	81.82
CTEG [58]	84.72	92.52	76.01	84.89
BERT-PAIR [58]	85.66	89.48	76.84	81.76
BERTEM+MTB [8]	88.9	-	-	-
baseline	73.64	86.93	60.91	76.88
local GloVe MASCOT-PA	76.43	86.01	62.31	74.84
MASCOT-PA	81.28	89.25	69.03	80.65

5.4. Ablation Study

To demonstrate the effect and importance of each component in MASCOT-PA, we performed the ablation study. Table 7 displays three ablation modules' performance, which includes our proposed MASCOT encoder, the paired function, and the input instance attention. We also used the standard FewRel 1.0 dataset to evaluate each module's effectiveness on this few-shot task. Given that the pair function depends on the enhancement and transformation that instance attention implements, it cannot be set apart from instance attention as a single module.

Table 7. Ablation study for our model, MASCOT is the MASCOT encoder, pair is the pair function, attention is the instance attention.

MASCOT	**Pair**	**Attention**	**No.**	**FewRel 5-w-1-s**
✗	✗	✗	1	73.64
✓	✗	✗	2	75.49
✗	✗	✓	3	77.67
✗	✓	✓	4	78.34
✓	✓	✓	5	81.28

The paired attention mentioned in Section 4.2 has achieved a remarkable performance, which means that this operation for query and support instances is suitable. The attention matrix collects similarities between query and support instances and magnifies the differences and noise in sentences and words will have less influence on class prediction by sharing this background information. The paired function also enhanced the model relation learning ability compared with the single instance attention module, and it increases the model performance by 0.67%. MASCOT has a stronger expression ability than CNN in the base model. However, in Nos. 2 and 5 in Table 7, MASCOT improves the model by 2.58% and 3.9% without and with pair attention, respectively. Therefore, only paired attention can learn and utilize the external features even though MASCOT generates richer expression features.

We also deeply analyzed the sentence encoders in Table 8. The MASCOT encoder adds a three-fold convolution layer to enhance the local relation between words compared with the standard transformer encoder. The convolutional DNN [31] shows that a proper

convolution network should have the ability to represent sentence-level features. In addition, MASCOT applies a parallel structure to avoid the information loss problem, which BRAN is mainly suffered from. MASCOT will be more expressive than BRAN and alleviate BRAN's overfitting problem. As a result, MASCOT increases the model's performance to 75.49%. By comparing the above three encoders, we can learn that transformer has a stable performance as the layer gets deeper by simultaneously computing the global and local features through multi-head attention and convolution layer, while it does not surpass MASCOT. From Figure 5, we can see that the residual connection structure for BRAN does not properly function after stacking six layers, which is probably due to the inner structure between the multi-head attention and convolution layer.

Table 8. Encoder performance comparison; BRAN represents the bi-affine relation attention networks, MASCOT is our proposed multi-head self-attention with convolution Encoder, and TE represents the standard Transformer Encoder.

Model	No.	5-Way-1-Shot (Single Layer)	5-Way-1-Shot (2 Layers)
Base model	1	73.64	
+BRAN encoder	2	74.06	73.43
+MASCOT	3	74.83	75.49
+TE	4	75.04	75.12

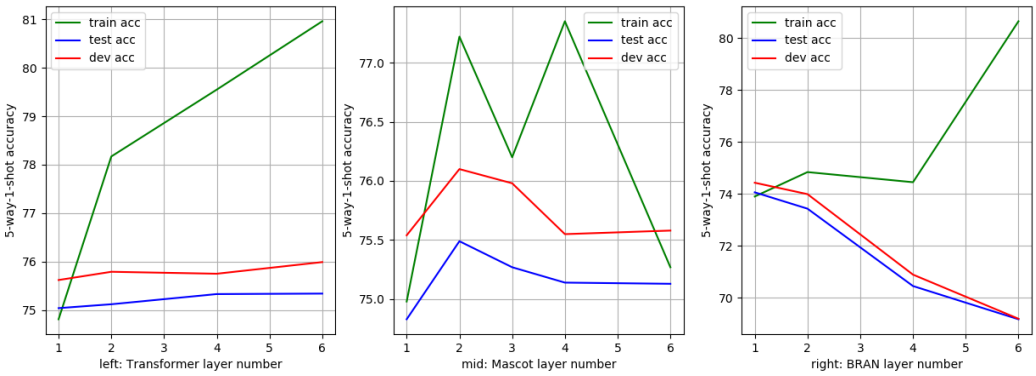

Figure 5. Comparison of the different model depth performance.

This result shows that an annex convolution module can capture additional information about the sentence. Nonetheless, the sequential structure may degrade the extraction performance due to the transformation between the self-attention and convolution layers. According to [61], the aggregation of convolution and self-attention regularly improves performance. The convolution layer is better at capturing local information through a constant weight owing to the limited size of the kernel. Vanilla self-attention will focus more on global information, and Longformer [62] shows that properly organized global and local information can lead to a significant improvement.

5.5. ShihChi Relation Extraction

We performed the few-shot relation extraction on the ShihChi document, which is designed to examine the models' capability on a digital humanity analytical task. As illustrated in Figure 6, our proposed model has the comparative advantage in handling complex sentences with scarce training corpus. When training with the 5-way-1-shot task, MASCOT-PA can extract the deep and implied meaning by the sufficient information representation structure, while baseline lack of this ability then easily overfits in the training procedure, which makes baseline lags 14.58% behind the MASCOT-PA.

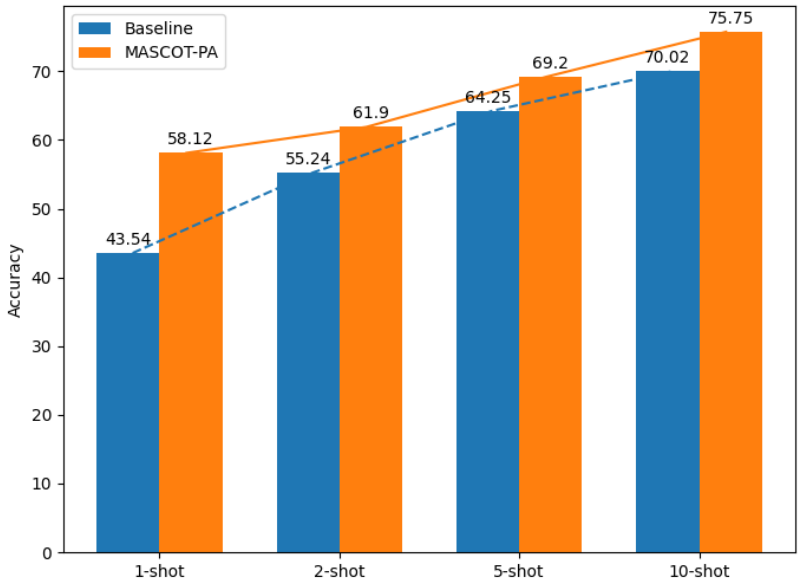

Figure 6. ShihChi document 5-way-*K*-shot relation extraction performance.

To analyze errors in the result of relation extraction, we draw two heatmaps for the baseline and MASCOT-PA in Figure 7. Where the left subgraph represents a simple baseline prediction, and the right subgraph is for the MASCOT-PA prediction. From the heatmap, we can see that MASCOT improves most of the relation predict performance. But there are a few relation classes that have no improvement or even get worse in prediction. For example, the give-to relation is predicted as the enfeoff relation both for these two models, this is because both of give-to and enfeoff relation contain the meaning of accepting something from a person, but due to the ellipsis of the sentence, most of the sentence will be like person A gives (an object) to person B and (person A) enfeoffs person B with place C. While the no pretraining-based model cannot easily distinguish whether an entity is a place or person, which exposes the drawbacks of the word embedding with limited corpus, and we believe this issue will be solved after collecting sufficient corpus for the word embedding model.

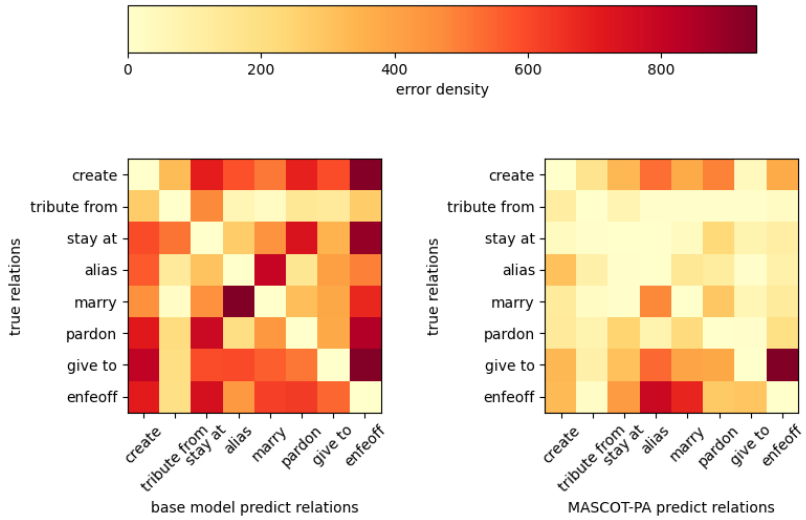

Figure 7. Wrong predicted relation distributions for base model and MASCOT-PA.

Besides we find that some of the predictions for enfeoff become worse in MASCOT-PA, the enfeoff is predicted as alias. We believe this problem has a connection with the polysemy for word enfeoff.

The word enfeoff in Chinese is not only the meaning of enfeoffing a person for a place, it also contains the meaning of conferring a title or an office on a person. And in the ShihChi, a large number of people is called by their offices or titles, which makes it easy to confuse the real referent for the word enfeoff. Moreover, as the baseline is simple enough to omit the minority implications, it can perform well to distinguish these two relations, while MASCOT-PA captures the minority meaning but cannot make a distinction between fief and office, then wrongly determines their relations.

6. Conclusions and Future Work

In this work, we presented a TinyACD-RC dataset for the ancient Chinese document analysis. In this no extra corpus task, we implemented several models to meet the no pretraining requirements for the few-shot relation classification and proposed a MASCOT-PA model. Our model outperforms satisfactory algorithms for this task in the FewRel dataset. We also applied this model to the TinyACD-RC dataset for training and test, which also shows our model's ability to handle scarce corpus.

Compared to an adhoc IR system, this model neither requires extra corpus nor has to build a complex parsing model for a specific language. When facing data scarcity or a scenario where the data needed to retrieve is little, our model can be considered as an alternative or auxiliary. Nevertheless, the other predictions require further study. We are aware of the pretraining models' advantages and will collect more of the ancient Chinese document corpus in future work to create the pretraining model for this task.

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

Funding: This research was funded by Shandong Provincial Natural Science Foundation, China (No. ZR2020MA064). The APC was funded by Shandong Provincial Natural Science Foundation, China (No. ZR2020MA064).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in MASCOT-PA repository at https://github.com/boss66757979/MASCOT-PA, accessed on 16 December 2021.

Acknowledgments: Thanks to the East Asia Digital Humanities Lab which annotates the TinyACD-RC dataset and supports this research.

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

References

1. Liu, A. The meaning of the digital humanities. *Pmla* **2013**, *128*, 409–423. [CrossRef]
2. Kaplan, F. A Map for Big Data Research in Digital Humanities. *Front. Digit. Humanit.* **2015**, *2*, 1. [CrossRef]
3. Alnazzawi, N.; Thompson, P.; Ananiadou, S. Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), Gothenburg, Sweden, 27–30 April 2014; pp. 69–74.
4. Ekstrom, J.A.; Lau, G.T. Exploratory text mining of ocean law to measure overlapping agency and jurisdictional authority. In Proceedings of the 2008 International Conference on Digital Government Research, Montreal, QC, Canada, 18–21 May 2008; pp. 53–62.
5. Krallinger, M.; Rabal, O.; Lourenco, A.; Oyarzabal, J.; Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **2017**, *117*, 7673–7761. [CrossRef] [PubMed]
6. 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. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 4803–4809.

7. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the ECMLPKDD'10 2010th European Conference on Machine Learning and Knowledge Discovery in Databases—Volume Part III, Ghent, Belgium, 14–18 September 2010; pp. 148–163.
8. Soares, L.B.; FitzGerald, N.A.; Ling, J.; Kwiatkowski, T. Matching the Blanks: Distributional Similarity for Relation Learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2895–2905.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K.N. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2018; pp. 4171–4186.
10. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 19–27.
11. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
12. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
13. Ye, Z.X.; Ling, Z.H. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2872–2881.
14. 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.
15. Satorras, V.G.; Estrach, J.B. Few-shot learning with graph neural networks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
16. Mishra, N.; Rohaninejad, M.; Chen, X.; Abbeel, P. A Simple Neural Attentive Meta-Learner. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
17. Munkhdalai, T.; Yu, H. Meta networks. In Proceedings of the ICML'17 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 2554–2563.
18. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-shot Learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4077–4087.
19. Hu, Y.; Gripon, V.; Pateux, S. Leveraging the feature distribution in transfer-based few-shot learning. In *International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 487–499.
20. Liu, J.; Song, L.; Qin, Y. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28. 2020, Proceedings, Part I 16*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 741–756.
21. Ziko, I.; Dolz, J.; Granger, E.; Ayed, I.B. Laplacian Regularized Few-Shot Learning. In Proceedings of the ICML 2020: 37th International Conference on Machine Learning, Online, 13–18 July 2020; Volume 1, pp. 11660–11670.
22. Kye, S.M.; Lee, H.B.; Kim, H.; Hwang, S.J. Meta-Learned Confidence for Few-shot Learning. *arXiv* **2020**, arXiv:2002.12017.
23. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–24 June 2018; pp. 1199–1208.
24. Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. Meta-learning with memory-augmented neural networks. In Proceedings of the ICML'16 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; Volume 48, pp. 1842–1850.
25. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016*; British Machine Vision Association: Durham, UK, 2016.
26. Graves, A.; Wayne, G.; Danihelka, I. Neural turing machines. *arXiv* **2014**, arXiv:1410.5401.
27. Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwinska, A.; Colmenarejo, S.G.; Grefenstette, E.; Ramalho, T.; Agapiou, J.P.; et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **2016**, *538*, 471–476. [CrossRef] [PubMed]
28. Girju, R.; Nakov, P.; Nastase, V.; Szpakowicz, S.; Turney, P.; Yuret, D. Semeval-2007 task 04: Classification of semantic relations between nominals. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007; pp. 13–18.
29. Hendrickx, I.; Kim, S.N.; Kozareva, Z.; Nakov, P.; Séaghdha, D.O.; Padó, S.; Pennacchiotti, M.; Romano, L.; Szpakowicz, S. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 33–38.
30. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; 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.
31. 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*; Dublin City University and Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 2335–2344.

32. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural relation extraction with selective attention over instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 2124–2133.
33. Feng, J.; Huang, M.; Zhao, L.; Yang, Y.; Zhu, X. Reinforcement learning for relation classification from noisy data. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
34. Gao, T.; Han, X.; Liu, Z.; Sun, M. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 27 January–1 February 2019; Volume 33, pp. 6407–6414.
35. Chen, Q.; Zhu, X.; Ling, Z.H.; Wei, S.; Jiang, H.; Inkpen, D. Enhanced LSTM for Natural Language Inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1657–1668.
36. Gao, T.; Han, X.; Zhu, H.; Liu, Z.; Li, P.; Sun, M.; Zhou, J. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6249–6254.
37. Geng, X.; Chen, X.; Zhu, K.Q.; Shen, L.; Zhao, Y. MICK: A Meta-Learning Framework for Few-shot Relation Classification with Small Training Data. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 415–424.
38. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
39. Gehring, J.; Auli, M.; Grangier, D.; Dauphin, Y. A Convolutional Encoder Model for Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 1–6 August 2017; pp. 123–135.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
41. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
42. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1243–1252.
43. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 24–29 June 2007; pp. 177–180.
44. Buck, C.; Heafield, K.; van Ooyen, B. N-gram Counts and Language Models from the Common Crawl. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 3579–3584.
45. Tang, G.; Muller, M.; Rios, A.; Sennrich, R. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 2–4 November 2018; pp. 4263–4272.
46. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
47. Alqushaibi, A.; Abdulkadir, S.J.; Rais, H.M.; Al-Tashi, Q. A Review of Weight Optimization Techniques in Recurrent Neural Networks. In Proceedings of the 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 8–9 October 2020; pp. 196–201.
48. Verga, P.; Strubell, E.; McCallum, A. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 872–884.
49. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented Transformer for Speech Recognition. *arXiv* **2020**, arXiv:2005.08100.
50. Mikolov, T.; tau Yih, W.; Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
51. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
52. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman Divergences. *Siam Int. Conf. Data Min.* **2005**, *6*, 1705–1749.

53. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18. [CrossRef]
54. Misra, D. Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv* **2019**, arXiv:1908.08681.
55. Parikh, A.P.; Tackstrom, O.; Das, D.; Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2249–2255.
56. Bodrunova, S.S.; Blekanov, I.S.; Kukarkin, M. Topics in the Russian Twitter and Relations between their Interpretability and Sentiment. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; pp. 549–554. [CrossRef]
57. Collier, M.; Beel, J. Memory-Augmented Neural Networks for Machine Translation. *arXiv* **2019**, arXiv:1909.08314.
58. Wang, Y.; Bao, J.; Liu, G.; Wu, Y.; He, X.; Zhou, B.; Zhao, T. Learning to Decouple Relations: Few-Shot Relation Classification with Entity-Guided Attention and Confusion-Aware Training. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 5799–5809.
59. Wen, W.; Liu, Y.; Ouyang, C.; Lin, Q.; Chung, T. Enhanced prototypical network for few-shot relation extraction. *Inf. Process. Manag.* **2021**, *58*, 102596. [CrossRef]
60. Yang, S.; Zhang, Y.; Niu, G.; Zhao, Q.; Pu, S. Entity Concept-enhanced Few-shot Relation Extraction. *arXiv* **2021**, arXiv:2106.02401.
61. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 3286–3295.
62. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.