Article

A New Entity Relationship Extraction Method for Semi-Structured Patent Documents

Liyuan Zhang [1,2,†] , Xiangyu Sun [3,†] , Xianghua Ma [3,*] and Kaitao Hu [3]

1 College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China; zly@icc.sh.cn
2 Shanghai IC Technology & Industry Promotion Center, Shanghai 201203, China
3 School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai 201418, China; 226101123@mail.sit.edu.cn (X.S.); 226102102@mail.sit.edu.cn (K.H.)
* Correspondence: xhuam@sit.edu.cn
† These authors contributed equally to this work as co-first authors.

Abstract: Aimed at mitigating the limitations of the existing document entity relation extraction methods, especially the complex information interaction between different entities in the document and the poor effect of entity relation classification, according to the semi-structured characteristics of patent document data, a patent document ontology model construction method based on hierarchical clustering and association rules was proposed to describe the entities and their relations in the patent document, dubbed as MPreA. Combined with statistical learning and deep learning algorithms, the pre-trained model of the attention mechanism was fused to realize the effective extraction of entity relations. The results of the numerical simulation show that, compared with the traditional methods, our proposed method has achieved significant improvement in solving the problem of insufficient contextual information, and provides a more effective solution for patent document entity relation extraction.

Keywords: semi-structured patent document; entity relationship extraction; hierarchical clustering; association rules; attention mechanism

1. Introduction

Through the analysis of patent documents, crucial information about the development status of technological and production processes in a research field can be obtained. However, due to the large number of patent documents, carrying out manual analysis and information extraction for each one would be an immense task. Additionally, this approach is influenced by the technical capabilities of the operators. Therefore, the automated retrieval of technical information becomes a critical factor in patent analysis [1]. Extracting entity relationships from patent documents involves a comprehensive process that begins with establishing an ontology model, followed by entity recognition, and culminating in entity relationship extraction. The ontology model [2] defines the concepts and relationships inherent in patent documents, such as inventors, applicants, patent classifications, and technical fields. By structuring these entities and their interconnections, the ontology provides a framework for understanding the complex information within patent documents. This structured representation is essential for accurately identifying and categorizing the various components of patent texts. In the realm of entity recognition [3], statistical characterization methods are commonly employed. These methods involve extracting representative features from patent documents, selecting the most pertinent features to enhance model accuracy, and training models on annotated data to learn the relationships between features. The trained models are then tested on unlabeled data to evaluate their accuracy and generalization capabilities. Convolutional Neural Networks (CNNs) play a significant role in this process by learning to map the characteristics of input data to output labels through layers of convolution, pooling, and fully connected nodes [4].

This deep learning approach allows for effective feature extraction and mapping, essential for identifying entities within the documents.

Entity relationship extraction is performed based on ontology models and entity recognition. Traditional methods for entity relationship extraction include rule-based and template-based approaches [5], as well as statistical machine learning methods [6]. These conventional approaches require extensive manual intervention for feature selection and struggle to scale to other domains. With the advancement of deep learning, leveraging deep neural network models for entity relationship extraction has become a trend. Convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) are two prominent networks used for entity relationship extraction. CNN-based networks [7,8] primarily use one-dimensional convolutions to extract local features and spatial information from the data, while they lack the ability to capture contextual information, which impacts the accuracy of entity relationship extraction. LSTM-based networks [9,10] can model long-range dependencies within the data, thereby handling complex entity relationships more effectively. However, they come with a larger number of parameters and longer training times. To address these issues, researchers have proposed integrating multiple neural network types to extract features at various levels [11,12]. Nevertheless, such methods can lead to feature redundancy and increased training complexity.

To enhance the representation of key features and improve the handling of contextual information regarding entity relationships, pre-training models and attention mechanisms are utilized to achieve accurate identification of entity relationships in this paper, dubbed MPreA. Specifically, the pre-training model RoBERTa [13] is used to enhance learning efficiency and accuracy in entity extraction by providing rich, high-dimensional feature spaces. Additionally, attention mechanisms are designed to improve the model's focus on important information within sequences, enabling a better contextual understanding and more precise entity boundary detection. By integrating these advanced techniques, the extraction process becomes more robust, resulting in accurate and comprehensive identification of entity relationships in patent documents.

2. Related Works

Information extraction (IE) technology aims to rapidly and efficiently extract valuable information from large datasets [14]. Entity relation extraction, as one of the core tasks in information extraction [15], has garnered widespread attention from both the academic and industrial sectors in recent years. By modeling document information, entity relation extraction aims to automatically identify entities, entity types, and specific types of relationships between entities, providing foundational support for reasoning in patent knowledge graphs. Currently, entity relation extraction methods can be broadly categorized into three types: supervised learning methods [16], semi-supervised learning methods [17], and unsupervised learning methods [18,19].

(1) Supervised learning methods: Supervised learning-based entity relation extraction methods treat relation extraction as a classification problem. By using labeled data for training, these methods input the labeled data into a constructed model for learning, resulting in a final model for entity recognition and entity relation extraction. Kambhatla et al. [20] studied features such as dependency syntactic analysis in document data and utilized a maximum entropy classification model for training, achieving good results in relation extraction tests on public datasets. Shan et al. [21] investigated features based on knowledge points, core predicates, and discourse differences, extracting key knowledge points in the sports domain using a relation model. Hou et al. [22] proposed a bootstrapping rule discovery method for robust relation extraction, validated through numerous experiments. The supervised entity relationship extraction algorithm, fused with deep learning, is primarily constructed using a convolutional neural network (CNN). Through the neural network, it learns the features of the relationships between relevant entities without the need for manually establishing a knowledge entity relationship library. Li et al. [23] built an entity relation training model using a convolutional neural network. Significantly

reducing the dependency on manual feature extraction. With the development of deep learning, researchers have made improvements to address various issues related to text entity relations. Zhou et al. [24] incorporated attention mechanisms into bidirectional long short-term memory networks, enhancing the network's extraction of key features. Schlichtkrull et al. [25] employed graph neural networks to accomplish entity relation prediction and classification using two standard public databases. Zhou et al. [26] proposed a globally contextualized graph convolutional network, using entities as nodes and the context between entity pairs as the edges, capturing rich global contextual information about entities in patent documents. The network was pre-trained using a certain distant supervision dataset, and the experimental results showed that the network could more effectively capture entity relations. Supervised entity relationship extraction methods can learn fine-grand features and contextual relationships from data, achieving high accuracy and F1 scores. However, these methods are highly dependent on labeled data and face significant challenges in extracting complex relationships.

(2) Semi-supervised learning methods: Semi-supervised learning-based entity relation extraction methods reduce the workload of data annotation by constructing entity relationship generation seeds and using pattern learning for iterative discovery of entity relations. Common semi-supervised learning-based entity relation extraction methods include co-training [27], bootstrapping [28,29], label propagation [30,31], and others. Yuan et al. [32] proposed an edge-enhanced graph alignment network and a word-to-relation tagging method, using edge information to assist alignment between objects and entities and finding correlations between entity–entity relations and object–object relations. The effectiveness of this model was demonstrated through experiments. Kamateri et al. [33] used deep neural networks to train different parts of the dataset and integrated the new model for entity relation extraction, showing that this method significantly outperformed models trained on a single dataset. Semi-supervised entity relationship extraction methods can reduce annotation costs and alleviate data imbalance issues, yielding higher recall values. However, the quality of the final model depends on the selection of seeds and is prone to noise, thereby reducing model accuracy.

(3) Unsupervised learning methods: Unsupervised learning-based entity relation extraction methods, taking a bottom-up approach, employ clustering principles to extract entity relations, overcoming the limitations of supervision and semi-supervision in terms of annotation. Chen et al. [34] utilized existing semantic and structural features to improve the accuracy of unsupervised learning-based entity relation extraction. Yan et al. [35] introduced pattern combination clustering into unsupervised learning, enhancing the accuracy of entity relation extraction. Unsupervised entity relationship extraction methods can greatly reduce annotation costs and automatically identify existing relationship types in the data, demonstrating strong transferability. However, they depend on the quality of the initial setting of relation seeds and generally perform less well in terms of evaluation metrics such as precision, *F*1 and recall compared to semi-supervised and supervised learning methods.

To achieve accurate patent entity relationship identification, we propose a supervised patent entity relationship extraction method tailored to the semi-structured nature of patent document data, building upon ontology modeling and entity recognition. Different from previous supervised learning approaches, we employ pre-trained models to effectively represent the intrinsic features of sentence sequences and utilize attention mechanisms to capture contextual information about entities and their relationships. This approach addresses issues such as relationship overlap and entity nesting in the process of extracting semi-structured data.

3. Methodology

3.1. Patent Document Ontology Modeling Method Based on Hierarchical Clustering and Association Rules

In this section, we address the complexity of relationships between hierarchical and non-hierarchical concepts in the ontology model. We propose an automatic ontology construction method based on hierarchical clustering and association rules, leveraging the sensitivity of hierarchical clustering to hierarchical structures and the rule-based association of non-hierarchical relationships. The method includes the following main steps.

3.1.1. Concept Acquisition

To better handle the textual data in patent documents, we converted them into txt document-type data. The Jieba segmentation tool was used to segment and annotate the text with part-of-speech tags due to its ability to efficiently handle large corpora and flexibility compared with other words segmentation tools. Subsequently, we further cleaned the segmentation results by removing semantically insignificant words such as prepositions, adverbs, and auxiliary words. The criteria for removing these words were based on a curated stop word list specifically designed for patent texts, ensuring that only words lacking substantive content were filtered out. Finally, we compiled the segmented results to create a relevant concept corpus. For concept extraction, we employed the LDA (Latent Dirichlet Allocation) topic model, which automatically discovers latent topics or concepts in the text. The model classified words in the text into these topics and extracted thematic information, achieving the goal of concept extraction.

3.1.2. Inter-Conceptual Relationship Extraction

(1) Hierarchical relationship extraction

In this paper, we adopt the hierarchical clustering method, which is widely used in the fields of natural language processing and ontology construction, to extract the relationships between concepts in patent documents. The basic idea behind this method is to cluster concepts according to their similarity to form a hierarchical structure of concepts.

Specifically, this method involves normalizing two vectors, calculating the cosine value of the angle between them to determine their similarity, and then computing inter-cluster distances to achieve the merging of similar clusters. For any two concepts, denoted as c_i and c_j, with word vector representations $\vec{c}_i$ and $\vec{c}_j$, respectively, the vectors are normalized. Since these vectors may contain different words, an extension operation is performed to balance the gap between them. The extension operation is represented by Equations (1) and (2):

$$\vec{c}'_i = ((w_1, a_1), (w_2, a_2), (w_3, a_3), (w_4, a_4), (w_5, 0)) \tag{1}$$

$$\vec{c}'_j = ((w1, 0), (w_2, b_2), (w_3, b_3), (w_4, b_4), (w_5, b_5)) \tag{2}$$

Among them w is the conceptual c_i, c_j is the context word, and a and b indicate the specific frequency of occurrence of the contextual words. The similarity between two concepts can be expressed as shown in Equation (3).

$$sim\left(\vec{c}'_i, \vec{c}'_j\right) = \cos\left(\vec{c}'_i, \vec{c}'_j\right) = \frac{\sum\limits_{a \in c_i, b \in c_j} ab}{\sqrt{\sum\limits_{a \in c_i} a^2 \sum\limits_{b \in c_j} b^2}} \tag{3}$$

included among these, $sim(\cdot)$ denotes the similarity between two concepts, and the n is the number of clusters.

To calculate the distance between two clusters, it is necessary to calculate the average value of the distance between all the elements between the two clusters, then the cluster P and Q. The formula for calculating the similarity between is shown in Equation (4).

$$sim(P, Q) = \frac{\sum\limits_{p \in P, q \in Q} sim(p, q)}{|P||Q|} \tag{4}$$

(2) Non-hierarchical relationship extraction

A non-hierarchical relationship between ontology concepts means that there is no obvious subclass or parent class relationship between two or more concepts, including the relationship between parts and the whole as well as the relationship between concepts and attributes, such as object attributes and data attributes. In this paper, the association rule method is used to discover the non-hierarchical relationship between concepts, and at the same time, in order to better mine the non-hierarchical relationship between concepts, the verb is used as the label of the non-hierarchical relationship between concepts, by calculating the correlation degree between the verb and the concept pairs, the verb that meets the conditions will be used as the label of the relationship between the concepts.

3.2. Patent Document Entity Identification Method Combining Statistical Learning and Deep Learning

This article addresses the diversity of entity types and the uncertainty of entity boundaries in patent documents. Leveraging the automatic rule-learning capability in statistical feature learning and the feature extraction capability for complex features in deep learning, we propose an integrated approach for entity recognition in patent documents that combines statistical learning and deep learning. The model framework, as illustrated in Figure 1, comprises key modules such as a rule dictionary, a BERT pre-trained model, an IDCNN (iterated dilated convolutional neural networks) layer, a bidirectional GRU network layer, and a Conditional Random Field (CRF).
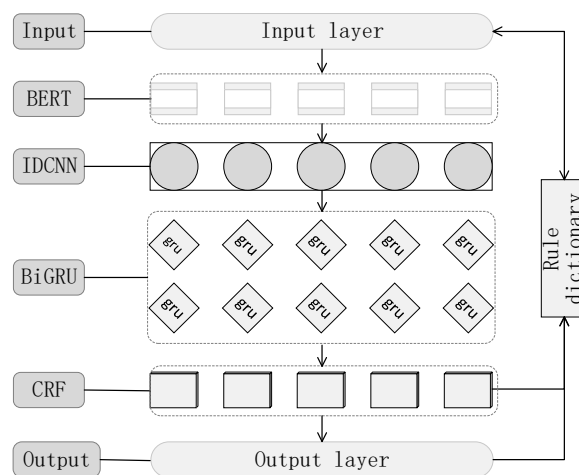

Figure 1. Frame diagram of an entity recognition model combining statistical learning and deep learning.

Initially, the text undergoes transformation using a rule-based dictionary layer. The transformed patent text data are then fed into a pre-trained model layer for vectorization. Subsequently, a dilated convolutional layer is employed to extract key features, which are encoded by a BiGRU network layer. Following this, the CRF inference layer is utilized to determine the most probable label sequence. Finally, a rule-based dictionary is used for correction.

3.2.1. Rule Dictionary

In patent documents, clearly named entities can be identified through regular expressions and artificially constructed dictionaries, the recognition results can be incorporated into dictionaries, the external knowledge base can be expanded, and the model output can be corrected. In addition, there are many English proper nouns in Chinese patents, which

need to be converted regularly, such as converting the word "Android" to "A7d", where "A" and "d" represent the first and last letters of the word, respectively, and "7" represents the length of the word, to avoid semantic confusion and loss of contextual information. At the same time, a dictionary containing commonly used ending words such as "system", "software" and "component" can be built to limit the range of neural network output results and improve the accuracy of tasks.

3.2.2. Vector Initialization

To process textual information for neural networks, patent data should first be converted into vector form. Traditional vector generation methods are insufficient for capturing the semantic information of the text. To better extract the contextual features of patent text, the BERT model is used to enhance the semantic representation of word vectors.

The Transformer encoding unit is the core of the BERT model architecture, comprising character representation, sentence representation, and positional representation. It is used to construct a vector matrix for input text. In the computation process, the Self-Attention mechanism is central to the Transformer. It begins by calculating the similarity between the feature vector Q and the context feature vector K, resulting in an attention weight vector. This vector is then utilized to enhance the feature representation in vector V. Feature expression in the calculation formula is shown in Equations (5) and (6):

$$\text{attention\_output} = \text{Attention}(Q, K, V) \tag{5}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

3.2.3. Hole Convolution Neural Network

In response to the issue of semantic confusion arising from the multimodal information in patent documents, we introduced a dilated convolutional neural network into the model, as illustrated in Figure 2. Compared to traditional convolutional neural networks, dilated convolutional neural networks can achieve a larger receptive field without relying on pooling operations. This approach mitigates the problem of internal data loss that typically occurs during convolution.

(a) Ordinary convolution   (b) Dilated convolution
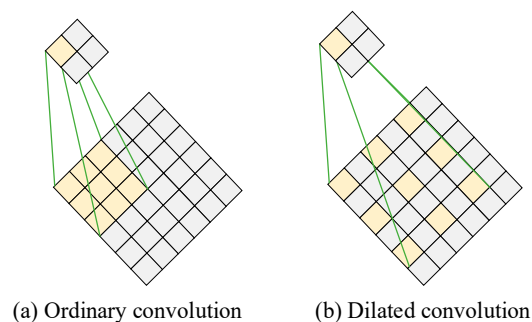
Figure 2. Convolution processes of convolutional neural networks and hole convolutional neural networks.

As shown in Figure 2, dilated convolution introduces a dilation rate d to the standard convolution, allowing it to capture information from a wider range of inputs without sliding over contiguous regions as traditional CNNs do. This results in an increased receptive field without requiring an increase in the size of the convolutional kernel.

3.2.4. BiGRU Network Layer

The BiGRU network includes both forward GRU and backward GRU networks. Simultaneously, leveraging the long short-term memory (LSTM) network, the feature extraction

and mapping processes are merged to address issues like gradient explosion and text dependencies, ensuring the effective retention of essential information.

As shown in Figure 3, the utilization of update gate and reset gate controls the reading and writing of information in the neural units.
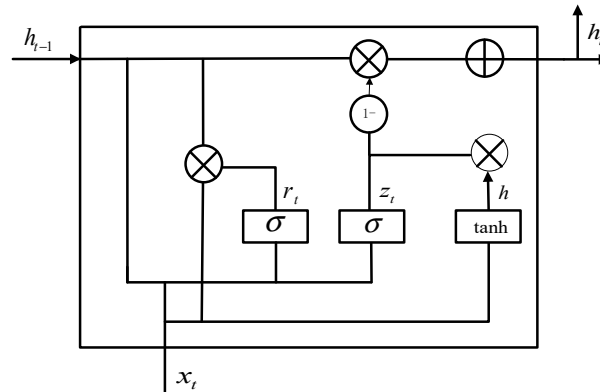

Figure 3. GRU Unit Structure.

The update gate influences the current stage's state based on the previous stage, while the reset gate represents the condition of the previous stage's content being written into the current content. The specific calculation formulas are as follows, as shown in Equations (7)–(11):

$$z_t = \sigma(W_z \times [h_{t-1}, x_t]) \tag{7}$$

$$r_t = \sigma(W_r \times [h_{t-1}, x_t]) \tag{8}$$

$$\widetilde{h}_t = \tanh(W_h \times [r_t \times h_{t-1}, x_t]) \tag{9}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widetilde{h}_t \tag{10}$$

$$y_t = \sigma(W_o \times h_t) \tag{11}$$

Here, h_t represents the activation value of the current network unit, y_t denotes the output at time t, and σ and tanh represent activation functions. w_z, w_r, and w_o, respectively, denote the update gate, reset gate, and continuously updated output parameters in the hidden layer. $\widetilde{h}_t$ represents, at time t, the activation value of the current network unit, controlling r_t, the next network unit h_{t-1}, and the current time step input x_t. To capture more contextual semantic information, reverse network learning is introduced into the text data by merging the forward GRU and backward GRU.

3.2.5. CRF Inference Layer

To mine the information about the association between patent document entities, a CRF reasoning layer is added behind the BiGRU layer to accurately identify patent document entities.

$x = (x_1, x_2, \ldots, x_n)$ is the character-level representation of the input patent document data, x_n denotes the input vector of the nth word. $y = (y_1, y_2, \ldots, y_n)$ is the tag sequence of x, all the tags corresponding to characters have corresponding scores, which are linearly increasing with the possibility of output results. From the whole reasoning process, the transfer phenomenon will appear between the front and back tags with a certain probability, and the comprehensive result is obtained by adding the tag score and the transfer score. The specific calculation methods are shown in Equations (12) and (13):

$$S(x, y) = \sum_{i=1}^{n} \left(W_{y_i-1, y_i} + P_{i, y_i} \right) \tag{12}$$

$$P_i = W_s + h^{(t)} + b_s \tag{13}$$

where W_s represents the transformation matrix, b_s represents the bias term in the computation, W_{y_i-1,y_i} represents the movement score of the data label, $h^{(t)}$ represents the hidden state of input x_t of the previous layer at time t, and P_{i,y_i} denotes the value of the y_i.

3.3. Patent Document Entity Relationship Extraction Method Integrating Attention Mechanism

To address the limited context information in extracting entity relationships from patent documents, we propose a method that combines attention mechanisms, utilizing pre-trained models for effective expression of intrinsic semantic features in sentence sequences and attention mechanisms for efficient extraction of entity information, dubbed as MPreA. The modeling formula of the relationship extraction task is shown in Equation (14). First, the head entity s extract, and then use the relationship between the main entity and the head entity to extract the tail entity o. Finally, the relationship between entities r is obtained.

$$P(s,r,o) = P(s)P(o|s)P(r|s,o) \tag{14}$$

The overall modeling formula for the feature extraction method in patent documents is:

$$P((s,r,o)|x) = P(s|x)P((r,o)|s,x) = P(s|x)\prod_{r\in T}P_r(o|s,x) \tag{15}$$

where T represents the entity relation type and x represents the network input. The specific steps of using Equation (15) to convert the patent document entity extraction task to a pointer annotation task, to realize the extraction of multiple inter-entity relationships, are as follows: first detect the head entity in the sentence, then use the detected head entity to search for the corresponding tail entity and inter-entity relationships, and finally make a judgment on the search results, if there is a corresponding tail entity, then keep or discard the head entity Finally, it traverses the all of the data and obtains all the inter-entity relationships.

The overall architecture of the proposed method is shown in Figure 4, including a RoBERTa coding layer, feature enhancement layer, head entity marking layer, and a head entity features fusion, relationship, and tail entity marker.
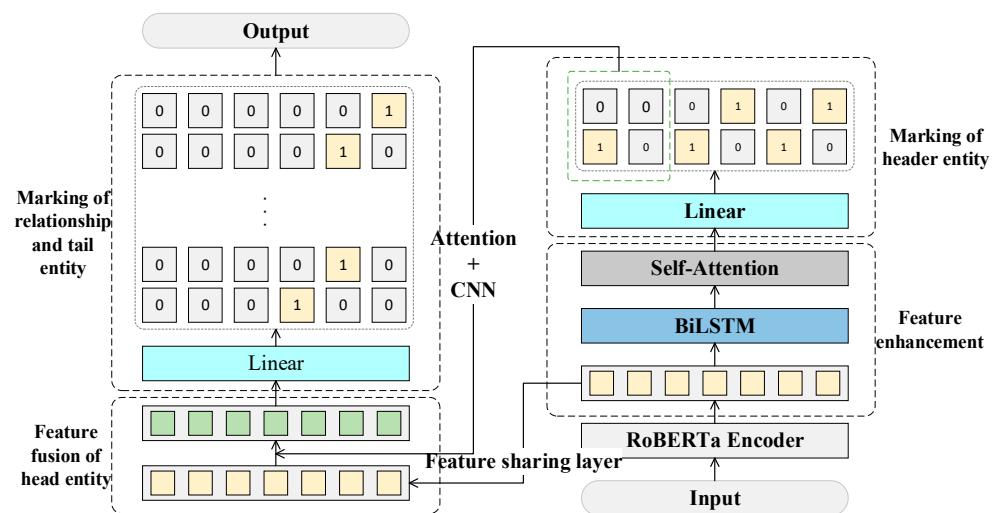

Figure 4. Entity Relationship Extraction Model Architecture, RoBERTa, encoding layer.

The RoBERTa model utilizes a large database as training material, allowing for the effective extraction of latent information from sentences and the acquisition of more contextual information through a multi-layer bidirectional Transformer architecture. Due to its excellent performance in numerous text processing tasks, the RoBERTa model has gained widespread adoption.

(1) Characteristic reinforcement layer

Bidirectional long short-term memory (BiLSTM) can realize forward and backward coding and can capture more feature information and context information through this mechanism to promote the extraction of text features. This module takes the feature vector encoded from the RoBERTa model as the input and conducts in-depth patent document information mining through the BiLSTM network.

In terms of the specific process, firstly, the processed vector matrix X is taken as the input, and the features are further encoded by the BiLSTM network. Then, the output of the previous moment and the word vector of the patent document are taken as the input of the current time t, and the input of the next time h_t is combined with the bidirectional encoding. The above formula is shown in Equation (16):

$$h_t = \mathrm{BiLSTM}(x_i, h_{t-1}) \tag{16}$$

Among them x_t expresses t which is the word vector input at any time. The vector obtained after the BiLSTM network coding is $H = \{h_1, h_2, \ldots, h_n\}$.

Although the BiLSTM network can better capture the features of longer text information, it has the defect of information loss when processing reverse semantic information and cannot accurately express text information. To solve this problem, the self-attention mechanism is used to allocate weights to help the model extract key features. The calculation formula is shown in Equation (17).

$$A(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

where Q, K, and V denote the query matrix, the key matrix, and the value matrix, respectively, and $\sqrt{d_k}$ is the square root of the first dimension of the key matrix, which is used to maintain the stability of the gradient.

Without increasing the computational complexity of the model, the attention mechanism can solve the defects of bidirectional long-short memory network in the long-distance feature information loss. Based on the advantages of the self-attention mechanism, this paper adds the self-attention mechanism module based on bidirectional long -short-term memory network, strengthens the correlation feature extraction between the data through the contextual information, and carries out the weight allocation. The specific steps are Q, K, V by means of the parameter matrix W_Q, W_K, W_V. A linear transformation is performed and then the attention size is calculated as shown in Equation (18).

$$M = \mathrm{Attention}(QW_Q, KW_K, VW_V)W^O \tag{18}$$

Here, $M = \{m_1, m_2, \ldots, m_n\}$ represents the feature vector after processing from the attention mechanism module.

(2) Header entity labeling layer

The header entity tagger consists of two identical binary classifiers, each independently processing positions using 0 and 1 for encoding. This enables the detection of the positions of header entities and the decoding of the output vectors from the previous layer, achieving accurate identification of header entities in patent document data.

(3) Head entity feature fusion

Firstly, the feature expressions between entities are obtained X_{head}, which are fed into a multilayer convolutional neural network to obtain the finally result x_{head}. The calculation process is shown in Equation (19).

$$x_{\mathrm{head}} = \mathrm{MaxPooling}(\mathrm{CNN}(X_{\mathrm{head}})) \tag{19}$$

Aimed towards the head entity features in the tail entity labeling task affected by the current text position, the feature fusion method based on the attention mechanism is adopted to achieve efficient tail entity labeling. The specific formula is shown in Equation (20).

$$T_i = \left[X_i; \left(X_i^T x_{\text{head}} \right) x_{\text{head}} \right] \tag{20}$$

After processing the vectors using the encoding X and X_{head} based on the dot product calculation, the coefficients in the result of the dot product operation in the previous step are multiplied with the corresponding head entity feature vectors, and then the result of the multiplication operation is combined with the word vectors at the current position to obtain the final result $T = \{T_1, T_2, \dots, T_n\}$.

(4) Relationship and Tail Entity Tagger

Patent inter-entity relationships and tail entities are processed through a multilayer binary classifier with the same number of layers as the number of predefined relationships. In the process of processing tail entities, the vector combines the information about the head entity T input with the labeler. In the case of the vector T, when decoding, the tagger will simultaneously tag the corresponding tail entity for each detected head entity.

$$P_i^{o\_start} = \sigma(W_{start}^o T_i + b_{start}^o) \tag{21}$$

$$P_i^{o\_end} = \sigma(W_{end}^o T_i + b_{end}^o) \tag{22}$$

T_i is the vector representation of the ith word's encoding vector after feature fusion. $P_i^{o\_start}$ and $P_i^{o\_end}$ express the first i the output values of the fusion vectors processed through the decoding layer, both of which are probabilistic values, of the $W(.)$ denotes the weight matrix, the $b(.)$ indicates the bias value, the σ represents sigmoid activation function, and o indicates the tail entity.

4. Experiments

4.1. Datasets and Implementation Details

Shanghai, as China's economic hub and a major manufacturing center, possesses a wealth of high-quality data on smart manufacturing technologies and patents. The data sourced from official patent databases and industry reports are comprehensive and accurate, ensuring the reliability and representativeness of the research. In this paper, two datasets from an intelligent manufacturing patent database in the Shanghai region are used for experiments, as shown in Table 1. Both datasets contain patent data in Chinese and English. Dataset 1 contains 67,071 sets of data, of which 61,530 sets are used as training data and 5541 sets are used as testing data. Dataset 2 contains 15,958 sets of data, of which 15,229 sets are used as training data and 729 sets are used as testing data.

Table 1. Dataset statistics.

Dataset	Training Set	Test Set
Dataset 1	61,530	5541
Dataset 2	15,229	729

The experiment uses the Windows11 operating system, and the processor, graphics card, programming language and optimizer are Intel i7-10700K, Nvidia GTX3060, python3.8, and Adam; the batch size was set to six, and the learning rate was set to 1×10^{-5}.

The accuracy (P), the recall rate (R) and the harmonic mean ($F1$) were used as the evaluation index of the algorithm, which is calculated as shown in Equations (23)–(25).

$$p = \frac{\text{Correctly identify the number of triples}}{\text{The number of identified triples}} \times 100\% \tag{23}$$

$$R = \frac{\text{The number of correctly identified triples}}{\text{The number of triples in the sample}} \times 100\% \tag{24}$$

$$F1 = \frac{2 \times P \times P}{P + R} \times 100\% \tag{25}$$

4.2. Comparative and Ablation Experiments

In order to verify the superiority of the proposed model, five methods were selected for comparative experimental analysis, and all models used the same database data as the input. At the same time, to verify the influence of different pre-training models on the extraction of patent entity relationships, this paper selected another three kinds of comparative experiments to test the training model. MPreA (BERT) presents the substitution of the pre-training model with BERT. MPreA (ALBERT) denotes the transition of the pre-training model to ALBERT. MPreA (ELECTRA) represents the substitution of the pre-training model with ELECTRA. The experiment uses the datasets mentioned in Section 4.1. The results of the comparison experiments are shown in Table 2.

Table 2. Experimental results of different models on two datasets.

Models	Dataset 1			Dataset 2		
	Accuracy/%	Recall Rate/%	*F1/%*	Accuracy/%	Recall Rate/%	*F1/%*
CopyRE	62	56.5	58.6	37.7	36.4	37.1
GraphRel	63.9	60.0	61.9	44.7	41.1	42.9
CopyRRL	77.8	68.1	72.1	63.3	59.9	61.6
ETL-Span	85.3	72.3	78.0	84.3	82.0	83.1
CasRel	88.7	88.2	89.5	93.4	90.1	91.8
MPreA	90.2	93.1	92.1	**93.8**	91.9	92.8
MPreA (BERT)	91.3	91.2	91.1	93.4	91.3	92.3
MPreA (ALBERT)	91.9	91.7	91.5	93.2	91.5	92.4
MPreA (ELECTRA)	**92.5**	**93.5**	**92.4**	93.6	**92.2**	**92.9**

Note: The optimal results are shown in bold.

To verify the effect of the attention mechanism in the network proposed in this paper, two groups of control experiments were set up to analyze the algorithm experiments carried out on the bidirectional long and short memory networks combined with a self-attention mechanism and the convolutional neural network combined with a self-attention mechanism. The experiment uses the datasets mentioned in Section 4.1. The experimental results are shown in Table 3.

Table 3. Entity relationship extraction model ablation experiments.

Model	Dataset 1			Dataset 2		
	Accuracy/%	Recall Rate/%	*F1/%*	Accuracy/%	Recall Rate/%	*F1/%*
Model-a	91.3	92.4	92.1	93.0	92.2	91.3
Model-b	92.3	89.1	89.3	93.1	90.6	90.4
Model-c	91.6	92.5	92.8	93.8	91.9	93.1

(1) Model-a removed the bidirectional long and short memory networks combined with a self-attention module from the MPreA basic model.

(2) Model-b removed the convolutional neural network combined with a self-attention module from the MPreA basic model.

(3) Model-c is the basic model of MPreA.

4.3. Validation

To validate the effectiveness of the proposed method on other patent datasets, we selected patent databases from the Yangtze River Delta region of China as the source for dataset 3. This region is one of the most economically developed and technologically advanced areas in China, boasting a wealth of high-quality patent databases. Dataset 3 contains patent data in Chinese and English. And dataset 3 contains 23,464 sets of data, of which 21,933 sets are used as training data and 1531 sets are used as testing data.

5. Results and Analysis

5.1. Experiment Results

From the results in Table 2, it can be seen that, in comparison with other patent entity relationship extraction methods, optimal results for the three evaluation indicators used for the method proposed in this paper were achieved for both datasets. CopyRE was selected as the benchmark data. For dataset 1, the method proposed in this paper had the following characteristics: the values of the accuracy, recall, and $F1$ indexes were improved 28.2, 36.6 and 33.5; in dataset 2, the index values were improved 56.1, 55.5 and 55.7, which proves that the patent document entity relationship extraction method proposed in this paper has a high precision and accuracy. When the pre-training model was replaced with BERT, the overall performance declined. This decline can be attributed to RoBERTa being trained on a larger dataset, allowing it to capture more entity relationship information. Similarly, when ALBERT was used as the pre-training model, there was also a decrease in performance, possibly due to ALBERT's reduced number of parameters, which limits its ability to learn sufficient key features. MPreA (ELECTRA) achieved a better performance on both datasets, possibly because ELECTRA employs a replaced token detection mechanism that enhances the model's learning capability. However, MPreA outperformed MPreA (ELECTRA) on dataset 2, which has a smaller data size, indicating that the proposed method can perform well even with limited data.

As shown in Figure 5, the comparative methods exhibited a better accuracy on dataset 1 than on dataset 2. However, the proposed method achieves promising accuracy scores on both datasets, with the best performance observed on dataset 2. This improvement is likely due to the use of pre-training models and attention mechanisms, which enhance the model's feature learning capabilities.
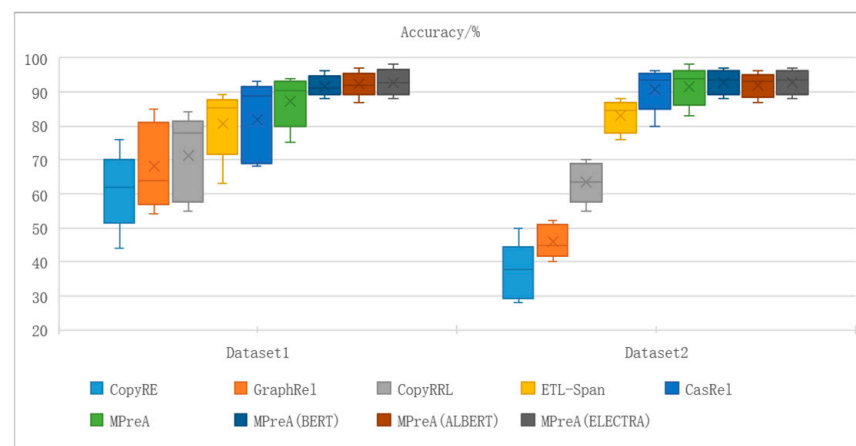

Figure 5. Box plots of accuracy of different models on two datasets.

As shown in Figure 6, both the comparative methods and the proposed methods have better recall performance on dataset 1 compared to dataset 2. The proposed methods achieve recall scores exceeding 90% on both datasets. This discrepancy may be attributed to dataset 1 containing a greater diversity of entity relationship types, which enhances the model's robustness.
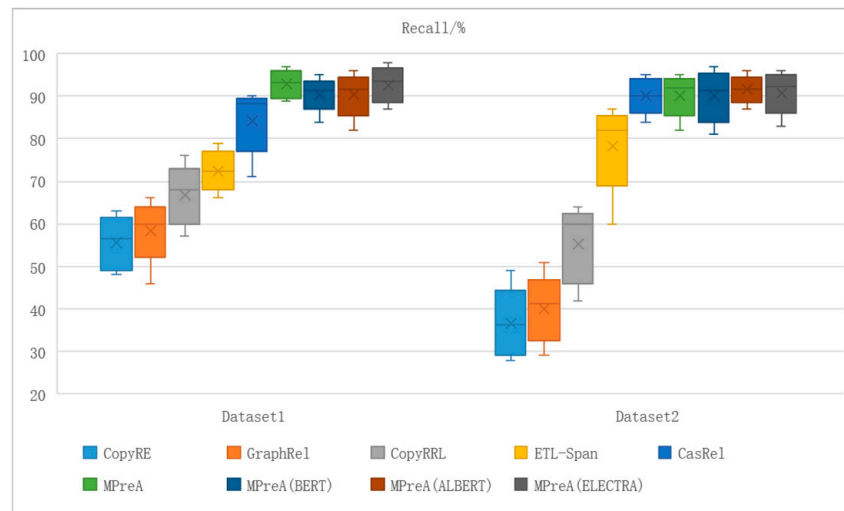
Figure 6. Box plots of recall for different models on two datasets.

As shown in Figure 7, the proposed method performs better on dataset 2 than on dataset 1 in terms of F1 score. This improvement could be attributed to dataset 2 containing less noise and fewer irrelevant features, thereby enhancing the model's recognition capability.
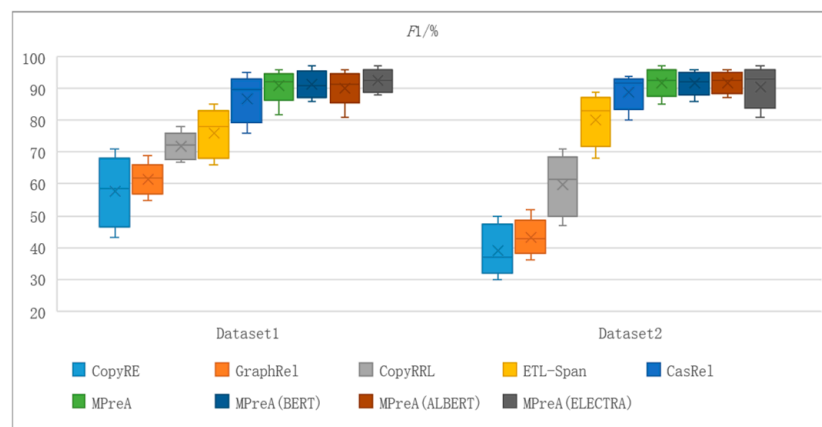
Figure 7. Box diagram of values of different models on different datasets.

From the results of ablation experiments, it can be observed that the inclusion of two self-attention modules enhances the model's ability to process data. The quasi-evaluation indexes of model-b in dataset 2 have decreased, indicating that the convolutional neural network combined with the attention mechanism module can effectively filter out the redundant information, and more accurately extract the entity relationships in text data. The $F1$ value of model-a is higher than that of model-b on both datasets, indicating that the fusion of bidirectional long and short network and attention mechanism strengthens the extraction of hidden information in the input vector, obtains more detailed features, and improves the recognition accuracy.

5.2. Validity Results

From the results of Table 4, it is clear that the proposed method achieves the promising performance in dataset 3. This indicates that the proposed approach exhibits a strong robustness and can be effectively applied to other patent document data processing tasks.

Table 4. Experimental results of the proposed models on dataset 3.

Models	Dataset 3		
	Accuracy/%	Recall Rate/%	*F1/%*
MPreA	92.9	91.2	93.7

6. Conclusions

To accurately extract entity relationships from semi-structured patent documents, this paper proposes a method that integrates pre-training models and attention mechanisms, building upon ontology model construction and patent entity recognition. More specifically, the pre-trained model was used to process the patent literature data to obtain a feature vector containing entity and relationship information. The accuracy of the proposed method was improved by extracting key features using a bidirectional long short-term memory network and attention mechanism module. Before the tail entity and the relationship information between entities are encoded and fused, the head entity vector and the input vector are fused with the features by using the convolutional neural network and the attention mechanism module to further improve the accuracy of the model. The experimental results show that the proposed method achieves a promising performance on dataset 1, dataset 2, and dataset 3. The accurate extraction and analysis of entity relationships in patent texts enable researchers to gain a comprehensive understanding of the technical content and innovative aspects of patents, thereby facilitating a better assessment of their technological contributions and market value. This has significant implications for the formulation of future development strategies and risk management for enterprises. Meanwhile, our method can be applied to other semi-structured document information analysis tasks, such as legal texts and medical records. However, we did not consider non-textual data in patent documents, such as flowcharts and structural diagrams, which are important for text analysis tasks. In future work, we intend to incorporate multimodal data from patent texts to further enhance the efficiency of patent analysis.

Author Contributions: Conceptualization, L.Z.; Software, X.S.; Validation, L.Z. and X.S.; Formal analysis, L.Z. and X.S.; Data curation, K.H.; Writing—original draft, L.Z.; Supervision, X.M. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data that support the findings of this study can be accessed upon reasonable request from the corresponding author.

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

References

1. Pejic-Bach, M.; Pivar, J.; Krstić, Ž. Big data for prediction: Patent analysis—Patenting big data for prediction analysis. In *Big Data Governance and Perspectives in Knowledge Management*; IGI Global: Hershey, PA, USA, 2019; pp. 218–240.
2. Ma, K.; Tian, M.; Tan, Y.; Qiu, Q.; Xie, Z.; Huang, R. Ontology-based BERT model for automated information extraction from geological hazard reports. *J. Earth Sci.* **2023**, *34*, 1390–1405. [CrossRef]
3. Puccetti, G.; Giordano, V.; Spada, I.; Chiarello, F.; Fantoni, G. Technology identification from patent texts: A novel named entity recognition method. *Technol. Forecast. Soc. Chang.* **2023**, *186*, 122160. [CrossRef]
4. Yang, G.; Niu, S.; Dai, B.; Zhang, B.; Li, C.; Jiang, Y. Named entity recognition method of blockchain patent text based on deep learning. In Proceedings of the Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024), Qingdao, China, 21 February 2024; Volume 13181.
5. Bhattacharya, K.; Chakrabarti, A. A Knowledge Graph and Rule based Reasoning Method for Extracting SAPPhIRE Information from Text. *Proc. Des. Soc.* **2023**, *3*, 221–230. [CrossRef]
6. Trappey, A.J.C.; Liang, C.-P.; Lin, H.-J. Using machine learning language models to generate innovation knowledge graphs for patent mining. *Appl. Sci.* **2022**, *12*, 9818. [CrossRef]
7. Yang, Y.; Li, S. Entity Overlapping Relation Extracting Algorithm based on CNN and BERT. *IEEE Access* **2024**. [CrossRef]

8. Bai, T.; Guan, H.; Wang, S.; Wang, Y.; Huang, L. Traditional Chinese medicine entity relation extraction based on CNN with segment attention. *Neural Comput. Appl.* **2022**, *34*, 2739–2748. [CrossRef]

9. Shi, M.; Huang, J.; Li, C. Entity relationship extraction based on BLSTM model. In Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; IEEE: Piscataway, NJ, USA, 2019.

10. Wei, M.; Xu, Z.; Hu, J. Entity relationship extraction based on bi-LSTM and attention mechanism. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems, Chongqing, China, 28–30 May 2021.

11. Liu, Y.; Zuo, Q.; Wang, X.; Zong, T. Entity relationship extraction based on a multi-neural network cooperation model. *Appl. Sci.* **2023**, *13*, 6812. [CrossRef]

12. Qiao, B.; Zou, Z.; Huang, Y.; Fang, K.; Zhu, X.; Chen, Y. A joint model for entity and relation extraction based on BERT. *Neural Comput. Appl.* **2021**, *34*, 3471–3481. [CrossRef]

13. Fan, C. The Entity Relationship Extraction Method Using Improved RoBERTa and Multi-Task Learning. *Comput. Mater. Contin.* **2023**, *77*, 1719–1738. [CrossRef]

14. Lin, Y.; Ji, H.; Huang, F.; Wu, L. A joint neural model for information extraction with global features. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.

15. Nasar, Z.; Jaffry, S.W.; Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.* **2021**, *54*, 1–39. [CrossRef]

16. Miric, M.; Jia, N.; Kenneth, G. Huang. Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategy Manag. J.* **2023**, *44*, 491–519. [CrossRef]

17. Lin, H.; Yan, J.; Qu, M.; Ren, X. Learning dual retrieval module for semi-supervised relation extraction. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019.

18. Shang, Y.; Huang, H.Y.; Mao, X.L.; Sun, X.; Wei, W. Are noisy sentences useless for distant supervised relation extraction? In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34.

19. Hong, Y.; Li, J.; Feng, J.; Huang, C.; Li, Z.; Qu, J.; Xiao, Y.; Wang, W. Competition or cooperation? exploring unlabeled data via challenging minimax game for semi-supervised relation extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37.

20. Kambhatla, N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, Barcelona, Spain, 22 July 2004.

21. Shan, Z.; Liang, F. Extraction of STEM Knowledge Relationship in Physical Education Course Textbooks Based on KNN. In Proceedings of the 2023 IEEE 6th Eurasian Conference on Educational Innovation (ECEI), Singapore, 3–5 February 2023; IEEE: Piscataway, NJ, USA, 2023.

22. Hou, W.; Hong, L.; Xu, H.; Yin, W. RoRED: Bootstrapping labeling rule discovery for robust relation extraction. *Inf. Sci.* **2023**, *629*, 62–76. [CrossRef]

23. Li, P.; Mao, K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Syst. Appl.* **2019**, *115*, 512–523. [CrossRef]

24. 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.

25. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In Proceedings of the Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Greece, 3–7 June 2018; proceedings 15. Springer International Publishing: Cham, Switzerland, 2018.

26. Zhou, H.; Xu, Y.; Yao, W.; Liu, Z.; Lang, C.; Jiang, H. Global context-enhanced graph convolutional networks for document-level relation extraction. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020.

27. Zhen, Y.; Zheng, L.; Chen, P. Constructing knowledge graphs for online collaborative programming. *IEEE Access* **2021**, *9*, 117969–117980. [CrossRef]

28. Zhao, T.; Yan, Z.; Cao, Y.; Li, Z. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021.

29. Oliveira, L.; Claro, D.B.; Souza, M. DptOIE: A Portuguese open information extraction based on dependency analysis. *Artif. Intell. Rev.* **2023**, *56*, 7015–7046. [CrossRef]

30. Bhatia, P.; Celikkaya, B.; Khalilia, M.; Senthivel, S. Comprehend medical: A named entity recognition and relationship extraction web service. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; IEEE: Piscataway, NJ, USA, 2019.

31. Berahmand, K.; Haghani, S.; Rostami, M.; Li, Y. A new attributed graph clustering by using label propagation in complex networks. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1869–1883. [CrossRef]

32. Yuan, L.; Cai, Y.; Wang, J.; Li, Q. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11051–11059. [CrossRef]

33. Kamateri, E.; Stamatis, V.; Diamantaras, K.; Salampasis, M. Automated single-label patent classification using ensemble classifiers. In Proceedings of the 2022 14th International Conference on Machine Learning and Computing, Guangzhou, China, 18–21 February 2022.

34. Chen, Y.; Yang, W.; Wang, K.; Qin, Y.; Huang, R.; Zheng, Q. A neuralized feature engineering method for entity relation extraction. *Neural Netw.* **2021**, *141*, 249–260. [CrossRef] [PubMed]

35. Yan, Y.; Okazaki, N.; Matsuo, Y.; Yang, Z.; Ishizuka, M. Unsupervised relation extraction by mining wikipedia texts using information from the web. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009.