IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 December 2022, accepted 1 March 2023, date of publication 6 March 2023, date of current version 16 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253128

RESEARCH ARTICLE

Multiple POS Dependency-Aware Mixture of Experts for Frame Identification

ZHICHAO YAN[1], XUEFENG SU[1,2], QINGHUA CHAI[3], XIAOQI HAN[1], YUNXIAO ZHAO[1], AND RU LI[1,4]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[2]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology, Jinzhong, Shanxi 030609, China
[3]School of Foreign Language, Shanxi University, Taiyuan, Shanxi 030006, China
[4]Key Laboratory of Computation Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China

Corresponding author: Ru Li (liru@sxu.edu.cn)

This work was supported in part by the Key Research and Development Project of Shanxi Province under Grant 202102020101008, in part by the Basic Research Program of Shanxi Province under Grant 202203021211286, in part by the China–Singapore International Joint Laboratory for Language Intelligence under Grant 202204041101016, in part by the Foundation of Shanxi Vocational University of Engineering Science and Technology under Grant KJ202203, and in part by the Graduate Education Innovation Project of Shanxi Province under Grant 2022Y146.

ABSTRACT Frame identification, which is finding the exact evoked frame for a target word in a given sentence, is a fundamental and crucial prerequisite for frame semantic parsing. It is generally seen as a classification task for target words, whose contextual representations are usually obtained using a neural network like BERT as an encoder, and enriched with a joint learning model or the knowledge of FrameNet. However, the distinction at a fine-grained level, such as the delicate differences in the information of syntax and PropBank roles caused by different parts-of-speech (POS) of target words, is neglected. We propose a **M**ultiple **POS** **D**ependency-**a**ware **M**ixture of **E**xperts(**MPDaMoE**) network that integrates five types of information, consisting of the syntactic information of target words whose POS are nominal, adjectival, adverbial, or prepositional, and the PropBank role information of target words whose POS are only verbal. To better learn such information, a Mixture of Experts network is employed, in which every expert is a Graph Convolutional Network, to incorporate the different dependency information of target words. Our model outperforms state-of-the-art models in experiments on two benchmark datasets, which shows its effectiveness.

INDEX TERMS Frame identification, semantic roles, syntactic information, BERT, mixture of experts, graph convolutional network.

I. INTRODUCTION

Frame **I**dentification (**FI**) is the process of searching for and determining the frame evoked by a target word in a sentence, and Frame-Semantic Role Labeling (**FSRL**) is the process of identifying participants and assigning them role labels licensed by the frame [1], [2], [3]. FI is a major premise of FSRL, which has been widely used in machine reading comprehension [4], text summarization [5], relation extraction [6], and other natural language processing tasks, as it can describe frame scenarios in a fine-grained manner, and whose challenge lies in classifying the massive number of labels. To simplify this task and enhance model performance, we can use FI to map thousands of labels to a smaller set [7]. The challenge of FI is to identify the exact frame when the target word is polysemous. A frame, which is described in the FrameNet knowledge base [8], [9], schematically represents a real scenario in our world and includes semantic roles (specifically, core and non-core roles). As shown in Fig. 1, the word *turned* evokes the two frames **Cause_change** and **Becoming**. In Fig. 1 (a) the role labels of *the programs that*, *your investment* and *into result* are *Agent*, *Initial_category* and *Final_category*, respectively; in Fig. 1 (c), *the islands that* and *against it* correspond to respective roles *Entity*

The associate editor coordinating the review of this manuscript and approving it for publication was Marta Cimitile.

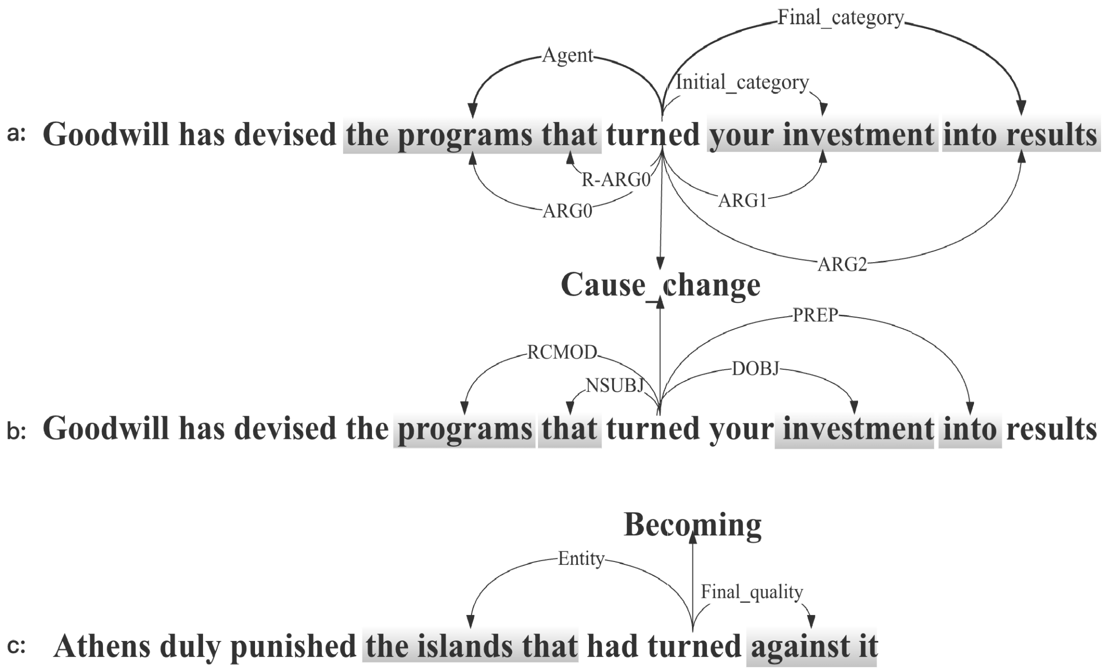
Final_category
Agent
Initial_category
a: Goodwill has devised the programs that turned your investment into results
R-ARG0
ARG0
ARG1
ARG2
Cause_change
PREP
RCMOD
DOBJ
NSUBJ
b: Goodwill has devised the programs that turned your investment into results
Becoming
Entity
Final_quality
c: Athens duly punished the islands that had turned against it

FIGURE 1. Three examples annotated with Frame-semantic (at the top of a and c), syntactic (b), and PropBank roles (at the bottom of a).

and *Final_quality*. The details of frames **Cause_change** and **Becoming** are shown in Table 1, which contains the frame definition (**Def**), Frame Elements (**FEs**) and Lexical Units (**LUs**). Due to space limitations, FEs and LUs are only partly shown.

FI is generally considered a token-level classification task, whose class labels are frame names, expressed as

$$f = \text{argmax}_{f_i \in F}\ p(f_i | r_t, C), \tag{1}$$

where f_i represents the i-th frame in frame database F, and C is the context that contains the target word r_t.

The deficiency of traditional machine learning methods [1], [10] lies in feature engineering, which usually learns the contextual representations of target words with manually-selected features. Although deep neural networks, especially large-scale pre-trained language models(PrLMs) like BERT [11], improve context-based representation by learning different knowledge in language, we are not sure about the extent of knowledge learned [12]. Swayamdipta et al. [13] and Peng et al. [14], respectively, used recurrent neural networks (RNN) and Long short-term memory network (LSTM) to learn the representation of target words, while Jiang and Riloff [15] used BERT to capture contextual features with frame definitions of target words. None of these methods have taken into account the variation in dependency information that arise due to differences in parts-of-speech (POS) of the target words.

To improve the performance of FI, it is essential to fuse the information of syntactic and PropBank roles, which is linguistic knowledge that reveals the relation between a target word and its semantic spans, and is conducive to enriching the representation of target words and disambiguating polysemy. For instance, in Fig. 1 (a) and (c), frame **Cause_change** and **Becoming** are evoked by *turned*, respectively. According to their definitions, **Cause_change** requires a semantic span acting as an **Agent** or a **Cause** in the context while **Becoming** does not. In Fig. 1 (a), the phrase *the programs that* semantically tends to express the meaning of an **Agent**, which guides the model to identify the target word evoking frame **Cause_change** rather than **Becoming**. Target words with various POSs may have different syntactic information, which undermines the effectiveness of a unified modeling method. Hence, we propose a **M**ultiple **POS** **D**ependency-**a**ware **M**ixture **o**f **E**xperts (**MPDaMoE**) Network, (MoE) layer in which we replace the traditional feedforward neural network with a Graph Convolutional Neural Network (GCN) [16]. The MoE layer incorporates multiple-POS linguistic knowledge in a unified module.

The main contributions of this paper are as follows:

- We propose a multiple POS dependency-aware Mixture of Experts network to capture the fine-grained dependency information of target words with different POSs and to alleviate the problem that a single GCN cannot well model this information;
- We observe that the dependency information of different POSs make different contributions to FI. We conduct extensive experiment to investigate the contributions of different information, and the results show

that the target words of adjectives make the greatest contribution;
- Extensive experiments on two FI benchmarks demonstrate that MPDaMoE outperforms state-of-the-art models.

The rest of this paper is organized as follows. The next section II reviews previous works related to this research. Section III gives a detailed description of MPDaMoE, which includes a dependency parse layer to get the dependency information of target words, a context encoder layer to gain the representation of the whole sentence, a Mixture of Experts layer to fuse dependency information into target words and a scoring layer to map textual features to labels. Section IV centers around the experimental setup, containing the analysis of the datasets, the baselines compared with our model, and parameter settings. Section V presents experiment results, ablation studies, and hyperparameter analysis. Section VI discusses the deficiency and future work. Finally, Section VII closes this research article by summarizing our research methods and pointing out our contribution.

II. RELATED WORK

A. FRAME IDENTIFICATION

The task of FI was first proposed in the Semeval-2007 [17]. Researchers initially approached the problem with machine learning methods, and Das et al. [1] proposed Semafor which employed a conditional log-linear model to calculate the probability of a frame, using manually-designed features like lemmatized target words, WordNet lexical-semantic relations and treatment of class labels as supervision signals. Johansson and Nugues [10] adopted an SVM classifier on all ambiguous words by utilizing information such as target-word dependency, lemmas and subcategorization frames.

The research of FI has begun to move toward distributed feature representation and neural networks, with three research trends. One is joint learning using distributed representation. Hartmann et al. [7] achieves higher scores for out-of-domain frame identification than previous systems via SimpleFrameId, which utilized SentBow (i.e., averaging the embeddings of all words in a sentence) after pointing out the sparsity of the feature space of Hermann's approach [18]. Hermann's system produced more interpretable output by taking the dependency information around the target word into a low-dimensional feature space and combining FI and FSRL with the WSABIE [19] algorithm. Another trend is to construct a multitask architecture to learn high-dimensional features of target words in context automatically. Swayamdipta et al. [13] proposed Open-Sesame, which employed bidirectional LSTM and dependency information to build a classification model, showing that syntax continues to be beneficial in frame-semantic parsing. Peng et al. [14] dealt with frame semantic parsing and semantic dependency parsing (FI is a subtask of frame semantic parsing) using LSTM to construct a joint learning model, which shows how joint learning and prediction can be done with scoring functions that explicitly relate spans and dependencies. The latest trend is to model the whole sentence via BERT and fuse external information to improve FI performance. Botschen et al. [20] proposed a multimodal method to capture image features to enrich text features, showing that accuracy can be improved by enriching the textual representations with multimodal ones for English FrameNet data. Jiang and Riloff [15] and Su et al. [21] employed some key information from the FrameNet knowledge base. Jiang proposed FIDO, which stitched together lemma, frame definition and context to acquire more information. The experimental results showed that this model performs better than previous systems on two versions of FrameNet data. Su came up with KGFI by constructing a frameGCN containing two subgraphs–FEsGCN (frame-elements-relation graph) and DefGCN (frame-definition graph via frame relations)– and mapping target words and frames into the same embedding space and gained the state-of-the-art performance of FI, demonstrating that all kinds of knowledge about frames are useful for FI task.

B. GRAPH CONVOLUTIONAL NEURAL NETWORKS

The GCN is the first neural network to transfer convolution operations to graph structures and can capture the structure and semantic features at the same time. The GCN has been adopted to learn dependency information. Li et al. [22] proposed DualGCN networks model which constructed a SynGCN and SemGGN utilizing syntactic and semantic information, respectively, in the aspect-based sentiment analysis task shown that DualGCN model outperforms baselines. Veyseh et al. [23] rethought the noise of dependency information and proposed a novel GCN-based gated mechanism for the same task. Zhang et al. [24] used joint learning for the tasks of semantic and opinion role labeling, using a GCN to share parameters. Narang et al. [25] and Goel and Sharma [26] used an LSTM and DepGCN to model the syntactic dependency information in the task of automated detection of abusive language. Hence, the GCN is well-suited to capture dependency features.

C. MIXTURE OF EXPERTS NETWORK

The Mixture of Experts network, first proposed in 1991 [27], it is based on the divide-and-conquer principle. Shazeer et al. [28] used the MoE, which was also adopted in recurrent neural networks, to increase the model parameters and proposed a gating network to select experts in training depending on inputs. Since different tasks require different skills, Liao et al. [29] and Zhang et al. [30] replaced the MLP in transformer with multiple experts to understand different knowledge. Due to the particularity of graph data, Zhou and Luo [31] explored the MoE in graph neural networks to solve the over-smoothing problem, and experiments showed that MoE has its greatest potential with very large datasets. In summary, MoE can be used to learn diverse knowledge and improve the model's awareness of fine-grained knowledge.

TABLE 1. Structure of frame **Cause_change** and **Becoming** in FrameNet knowledge base.

Frame Name	Cause_change	Becoming
Def	*An Agent or Cause causes an Entity to change, either in its category membership or in terms of the value of an Attribute. In the former case, an Initial_category and a Final_category may be expressed. In the latter case, an Initial_value and a Final_value can be specified.*	*An Entity ends up with some Final_quality–a new fact about the Entity. Alternatively, based on a cluster of changes of characteristics, the Entity newly meets the conditions for being a member of a Final_category.*
FEs(roles)	Core:*Agent, Entity, Final_category...* Non-Core:*Containing _event,Degree...*	Core:*Entity,Final_category...* Non-Core:*Circumstances,Explanation...*
LUs	assemble.v create.v generate.v **produce.v**...	become.v **turn.v...**

In short, the previous work concerning FI, with the assistance of supervision signals, has led to great progress. However, they still have the following two shortcomings: (1) Traditional machine learning approaches fed manually designed features into a traditional machine learning model to learn a classifier, the main drawback of which is that they rely on hand-designed features too much, which further leads to the failure of capturing high-dimensional features in the text; (2) Although some studies utilized the information of syntactic dependency via graph neural networks, they neglected the particularity of multiple POS dependency information and such particularity are difficult to capture with a uniform model. To overcome these two deficiencies, we utilize GCN to automatically learn the high-dimensional dependency features of target words to relieve the inadequacy of feature selection caused by hand-designed features and employ the MoE network to alleviate the problem that the single model cannot integrate the multi-part-of-speech dependency features. To sum up, this paper explores the impact of different dependency information and POSs on the FI task in a fine-grained manner.

III. METHODOLOGY

To enrich the representation of target words, we propose an **MPDaMoE** Network based on multiple POS dependency information. The network consists of four layers (Fig. 2): **dependency parse**, **context encoder**, **Mixture of Experts** and **scoring**.

The dependency parse layer parses the dependency relations between target words and other semantic spans, and outputs an adjacency matrix for the context encoder layer and a position list with spans for the mixture of experts layer. The context encoder layer models the context via BERT and extracts the corresponding embeddings according to the position of a target word and its semantic spans' positions. The Mixture of Experts layer contains a number of GCNs (the number of experts is a hyperparameter) and a gating layer allocating different expert modules according to the embeddings from the context encoder layer and the adjacency matrix. The score layer calculates the probability of frame and the loss using a Focal Loss (FL) [32] function.

A. DEPENDENCY PARSE LAYER

To get the target word's dependency information containing syntactic and PropBank roles (for verbs), we employ Allennlp [33], which can provide both dependency parser [34] and PropBank role labeling tools [35]. The dependency parser in Allennlp using the biaffine classifiers on top of a bidirectional LSTM achieves 95.57% and 94.44% UAS(unlabeled attachment score) and LAS(labeled attachment score) using gold POS tags. The PropBank role labeling tool is based on BERT with some modifications and is currently the state-of-the-art single model for English PropBank semantic role labeling task. The tool achieves 86.49% F1 on the Ontonotes 5.0 dataset.

1) ANALYZING DEPENDENCY INFORMATION

The information of frame roles is crucial to identify the frame because it is associated with its core elements(roles). Unfortunately, it cannot be used because it is unknown in FI. To solve this problem, we can draw on syntactic information and PropBank role information, which we collectively refer to as dependency information. It is also conducive to finding the semantic spans of target words, because semantic spans identified using syntactic information and those using PropBank role information have a full or partial overlapping relationship with the semantic spans of frame roles. For instance, the dependency information of the target word *turned* evokes the frame **Cause_change** in the sentence *Goodwill has devised the problems that turned your investment into result*, as shown in Fig. 1(a). The spans of *the programs that*, *your investment* and *into result* are the constituents of the roles **Agent**, **Initial_category** and **Final_category**. In Fig. 1(a), the PropBank role information of **ARG0**, **R-ARG0**, **ARG1** and **ARG2** is the same as the semantic spans of frame roles of the target word *turned*; in Fig. 1(b), the syntactic information of *turned* consists of **RCMOD**, **NSUBJ**, **PREP** and **DOBJ**, which partially match the semantic spans of frame roles. Therefore, the performance can be improved if the representations of target words and their dependency information are combined.

2) CONSTRUCTING DEPENDENCY GRAPHS

We regard the spans directly connected with a target word and provided by the dependency information as the nodes in the graph (the direction of relation is not taken into consideration). $A = (a_{ij})_{s\times s}$ denotes the adjacency matrix of a graph, where s is the number of spans, R denotes a relation set, which consists of the relations between the semantic spans and the

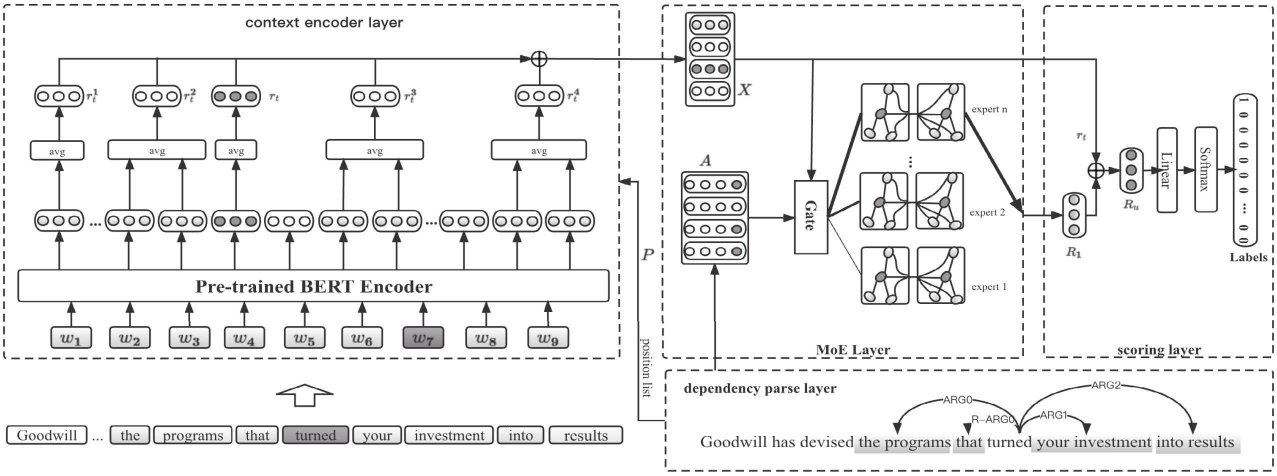

FIGURE 2. Overall architecture of MPDaMoE.

target word, and

$$a_{ij} = \begin{cases} 1 & w_i \in R \\ 0 & other. \end{cases} \tag{2}$$

$P = [[p_1, l_1], \ldots, [p_i, l_i], \ldots, [p_s, l_s]]$ denotes a list consisting of all span positions, which will be delivered to the context encoder layer. p_i and l_i denote the start and the end position, respectively, of a given span.

B. CONTEXT ENCODER LAYER

To get a target word's embedding that contains the contextual information, we employ BERT as an encoder, whose pre-training procedures is shown in Fig. 3. Sentences were input to the encoder with *[CLS]* signaling the beginning of the input, and with *[SEP]* discriminating different sentences. Specifically, *[CLS]* represents the embeddings of the input. BERT, preferring natural language understanding to language generation, is composed of the multiple encoders of bidirectional transformers [36] and based on two unsupervised tasks, pre-training objectives using a Masked Language Modeling and predicting the next sentence. The BERT model has inputs of word, segmentation, and position embeddings, which are used to calculate the output embedding. It has two versions: BERT-base and BERT-large, the first with 12 layers, 768 hidden sizes, 12 Heads and 110 million parameters, and the second with 24 layers, 1024 hidden sizes, 16 Heads and 340 million parameters.

The self-attention mechanism is the core of transformer, which aggregates and filters the information of the other words with the target words. Although the problem of long-distance dependence is alleviated, the self-attention mechanism sets a limit on BERT, for it cannot embed the length of a series longer than 512. As shown in Fig. 4, the structure of transformer contains multiple Multi-Head Attentions, each composed of multiple self-attentions, and each head of attention catches different information, which enriches the token embeddings with multidimensional characteristics. Self-attention and Multi-Head Attention can be expressed as

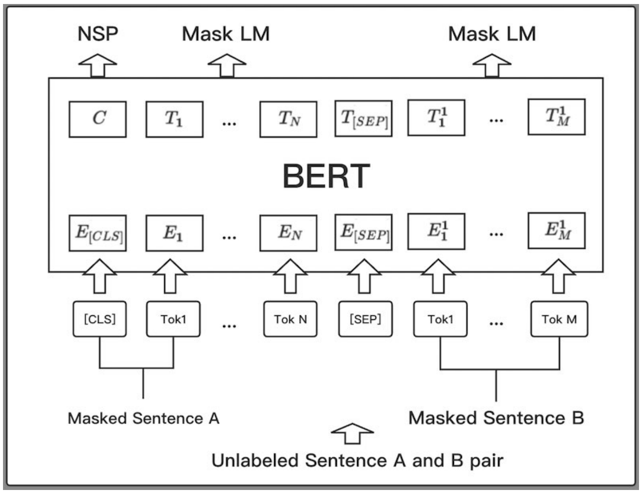

FIGURE 3. Overall pre-training procedures for BERT.

$$\text{Attention} = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}V) \tag{3}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \ldots head_h)W^O, \tag{4}$$

where Q, K and V can be calculated with the self-input X, and $head_i$ can be expressed as

$$Q = W_Q X \tag{5}$$

$$K = W_K X \tag{6}$$

$$V = W_V X \tag{7}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{8}$$

The list of span positions P is known in the dependency parse layer. The input sentence C, target word t, start position

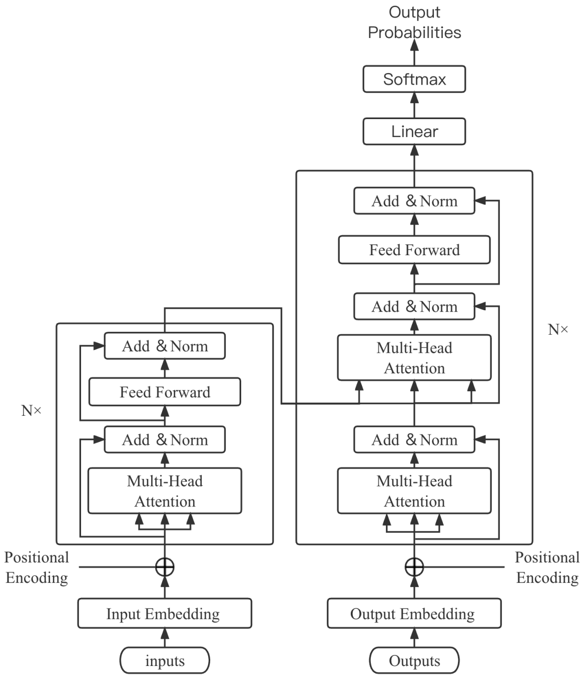

FIGURE 4. Transformer architecture, consisting of encoder and decoder partitions.

st and end position en of the target word are input to the encoder E_s, and the last layer of BERT output is H_t. Eventually, the embedding of target word r_t and the i-th spans of dependency information embedding r_{st}^i can be expressed as

$$r_t = E_s(C, t, st, en) = W_s^T H_t + b_s \tag{9}$$

$$r_{st}^i = E_s(C, t, P_i) = W_s^T H_{sk} + b_s. \tag{10}$$

where $W_s \in \mathcal{R}^{n \times m}$ and $b_s \in \mathcal{R}^m$ are input weight matrices to be learned. Because the target word and its dependency information of spans generally have more than one token, H_t and H_{st} can be calculated as

$$H_t = \frac{1}{en + 1 - st} \sum_{i=st}^{en} (H_t[i]) \tag{11}$$

$$H_{sk}^i = \frac{1}{P_i[1] + 1 - P_i[0]} \sum_{i=P_i[0]}^{P_i[1]} (H_t[i]), \tag{12}$$

where $P_i[1]$ and $P_i[0]$, respectively, represent the start and end positions of the i-th span of its dependency information. Eventually, we obtain the node feature matrix X, which consists of $r_{st}^1, \ldots, r_{st}^m, r_{st}$. X and r_t will be, respectively, delivered to the next and last layer.

C. MIXTURE OF EXPERTS LAYER

MoE was originally designed to increase model parameters under the hypothesis that datasets can be naturally divided into several subsets, which may come from different data (e.g., domains and topics) and may interfere with each other when dealing with a single model. It consists of a number of expert networks $E_1, \ldots, E_k$ allocated by a gating mechanism according to the weight of different information. Such a strategy transforms the structure of a multilayer network into modular construction. Structurally, MoE has two parts: (1) an experts network, in which each expert is a feedforward neural network used to learn different knowledge [28], [29], [30], [37]; and (2) a Gating Network that usually adopts the approaches of no-sparse gating [38] and noisy top-K gating [28]. We use no-sparse gating because it is superior [31] when the number of experts is smaller. The architecture of MoE is shown in Fig. 5.

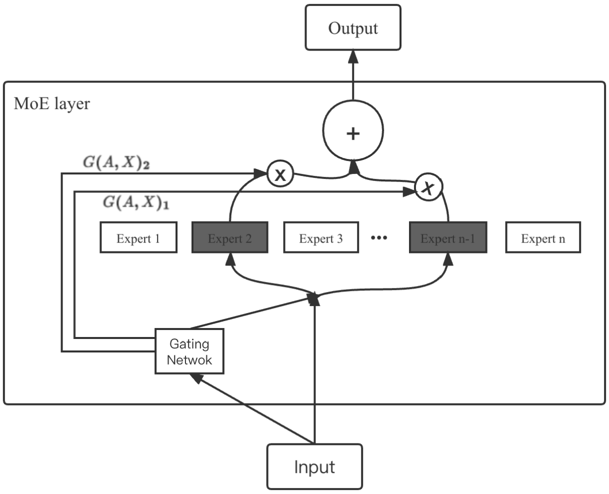

FIGURE 5. Detail of MoE layer, consisting of Gating network and number of experts.

Considering that our knowledge is dependency information, we construct it with a graph because it can better reflect the relations between the target word and its semantic spans. As Fig. 11 (b) shows, the sentence *Finally, Greece was ousted from its new territory in Asia Minor, which became part of the new Turkish state* has eight target words–*Finally*, *ousted*, *new*, *territory*, *in*, *became*, *Turkish* and *state*–evoking respective frames **Time_vector**, **Removing**, **Age**, **Political_locales**, **Interior_profile_relation**, **Becoming**, **Origin** and **Political_locales**. The information of the PropBank role of *ousted* and *became* is split into two partitions (Figs. 11 (a) and (c)) because the syntactic information is different from that of the PropBank role.

Actually, target words with different POS have different syntactic information. A target word with certain POS usually has multiple semantic spans, with different contributions to the target word, leading to different graph features. Considering verbs, the information of the PropBank role is used instead of syntax, because **ARG1** and **ARG2** are usually important to a target word. For nouns, **AMOD** and **DEP** are two core relations. For adjectives, syntactic positions of target words may have different types of dependency graphs. Specifically, some adjectives just have an **AMOD** relation to modify a noun, while others have multiple relations, such as **NSUBJ**, **DEP** and **AMOD**. Adverbs generally have

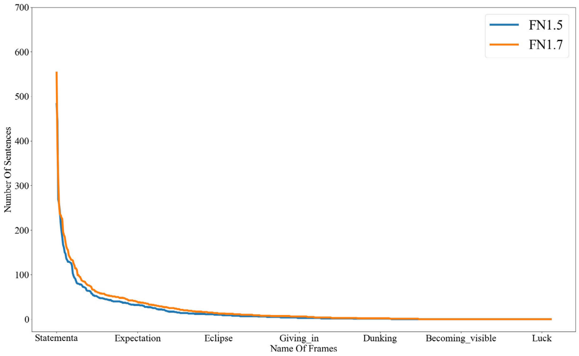

FIGURE 6. The distribution of frame classes in FN1.5 and FN1.7.

ADVMOD or **AMOD** relations with other spans. Prepositions have two important relations, **PREP** and **POBJ**, with other spans. Therefore, we focus on different dependency information of target words that may possess various POSs, which are employed to construct adjacency matrix A, and an MoE Network is used to model it.

A GCN is useful for extracting high-dimensional features from non-Euclidean space using convolution. It is described as a function map $f(.)$, fitted such that every node v_i aggregates its own features x_i and its neighbor's features x_j. We use a GCN acting as an expert, instead of the feedforward neural network adopted by the classical MoE, to fuse dependency information into target words. We construct the GCN with only one layer to aggregate the information of first-order neighbors because we only consider the direct relations between target words and their dependency information.

Finally, the MoE layer can be expressed as

$$y = \sum_{i=1}^{k} G(x)_i E_i(x_i), \tag{13}$$

where $G(x)_i$ is the weight of E_i. In each expert, the node features of the target word are calculated as

$$R_t = \text{ReLu}(\text{GCN}(A, X))[-1], \tag{14}$$

where –1 represents the last position of output embeddings.

D. SCORING LAYER

In this layer, the ultimate representation of a target word is concatenated with r_t and R_t, fusing both dependency and semantic information as

$$R_u = \text{Concat}(r_t, R_t). \tag{15}$$

Then R_u is fed into classification layer L, and softmax is used to calculate the probability of belonging to frame $f_j \in F$,

$$p_t = \text{softmax}(\text{L}(R_u)). \tag{16}$$

Finally, the position of max probability is regarded as the predicted class,

$$f = \text{argmax}(p_t). \tag{17}$$

TABLE 2. Statistics for FrameNet datasets.

Dataset	Train	Dev	Test	|F|
FN1.7	19391	2272	6714	1221
FN1.5	16621	2284	4428	1019

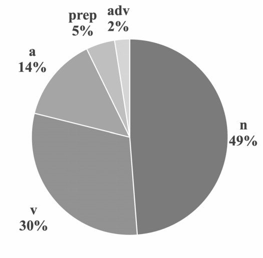

FIGURE 7. Proportion of five POSs in train datasets of FN1.7.

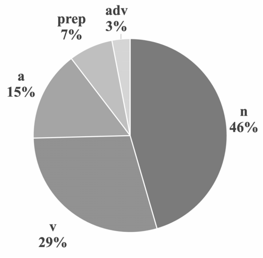

FIGURE 8. Proportion of five POSs in test datasets of FN1.7.

A long-tailed distribution of datasets is a common phenomenon, the statistical results of the number of every classification of FrameNet1.5 (FN1.5) and FrameNet1.7 (FN1.7) are shown in Fig. 6, and it shows the same distribution. Therefore, Focal Loss [32] replaces cross-entropy, and the loss is defined as

$$FL = -(1 - p_t)^{\gamma} \log(p_t), \tag{18}$$

where $\gamma \geq 0$ is a focusing parameter, and $1 - p_t$ is used to decrease the weight of the sample, which is easy to learn.

IV. EXPERIMENTAL SETTINGS

A. DATASETS

We employed two benchmark datasets, FN1.5 and FN1.7, to test our model's performance. Because each sentence may contain multiple target words in datasets, we consider it as multi-pairs between target words and the corresponding sentences. The number of Train, Development (Dev), Test and the number of frames |F| are shown in Table 2.

The distributions of target words with different POS (i.e., verbs, nouns, adjectives, adverbs and prepositions) in FN1.7 and FN1.5 (train and test) are illustrated in Figs. 7, 8, 9 and 10. This shows that the proportion of nouns is the largest, whether in FN1.7 or FN1.5, and the distributions of various POSs are relatively consistent in the train and test sets.

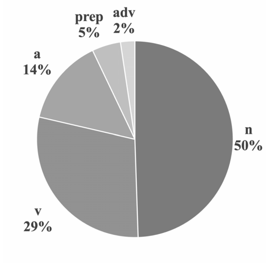

FIGURE 9. Proportion of five POSs in train datasets of FN1.5.

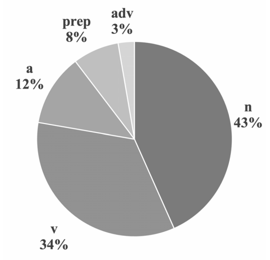

FIGURE 10. Proportion of five POSs in test datasets of FN1.5.

B. EVALUATION METRICS

We use accuracy,

$$\text{Acc} = \frac{1}{N'}\sum_{i=1}^{N'} I(y_i = \hat{f}(x_i)), \tag{19}$$

as the evaluation metric, where N' is the sum of samples in test datasets, f_i is the true label, x_i is the i-th sample and $\hat{f}(x_i)$ is the predicted result of the model for each sample.

C. BASELINES

We compared seven existing models with MPDaMoE. The first model, proposed by Das et al. [1], predicts target-word labels using a log-linear model based on statistical features. Hartmann et al. [7] average the word embeddings of the entire sentence and regard them as the target word representation. Swayamdipta et al. [13] construct a classifier with bidirectional LSTM. Peng et al. [14] use a joint model to mutually promote the performance for the model of FI and FSRL by adding frame-semantic role labels and frame labels. Jiang and Riloff [15] concat the frame definition and other messages of lexical units and fuse such extra information with the target word using BERT. Hermann et al. [18] use syntactic information of a target word in the multitask structure. Botschen et al. [20] map the picture and text information into the same feature space. Su et al. [21], the previous state-of-the-art model of the FI, incorporate frame knowledge using GCN-based methods and select an appropriate frame with an attention mechanism,

$$f = \text{argmax}_{f_i \in F_t}\ p(f_i | r_t, C), \tag{20}$$

where F_t represents the new set that has been changed and is a subset of F.

TABLE 3. Accuracy of frame identification with filtering in FN1.5 and FN1.7.

Model	FN1.7	FN1.5
Semafor	-	83.60
Hermann-14	-	88.41
SimpleFrameId	83.00	87.63
Open-Sesame	86.55	86.40
Peng	89.10	90.00
Botscehn	-	88.82
FIDO	92.10	91.30
KGFI(1-layers)	91.17	92.13
KGFI(2-layers)	92.40	91.91
MPDaMoE(BERT-base)	**93.44**	**93.29**
MPDaMoE(BERT-large)	**94.20**	**93.71**

TABLE 4. Accuracy of frame identification without filtering in FN1.5 and FN1.7.

Model	FN1.7	FN1.5
SimpleFrameId	76.10	77.49
Botscehn	-	81.21
KGFI(1-layers)	84.95	85.63
KGFI(2-layers)	85.81	85.00
MPDaMoE(BERT-base)	**86.73**	**86.12**
MPDaMoE(BERT-large)	**86.92**	**86.33**

D. PARAMETER SETTINGS

We set the learning rate of models to 5e-5, the batch_size to 64, the GCN module dimension to 128 and max_seq_length to 128. For the MoE layer, we set the number of experts to 6, and the dropout of the Gating Network to 0.3. For Focal Loss, we set γ to 2.

V. EVALUATION

A. OVERALL RESULTS

We tested our model in FN1.5 and FN1.7, and performed filtering with FrameNet knowledge during the testing process because the FrameNet knowledge base includes the mapping between target words and their candidate frames, which can change the search space from the whole set of frames to those evoked with the target word. Therefore, function (1) was modified to (20). The experimental results are shown in Tables 3 and 4.

MPDaMoE, FIDO [15] and KGFI [21] used PrLMs as the encoder. Obviously, models that adopted PrLMs were superior to other existing methods. Our MPDaMoE model performed best on the FN1.7 and FN1.5 datasets. Compared with the best state-of-the-art model, KGFI, our model with filtering achieves 1.04% and 1.16% absolute rising, and our model without filtering gains 0.92% and 0.49% improvements in two datasets. The model achieved 94.20% and 93.71% in terms of accuracy with filtering and 86.92% and 86.33% in terms of accuracy without filtering when BERT-base was replaced with BERT-large as the encoder. The experimental results show that dependency information is helpful for FI.

In terms of dependency information, previous work seldom focused on fusing this fine-grained information. We fuse

TABLE 5. Result of only incorporating dependency information of one POS.

Model	FN1.7
KGFI(2-layers)	85.81
MPDaMoE(all)	86.73
$w/\ n\_s$	86.10
$w/\ v\_s$	85.81
$w/\ v\_p$	85.95
$w/\ adv\_s$	86.27
$w/\ adj\_s$	86.47
$w/\ prep\_s$	86.32

just one type of dependency information in a POS into the corresponding target words, while other POS target words use their semantic embeddings from BERT. Because this method uses just one type of dependency information, we set the number of experts to 1. The experimental results of this manipulation in FN1.7 are shown in Table 5. *w/ v_s* and *w/ v_p*, respectively, denote only syntactic information or only PropBank role information of verbs is used, and *w/ n_s*, *w/ adv_s*, *w/ adj_s* and *w/ prep_s* represent that only syntactic information of these POSs is used. According to the results, we conclude that using only one type of dependency information promotes performance. It is noteworthy that *w/ adj_s* achieved 86.47% accuracy, 0.66% better than KGFI(2-layers), and *w/ v_s*, which is equal to the stronger baseline KGFI, enhanced our performance. Comparing the performance in *w/ v_s* and *w/ v_p*, we find that the PropBank role information is more valuable than syntactic information for target words with the POS of a verb.

B. ABLATION EXPERIMENTS

To test the impact of the dependency information of each POS on MPDaMoE, we successively removed one type of information. As shown in Table 6, w/o means "without". The results demonstrate that all five types of information, i.e., *noun, verb, adjective, adverb*, and *preposition*, are helpful for promoting the model's performance and have different weights for this task, especially in w/o adj_s, whose performance declines by 0.58%. In other words, this information is more important than other information.

In addition, we replaced FL with the traditional cross-entropy loss function, whose result shows the efficiency of FL for long-tailed distributions of datasets. Finally, we replaced the MoE structure with a single-layer GCN to test the impact of MoE, which gained 86.11% in terms of accuracy. The result shows that MoE is an effective method for incorporating the dependency information of different POSs, and the MoE layer incorporating the information of fine-grained multiple POSs performs better than a single GCN layer.

C. HYPERPARAMETER ANALYSIS

In our model, the core module is an MoE layer that sets a number of experts handling dependency information with different POSs and calculates the weight of every expert in the Gating Network using a dropout strategy to prevent overfitting. We analyze the number of experts and the dropout, with experimental results as shown in Fig. 12 and Fig. 13, based on which we conclude that the number of experts should be set according to the dataset. Because the FrameNet knowledge base has five major POSs, we set the number of experts to 6, and the model achieves the best performance. Another conclusion is that our model has the best generalization ability when setting the dropout ratio to 0.3.

In addition, we analyze the impact of γ in Focal Loss. When γ is set to 2, our model gains the best FI performance; experimental results are shown in Fig. 14.

TABLE 6. The results of the ablation experiments.

Model	FN1.7
KGFI(2-layers)	85.81
MPDaMoE(all)	86.73
w/o n_s	86.36
w/o v_p	86.30
w/o adv_s	86.23
w/o adj_s	86.15
w/o prep_s	86.18
w/o FL	86.27
w/o MoE	86.11

D. CASE STUDY

As shown in Fig. 11, each target word corresponds to different dependency information, which is modeled by different experts in the MoE layer. One of the experts models the PropBank role information of verbs. For instance, the spans of main roles are *its new territory in Asia Minor* and *part of the new Turkish state*, which both refer to countries, for **ARG1** and **ARG2**, respectively. This information, together with the meaning of target word *became*, guides the model to recognize the frame **Becoming** rather than **Suitability** (another frame evoked by **became** in the FrameNet knowledge base). Another expert models the graph of syntactic information to enhance representations of target words. For example, *state* can evoke the frames **Polical_locales**, **Thermodynamic_phase** and **Leadership** in different context, the word *state* has multiple meanings, such as: 1) *the territory, or one of the territories, of a government*; 2) *the condition of matter with respect to structure*; and 3) *a politically unified people occupying a definite territory*. The dependency information of *state* shows it is connected with *new Turkish*, whose meaning clear, which helps identify the meaning of *state*. However, if *new Turkish* is connected with 2) or 3), then nonconformity of the language expression occurs. It is such dependency information that guides our model to predict the frame **Political_locales** correctly.

VI. DISCUSSION

We use a trained dependency parsing model and a PropBank semantic role annotating tool to analyze the context of a target word to obtain dependency information. Although the two

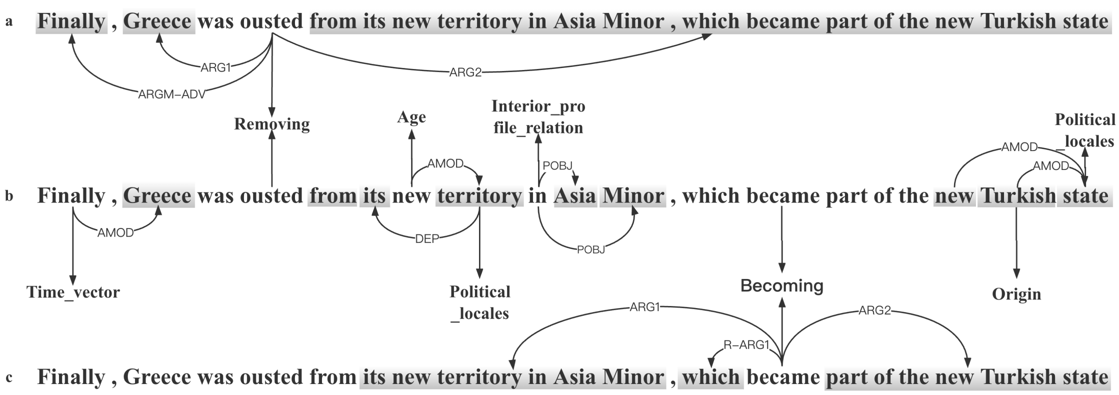
a
Finally , Greece was ousted from its new territory in Asia Minor , which became part of the new Turkish state
ARG1
ARGM–ADV
ARG2
Removing
Age
Interior_pro
file_relation
Political
_locales
AMOD
POBJ
AMOD
AMOD
b
Finally , Greece was ousted from its new territory in Asia Minor , which became part of the new Turkish state
AMOD
DEP
POBJ
Time_vector
Political
_locales
Becoming
Origin
ARG1
R–ARG1
ARG2
c
Finally , Greece was ousted from its new territory in Asia Minor , which became part of the new Turkish state

FIGURE 11. Dependency information of multiple POS in the same sentence, in which (a) and (c) show the PropBank role information, and (b) shows the syntactic information.

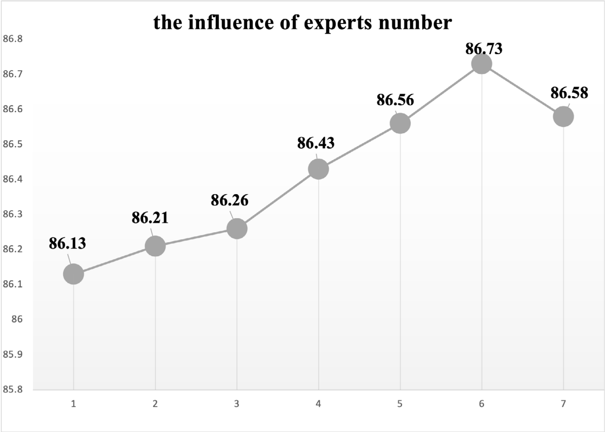

FIGURE 12. Influence of number of experts.

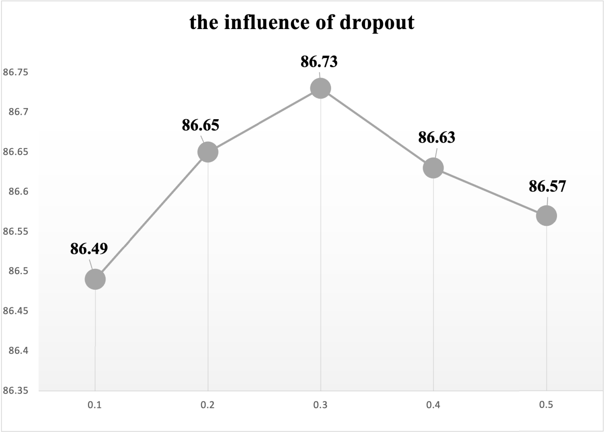

FIGURE 13. Influence of dropout.

models have achieved a good performance in the corresponding benchmark datasets, there are still two problems when we employ the two models: (1) Error probability in the two models may impact the downstream model; (2) FrameNet datasets may become cross-domain data for the two models, which probably decreases the accuracy of the analysis result.

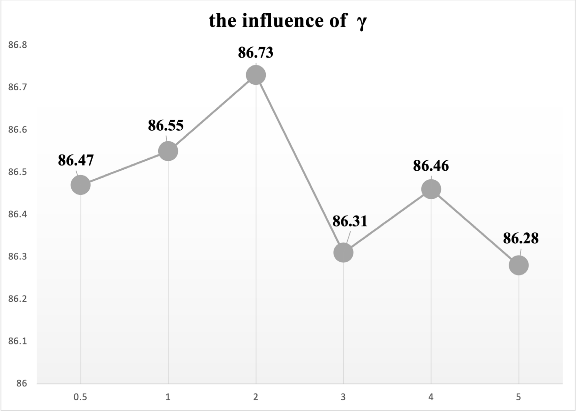

FIGURE 14. Influence of γ.

Both problems lead to error propagation, but when using this non-standard information, we still improve the performance of FI, which demonstrates the usefulness of dependency information for FI. Discussing the impact of such errors on FI will be our future work.

We consider the fine-grained dependency information of different POSs and integrate such fine-grained dependency information using the MoE network, which improves the model's performance. Specifically, for each and every target word, its dependency information may be represented by multiple syntactic dependency relations, such as NSUBJ, RCMOD and DOBJ, etc. In this paper, we assume all these relations are helpful for FI, so we adopt a weighting strategy with uniform distribution for different relations. In addition, We also believe different relations may have different importance for FI, and distinguishing the importance of different relations will be our future work.

VII. CONCLUSION

To solve the problem of incorporating dependency information of target words with different POSs, we proposed the MPDaMoE network, using a BERT encoder to obtain the embeddings of target words and its semantic spans via dependency information, where every expert allocates one type of information by a Gating Network, which is a Graph Convolutional Network to incorporate the different dependency information of target words and gain the embeddings of dependency features of target words. These two types of embeddings are concatenated as the ultimate embeddings of target words. In this way, we gain state-of-the-art FI performance. We compared the impact of PropBank role information and syntactic information of verbs and found that the information of PropBank roles is superior.

REFERENCES

[1] D. Das, D. Chen, A. F. T. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Comput. Linguistics*, vol. 40, no. 1, pp. 9–56, Mar. 2014.

[2] S. Swayamdipta, S. Thomson, K. Lee, L. Zettlemoyer, C. Dyer, and N. A. Smith, "Syntactic scaffolds for semantic structures," 2018, *arXiv:1808.10485*.

[3] A. Kalyanpur, O. Biran, T. Breloff, J. Chu-Carroll, A. Diertani, O. Rambow, and M. Sammons, "Open-domain frame semantic parsing using transformers," 2020, *arXiv:2010.10998*.

[4] S. Guo, Y. Guan, H. Tan, R. Li, and X. Li, "Frame-based neural network for machine reading comprehension," *Knowl.-Based Syst.*, vol. 219, May 2021, Art. no. 106889.

[5] Y. Guan, S. Guo, R. Li, X. Li, and H. Zhang, "Integrating semantic scenario and word relations for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2522–2529.

[6] H. Zhao, R. Li, X. Li, and H. Tan, "CFSRE: Context-aware based on frame-semantics for distantly supervised relation extraction," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106480.

[7] S. Hartmann, I. Kuznetsov, T. Martin, and I. Gurevych, "Out-of-domain FrameNet semantic role labeling," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 471–482.

[8] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *Proc. 17th Int. Conf. Comput. Linguistics*, 1998, pp. 1–5.

[9] C. F. Baker, "The structure of the FrameNet database," *Int. J. Lexicography*, vol. 16, no. 3, pp. 281–296, Sep. 2003.

[10] R. Johansson and P. Nugues, "LTH: Semantic structure extraction using nonprojective dependency trees," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 227–230.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[12] W. G. Lycan, *Philosophy of Language: A Contemporary Introduction*. Routledge, 2018.

[13] S. Swayamdipta, S. Thomson, C. Dyer, and N. A. Smith, "Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold," 2017, *arXiv:1706.09528*.

[14] H. Peng, S. Thomson, S. Swayamdipta, and N. A. Smith, "Learning joint semantic parsers from disjoint data," 2018, *arXiv:1804.05990*.

[15] T. Jiang and E. Riloff, "Exploiting definitions for frame identification," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 2429–2434.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[17] C. Baker, M. Ellsworth, and K. Erk, "SemEval'07 task 19: Frame semantic structure extraction," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 99–104.

[18] K. M. Hermann, D. Das, J. Weston, and K. Ganchev, "Semantic frame identification with distributed word representations," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1448–1458.

[19] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–7.

[20] T. Botschen, I. Gurevych, J.-C. Klie, H. M. Sergieh, and S. Roth, "Multimodal frame identification with multilingual evaluation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Language Technol., Volume*, 2018, pp. 1481–1491.

[21] X. Su, R. Li, X. Li, J. Z. Pan, H. Zhang, Q. Chai, and X. Han, "A knowledge-guided framework for frame identification," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5230–5240.

[22] R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, and E. Hovy, "Dual graph convolutional networks for aspect-based sentiment analysis," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6319–6329.

[23] A. P. B. Veyseh, N. Nour, F. Dernoncourt, Q. H. Tran, D. Dou, and T. H. Nguyen, "Improving aspect-based sentiment analysis with gated graph convolutional networks and syntax-based regulation," 2020, *arXiv:2010.13389*.

[24] B. Zhang, Y. Zhang, R. Wang, Z. Li, and M. Zhang, "Syntax-aware opinion role labeling with dependency graph convolutional networks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3249–3258. [Online]. Available: https://aclanthology.org/2020.acl-main.297

[25] K. Narang and C. Brew, "Abusive language detection using syntactic dependency graphs," in *Proc. 4th Workshop Online Abuse Harms*, 2020, pp. 44–53. [Online]. Available: https://aclanthology.org/2020.alw-1.6

[26] D. Goel and R. Sharma, "Leveraging dependency grammar for fine-grained offensive language detection using graph convolutional networks," in *Proc. 10th Int. Workshop Natural Lang. Process. Social Media*. Seattle, Washington: Association for Computational Linguistics, 2022, pp. 45–54.

[27] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991.

[28] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017, *arXiv:1701.06538*.

[29] J. Liao, D. Tang, F. Zhang, and S. Shi, "SkillNet-NLG: General-purpose natural language generation with a sparsely activated approach," 2022, *arXiv:2204.12184*.

[30] F. Zhang, D. Tang, Y. Dai, C. Zhou, S. Wu, and S. Shi, "SkillNet-NLU: A sparsely activated model for general-purpose natural language understanding," 2022, *arXiv:2203.03312*.

[31] X. Zhou and Y. Luo, "Explore mixture of experts in graph neural networks," Tech. Rep., 2019.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[33] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," 2017, *arXiv:1803.07640*.

[34] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," 2016, *arXiv:1611.01734*.

[35] P. Shi and J. Lin, "Simple BERT models for relation extraction and semantic role labeling," 2019, *arXiv:1904.05255*.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[37] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8583–8595.

[38] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, no. 2, pp. 181–214, 1994.

ZHICHAO YAN was born in Datong, Shanxi, China. He is currently pursuing the master's degree with Shanxi University. During his master's study, he was mainly responsible for constructing the knowledge base of the Chinese FrameNet Project. He has participated in the project of research on Chinese frame semantic computing based on cognitive (key research project supported by the National Natural Science Foundation) mechanism behind language facts. His research interests include nature language process and deep learning.

XUEFENG SU received the master's degree from the Taiyuan University of Technology, Shanxi, China. He is currently pursuing the Ph.D. degree with Shanxi University (SXU). He is currently an Associate Professor with SXU. His main research interests include deep learning and natural language processing. He is currently responsible for the funded projects related to semantic computing theory, method, and application. He is also participating in the National Key Foundation Project of Natural Science, and relevant achievements were published in ACL, NLPCC, and other international conferences.

QINGHUA CHAI received the M.A. degree from Shanxi University (SXU), China, in 2010. He is currently a Lecturer with the School of Foreign Languages, SXU. His current research interests include frame semantics, cognitive linguistics, and contrastive linguistics.

XIAOQI HAN is currently pursuing the Ph.D. degree with Shanxi University. During his Ph.D. study, he took part in the Research Chinese FrameNet Computing Based on Language Cognitive Mechanism Project, mainly responsible for the semantic analysis. His current research interests include natural language processing and semantic parsing.

YUNXIAO ZHAO was born in Linfen, Shanxi, China. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Shanxi University, Shanxi. His current research interests include natural language processing and machine reading comprehension.

RU LI received the Ph.D. degree. She is a second-level professor and an expert who enjoys special allowances from the State Council of China. She has published more than 100 papers in top international academic conferences and journals and holds nine invention patents. Her research interest includes natural language processing.

She has presided over and completed two National High-Tech Research and Development Programs of China (863 Program) Projects, one sub-project of the National Key Basic Research and Development Program, and four National Natural Science Foundation Projects (including one major project). She has received two second prizes for scientific and technological progress in Shanxi. She received the special prize and second prize for teaching achievements in Shanxi. She was the Executive Director of the Chinese Information Society of China, the Vice-Chairperson of the Shanxi Internet of Things and Artificial Intelligence Standardization Technical Committee, a Teaching Celebrity in Shanxi Higher Education, and a Leader of Academic Technology in Shanxi.