More Than Syntaxes: Investigating Semantics to Zero-shot Cross-lingual Relation Extraction and Event Argument Role Labelling

KAIWEN WEI, LI JIN, ZEQUN ZHANG, ZHI GUO, XIAOYU LI, and QING LIU, Aerospace Information Research Institute, China
WEIMIAO FENG, Institute of Information Engineering, Chinese Academy of Sciences, China

Syntactic dependency structures are commonly utilized as language-agnostic features to solve the word order difference issues in zero-shot cross-lingual relation and event extraction tasks. However, while sentences in multiple forms can be employed to express the same meaning, the syntactic structure may vary considerably in specific scenarios. To fix this problem, we find *semantics* are rarely considered, which could provide a more consistent semantic analysis of sentences and be served as another bridge between different languages. Therefore, in this article, we introduce Syntax and Semantic Driven Network (SSDN) to equip syntax and semantic knowledge across languages simultaneously. Specifically, predicate–argument structures from semantic role labelling are explicitly incorporated into word representations. Then, a semantic-aware relational graph convolutional network and a transformer-based encoder are utilized to model both semantic dependency and syntactic dependency structures, respectively. Finally, a fusion module is introduced to integrate output representations adaptively. We conduct experiments on the widely used Automatic Content Extraction 2005 English, Chinese, and Arabic datasets. The evaluation results demonstrate that the proposed method achieves the state-of-the-art performance. Further study also indicates SSDN could produce robust representations that facilitate the transfer operations across languages.

CCS Concepts: • **Computing methodologies → Information extraction**; *Natural language processing*; Semantic networks;

Additional Key Words and Phrases: Cross-lingual relation and event extraction, zero-resource transfer, semantic parsing, relational graph convolutional network

ACM Reference format:
Kaiwen Wei, Li Jin, Zequn Zhang, Zhi Guo, Xiaoyu Li, Qing Liu, and Weimiao Feng. 2024. More Than Syntaxes: Investigating Semantics to Zero-shot Cross-lingual Relation Extraction and Event Argument Role Labelling. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 23, 5, Article 61 (May 2024), 21 pages.
https://doi.org/10.1145/3582261

The work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (no. Y835120378) and the National Natural Science Foundation of China (62206267).

Authors' addresses: K. Wei, L. Jin (corresponding author), Z. Zhang, Z. Guo, X. Li, and Q. Liu, Aerospace Information Research Institute, Beijing, China, 100190; emails: weikaiwen19@mails.ucas.ac.cn, jinlimails@gmail.com, zqzhang1@mail.ie.ac.cn, guozhi@mail.ie.ac.cn, lixy01@aircas.ac.cn, liuqing1@aircas.ac.cn; W. Feng, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China; email: fengweimiao@iie.ac.cn.

2375-4699/2024/05-ART61 $15.00
https://doi.org/10.1145/3582261

1 INTRODUCTION

Relation and event extraction are two key components of information extraction that provide useful information for many **natural language processing (NLP)** downstream tasks, such as question answering [3, 47], document summarization [37, 39], and knowledge base construction [20, 53]. **Relation extraction (RE)** hopes to classify the relation type from pairs of entity mentions. Given a sentence "Chris hit Scott with a baseball," a RE system seeks to find the tuple such as (*Chris, PER-SOC:Lasting-Personal, Scott*). In addition, event extraction can be divided into the following two sub-tasks: event detection and **event argument role labelling (EARL)**, where the first one refers to identifying the event triggers (e.g., *hit* as *Injure* type), and the other attempts to extract the (trigger, argument role, argument) triples based on given event trigger (e.g., triple (*hit, victim, Scott*)). As a more challenging task, zero-shot cross-lingual relation and event extraction is conducted in multi-lingual situations and has achieved promising progress [1, 33, 44]. The workflow is illustrated in Figure 1, where the zero-shot setting refers to the state transformation of training on the *source language* and directly testing on another *target language*. Meanwhile, following the settings in Reference [1], we hypothesize all entities, event mentions (including event triggers and event arguments) are provided, and we focus on the study of RE and EARL.

Recently, lots of the existing pre-training models are trained in monolingual settings, but there is a huge gap between the features of words in different languages. If the models trained on the source language are directly applied to the target language tasks, then the performance is typically under satisfactory. To alleviate this problem, there has been a trend to utilize a universal encoder such as **multilingual BERT (mBERT)** [8] or XLM-R [7] to produce cross-language contextualized representations, and thus the model that learned on one language can be easily transferred to others. Moreover, in the zero-shot cross-lingual scenarios, it is crucial to discover features that do not change heavily with different languages, which we refer to *language-agnostic features*. On the one hand, the features of the target language are not available. On the other hand, since the distributions of data during training and testing are quite different, heavily relying on language-dependent features from the source language may result in an over-fitting dilemma. Research from References [1, 33, 44] demonstrate that dependency structures extracted from syntactic **dependency parsing (DP)** could be regarded as language-agnostic knowledge and effectively boost cross-lingual RE and EARL performance. As shown in the upper portion of Figure 2, DP seeks to find the grammatical relations between phrases. Rather than leveraging the features of the whole sentence, dependency structures could effectively reduce the distance between words, since words are skip-connected. In addition, it could mitigate the word order difference issue [2] in diverse languages to some extent.

However, since various languages have different styles of expression and grammatical structures, there may exist variances in sentences stating the same meaning. Take the sentences in the middle of Figure 2 as an example, where an active English sentence is translated into Chinese and turn it into passive order. We discover the following: Even if two sentences that express the same message in different languages, the syntactic structures have changed. Despite there being some implicit correlation between active and passive sentences, the features to be learned during training are still various. This phenomenon is not conducive to leveraging the knowledge from the source language and migrating such knowledge to other languages, especially in cross-lingual zero-shot settings, where samples from the target language are not available. In addition, this finding also indicates that it is not enough to merely consider syntactic structures in RE and EARL tasks. To alleviate such a problem, we find semantic knowledge is rarely considered compared with utilizing syntactic dependency information in cross-lingual tasks. As shown in the lower part of Figure 2, the semantics basically remain the same; for instance, in Chinese and English situations, *Chris* consistently plays the **semantic role labelling (SRL)** role as *ARG0*, and *Chris* is still the *Arg1*

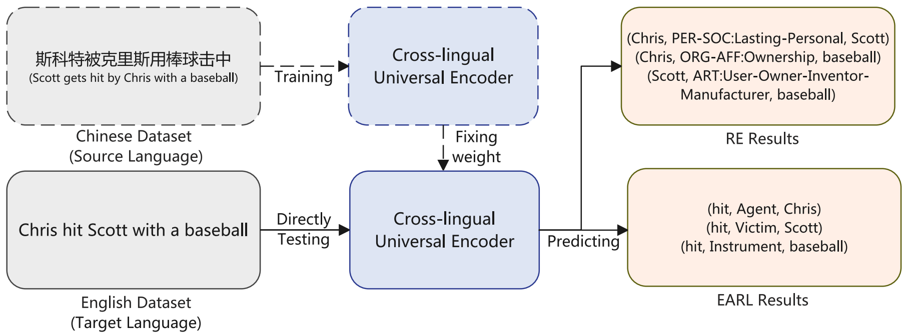

Fig. 1. The workflow of zero-shot cross-lingual relation and event extraction. The cross-lingual universal encoder is first trained on the source language dataset and then directly tested on the target language dataset. The goals of RE and EARL tasks are extracting the (entity, relation, entity) or (trigger, argument role, argument) triples, respectively.

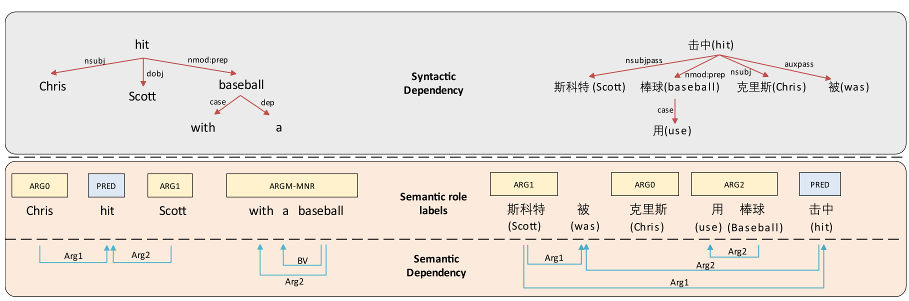

Fig. 2. Example of the syntactic dependency, semantic role labels, and semantic dependency. These elements are marked in black font in the figure. We translate an active English sentence into Chinese and turn it into passive order. The top part shows the syntactic dependency after DP, and the middle coloured boxes refer to the results after SRL (the blue boxes denote predicate, and the yellow ones indicate arguments). In addition, the lower portion illustrates the semantic dependency after SDP. The sentences in different languages express the same meaning, and syntactic structures are distinct, but the results of SRL and SDP basically keep the same. All the examples are parsed from HanLP online demo website at https://hanlp.hankcs.com/.

semantic dependency parsing (SDP) argument of *hit*. It illustrates that semantics are consistent and could be served as extra language-agnostic features.

The semantic analysis attempts to discover *who did what to whom, when, and why* with respect to the central meaning of the sentence, which provides an in-depth parsing of sentences. There are two practical approaches to extracting semantic information: SRL [58] and SDP [14]. As shown in the middle of Figure 2, SRL tries to extract the predicate–argument structures (e.g., *hit* as *predicate* and *Chris* as one *argument*). Meanwhile, SDP seeks to find the semantic factual or logical relationship between words (e.g., *baseball* is an *argument* of *Chris*). There are two advantages to introducing SRL and SDP: First, SRL and SDP provide more consistent parsing results and could be regarded as a bridge between languages. Utilizing such knowledge could effectively handle the expression difference issues (e.g., active-passive sentence problem), which cannot be appropriately solved by syntactic analysis. Second, since SRL and event argument role classification have similar task settings (predicate–argument and trigger–argument structures are analogous), SRL could be performed as a supplemental task and provides more prior knowledge.

In this article, we propose **Syntax and Semantic Driven Network (SSDN)** to equip syntax and semantic information simultaneously. To utilize SRL information, we map the discrete word-level parsing results to continuous representations and then integrate them into the word embeddings. For the results from SDP, not only is the category of each word critical, but also the labelled edges connecting the words could reflect the semantic relations. Therefore, we introduce a **semantic-aware relational graph convolutional network (Sem-RGCN)** to model the treelike semantic dependency structures and the corresponding semantic dependency relation types. In addition, a transformer-based architecture is used for encoding syntactic information, where the multi-head attentions are constrained by dependency tree distance to control the distance of information spreading. A fusion module is leveraged to adaptively choose the output syntax and semantic representations. In addition, since the soft labels and features from trained models bring useful information, a knowledge distillation mechanism is further introduced to boost performance by transferring knowledge from a well-trained teacher model to a student model at logit level and feature level.

Extensive experiments are conducted to evaluate the proposed SSDN on English, Chinese, and Arabic languages from **Automatic Content Extraction (ACE)** 2005 datasets. The results show that SSDN model achieves state-of-the-art performance. In addition, SSDN performs well in different languages, demonstrating the robustness and superiority of semantic information. The contributions of the article can be summarized as follows:

- In addition to syntax information, we further propose to leverage semantic features to enhance the migration capability of cross-language models. To the best of our knowledge, this is the first work to simultaneously consider semantic and syntax information in zero-shot cross-lingual relation and event extraction tasks.
- We adopt SSDN to explicitly incorporate syntaxes and semantics. Discrete predicate–argument structures are integrated into the word representation after semantic role labelling. In addition, semantic dependency structures and the corresponding semantic dependency relations are fused by a Sem-RGCN.
- Experiments on the widely used ACE2005 English, Chinese, and Arabic datasets showcase that the proposed method achieves state-of-the-art performance in most single-source and multi-source transfer scenarios. The further study illustrates that SSDN is less sensitive to the source language, indicating the robustness of semantics.

2 RELATED WORK

2.1 Relation and Event Extraction

In recent decades, relation and event extraction has achieved promising performance and received increasing attention. Early approaches usually utilize symbolic features [22, 26] to mine relational knowledge. And recent methods use continuous vector representations [27, 34, 40, 51], which leverage convolutional neural networks [25], attention mechanism [45], and **graph convolutional network (GCN)** [24, 50, 52] to promote experimental performance. Later, several joint learning or inference methods [11, 12] are proposed by benefiting from other relevant tasks. Cross-lingual relation and event extraction attempt to boost performance from other languages. However, different languages have various ways of expression and grammar structure, making it more challenging than single-lingual tasks. To resolve this problem, earlier methods use manually designed rules to establish links between different languages, including manually annotating on parallel data [4, 38], calculating annotation projection [23], or making bilingual dictionaries [17, 36]. The most intuitive approach is leveraging machine translation [10, 59] tools to get the parallel data. Nevertheless, those methods have the problem of error accumulation and are not applicable for languages not widely used. As a result, References [31, 32] proposed to explore common patterns across languages and acquire successful results.

Table 1. Comparison of Main Methods between our Proposed SSDN on Zero-shot Cross-lingual Relation and Event Extraction Tasks

Methods	CL_GCN	GATE	SSDN (ours)
Syntactic Dependency	✓	✓	✓
Semantic Role Labelling	×	×	✓
Semantic Dependency	×	×	✓
Graph Convolutional Network	✓	×	✓
Transformer	×	✓	✓

CL_GCN denotes the model in Reference [44] and GATE is from Reference [1].

However, those mentioned approaches are trained in a supervised learning setting, which relies on high-quality labelled data, and the performance of models typically degrade in low-resource situations. To solve this problem, Reference [33] and Reference [44] proposed to use GCN models to learn language-agnostic information from dependency parsing [35] results. Considering that the closer tokens in the parse tree should be paid more attention than the faraway ones, Reference [1] further introduced a transformer-based encoder to weight the syntactic distance attention. Nevertheless, those methods do not explicitly consider the semantic information and the relation types among words [41]. In this article, we introduce a Sem-RGCN to incorporate the semantic dependency structures.

2.2 Language-agnostic Information and Knowledge Distillation

It has been widely accepted that dependency structures obtained from DP tools can mitigate the word order difference issues [2] from diverse languages and could be served as language-agnostic information. References [1, 33, 44] successively propose GCN models and transformer-based encoder to equip such syntactic knowledge. However, as mentioned in the above introduction, syntactic structures of sentences with the same meaning could be slightly different. But semantics could provide a more in-depth and consistent semantic analysis of sentences, which could also be served as an effective bridge between different languages. Therefore, in this article, our SSDN additionally leverages SRL [58] and SDP [14] to obtain semantic knowledge. Specifically, we integrate the discrete semantic results from SRL to word representation and then leverage sem-RGCN to fuse semantic dependency features and the dependency type information. Table 1 illustrates the comparison between mainstream methods and SSDN.

SRL generally presents the semantic relationship as a predicate–argument structure, which is beneficial for event argument role labelling, since they have similar mission settings. Unlike DP, which emphasizes the role of prepositions, auxiliaries, and so on, SDP focuses on the semantic factual or logical relationship between words and provides more consistent language-independent results. There are some researches [55, 57] working on integrating semantic information, but those methods are performed in monolingual settings, and there is no word order difference problem across languages. Pre-trained language models such as BERT [8] and ERNIE [56] contain semantic information, but the knowledge is implicit and hard to be explained. To the best of our knowledge, we are the first to simultaneously consider semantic and syntax information in zero-shot cross-lingual relation and event extraction tasks.

The core ideas of knowledge distillation is to guide a student model to imitate the behaviour of well-trained teacher models, which is first proposed in Reference [15] and has been widely used in the natural language processing field [43, 48]. It is mainly used to compress model size [30, 49] or ensemble of models [15] via transferring knowledge from larger models (teachers) to a smaller

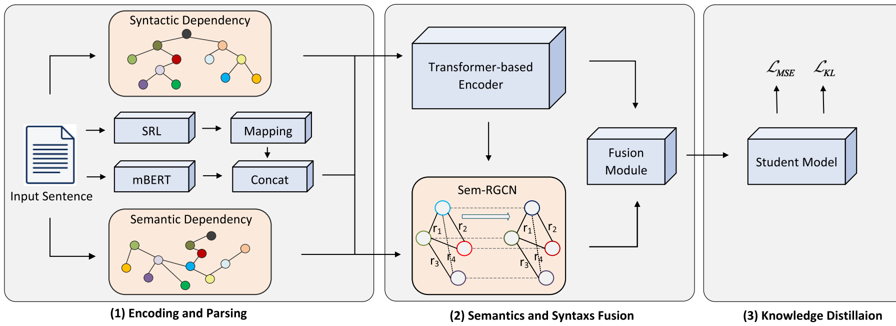

Fig. 3. The overview of our proposed approach, which consists of three stages: (1) Encoding and parsing stage, where BERT and parsers are leveraged to obtain the word representations and DP, SRL, and SDP parsing results. (2) Semantic and syntactic fusion stage, where a transformer-based encoder and Sem-RGCN are utilized to incorporate DP and SDP. (3) Knowledge distillation stage, where the knowledge distillation mechanism is used to improve the model performance further.

model (student). Since the teacher models contain valuable prior information, in this article, we distill the the output features from a well-trained teacher model to a student model. Moreover, we also adopt knowledge distillation mechanism to leverage the predicted logits as "soft pseudo-labels."

3 METHODOGY

This section describes the architecture of our model SSDN in detail, which explicitly incorporates semantic and syntactic information. The framework of the model is illustrated in Figure 3. Specifically, we first obtain the SRL, syntactic DP, and SDP and concatenate the mapped continuous SRL to the multi-lingual word embeddings. Then, a transformer-based encoder is utilized to incorporate syntactic dependencies, and we introduce Sem-RGCN to encode semantic dependency structures with the corresponding relation types. A fusion module is leveraged to adaptively select semantic and syntactic output representations. In addition, a knowledge distillation mechanism further transfers knowledge to a student model, which has the same architecture as the teacher model.

In the experiments, we hypothesize all entities, event mentions are provided, and we focus on the RE and EARL tasks. Formally, given a pair of entities e_s and e_o from a sentence, the RE task seeks to classify the relation $r_r \in \mathcal{R}_r \cup \{None\}$, where the subscript r indicates the RE task, r_r is the golden-standard category label, and $\mathcal{R}_r$ is the pre-defined set of relation labels. Likewise, given an event trigger e_t and a candidate event argument e_a, EARL refers to predicting the argument role $r_a \in \mathcal{R}_a \cup \{None\}$, where the subscript a denotes EARL task.

3.1 Parsing and Encoding

Our preliminary analysis illustrates that not only syntactic dependencies but also semantic knowledge could be served as language-agnostic information and provide more stable results across languages. To obtain such information, we use open source tools to acquire the SRL, DP, and SDP results. Specifically, *Stanford CoreNLP Toolkit* [35] is utilized to parse the syntactic dependencies. Meanwhile, such a tool is also employed to obtain the part-of-speech, entity type, and syntactic dependency type. In addition, the SRL and SDP output could also be obtained from the Electra small model [6] and Biaffine SDP model [13] integrated in the *HanLP* toolkit, respectively. Because the semantic labels for diverse languages are slightly different, we select uniform labels across all languages.

At the encoding stage, for the input sentence, we convert it into an embedding matrix and use it as the input of the transformer-based encoder. mBERT [8] are leveraged to build semantic representation for each word in the context. After the encoding stage, we utilize hidden states from the last layer of mBERT to represent each token. Formally, given a sentence with N words $\{x_1, x_2, \ldots, x_i, \ldots, x_N\}$, where the x_i indicates the ith word, the encoded features can be calculated as follows:

$$h_1^{bert}, h_2^{bert}, \ldots, h_i^{bert}, \ldots, h_N^{bert} = mBERT(x_1, x_2, \ldots, x_i, \ldots, x_N). \tag{1}$$

Considering the phenomenon that the predicate–argument structure extracted from SRL tools may cross many words, the BIO tagging mechanism [9] is leveraged from sequence tagging to model the connection between words. For instance, given the sentence "with a baseball" with the semantic role as "ARGM-MNR," we further pre-process it to "B-ARGM-MNR I-ARGM-MNR I-ARGM-MNR." After acquiring the labels for word granularity from SRL, since the semantic labels are discrete, we construct a mapping dictionary. The keys of the dictionary are semantic labels, and the corresponding values are randomly initialized continuous embeddings. Formally, given the ith word with discrete semantic role label name l_i^{srl}, we could get the mapped continuous semantic-aware embedding v_i^{srl} by looking up the mapping dictionary:

$$h_i^{srl} = mapping(l_i^{srl}), \tag{2}$$

where *mapping* is the operation of finding continuous embeddings from the dictionary. The semantic-aware embedding is obtained by looking up the randomly initialized mapping dictionary, and it does not make any sense at the beginning. As a result, we hope to keep it updated during training so that embedding could get a better experiment result. It should be noted that a sentence may yield many pairs of SRL results, only the top-S results that have the most SRL labels are chosen in our experiments, where S is a hyper-parameter that controls the number of SRL labels to be chosen.

Then we concatenate the embedding h_i^{srl} to the corresponding word-level BERT encoded features h_i^{bert}. Similarly, we also get the ith continuous embedding h_i^{pos}, h_i^{et}, and h_i^{dr} of part-of-speech, entity type, dependency relation, respectively. And concatenate them to the mBERT output, getting the word representation:

$$h_i^{in} = \left[h_i^{bert}, h_i^{srl}, h_i^{pos}, h_i^{et}, h_i^{dr}\right], \tag{3}$$

where [.., .., ..] is the concatenation operation. And $h_i^{in} \in \mathbb{R}^d$, where d is the dimension of the ith concatenated word embedding.

3.2 Transformer-based Encoder

Transformer-based Encoder [1] leverages self-attention mechanism to consider syntactic structure and distances between words simultaneously. The key idea is to manipulate the mask matrix to impose the graph structure and retrofit the attention weights based on pairwise syntactic distances. Specifically, we concatenate the input word embeddings $\{h_1^{in}, h_2^{in}, \ldots, h_N^{in}\}$ and form the sentence embedding matrix $H^{in} \in \mathbb{R}^{N \times d}$. And the self-attention mechanism is formulated as follows:

$$\begin{gathered} H^{in} = \left[h_1^{in}, h_2^{in}, \ldots, h_N^{in}\right] \\ Q = H^{in}W_q, K = H^{in}W_k, V = H^{in}W_v, \end{gathered} \tag{4}$$

where $W_q \in \mathbb{R}^{d \times d_q}$, $W_k \in \mathbb{R}^{d \times d_k}$, and $W_v \in \mathbb{R}^{d \times d_v}$ are the projection matrices and d_q, d_k, and d_v are dimensions of each self-attention head.

The output attention A can be calculated as follows:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_v}} + M\right)V, \tag{5}$$

where $softmax(\cdot)$ denotes the softmax activation function. The mask matrix $M \in \mathbb{R}^{N\times N}$ is based on the syntactic dependency distance and formulated by

$$M_{ij} = \begin{cases} 0, & D_{ij} \le \delta \\ -\infty, & \text{otherwise} \end{cases}, \tag{6}$$

where D_{ij} is the syntactic distance between the ith and the jth words. δ is a hyper-parameter that controls after how many hops of dependencies will be kept for diverse self-attention heads. For example, if δ is 4 and the dependency distance is 1, then the element would be set as 0.

Assuming the transformer-based encoder consists of L^{syn} self-attention layers, after successively undergoing the above operation L^{syn} times, we could finally obtain a syntactic-aware sentence representation $H^{syn} \in \mathbb{R}^{N\times d}$.

3.3 Semantic-aware Relational Graph Convolutional Network

Different from previous works that encode *unlabelled* syntactic structure via GCN [32, 44], we further introduced a Sem-RGCN module to explicitly encode semantic dependency structures and the corresponding dependency relations. Sem-RGCN could effectively incorporate the *labelled* semantic graphs by learning a separate projection matrix for each semantic relation.

We treat the words in a sentence as the nodes, and the dependency structures as the edges of Sem-RGCN. Formally, the semantic dependency structures for a sentence is defined as a directed and labelled graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where $\mathcal{V} = \{x_1, x_2, \ldots, x_N\}$ is a set of nodes, $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$ is a set of language-universal semantic dependency relations, and $\mathcal{R}$ is the corresponding relation type between two nodes (including inverse relations for inverse edges). N is the number of words in the sentence, and M is the number of dependency relations between words. It should be noted that $\mathcal{R}$ could be either the pre-defined relation class set $\mathcal{R}_r$ or the event argument class set $\mathcal{R}_a$.

The semantic dependency structure of a sentence with N words is converted into a $N \times N$ adjacency matrix A, where the element $a_{i,j}$ from ith row and jth column is set as 1 if there is a connection in the parsed semantic dependency graph. In addition, self-connections at each node are adopt to help capture information about the current node itself. The node embedding is initialized as the concatenation of word embeddings h^{in} and syntactic output h^{syn}, followed by a learned affine transformation and a ReLU nonlinearity:

$$\begin{aligned} h^{concat} &= [h^{in}, h^{syn}] \\ o^{(0)} &= \mathrm{ReLU}\left(W_e h^{concat}\right), \end{aligned} \tag{7}$$

where $W_e \in \mathbb{R}^{d\times 2d}$ is a learned matrix. The superscript of o denotes the layer number, and $o^{(0)}$ refers to the input of Sem-RGCN.

Assuming there are L^{sem} Sem-RGCN layers, for each Sem-RGCN layer l, the node's hidden representations are propagated with their direct neighbors:

$$o_i^{(l)} = \mathrm{ReLU}\left(\sum_{r\in\mathcal{R}} \sum_{j\in\mathcal{N}_i^r} \frac{1}{\left|\mathcal{N}_i^r\right|} W_r^{(l)} o_j^{(l)} + W_0^{(l)} o_i^{(l)}\right), \tag{8}$$

where $r \in \mathcal{R}$ is the pre-defined relation type of RE or EARL. $\mathcal{N}_i^r$ refers to the neighbors of the ith node with relation r. W_r and W_0 are learnable parameters indicating relation-specific transformation and a self-loop transformation, respectively. In the experiment, only the one-hop neighbour of each node at each iteration is calculated.

After L^{sem} layers of iteration, we take the final layer of Sem-RGCN model as the semantic-aware sentence representation $H^{sem} \in \mathbb{R}^{N \times d}$.

3.4 Fusion Module

A fusion module is adopted to incorporate syntactic and semantic knowledge. Motivated by Reference [16], a gate mechanism is utilized to dynamically integrate the syntactic representation H^{syn} after transformers-based encoder and semantic representation H^{sem} generated from Sem-RGCN module. Specifically, the gate G is calculated as follows:

$$G = sigmoid(W_g[H^{syn}, H^{sem}] + b_g), \tag{9}$$

where $W_g \in \mathbb{R}^{1 \times 2d}$ and b_g are trainable variables of the gate, $sigmoid(\cdot)$ is the sigmoid activation function, G is a 1-d vector, and each element is $g_i \in [0, 1]$. We leverage the gate G to form the final representation as follows:

$$H = G \odot H^{syn} + (1 - G) \odot H^{sem}, \tag{10}$$

where $\odot$ denotes the element-wise production operation. In this way, G controls the proportion of each input, and the output sentence representation $H \in \mathbb{R}^{N \times d}$ could be the final result that adaptively integrates the syntactic and semantic knowledge.

3.5 Training and Knowledge Distillation

Given the fused representation H, we aim to identify the label from pre-defined categories. For the relation extraction task, considering that entities may have different lengths, given pairs of subject entity e_s and object entity e_o, we perform a max-pooling operation to get fixed-length entity representations h_s and h_o. Following Reference [44], we also obtain sentence representation h_s by conducting a max-pooling operation over the encoding sequence H of every sentence. Finally, the concatenated vectors $[h_s, h_o, h_s]$ are fed to a linear classifier followed by a softmax layer to predict the relation type,

$$y_r = softmax\left(\frac{W_r^T[h_s, h_o, h_s] + b_r}{T}\right), \tag{11}$$

where y_r is the predicted logit probability, $W_r^T \in \mathbb{R}_r^{d \times |\mathcal{R}|}$ and b_r are learnable parameters of the last feed forward layer, and $|\mathcal{R}_r|$ indicates the number of pre-defined categories of relation extraction task. In addition, T is a temperature factor to control the smoothness.

Likewise, for the event argument role labelling task, we conduct max-pooling over argument candidate, event trigger, and sentence and get the vectors e_a, e_t, and e_s. Then we concatenate the vectors $[e_a, e_t, e_s]$ and pass them through a linear classifier and softmax layer to predict the argument role label,

$$y_a = softmax\left(\frac{W_a^T[e_a, e_t, e_s] + b_a}{T}\right), \tag{12}$$

where y_a is the distribution of prediction probability output.

The relation extraction and event argument role labelling models are optimized by minimizing **Cross Entropy (CE)** loss as follows:

$$\mathcal{L}_{CE} = -\frac{1}{K}\sum_{k}^{K} y_k \, log(\hat{y_k}) + (1 - y_k) \, log(1 - \hat{y_k}), \tag{13}$$

where K denotes the number of training samples, y_k could be the predicted logits of the kth sample from either RE or EARL models, and $\hat{y_k}$ is the golden-standard label.

Knowledge distillation focuses on utilizing the final output layer or the intermediate layers of the teacher model. The hypothesis is that the student model will learn to mimic the predictions of the teacher model. The knowledge distillation process can be achieved by using a loss function, termed the distillation loss, that captures the difference between the logits of the student and the teacher model, respectively. As this loss is minimized during training, the student model will become better at making the same predictions as the teacher. Suppose there is a classification task, and the original label can be [0,1,0], which is the hard target. While for the soft targets, the label could be [0.35, 0.6, 0.05], which considers the implicit relationship between the labels. When a neural network is hard labelled, it actually loses the information of the original data and reduces the difficulty of fitting the model to the data, making it easier to fit the model, which may produce overfitting and lead to a decrease in the generalization ability of the model. When soft labelling is used, the model needs to learn more knowledge, such as the similarity and difference between two close probabilities, thus creating a challenge for the model's fitting ability and enhancing the model's generalization ability.

To make the most of the data from the source language, because a lot of helpful information can be carried in soft targets instead of hard targets [15], a knowledge distillation mechanism is further utilized. The workflow is as follows: First, we train a teacher model M^{tea} from the source language and fix its weights. By feeding the data from source language to the teacher model, the corresponding output logits y^{tea} and fused features h^{tea} are obtained. Second, we re-train a student M^{stu} sharing the same architecture as teacher M^{tea} but with diverse parameters, where M_{stu} is initialized by the pre-trained weights from mBERT. And we could obtain the output logits y^{stu} and features h^{stu} by inputting the source language data to the student model. Finally, **mean squared error (MSE)** and **Kullback–Leibler (KL)** divergence are utilized as loss functions to minimize the difference between teacher and student models. The KL divergence is a widely used metric in the knowledge distillation process to measure the difference between two probability distributions (i.e., the distributions of the teacher model and the student model). The essence of KL divergence loss is cross entropy minus information entropy, that is, the difference between the number of bits required to encode the true distribution using the estimated distribution and minus the number of bits required to encode the true distribution. By minimizing Kullback–Leibler divergence during training, the distribution of the student model gradually converges to that of the teacher model, thus allowing the student model to "learn" knowledge from the teacher model.

Formally, the MSE loss is utilized to minimize the gap between teacher model M^{tea} and student model M^{stu} in the fused output features and could be calculated as follows:

$$\mathcal{L}_{MSE} = \frac{1}{K}\sum_{k=1}^{K}\left(h_k^{tea} - h_k^{stu}\right)^2. \tag{14}$$

For output logits, the KL loss between teacher and student models could be obtained as follows:

$$\mathcal{L}_{KL} = \sum_{k=1}^{K}\left(y_k^{tea} log y_k^{tea} - y_k^{tea} log y_k^{stu}\right). \tag{15}$$

During the training process, in addition to considering the ground-truth labels, the student model is also influenced by the soft labels and fused output features from the teacher model. The overall loss of the student model under the knowledge distillation framework is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{MSE}, \tag{16}$$

where α and β are two weight coefficients.

Table 2. Statistics of ACE2005

	Relation Mentions	Event Mentions	Event Arguments
Chinese	9,317	3,333	8,032
English	8,738	9,317	4,731
Arabic	4,731	2,270	4,975

4 EXPERIMENTS

4.1 Datasets and Evaluations

We conduct the experiments on the widely used ACE2005 [5] corpus. It annotated relation mentions (entities with their relations) and event mentions (including event trigger and event arguments) in Chinese (Ch), English (En), and Arabic (Ar). The statistic of the dataset is illustrated in Table 2. In addition, ACE defines an ontology that includes 7 entity types, 18 relation subtypes, and 33 event subtypes. Extra *None* subtype is appended to indicate that there is no relation between entities or one candidate argument is not an argument of the event trigger. Following Reference [44], we randomly choose 80% of the corpus for training, 10% for development, and 10% for blind test. We downsample the negative training instances by limiting the number of negative samples to be no more than the number of positive samples for each document.

We leverage Precision (P), Recall (R), and F1-measure (F1) as the evaluate standards. We follow the criteria in References [28, 29], and relation extraction is considered correct if its relation type is correct. And an event argument is correctly labelled if its event type, offsets, and role label match any of the golden-standard event arguments.

4.2 Experiment Settings

We adopt the pre-trained language model mBERT [8] as the multi-lingual feature extractor and use the corresponding model weights from huggingface,[1] which supports 104 languages. But in this article, we only test the model in English, Chinese, and Arabic. Meanwhile, following Reference [1], syntactic dependency, part-of-speech, entity type, and dependency relation are parsed from *Stanford CoreNLP Toolkit* [35]. Since the experiments are conducted in multilingual situation, we use the open source multilingual tool *HanLP*[2] to obtain the semantic role label and semantic dependency results. The semantic role label parser utilizes Electra small model [6], which is trained on CPB3 and follow *Chinese Proposition Bank* annotation rules. In addition, the semantic dependency parser leverage Biaffine SDP model [13], which is trained on SemEval2016 and complies with CSDP specifications.

The hyper-parameter details of our model during training is as follows: The word embedding size is set as 768. In addition, the part-of-speech embedding size, entity type embedding size, dependency relation embedding size, and continuous semantic role label embedding are set as 10, 10, 10, and 30, respectively. The number of chosen SRL results S are set as 3. The transformer/GCN layers of L^{syn} and L^{sem} are set as 4 and 2. We set $\delta = [2, 2, 4, 4, \infty, \infty, \infty, \infty]$ and $\delta = [8, 8, 8, 8, \infty, \infty, \infty, \infty]$ for EARL and RE tasks on eight attention-heads. In addition, we leverage transformer-based encoder [1] as a feature extractor and SGD optimizer with a learning rate of 0.1. Moreover, the temperature T of knowledge distillation is 1. The coefficients α, β of KL loss and MSE loss are set as 0.5 and 0.5, respectively. We leverage grid search to select the best hyper parameters.

[1] https://github.com/huggingface/transformers.
[2] https://github.com/hankcs/HanLP.

4.3 Baseline Models

CL_Trans_GCN: This is proposed in Reference [32], where a sentence from the source language is mapped to the best-suited translation in the target language. In addition, a GCN module is leveraged to capture the syntactic dependency structures.

CL_GCN: This is proposed in Reference [44], which uses a GCN module to learn structured common space representation.

CL_RNN: This is proposed in Reference [36], which uses a bidirectional Long Short-Term Memory–type recurrent neural networks to learn contextual representation.

Transformer: This is proposed in Reference [46], which leverages multi-head self-attention mechanism to learn contextual representation.

Transformer_RPR: This is proposed in Reference [42], which uses relative position representations to encode the structure of the input sequences.

GATE: This is proposed in Reference [1], which is a modified version of multi-head self-attention mechanism. It introduces distance-based attention modelling strategy to weight the syntactic dependencies.

X-GEAR*: This is proposed in Reference [19], which is a generation-based framework for the event argument extraction task. In the encoding phase, it provides the template needed for decoding, and in the decoding phase, the model fills in the template with the appropriate results. However, event argument extraction task has some task-setting differences between the RE and the EARL tasks. To make X-GEAR applicable for the former tasks, we modify the input templates. Concretely, in the EARL task, for each trigger, all the candidate arguments are concatenated in the input template. In the RE task, for each subject entity, we enumerate all the possible object entities and the possible relations in the input template. For a fair comparison, we leverage the mT5-base [54] pre-trained language model from the huggingface library as the backbone.

4.4 Main Results

We conducted two kinds of experiments to illustrate the superiority of the proposed model SSDN, including single-source transfer and multi-source transfer. For single-source transfer experiments, we train the model on one single source language and directly evaluate it on another target language. For instance, train the model in English and directly test it in Chinese. For the experiments of multi-source transfer, the models are trained in order on a pair of source languages (e.g., {English, Chinese}, {English, Arabic}, and {Chinese, Arabic}) and directly tested on the other target language; for example, train the model in English and Chinese and test in Arabic.

The single-source transfer experiment results of RE and EARL tasks are illustrated in Table 3 and Table 4, respectively. We observe that (1) SSDN outperforms many strong baseline models in most situations, indicating the superiority of the proposed method. (2) How to encode syntactic dependency structures has a considerable impact on the results: The methods that leverage GCN to model syntactic dependency structures (e.g., CL_Trans_GCN and CL_GCN) perform worse than those methods that utilize transformer-based model (e.g., GATE and SSDN model). This finding proves the effectiveness of transformer-based models in capturing syntactic knowledge. (3) On the EARL task, X-GEAR has achieved relatively good experimental results, proving the effectiveness of the generative model. However, X-GEAR does not perform well on the RE task. A possible reason could be that the decoding template is too sparse (since a head entity usually has relationships with only a few tail entities, most of the relationships in the template are still marked as None), which increases the difficulty of generating correct answers. (4) Comparing with the state-of-the-art model GATE that only leverages syntactic information, SSDN further introduces semantic knowledge (from SRL and SDP) and achieves stable promotion in most circumstances. These results indicate

Table 3. Single-source Transfer Results on Relation Extraction

Model	En ⇒ Zh	En ⇒ Ar	Zh ⇒ En	Zh ⇒ Ar	Ar ⇒ En	Ar ⇒ Zh	Avg
CL_Trans_GCN	56.7	65.3	65.9	59.7	59.6	46.3	58.9
CL_GCN	49.4	58.3	65.0	55.0	56.7	42.4	54.5
CL_RNN	53.7	63.9	70.9	57.6	67.1	55.7	61.5
Transformer	57.1	63.4	69.6	60.6	67.0	52.6	61.7
Transformer_RPR	58.0	59.9	70.0	55.6	66.5	56.5	61.1
GATE	55.1	66.8	71.5	61.2	69.0	54.3	63.0
X-GEAR*	56.0	65.8	70.3	60.5	68.2	55.3	62.6
SSDN (ours)	**58.2**	**67.9**	**72.4**	**62.0**	**69.1**	**57.3**	**64.5**

⇒ denotes transferring knowledge from the source to the target language. Avg denotes the average results of the six single-source transfer scenarios of models. The * indicates the models are modified from the original paper to make it applicable for the relation extraction task.

Table 4. Single-source Transfer Results on Event Argument Role Labelling

Model	En ⇒ Zh	En ⇒ Ar	Zh ⇒ En	Zh ⇒ Ar	Ar ⇒ En	Ar ⇒ Zh	Avg
CL_Trans_GCN	41.8	55.6	41.2	52.9	39.6	40.8	45.3
CL_GCN	51.9	50.4	53.7	51.5	50.3	51.9	51.6
CL_RNN	60.4	53.9	55.7	52.5	50.7	50.9	54.0
Transformer	61.5	55.0	58.0	57.7	54.3	57.0	57.3
Transformer_RPR	62.3	60.8	57.3	66.3	**57.5**	59.8	60.1
GATE	63.2	68.5	59.3	69.2	53.9	57.8	62.0
X-GEAR*	64.1	69.8	**61.5**	69.8	52.1	59.4	62.8
SSDN (ours)	**64.8**	**70.5**	61.4	**70.1**	51.3	**60.1**	**63.0**

⇒ denotes transferring knowledge from the source to the target language. Avg denotes the average results of the six single-source transfer scenarios of models. The * indicates the models are modified from the original paper to make it applicable for the event argument role labelling task.

the superiority of semantic knowledge and verify our intuition that semantic knowledge provides an in-depth and consistent analysis of sentences.

The multi-source transfer experiment results are listed in Table 5. We find (1) comparing with single-source transfer situations, multi-source transfer experiments obtain significantly better experimental results. For instance, in the EARL task, SSDN trained on an English and Chinese corpus then tested on Arabic has a 13.9 points promotion over the situation where it is only trained on Chinese. This finding is similar to the previous study [18] and proves that training on the bilingual corpus often leads to better performance in cross-lingual transfer tasks. (2) Even trained on multiple single-lingual corpus, SSDN still achieves promising results than other baselines. A significant reason is that SSDN leverages semantic knowledge. Such results illustrate that semantic information could be served as extra language-agnostic information for both RE and EARL tasks.

4.5 Ablation Study

We conduct ablation studies on SSDN. In the experiment, we treat English as the source language, and Chinese and Arabic are utilized as the target languages, respectively. The ablation studies are conducted by incorporating (1) the part-of-speech (+POS tag), (2) the syntactic dependency relation label (+Dep. label), (3) the named entity type (+Entity type), (4) the transformer-based encoder to incorporate syntactic dependencies (+Syntactic), (5) the semantic dependency parsing and semantic role labelling results (+Semantic), and (6) the knowledge distillation mechanism (+KD). From

Table 5. Multi-source Transfer Results on EARL and RE

Model	{En, Zh} ⇒ Ar	{En, Ar} ⇒ Zh	{Zh, Ar} ⇒ En	Avg
Event Argument Role Labelling				
CL_Trans_GCN	57.0	44.5	44.8	48.8
CL_GCN	58.9	56.2	57.9	57.7
CL_RNN	53.5	62.5	60.8	58.9
Transformer	59.5	62.0	60.8	60.8
Transformer_RPR	71.1	**68.4**	62.2	67.2
GATE	73.9	65.3	61.3	66.8
X-GEAR*	72.2	67.2	62.8	67.4
SSDN (ours)	**74.4**	66.9	**63.1**	**68.1**
Relation Extraction				
CL_Trans_GCN	66.8	54.4	69.5	63.6
CL_GCN	64.0	46.6	65.8	58.8
CL_RNN	66.5	60.5	73.0	66.7
Transformer	68.3	59.3	73.7	67.1
Transformer_RPR	65.0	**62.3**	73.8	67.0
GATE	67.0	57.9	74.1	66.3
X-GEAR*	66.3	57.1	72.9	65.4
SSDN (ours)	**68.5**	59.1	**74.5**	**67.4**

⇒ denotes transferring knowledge from the source to the target language. Avg denotes the average results of the three multi-source transfer scenarios of models. The * indicates the models are modified from the original paper to make it applicable for the EARL and RE tasks.

Table 6. Ablation Study on the Use of Part-of-speech (POS) Tag, Syntactic Dependency Relation Label, Entity Type, Syntactic Knowledge, Semantic Knowledge, and Knowledge Distillation (KD)

Input features	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
mBERT	52.5	47.4	44.0	49.7
+ POS tag	49.3	47.5	44.1	47.0
+ Dep. label	49.7	51.0	48.6	47.0
+ Entity type	57.8	60.2	56.3	63.0
+ Syntactic	63.2	68.5	55.1	66.8
+ Semantic	64.5	69.9	57.6	67.8
+ KD	64.8	70.5	58.2	67.9

We leverage English as the source language and Chinese, Arabic as the target languages, respectively.

the experiment results listed in Table 6, we obtain the following observations: (1) Each component of the proposed SSDN has a specific level of improvement, demonstrating the effectiveness of each component. (2) The symbolic features (including part-of-speech and dependency path) and distributional information (including type representation and contextualized representation) mentioned in Reference [44] play an important role in RE and EARL. Especially for the entity type, which has 10.25 points averaged promotion on both EARL and RE tasks. (3) Syntactic and semantic information could further promote experiment results. Such observation proves that both syntactic information and semantic knowledge could be served as language-agnostic features simultaneously.

Table 7. Contribution of Multi-WE, mBERT, and XLM-R as Word Features on GATE and SSDN (Ours)

Word features	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
Multi-WE-GATE	35.9	43.7	41.0	54.9
mBERT-GATE	57.1	54.8	55.1	66.8
XLM-R-GATE	51.8	61.7	51.4	68.1
XML-R-SSDN	59.2	69.3	51.4	65.2
mBERT-SSDN	64.8	70.5	58.2	67.9

English is utilized as the source language, and Chinese and Arabic are adopted as target languages.

Table 8. Ablation of Semantic Information, Including SRL and SDP

Input features	EARL		RE	
	Chinese	Arabic	Chinese	Arabic
SSDN (ours)	64.8	70.5	58.2	67.9
- KD	64.5	69.9	57.6	67.8
- SDP	63.8	69.2	56.9	67.1
- SRL	63.2	68.5	55.1	66.8

We leverage English as the source language and Chinese and Arabic as the target languages.

(4) Knowledge distillation mechanism could also improve the performance by leveraging the "soft-label" generated from the well-trained teacher model.

We also conduct experiments to investigate the influence of different word embeddings. We utilize **multilingual word embeddings (Multi-WE)** [21], mBERT [8], and XLM-R [7] as word features, and GATE [1] as the baseline model. We choose to leverage English as the source language. The experiment results are illustrated in Table 7, and we discover that there is a massive gap between the pre-trained language models and traditional static word embeddings that illustrates the strength of pre-trained language models. In addition, the mBERT model generally has a comparable performance to XLM-R models in Arabic, but there was a significant drop in Chinese for XML-R for both GATE and SSDN models. As a result, we adopt mBERT as the feature extractor of SSDN.

4.6 Effective of Semantic Information

To verify the effectiveness of semantic information (including SRL and SDP), we also conduct experiments by progressively reducing the semantic information. Specifically, we gradually remove (1) the knowledge distillation mechanism (-KD), (2) the semantic dependency parsing information incorporated by the introduced Sem-RGCN (-SDP), and (3) the semantic role labelling information that is first mapped to a continuous embedding and then concatenated to the word embedding (-SRL). The experiment results are illustrated in Table 8. We find that both SRL and SDP have a positive impact on the final results. Meanwhile, the lifting is pretty steady. The reasons could be as follows: First, the mapping representation from SRL concatenated to the input embedding learns during the training process, which is beneficial for thoroughly mixing up with the other features of the input embedding (e.g., pos and dependency type). Second, we simultaneously feed the word embeddings and the syntactic-aware representations from transformer-based encoder to Sem-RGCN model; as a result, the syntactic information will not be lost. Third, the Sem-RGCN

Table 9. Experiments to Test the Sensitivity of the Model toward Source Language

Word features	EARL		RE	
	English	Chinese*	English	Chinese*
CL_GCN	51.5	56.3	46.9	50.7
CL_RNN	55.6	59.3	56.8	62.0
GATE	59.3	61.3	71.5	72.9
SSDN (ours)	61.4	62.3	72.4	73.5

We use Chinese as the source and English as the target language. * means the English samples are translated into Chinese by *Google Cloud Translate*.

Table 10. The F1-score under Different Fusion Operations on the Syntactic Representation and Semantic Representation, Where "Concat" Means Concatenate the Two Kinds of Representations

Number	Sum	Max	Concat	ours
F1	57.4	57.2	57.9	58.2

We use English as the source language and Chinese as the target language for the relation extraction task.

model could effectively encode the semantic dependencies and the corresponding relation with the relation type. Finally, thanks to the fusion module, SSDN adaptively choose the balance between semantic information and syntactic knowledge.

4.7 Sensitivity Toward Source Language

In this section, we investigate the sensitivity of models toward source language. The intuition of the experiments is that if a model performs significantly well on the translated source language sentences, the model is more sensitive toward the source language and may not be ideal for cross-lingual transfer. As a result, we leverage *Google Cloud Translate* to translate Chinese into English. In the experiment, we train the models in Chinese and then test on English and their Chinese translations. The experiment results are illustrated in Table 9, we observe that CL_GCN and CL_RNN have much higher accuracy on the translated (Chinese) sentences than the target language (English) sentences, and GATE obtains a relatively small disparity. However, SSDN achieves a minor disparity between the target language and translated corpus. A possible reason is that SSDN leverages semantic information as a bridge to connect different languages, and such information is language agnostic and insensitive to the source language.

4.8 The Influence of Fusion Strategies

We conduct experiments to illustrate the influence of diverse feature fusion strategies after obtaining the syntactic features and semantic features from the transformer-based encoder and the Sem-RGCN module, respectively. We utilize English as the source language, and the Chinese as the target language for the relation extraction task. Specifically, we provide four options: (1) Sum, which stands for summing the two kinds of features; (2) Max, which stands for maximizing the two kinds of features; (3) Concat, which stands for concatenating the two kinds of features; and (4) ours, which stands for leveraging the fusion module to fuse the two kinds of features. The experiment results are illustrated in Table 10, and we can observe that the fusion module

Table 11. The Influence of the Max Number of Predicate–Argument Structures from SRL

Number	0	1	2	3	4	5
F1	55.1	57.7	58.1	58.2	58.0	57.9

We use English as the source language and Chinese as the target language for the relation extraction task.

achieves the best experimental results. An explanation is that the Sum and the Max methods may lose some of the information. The Concat method takes all the information into consideration and may introduce unnecessary noise. But with the help of the gate mechanism, the final result could adaptively integrate the syntactic and semantic knowledge, thus achieving better performance.

4.9 The Influence of Chosen Semantic Role Number

This section introduces the influence of the hyper-parameter of the max number of predicate–argument structures S from SRL by setting it from 0 to 5, where we use English as the source language, Chinese as the target language. The experiment results are illustrated in Table 11. We observe that the modest number of S is the best. When S equals 0 (without SRL information), there is a considerable decrease, indicating the effectiveness of SRL knowledge. In addition, when S is set as 1 or 2, since SSDN cannot obtain enough semantic information, the performance is relatively poor. Meanwhile, because the results obtained by the SRL open source tool may produce errors, overusing the results from the SRL tool may lead to error propagation. As a result, when S is set to a relatively larger number (e.g., 4 or 5), the performance decreases slightly.

4.10 Case Study

In this section, we conduct case studies to further explore the effectiveness of our model. We choose English as the source language and Chinese as the target language for the relation extraction task. Table 12 shows the related training and test examples, where the ground-truth labels of training/test samples are all *ART:User-Owner-Inventor-Manufacturer* and subject/object entities are marked in [] and {}, respectively. We could observe that although all the samples express similar content (users own some manufacturers), all the training examples are active sentences, while the test sample is in a passive way. As a result, the syntactic dependency structures between training and test are significantly different and the state-of-the-art GATE model fails to identify the correct answer. However, not only based on syntactic information, SSDN further leverages semantic knowledge as another language-agnostic knowledge. Such knowledge provides a more in-depth and consistent result on active and passive sentences. Consequently, SSDN successfully recalls the relation *ART:User-Owner-Inventor-Manufacturer*.

5 CONCLUSION

In this article, we propose an SSDN to simultaneously consider syntaxes and semantics. Experiments from the widely used ACE2005 English, Chinese, and Arabic corpus show that our method achieves state-of-the-art performance in most single-source and multi-source language transfer scenarios. Further studies also illustrate the effectiveness and robustness of semantic knowledge space. This work demonstrates that in addition to syntactic knowledge, semantic information could be served as another language-agnostic features for zero-shot cross-lingual relation and event extraction tasks. In the future, we hope to shed some lights on incorporating semantic information into more zero-shot cross-lingual information extraction tasks, such as named entity recognition

Table 12. Case Study Experiment on Model GATE and SSDN (ours)

Input Sentence	GATE	SSDN (ours)
English Training Samples		
...concealed {weapons} permit [holders] respond faster than police because we are in greater numbers...	–	–
Ulster Town Supervisor Fred Wadnola identified the {gunman}'s [weapon] ...	–	–
Chinese Test Sample		
...美国亚里桑那州一名男子，被[警员]连开80发{胡椒弹}制服... (A man in Arizona, USA, was subdued by the [police] who fired 80 {pepper bombs}))	no_relation	ART:User-Owner-Inventor-Manufacturer
..警察在1名被击毙的[逃犯]手中发现了1枚{手榴弹}... (Police found a {hand grenade} in the hands of a [fugitive])	ORG-AFF: Ownership	ART:User-Owner-Inventor-Manufacturer
...有 5 名[小学生]因玩耍一枚被带入学校的{炸弹}而被炸伤... (Five [elementary school students] were injured playing with a {bomb} ...)	PHYS: Located	ART:User-Owner-Inventor-Manufacturer

We choose English as source language, and Chinese as the target language on relation extraction task. The ground-truth label of all the mentioned examples are *ART:User-Owner-Inventor-Manufacturer*. In addition, subject and object entities are marked in [] and {}, respectively.

and entity typing, and so on. In addition, we seek to discover suchmore approaches to incorporate accurate semantic signals for deeper comprehension.

REFERENCES

[1] Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. GATE: Graph attention transformer encoder for cross-lingual relation and event extraction. In *National Conference on Artificial Intelligence*.

[2] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard H. Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 2440–2452. https://doi.org/10.18653/v1/n19-1253

[3] Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. UHop: An unrestricted-hop relation extraction framework for knowledge-based question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 345–356. https://doi.org/10.18653/v1/n19-1031

[4] Zheng Chen and Heng Ji. 2009. Can one language bootstrap the other: A case study on event extraction. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'09)*.

[5] Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2006. ACE 2005 multilingual training corpus. In *Linguistic Data Consortium.*

[6] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*. OpenReview.net.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423.

[9] Benjamin Farber, Dayne Freitag, Nizar Habash, and Owen Rambow. 2008. Improving NER in Arabic using a morphological tagger. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08).*

[10] Manaal Faruqui and Shankar Kumar. 2015. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'15)*, Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar (Eds.). The Association for Computational Linguistics, 1351–1356. https://doi.org/10.3115/v1/n15-1151

[11] Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 434–444. https://doi.org/10.18653/v1/D19-1041

[12] Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 5717–5729. https://doi.org/10.18653/v1/2020.emnlp-main.461

[13] Han He and Jinho Choi. 2020. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. In *Proceedings of the 33rd International Flairs Conference.*

[14] Han He and Jinho D. Choi. 2019. *Establishing Strong Baselines for the New Decade: Sequence Tagging, Syntactic and Semantic Parsing with BERT*. The Florida AI Research Society.

[15] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. arXiv:1503.02531. Retrieved from http://arxiv.org/abs/1503.02531.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9 (1997), 1735–1780.

[17] Andrew Hsi, Yiming Yang, Jaime G. Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of the International Conference on Computational Linguistics.*

[18] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4411–4421.

[19] Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2022)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 4633–4646. https://doi.org/10.18653/v1/2022.acl-long.317

[20] Ali Hürriyetoglu, Erdem Yörük, Deniz Yüret, Osman Mutlu, Çagri Yoltar, Firat Durusan, and Burak Gürel. 2020. Cross-context news corpus for protest events related knowledge base construction. In *Proceedings of the Conference on Automated Knowledge Base Construction (AKBC'20)*, Dipanjan Das, Hannaneh Hajishirzi, Andrew McCallum, and Sameer Singh (Eds.). https://doi.org/10.24432/C5D59R

[21] Armand Joulin, Piotr Bojanowski, Tomás Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2979–2984. https://doi.org/10.18653/v1/d18-1330

[22] Eunju Kim, Yu Song, Cheongjae Lee, Kyoungduk Kim, Gary Geunbae Lee, Byoung-Kee Yi, and Jeongwon Cha. 2006. Two-phase learning for biological event extraction and verification. *ACM Trans. Asian Lang. Inf. Process.* 5, 1 (March 2006), 61–73. https://doi.org/10.1145/1131348.1131353

[23] Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Trans. Asian Lang. Inf. Process.* 13, 1, Article 3 (February 2014), 26 pages. https://doi.org/10.1145/2529994

[24] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations (ICLR'17, Toulon, France, April 24–26, 2017, Conference Track Proceedings)*, OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114.

[26] Lishuang Li, Jing Zhang, Liuke Jin, Rui Guo, and Degen Huang. 2015. A distributed meta-learning system for Chinese entity relation extraction. *Neurocomputing* 149 (2015), 1135–1142. https://doi.org/10.1016/j.neucom.2014.07.016

[27] Peifeng Li, Guodong Zhou, and Qiaoming Zhu. 2016. Minimally supervised chinese event extraction from multiple views. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16, 2, Article 13 (November 2016), 16 pages. https://doi.org/10.1145/2994600

[28] Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'14)*. 402–412.

[29] Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 73–82.

[30] Shih-Hsiang Lin, Berlin Chen, and Hsin-Min Wang. 2009. A comparative study of probabilistic ranking models for chinese spoken document summarization. *ACM Trans. Asian Lang. Inf. Process.* 8, 1, Article 3 (March 2009), 23 pages. https://doi.org/10.1145/1482343.1482346

[31] Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[32] Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018. Event detection via gated multilingual attention mechanism. In *Proceedings of the National Conference on Artificial Intelligence*.

[33] Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[34] Jianwei Lv, Zequn Zhang, Li Jin, Shuchao Li, Xiaoyu Li, Guangluan Xu, and Xian Sun. 2021. HGEED: Hierarchical graph enhanced event detection. *Neurocomputing* 453 (2021), 141–150. https://doi.org/10.1016/j.neucom.2021.04.087

[35] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Meeting of the Association for Computational Linguistics*.

[36] Jian Ni and Radu Florian. 2019. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 399–409. https://doi.org/10.18653/v1/D19-1038

[37] Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), the 33rd Conference on Innovative Applications of Artificial Intelligence (IAAI 2021), the 11th Symposium on Educational Advances in Artificial Intelligence (EAAI'21)*. AAAI Press, 13666–13674.

[38] Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. Bilingual active learning for relation classification via pseudo parallel corpora. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[39] Xueming Qian, Mingdi Li, Yayun Ren, and Shuhui Jiang. 2019. Social media based event summarization by user-text-image co-clustering. *Knowl. Based Syst.* 164 (2019), 107–121. https://doi.org/10.1016/j.knosys.2018.10.028

[40] Yanxia Qin, Zhongqing Wang, Yue Zhang, Kehai Chen, and Min Zhang. 2022. Advancing chinese event detection via revisiting character information. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 4, Article 78 (Februrary 2022), 9 pages. https://doi.org/10.1145/3502197

[41] Michael Sejr Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *Proceedings of the European Semantic Web Conference*.

[42] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18), Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 464–468. https://doi.org/10.18653/v1/n18-2074

[43] Shumin Shi, Xing Wu, Rihai Su, and Heyan Huang. 2022. Low-resource neural machine translation: Methods and trends. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (January 2022). https://doi.org/10.1145/3524300

[44] Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare R. Voss. 2019. Cross-lingual structure transfer for relation and event extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[45] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 3104–3112.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[47] Pulkit Verma, Shashank Rao Marpally, and Siddharth Srivastava. 2021. Asking the right questions: Learning interpretable action models through query answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21), the 33rd Conference on Innovative Applications of Artificial Intelligence (IAAI'21), the 11th Symposium on Educational Advances in Artificial Intelligence (EAAI'21)*. AAAI Press, 12024–12033. https://ojs.aaai.org/index.php/AAAI/article/view/17428.

[48] Kaiwen Wei, Xian Sun, Zequn Zhang, Jingyuan Zhang, Zhi Guo, and Li Jin. 2021. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP'21), Volume 1: Long Papers*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4672–4682. https://doi.org/10.18653/v1/2021.acl-long.360

[49] Chung-Hsien Wu and Haizhou Li. 2009. Introduction to the special issue on recent advances in Asian language spoken document retrieval. *ACM Trans. Asian Lang. Inf. Process.* 8, 1, Article 1 (March 2009), 3 pages. https://doi.org/10.1145/1482343.1482344

[50] Xiaohua Wu, Tengrui Wang, Youping Fan, and Fangjian Yu. 2022. Chinese event extraction via graph attention network. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 4, Article 71 (January 2022), 12 pages. https://doi.org/10.1145/3494533

[51] Jing Xia, Xiaolong Li, Yongbin Tan, Wu Zhang, Dajun Li, and Zhengkun Xiong. 2022. Event detection via context understanding based on multi-task learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (March 2022). https://doi.org/10.1145/3529388 Just Accepted.

[52] Yan Xiang, Zhengtao Yu, Junjun Guo, Yuxin Huang, and Yantuan Xian. 2021. Event graph neural network for opinion target classification of microblog comments. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 1, Article 17 (November 2021), 13 pages. https://doi.org/10.1145/3469725

[53] Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, and Changsheng Xu. 2020. Knowledge-based topic model for multi-modal social event analysis. *IEEE Trans. Multimedia* 22, 8 (2020), 2098–2110. https://doi.org/10.1109/TMM.2019.2951194

[54] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

[55] Shuailiang Zhang, Hai Zhao, Junru Zhou, Xi Zhou, and Xiang Zhou. 2021. Semantics-aware inferential network for natural language understanding. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21), the 33rd Conference on Innovative Applications of Artificial Intelligence (IAAI'21), the 11th Symposium on Educational Advances in Artificial Intelligence (EAAI'21)*. AAAI Press, 14437–14445.

[56] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[57] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *Proceedings of the National Conference on Artificial Intelligence*.

[58] Hai Zhao, Wenliang Chen, and Chunyu Kit. 2009. Semantic dependency parsing of NomBank and PropBank: An efficient integrated approach via a large-scale feature selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[59] Zhu Zhu, Shoushan Li, Guodong Zhou, and Rui Xia. 2014. Bilingual event extraction: A case study on trigger type determination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Received 10 May 2022; revised 6 January 2023; accepted 21 January 2023