Constructing a Silver Corpus for Weakly Supervised Vietnamese Event Extraction using Cross-Document N-ary Relation Filtering

Xuan-Hieu Pham*, **Minh-Tuan Vu***, **Mai-Vu Tran, Hoang-Quynh Le**†
Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam
{20020125, 21020664, vutm, lhquynh}@vnu.edu.vn

Abstract

Event extraction for low-resource languages such as Vietnamese is limited by the lack of large-scale annotated data. To address this, we propose a weakly supervised framework that constructs a silver corpus via pseudo-labeling. We introduce a cross-document n-ary relation filtering strategy to reduce noise by leveraging consistency across multiple articles describing the same event, and further enhance data diversity with schema-based augmentation. Experiments on the BKEE benchmark show consistent improvements, demonstrating the effectiveness of our approach. Data is available at: https://github.com/Larken1612/VietEE2.

1 Introduction

Event Extraction (EE) is a fundamental and challenging task in Information Extraction, aiming to identify structured representations of events described in unstructured text (Xiang and Wang, 2019). An event is typically defined as an occurrence that takes place at a specific time and location, involving one or more entities and often associated with a change of state. Accordingly, EE seeks to detect and structure event-related information from text, including event triggers and their associated arguments (Jurafsky and Martin, 2026).

A common end-to-end formulation of EE decomposes the task into three subtasks (Walker et al., 2006; Liu et al., 2020; Xiang and Wang, 2019): (i) *Entity Mention Detection (EMD)*, which identifies and classifies mentions of real-world entities such as persons, organizations, locations, and temporal expressions; (ii) *Event Detection (ED)*, which identifies event triggers-words or phrases that indicate the occurrence of events, and classifies them into predefined event types. (iii) *Event Argument Extraction (EAE)*, which identifies entities participating in each event and assigns them semantic roles. Figure 1 illustrates an example of EE.

EE remains challenging due to the complexity of event structures and the diverse interactions among their components, motivating a variety of approaches (Kontostathis et al., 2004; Xiang and Wang, 2019). Traditional pipeline-based methods decompose EE into sequential subtasks, i.e., EMD, ED and EAE, where each component is modeled independently. While this modular design allows for task-specific optimization, it suffers from error propagation, as mistakes in earlier stages (e.g., incorrect entity or trigger detection) can adversely affect downstream predictions. To mitigate this issue, recent studies have explored joint learning approaches that model entities, event triggers, and arguments simultaneously, thereby capturing their interdependencies and reducing cascading errors (Nguyen et al., 2021; Wadden et al., 2019; Lin et al., 2020). In this work, we follow this line of research as a strong EE baseline.

Vietnamese EE remains challenging due to its linguistic characteristics, including the lack of explicit word boundaries, strong contextual ambiguity, and the prevalence of multi-word event triggers. These factors make accurate detection of entities and events more difficult. Beyond linguistic challenges, a major bottleneck for Vietnamese EE lies in the scarcity of large-scale annotated data. The recently proposed BKEE dataset (Nguyen et al., 2024), although pioneering, is relatively small and lacks sufficient diversity to cover complex real-world event structures. This limitation significantly restricts the performance of data-driven approaches and motivates the need for scalable alternatives.

Our contributions are as follows. First, we construct a large-scale silver corpus for Vietnamese EE using pseudo-labeling, enhanced by a cross-document n-ary relation filtering strategy to improve label quality. Building upon this resource, we propose a weakly supervised joint learning frame-

* Co-first authors.
† Corresponding authors.

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 26–37

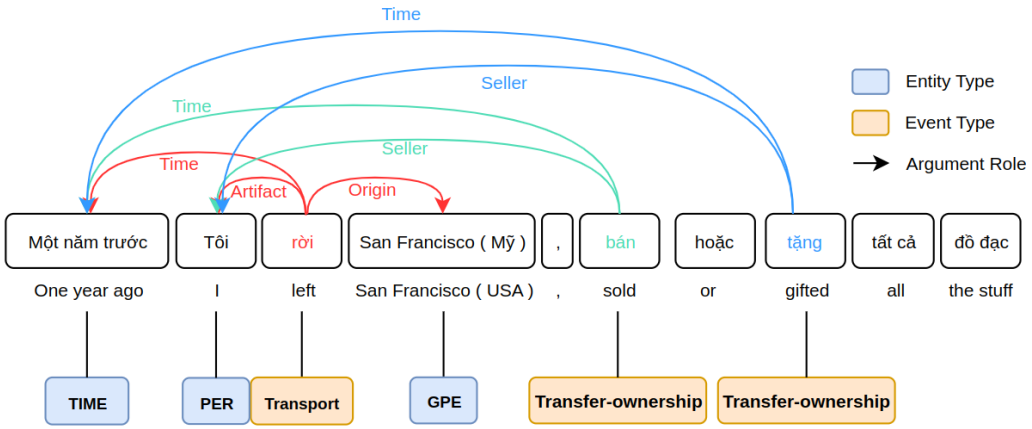

Figure 1: An example of event extraction.

work and demonstrate its effectiveness through baseline experiments on the BKEE benchmark.

2 Related Work

Existing approaches for EE can be broadly categorized into pipeline and joint learning methods (Xie et al., 2021). Pipeline approaches decompose the task into subtasks such as entity mention detection, event detection, and argument extraction, but suffer from error propagation across stages. To address this issue, joint learning models such as DyGIE++ (Wadden et al., 2019), OneIE (Lin et al., 2020) have been proposed to jointly model multiple components of EE. However, these models still have limitations in capturing complex dependencies across tasks and instances. FourIE (Nguyen et al., 2021) further improves joint learning by explicitly modeling inter-task dependencies through instance interaction and type dependency graphs. In this work, we adopt FourIE as the backbone model due to its strong performance and ability to capture structured dependencies, and focus on improving data quality via weak supervision.

Although EE has been extensively studied, most prior work focuses on high-resource languages such as English and Chinese, supported by large-scale datasets like MAVEN (Wang et al., 2020), RAMS (Ebner et al., 2020), and WikiEvents (Li et al., 2021). Multilingual benchmarks such as ACE 2005 (Walker et al., 2006) and TAC KBP (Mitamura et al., 2017) have further facilitated cross-lingual research. However, resources for low-resource languages remain limited. In particular, Vietnamese lacks large-scale annotated datasets for EE, with BKEE (Nguyen et al., 2024) being one of the few available resources, which significantly restricts the development of data-driven approaches.

Weakly supervised learning has emerged as a promising paradigm for EE, leveraging various forms of incomplete, inexact, or imprecise supervision signals to reduce annotation costs. *Distant supervision* generates training data by aligning text with external knowledge sources (e.g., Araki and Mitamura (2018)). While scalable, this approach depends on the availability of knowledge bases and often introduces noisy labels, as the presence of trigger words does not necessarily imply actual events. *Semi-supervised approaches* leverage both labeled and unlabeled data to improve EE. Huang and Ji (2020) learn latent representations for unseen event types, Ferguson et al. (2018) use self-training methods to generate pseudo-labels for unlabeled data. However, these methods are sensitive to pseudo-label quality, as errors can be reinforced during training. Recent work explores *pseudo-labeling strategies* to construct silver corpus. Yao et al. (2020) generate seed event pairs using heuristic patterns and refine them via semantic consistency before expanding with a trained classifier. However, these approaches mainly rely on local contextual signals, which may introduce noise for complex event structures. *Data augmentation* techniques have also been explored to improve model robustness and data diversity, including back-translation (Xie et al., 2020), synonym replacement (Dai and Adel, 2020), contextual rewriting (Yang et al., 2019), and schema-based generation (Jin and Ji, 2024). These methods typically preserve existing labels and operate at the sentence level, without explicitly addressing the quality of supervision signals. Our proposed method combines pseudo-labeling and self-training idea, reduce noise by exploiting cross-document consistency

and further incorporate a schema-based data augmentation strategy to improve data diversity while preserving structural validity.

3 Silver Corpus Construction

The silver corpusset construction process is illustrated in Figure 2. It consists of three main phases: data preparation, cross-document filtering and schema-based data augmentation.

3.1 Data Preparation

We collect a large-scale corpus of unlabeled Vietnamese news articles and organize them into groups based on shared topics or underlying events to support silver corpus construction. We leverage two widely used news aggregation platforms, *Báo Mới*[1] and *Google News*[2], which continuously collect and categorize news articles from multiple sources, providing topic-level grouping of semantically related documents.

We retain their topic assignments to form document groups and segment each article into sentences. As a result, we obtain a collection of sentence sets, where each set corresponds to a group of documents discussing a similar topic. In total, our dataset comprises approximately 72,000 articles, organized into more than 300 topic-based groups, yielding 2,673,796 unlabeled sentences.

To obtain initial annotations, we use BKEE event extraction model (Nguyen et al., 2024) to produce *coarse-grained annotations* for each sentence, including candidate event triggers and argument roles. This step results in a pseudo-labeled corpus that serves as input for subsequent refinement via cross-document filtering.

3.2 Cross-document Filtering

Pseudo-labeled data obtained from the previous stage inevitably contains noise due to model errors and domain mismatch. Based on the observation that real-world events are often reported by multiple news sources, resulting in multiple mentions of the same event across different articles, we propose a cross-document filtering strategy to improve annotation quality (see Table 5 in Appendix A for an example of sentences from different articles within the same topic may describe the same event with consistent triggers and arguments). This cross-document consistency provides a strong signal for distinguishing reliable event structures from noisy predictions.

Given a group of documents discussing the same topic, we first extract *n-ary relations* from pseudo-labeled sentences. Each relation is defined by an event trigger and its associated arguments (e.g., time, location, participants), as predicted by the BKEE model (see Table 6 in Appendix A for examples of n-ary relations extracted from pseudo-labeled sentences within a topic).

We aggregate these relations across documents within the same group and retain those that appear frequently. To implement this approach, we apply *frequency-based filtering* at two levels. First, for each event type et_j, we retain only those appearing in at least μ sentences within the group, i.e.,

$$count(|R_j|, m) \geq \mu. \tag{1}$$

Second, for each n-ary relation r_i associated with event type et_j, we retain it only if its occurrence count exceeds an adaptive threshold θ_j:

$$count(|r_i|, R_j) \geq \theta_j. \tag{2}$$

Let A_j denote the occurrence counts of all candidate n-ary relations associated with et_j. We compute the interquartile range (IQR) of A_j to measure the variability of relation frequencies. If the IQR is sufficiently small relative to the minimum count, i.e.,

$$IQR(A_j) \leq \frac{\min(A_j)}{\lambda}, \tag{3}$$

where $\lambda = 3$, we consider the relation frequencies to exhibit low variability and set $\theta_j = 0$. Otherwise, the threshold is defined as:

$$\theta_j = \frac{\min(A_j) + \max(A_j)}{2}. \tag{4}$$

Finally, we select all sentences that contain at least one such relation to construct the *filtered corpus*. This process effectively filters out noisy pseudo-labels while preserving frequently observed and contextually consistent event structures, resulting in a higher-quality silver-standard dataset. After cross-document filtering phase, we obtain a total of 15,260 qualified sentences in the filtered corpus.

3.3 Schema-based Data Augmentation

While filtering improves data quality, it does not increase structural diversity. We therefore introduce schema-based augmentation to generate diverse yet structurally valid event instances. This

[1] https://baomoi.com
[2] https://news.google.com/home?hl=vi&gl=VN

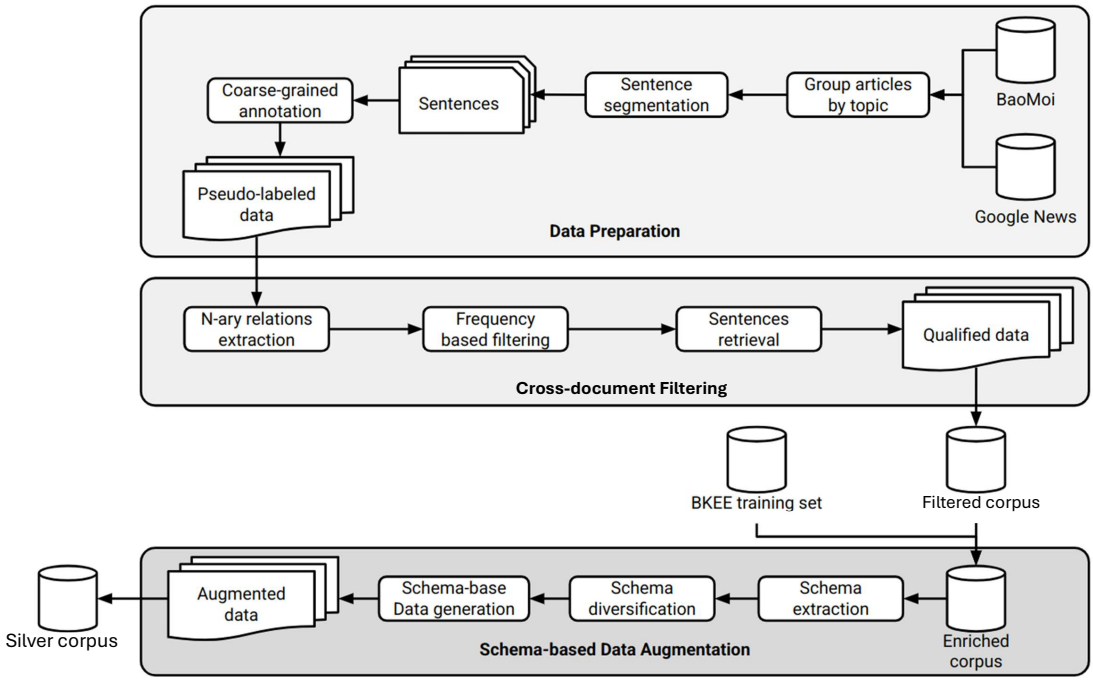

Figure 2: Silver corpus construction process.

phase adopts a sentence-level schema-based data augmentation strategy that introduces structural constraints during generation. Inspired by prior work (Jin and Ji, 2024), we adapt it to a setting where events are expressed within isolated sentences rather than across consecutive sentences.

Schema extraction: Since a sentence may contain multiple event triggers, we extract an event schema based on the triggers and their associated arguments. Figure 3 illustrates an example of such a schema, capturing the structural pattern of events and their contextual arguments, serving as an abstract template for data augmentation.

Schema diversification: To increase diversity, we construct new schemas from existing ones. For sentences containing multiple events, we decompose them into substructures and recombine them to form new event configurations. For sentences with a single event, we extend the schema by introducing additional arguments (e.g., *Time*, *Place*) when they are absent, ensuring that the resulting structures remain contextually valid. We represent schemas as structured graphs, where nodes correspond to event triggers and argument roles. To instantiate these schemas, we build a mapping M from event and entity types to candidate surface forms, collected from both internal (labeled data) and external sources. We sample from this pool to assign concrete values to each node, producing diverse realizations of the same structural pattern.

Schema-based data generation: The instantiated schema is serialized into a structured format (e.g., JSON) and used as input to an LLM (GPT-4o) for sentence generation. For schemas containing multiple events, the order of event instances is randomized to increase diversity. The generated sentences are then automatically annotated by aligning them with the schema, where matched spans are assigned their corresponding event and argument labels, ensuring structural consistency. This process ensures that the generated data remains structurally valid while introducing diverse surface realizations. The final silver corpus is expanded to 46,240 sentences.

4 Proposed Weakly Supervised Event Extraction Model

We build our framework upon the FourIE architecture (Nguyen et al., 2021), which jointly performs event mention detection (EMD), event detection (ED), and event argument extraction (EAE) over an input sentence $\mathbf{w} = [w_1, \ldots, w_n]$. We adopt this model as a backbone without modifying its core architecture, and focus on improving performance through enhanced training data constructed via weak supervision. The architecture comprises three phases: Span Detection, Instance Interaction, and Type-aware Regularization, as illustrated in Figure 4.

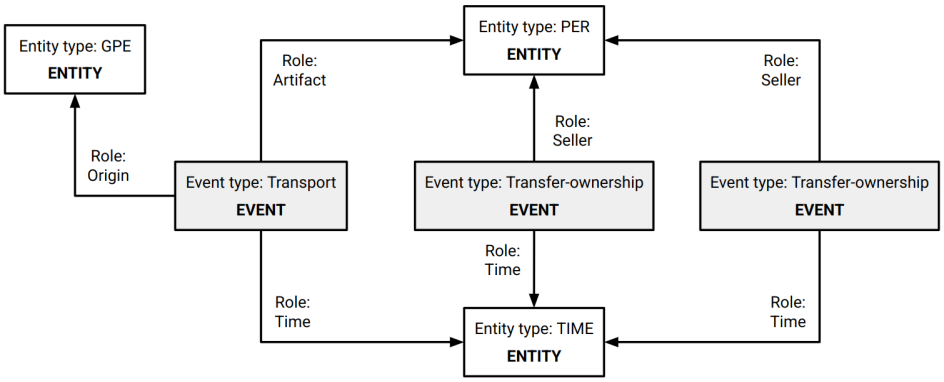

Figure 3: An example of an event schema extracted from a sentence. The schema represents event triggers and their associated arguments as a structured graph.

4.1 Span Detection

This module identifies entity mentions and event triggers from an input sentence to form nodes for the subsequent interaction graph. We formulate this step as a sequence labeling task using the BIO scheme, where each token is assigned one of three tags: B (Begin), I (Inside), or O (Outside). Unlike conventional named entity recognition, this stage does not predict specific entity or event types.

Given an input sentence $\mathbf{w}$, a pretrained encoder (i.e, PhoBert (Nguyen and Nguyen, 2020) or XLM-RoBERTa (Conneau et al., 2019)) produces contextual representations $X = [x_1, \ldots, x_n]$. These representations are then fed into two Conditional Random Fields layers to decode the optimal BIO tag sequences for entity mentions and event triggers, respectively. The model is trained by minimizing the negative log-likelihood losses $L_{\text{span}}^{\text{entity}}$ and $L_{\text{span}}^{\text{trigger}}$.

4.2 Instance Interaction

Given two span sets (entities and event triggers), the Instance Interaction module captures and enhances interactions between instances across tasks to improve prediction accuracy.

First, a representation vector for each span (i, j) $(1 \leq i \leq j \leq n)$ in these two sets is computed using the representation vectors $x_i, \ldots, x_j$. Let $R^{\text{entity}} = \{e_1, e_2, ..., e_{n_{\text{entity}}}\}$ $(n_{\text{entity}} = |R^{\text{entity}}|)$ and $R^{\text{trigger}} = \{t_1, t_2, ..., t_{n_{\text{trigger}}}\}$ $(n_{\text{trigger}} = |R^{\text{trigger}}|)$ denote the sets of span representation vectors for entities and event triggers in $\mathbf{w}$, respectively. The use of these two sets will be described in the following sub-sections.

Once R^{entity} and R^{trigger} are formed, we construct instance representations for the three IE tasks (EMD, ED, and EAE). Entity and trigger instances are directly derived from R^{entity} and R^{trigger}, respectively.

For argument prediction, which involves both a trigger and an entity, argument instances are defined as:

$$R^{\text{argument}} = \{\, arg_{ij} = [t_i, e_j] \mid t_i \in R^{\text{trigger}}, \; e_j \in R^{\text{entity}} \,\}. \tag{5}$$

The initial representation vectors for argument instances are constructed accordingly.

To model interactions between related instances, a graph G^{inst} is constructed, consisting of nodes N^{inst} and edges E^{inst}. The node set is defined as $N^{\text{inst}} = R^{\text{entity}} \cup R^{\text{trigger}} \cup R^{\text{argument}}$. Each entity node e_i is connected to all argument nodes $arg_{ij} = [t_j, e_i]$, and each trigger node t_j is also connected to these argument nodes, enabling information sharing among related instances.

This graph is then processed by a Graph Convolutional Network (GCN) to enrich instance representations. Let the initial node representations be $\{r_1, r_2, ..., r_{n_i}\}$ and the adjacency matrix be A^{inst}, where $A_{ij}^{\text{inst}} = 1$ indicates a connection between nodes r_i and r_j. The enriched representations are computed as:

$$r_1^{\text{inst}}, r_2^{\text{inst}}, ..., r_{n_i}^{\text{inst}} = \text{GCN}(A^{\text{inst}}; r_1, r_2, ..., r_{n_i}; N_i). \tag{6}$$

Finally, the enriched vectors are used to perform EMD, ED, and EAE. Let $\mathcal{T} = \mathcal{T}^{\text{entity}} \cup \mathcal{T}^{\text{trigger}} \cup \mathcal{T}^{\text{argument}}$, where $\mathcal{T}^{\text{entity}}$ denotes the set of entity types, and similarly for triggers and arguments. Let $t_k \in \{\text{entity}, \text{trigger}, \text{argument}\}$ be the task index and y_k the ground-truth label. Each type is associated with an embedding vector v, forming the set $\mathcal{V}$, with $\mathcal{V}^{t_k}$ corresponding to task t_k.

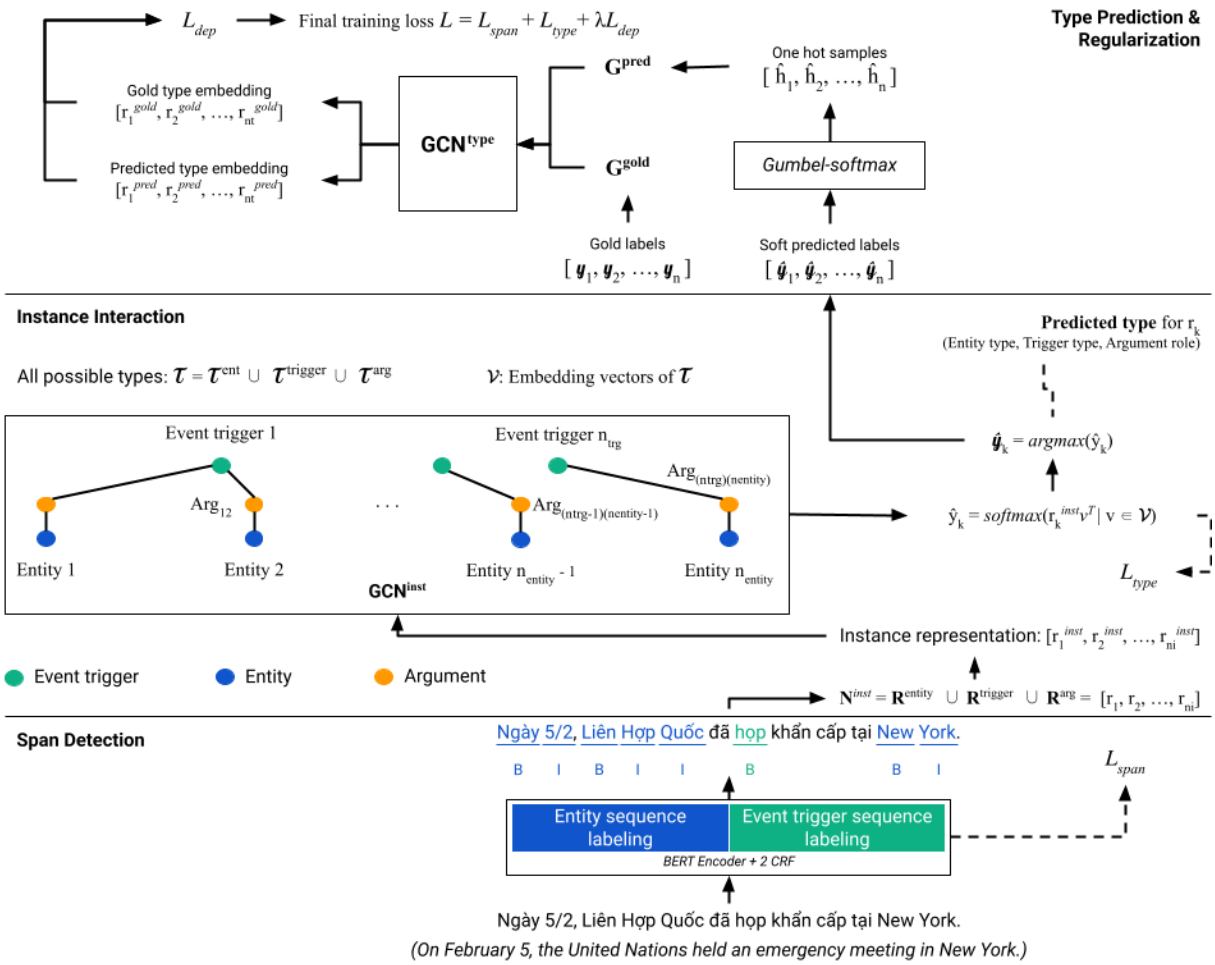

Figure 4: Overall architecture of the event extraction model.

For each instance $r_k \in N^{\text{inst}}$, the probability distribution over types is computed as:

$$\hat{y}_k = \text{softmax}(r_k^{\text{inst}} v^T \mid v \in \mathcal{V}^{t_k}), \tag{7}$$

and the predicted label is:

$$\hat{y}_k = \text{argmax}(\hat{y}_k). \tag{8}$$

4.3 Type Prediction and Regularization

This component models global type dependencies across the three IE tasks (EMD, ED, and EAE) to refine instance representations and improve prediction consistency.

Two dependency graphs, G^{gold} and G^{pred}, are constructed based on the gold types $y = \{y_1, y_2, ..., y_{n_i}\}$ and predicted types $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{n_i}\}$. The nodes correspond to types in $\mathcal{T}$, while edges encode relations between types (e.g., entity–argument compatibility). Their adjacency matrices are denoted as A^{gold} and A^{pred}, respectively.

Type representations are computed via a GCN over the initial type embeddings $\mathcal{T} = [v_1, ..., v_{n_t}]$:

$$r_1^{\text{gold}}, r_2^{\text{gold}}, ..., r_{n_t}^{\text{gold}} = \text{GCN}(A^{\text{gold}}; v_1, ..., v_{n_t}; N_t), \tag{9}$$

and similarly for the predicted representations. The dependency loss is defined as:

$$L_{\text{dep}} = \sum_{i=1}^{n_t} \| r_i^{\text{gold}} - r_i^{\text{pred}} \|_2^2. \tag{10}$$

Since G^{pred} is derived from discrete predictions, direct backpropagation is not feasible. To address this, A^{pred} is approximated by a differentiable matrix $\hat{A}^{\text{pred}}$:

$$\hat{A}^{\text{pred}} = \sum_{(i,j) \in I^{\text{inst}}} \exp\left(-\beta \left(B - \text{int}_t - j\right)^2\right), \tag{11}$$

where $I^{\text{inst}} = \{(i,j) \mid A_{ij}^{\text{pred}} = 1\}$, $B = \{b_{ij}\}_{i,j=1,...,n_t}$, and β is a large constant. In addition, we employ the Gumbel-Softmax trick to

approximate categorical predictions with continuous relaxations for gradient-based optimization.

The final training objective is:

$$L = L_{\text{span}}^{\text{entity}} + L_{\text{span}}^{\text{trigger}} + L_{\text{type}} + \lambda L_{\text{dep}}. \quad (12)$$

where λ balances the regularization term.

5 Experimental Results and Discussion

5.1 Experimental Settings

Dataset. We conduct experiments on the BKEE dataset, a benchmark for Vietnamese event extraction collected from 11 news domains. The dataset covers 12 entity types, 8 event types, 33 event subtypes, and 28 argument roles, with nearly 9,000 annotated event mentions and over 16,000 annotated entity mentions and arguments (Nguyen et al., 2024). Following the standard data split, it contains 10,959 training instances, 4,301 development instances, and 3,736 test instances.

Metrics. We report F1 scores for the three tasks, including entity mention detection (EMD), event detection (ED), and event argument extraction (EAE).

Environments and Configurations. We implement our model using Python 3.10.16, PyTorch 2.0.1+cu117, and Transformers 4.47.1. Experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU with 12GB VRAM. We use the pretrained FacebookAI/xlm-roberta-large model (Conneau et al., 2019) and PhoBert (Nguyen and Nguyen, 2020) for encoding and VnCoreNLP (Vu et al., 2018) for word tokenization.

The model is configured with a dropout rate of 0.4, sigmoid activation, and $N_i = 2$ hidden layers. We set $\beta = 1000$, $\lambda = 0.5$, and the learning rate to $5\text{e}{-6}$. The maximum number of training epochs is 25, and the batch size is 1 due to resource constraints.

5.2 Model Performance and Comparisons

We compare our method with three baseline models: (i) Pipeline models that perform each task independently, (ii) FourIE (Nguyen et al., 2021) as a joint learning baseline originally designed for English event extraction, and (iii) OneIE (Nguyen et al., 2024), an extension of joint learning approaches adapted for Vietnamese. We also evaluate two embedding settings, including multilingual XLM-RoBERTa (Conneau et al., 2019) and Vietnamese-specific PhoBERT (Nguyen and Nguyen, 2020).

Model	EMD	ED	EAE
(1) Pipeline + XLM-RoBERTa	55.0	60.3	44.9
(2) FourIE + XLM-RoBERTa	56.4	61.5	51.6
(3) OneIE + XLM-RoBERTa	56.3	60.0	51.7
(4) Pipeline + PhoBERT	54.4	61.8	44.4
(5) FourIE + PhoBERT	57.6	61.9	53.4
(6) OneIE + PhoBERT	55.8	**62.8**	53.0
(7) Proposed model	**59.1**	62.5	**55.2**

Table 1: Model performance comparison. Results are reported in F1 (%). The best results are highlighted in bold.

Table 1 shows that joint learning approaches (FourIE and OneIE) consistently outperform pipeline models across all tasks, highlighting the importance of modeling interdependencies between EMD, ED, and EAE. In addition, PhoBERT-based models generally achieve better performance than XLM-RoBERTa, confirming the advantage of language-specific representations for Vietnamese.

Compared to these baselines, our proposed model achieves the best overall performance, obtaining the highest F1 scores on EMD (59.1) and EAE (55.2), while remaining competitive on ED (62.5). In particular, compared to OneIE with PhoBERT, our model improves EMD and EAE by $+3.3$ and $+2.2$ F1, respectively, with a slight decrease of 0.3 F1 on ED. These results suggest that the proposed weak supervision strategy is especially effective for entity and argument extraction, where richer contextual and structural signals are required, while offering limited gains for trigger detection, which is less dependent on additional data. Overall, this demonstrates the effectiveness of leveraging enhanced training data for improving joint event extraction performance.

5.3 Impact of Silver Corpus Construction and Training Strategies

We evaluate different strategies for constructing and utilizing Silver corpus:

- (0) *w/o Silver corpus*: the model is trained only on the golden BKEE training set.
- (1) *w/o augmentation*: silver corpus is used but without schema-based augmentation and LLM-based generation.
- (2) *w/o filtering*: silver corpus is constructed without applying cross-document n-ary relation filtering.

Table 2: Impact of silver corpus Construction and Training Strategies. Results are reported in F1 (%). The best results are highlighted in bold.

Training setting	EMD	ED	EAE
(0) w/o silver corpus	55.5	**62.6**	52.9
(1) w/o augmentation	56.59	60.33	51.32
(2) w/o filtering	57.46	60.83	52.80
(3) Full corpus (scratch)	57.43	62.22	53.09
(4) Full + PM_1	56.88	61.24	52.50
(5) Full + PM_2	**59.11**	61.73	**55.17**

Full corpus: BKEE + silver corpus. PM_1: pretrained on BKEE. PM_2: pretrained on BKEE + 1/5 Silver corpus.

- (3) *Full corpus (scratch)*: the golden training data and the full Silver corpus are combined and used to train the model from scratch.
- (4) *Full + PM1*: the model is first pre-trained on the golden training set, and then further trained on the full Silver corpus.
- (5) *Full + PM2*: a small corpus is first constructed by combining the golden training set with 20% of the Silver corpus to pre-train the model, which is then further trained on the remaining silver corpus.

The results on Table 2 shows that incorporating silver corpus consistently improves performance on EMD and EAE compared to training only on the golden data (row (0)), demonstrating the effectiveness of weakly supervised data for span and argument extraction. However, ED does not benefit as much, indicating that trigger detection is less sensitive to additional data. Both schema-based augmentation and filtering play important roles in improving data quality. Removing augmentation (row (1)) or filtering (row (2)) leads to noticeable drops in performance compared to the best setting (row (5)), showing that both diversity and noise reduction are crucial for constructing effective silver corpus. Finally, training strategy has a significant impact on performance. Simply training on the combined data from scratch (row (3)) does not yield the best results, and pretraining only on the golden data (row (4)) is also suboptimal. The best performance is achieved by progressively leveraging silver corpus (row (5)), where the model is first exposed to a smaller, mixed corpus before being trained on the full dataset. This suggests that a curriculum-style training strategy is more effective than directly mixing all data or relying solely on golden pretraining.

5.4 Error Analysis

To improve the proposed model and incentivize future research, the model output has been analyzed to find out errors that need to be taken into account. For errors examples, please refer to Appendix B.

Span errors. In EMD and ED, the model sometimes predicts incomplete or over-extended spans, especially for long or nested mentions. For example, *"Ngân hàng Nông nghiệp và Phát triển nông thôn Việt Nam Agribank"* may be partially detected as *"Agribank"*.

Isolated entity detection failure. Entities without explicit trigger associations are occasionally missed, suggesting that the model relies heavily on trigger-aware context for entity recognition.

Polysemy. Words with multiple meanings may lead to incorrect entity type predictions, e.g., *"Jordan"* being classified as PER instead of GPE.

Ambiguous context in EAE. In complex sentences with overlapping contextual cues, the model may assign incorrect semantic roles, indicating limitations in contextual reasoning.

Overall, these errors suggest that the main limitations lie in boundary detection and contextual reasoning. Future work may benefit from stronger structural modeling and more fine-grained semantic supervision.

6 Conclusion

In this paper, we address the data scarcity problem in Vietnamese event extraction by proposing a weakly supervised framework for constructing a qualified silver corpus. Our approach combines pseudo-labeling with a cross-document n-ary relation filtering strategy to improve annotation quality, and a schema-based data augmentation method to enhance data diversity. Built upon a strong joint learning backbone, the proposed framework effectively leverages the constructed silver corpus of 46,240 sentences to improve event extraction performance. Experimental results on the BKEE benchmark demonstrate consistent improvements, achieving gains of +3.3% F1 on EMD and +2.2% F1 on EAE compared to strong baselines, while maintaining competitive results on ED. Overall, our findings suggest that a carefully constructed silver corpus, together with appropriate training strategies, can serve as an effective alternative to costly manual annotation for low-resource event extraction.

Limitations

Despite the effectiveness of our approach, several limitations remain.

First, the quality of the constructed silver corpus still depends on the initial pseudo-labeling model. Errors introduced in this stage may propagate through subsequent filtering and training, especially for rare or complex event types.

Second, the cross-document n-ary relation filtering strategy relies on the assumption that important events are reported multiple times across different documents. As a result, infrequent or emerging events may be underrepresented or filtered out, limiting coverage.

Third, the schema-based data augmentation process depends on predefined schema structures and LLM-based generation, which may introduce noise or generate less natural sentences in some cases.

Finally, our framework builds upon an existing backbone model without modifying its architecture. While this allows us to focus on data-centric improvements, it may limit the potential gains achievable through model-level innovations.

Addressing these limitations, particularly improving pseudo-label quality and better handling rare events, remains an important direction for future work.

References

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867. International Committee on Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077. Association for Computational Linguistics.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 359–364. Association for Computational Linguistics.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724. Association for Computational Linguistics.

Xiaomeng Jin and Heng Ji. 2024. Schema-based data augmentation for event extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14382–14392. ELRA and ICCL.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. 2004. *A Survey of Emerging Trend Detection in Textual Data Mining*, pages 185–224. Springer New York.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 894–908.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009. Association for Computational Linguistics.

Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. 2020. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2017. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *TAC*.

Dat Quoc Nguyen and Anh-Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 1037–1042.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38. Association for Computational Linguistics.

Thi-Nhung Nguyen, Bang Tien Tran, Trong-Nghia Luu, Thien Huu Nguyen, and Kiem-Hieu Nguyen. 2024. BKEE: Pioneering event extraction in the Vietnamese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2421–2427. ELRA and ICCL.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. (URL: https://catalog. ldc. upenn. edu/LDC2006T06).

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1652–1671.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training.

QunLi Xie, JunLan Pan, Tao Liu, BeiBei Qian, XianChuan Wang, and Xianchao Wang. 2021. A survey of event relation extraction. In *International Conference on Frontier Computing*, pages 1818–1827. Springer.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly supervised subevent knowledge acquisition. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 5345–5356.

#	Vietnamese	English
1	Ngày 12/02/2025, Bộ Tư pháp đã tổ chức **họp thẩm định** đề nghị xây dựng Luật thuế TNCN.	(On February 12, 2025, the Ministry of Justice held an **appraisal meeting** on the proposal to draft the Personal Income Tax Law.)
2	Trên cơ sở đó, ngày 12/2 Bộ Tư pháp đã tổ chức **họp thẩm định** đề nghị xây dựng Luật thuế thu nhập cá nhân.	(On that basis, on February 12, the Ministry of Justice held an **appraisal meeting** on the proposal to draft the Personal Income Tax Law.)
3	Ngày 12/2/2025, Bộ Tư pháp đã tổ chức **họp thẩm định** với mục đích đề nghị xây dựng Luật Thuế thu nhập cá nhân.	(On February 12, 2025, the Ministry of Justice held an **appraisal meeting** with the aim of proposing the drafting of the Personal Income Tax Law.)

Figure 5: Sentences from different documents describing the same event.

A Cross-document Filtering Examples

Figure 5 illustrates an example of sentences from different articles within the same topic may describe the same event with consistent triggers and arguments.

Figure 6 presents examples of n-ary relations extracted from pseudo-labeled sentences within a topic group. Each relation is characterized by an event type and its associated arguments, along with its occurrence count across sentences.

B Model Error Examples

Some representative examples of model errors are provided in Table 3.

Span errors. This type of error occurs in both EMD and ED when the predicted span does not exactly match the gold annotation. The model may either partially detect a mention or over-extend the span beyond its correct boundary. For instance, the entity *"Ngân hàng Nông nghiệp và Phát triển nông thôn Việt Nam Agribank"* is only partially detected as "*Agribank*", while complex mentions such as *"Bí thư tỉnh ủy và Chủ tịch Hội đồng nhân dân tỉnh Bến Tre Hồ Thị Hoàng Yến"* should be split into multiple entities but are instead merged into a single span. These errors typically arise when entity mentions or event triggers have long and nested structures, making it difficult for the model to accurately determine span boundaries.

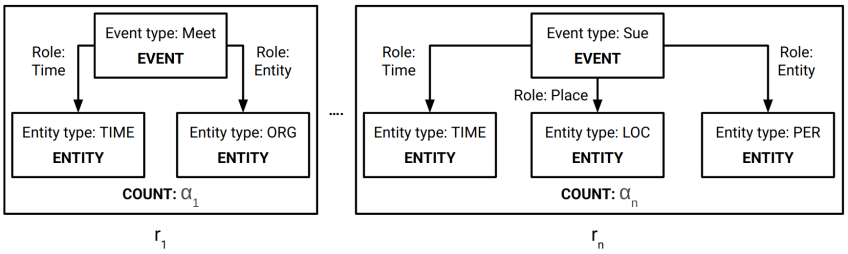

Figure 6: Examples of n-ary relations extracted from sentences within the same topic group. Each relation consists of an event type and its associated arguments with occurrence counts.

Isolated entity detection failure. This error occurs when entities that are not explicitly associated with an event trigger are not detected. For example, in sentence (3), entities such as *"ông Hùng"* and *"năm 2023"* are missed because they do not directly participate in a clearly expressed event. This suggests that the model relies heavily on trigger-aware context to identify relevant entities, and struggles to recognize standalone entities when contextual cues are limited.

Polysemy. Ambiguous words with multiple semantic meanings can lead to incorrect entity type predictions. For example, in sentence (4), the word *"Jordan"* is incorrectly classified as a person (PER) instead of a geo-political entity (GPE). Such errors occur when the model fails to effectively leverage contextual signals to disambiguate between different meanings of the same surface form. Although relatively less frequent, these errors highlight limitations in semantic understanding.

Ambiguous context in EAE. In the EAE task, errors often arise from complex or ambiguous contexts where multiple entities and actions are present within the same sentence. In sentence (5), the entity "*An*" is incorrectly assigned the Agent role in a Transport event, even though it does not perform the action of being transported. This type of error indicates that the model may misinterpret semantic roles when contextual signals are dense or overlapping, leading to incorrect argument-role assignments.

Table 3: Some notable errors on the test set output of the model.

#	Vietnamese	English	Error
1	*Lãnh đạo Ngân hàng Nông nghiệp và Phát triển nông thôn (**Agribank**) Chi nhánh Bắc Yên Bái trao tiền hỗ trợ gia đình chị Nông Thị Đông ở thị trấn Yên Bình.*	*Leaders of the Agricultural Bank and Rural Development (**Agribank**) North Yen Bai Branch donated financial support to the family of Ms. Nong Thi Dong in Yen Binh Town.*	**Wrong entity detected**: Agribank
2	***Bí thư tỉnh ủy và Chủ tịch Hội đồng nhân dân tỉnh Bến Tre Hồ Thị Hoàng Yến** chủ trì phiên họp.*	***Secretary of the Provincial Party Committee and Chairwoman of the People's Council of Ben Tre Province, Ho Thi Hoang Yen** chaired the meeting.*	**Wrong entity detected**: full span incorrectly merged
3	*Trước đây, gia đình ông **Hùng** thuộc diện hộ nghèo, đến năm 2023 là hộ cận nghèo.*	*In the past, Mr. **Hung**'s family was classified as a poor household, but by 2023, they became a near-poor household.*	**Missing entity**: Hung
4	***Jordan** bày tỏ lo ngại về những căng thẳng gia tăng trong khu vực.*	***Jordan** expressed concerns over the escalating tensions in the region.*	**Wrong entity type**: Jordan (PER → LOC)
5	*Hải đến nhà **An**, và **An** dùng điện thoại để giữ liên lạc với Hải trong quá trình di chuyển.*	*Hai went to **An**'s house, and **An** used his phone to stay in touch with Hai during the journey.*	**Wrong argument**: An incorrectly labeled as *Agent* in *Transport* event