Biomedical Entity-Aware Semantic Role Labelling via Model Merging

Hoai-Duc Tuan-Nguyen[1,2]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Correspondence: Hoai-Duc Tuan-Nguyen. Email: tnhduc@fit.hcmus.edu.vn
Communication: received 09/05/2025, revised 03/06/2025, accepted 18/06/2025
DOI: 10.32913/mic-ict-research.v2025.n2.1368

***Abstract*: Semantic Role Labeling (SRL) is essential for extracting structured knowledge from biomedical texts, yet it remains challenging due to the complexity of biomedical terminology. This paper introduces a novel method that enhances biomedical SRL by directly integrating Named Entity Recognition (NER) knowledge using model merging. NER plays a crucial role in anchoring domain-specific arguments, allowing the model to better distinguish semantic roles in biomedical contexts. Unlike prior approaches that rely on multitask learning or syntax-aware features, our method merges pretrained SRL and NER models in a way that preserves entity knowledge while aligning with semantic role structures. Experimental results demonstrate that our approach surpasses state-of-the-art multitask and syntax-aware SRL models in both accuracy and training speed. The integration of biomedical entity knowledge proves especially effective for predicates with high proportions of entity-based arguments, highlighting the value of direct NER-to-SRL knowledge transfer in advancing biomedical semantic understanding.**

Keywords: *Model merging, NER, NLP, SRL.*

I. Introduction

Biomedicine is a field that applies biotechnology to human healthcare. It plays an important role in diagnosing and treating diseases. In recent years, research in biomedicine has grown rapidly, leading to a vast amount of knowledge being stored in scientific texts. However, the increasing volume of biomedical literature makes it impossible for humans to manually process all this information [1]. To make full use of this knowledge, automated methods are needed to extract meaningful information efficiently [2, 3].

A key challenge in biomedical text mining is extracting structured knowledge, which requires identifying the relationships between actions (predicates) and the entities involved in those actions (arguments). A method to achieve this is through Semantic Role Labeling (SRL), a technique that helps computers recognize who is doing what to whom in a sentence. SRL is based on Predicate Argument Structure (PAS), where the predicate represents the main verb describing an event or action, and the arguments provide the key details related to that action [4]. For example, in the sentence "Aspirin inhibits platelet aggregation", the predicate is inhibits, while Aspirin is the agent (the entity performing the action), and platelet aggregation is the affected entity (the patient of the action).

Although SRL has been widely studied in general texts, applying it to biomedical literature presents new challenges. Biomedical sentences tend to be more complex, with specialized terminology and intricate sentence structures that make it harder for computers to recognize arguments correctly [5]. Another difficulty is the lack of annotated training data, which limits the ability of machine learning models to learn from examples [6]. Because of these challenges, adapting SRL models for biomedical texts requires special techniques, such as fine-tuning pre-trained language models with biomedical-specific knowledge [7].

One important factor in improving SRL for biomedical texts is the use of biomedical entities, such as proteins, DNA, RNA, cell lines, and cell types 5. These entities often appear as arguments in PAS, and recognizing them correctly helps improve the accuracy of SRL [8]. Therefore, we propose an improved approach that combines biomedical entity knowledge with a pre-trained language model to enhance SRL performance on biomedical text. As SRL focuses on identifying arguments within PAS, recognizing these domain-specific entities provides additional context that aids in distinguishing arguments and assigning their correct semantic roles.

To evaluate our approach, we conducted experiments using the PASBio+ corpus, a comprehensive biomedical SRL dataset [6]. Experimental results demonstrate that our method of incorporating biomedical entity knowledge enhances SRL performance, outperforming both multitask

learning and syntax-aware SRL approaches tested on the same dataset, as well as the baseline model that lacks entity information. These findings demonstrate the effectiveness of integrating entity knowledge in SRL for biomedical text mining and highlight a promising direction for future research in the field. While our method is evaluated in a biomedical context, the model merging strategy is domain-agnostic and can be extended to other domains in future work. We explicitly scope this study to biomedical predicates, where entity-based argument structures are prominent and well-defined.

This paper is structured as follows: Section 1 introduces the significance of the SRL task and provides an overview of our proposed approach. Section 2 provides background on PAS, contrasts general and biomedical domains, and reviews existing datasets. Section 3 discusses related work, emphasizing the role of entity knowledge in SRL. Section 4 details our proposed model, while Section 5 presents experimental results and analysis. Finally, Section 6 summarizes key contributions and suggests directions for future research.

II. Background

In NLP, Predicate-Argument Structure (PAS) represents how a predicate (the main verb of a sentence that describes an action or state) connects to its arguments (entities involved in that action or state). Each argument is assigned a semantic role, which explains its function in relation to the predicate.

For example, consider the sentence: "The TP53 gene encodes a tumor suppressor protein in human cells". Here, the PAS consists of the predicate "encode" and two arguments:

- Arg0: "the TP53 gene" (semantic role: gene or RNA)
- Arg1: "a tumor suppressor protein in human cells" (semantic role: gene product)

PAS is essential for extracting meaningful biomedical relationships, such as identifying which drugs interact with specific diseases or which proteins regulate particular biological processes. Therefore, understanding PAS helps machines capture the main meaning of a sentence. In general-domain NLP, several PAS-annotated datasets exist, such as FrameNet [9], VerbNet [10], and PropBank [11], with PropBank providing the most detailed argument structures.

In biomedical texts, PAS works differently from general-domain PAS, mainly because the roles of arguments change. As exemplified in Table I, the verb mutate has different meanings depending on the context. In general texts, it describes a change, focusing on who or what caused the change and the states before and after [12]. However, in biomedical texts, it describes genetic mutations, where arguments refer to the mutation site, the affected gene, and the resulting molecular and phenotypic changes [5].

Table I
BIOMEDICAL AND GENERAL ARGUMENT FRAMES OF THE PREDICATE MUTATE.

Biomedical PAS (PASBio)	**General PAS (PropBank)**
Mutate = Exhibit a phenotype • Arg0: Physical location where mutation happen (exon, intron) • Arg1: Mutated entity (gene) • Arg2: Changes at molecular level • Arg3: Changes at phenotype level	Mutate = Changing state, morphing • Arg0: Causer of change • Arg1: Entity undergoing mutation • Arg2: End state • Arg3: Start state

Because of such differences, specialized biomedical PAS datasets have been developed. Unlike general-domain PAS resources, which define argument structures for all verbs, biomedical PAS datasets focus only on key biomedical verbs, such as mutate, encode, express, catalyse,... However, researchers have not yet agreed on a standard list of biomedical predicates. Some of the most well-known biomedical PAS datasets are:

- BioProp: Comprising 1,635 sentences extracted from 500 biomedical papers, this dataset adopts general-domain PAS structures from PropBank [1]. Therefore, its argument roles are not fully adapted to biomedical texts.
- GREC: Containing 1,489 sentences from 677 biomedical abstracts, GREC improves upon BioProp by defining biomedical-specific argument roles [3]. However, it does not explicitly link predicates and arguments, instead treating them as independent elements.
- BioVerbNet : With 521 annotated sentences, BioVerbNet extends VerbNet by incorporating missing biomedical verbs [13]. However, it suffers from limited annotations and retains general-domain PAS structures, making it less suitable for biomedical SRL.
- PASBio+: PASBio+ is the most comprehensive dataset [6]. Unlike BioProp and BioVerbNet, PASBio+ defines argument frames specifically tailored for biomedical predicates [5]. Unlike GREC, PASBio+ explicitly links each argument to its predicate. PASBio+ is also larger than both BioProp and GREC, making it more effective for training biomedical SRL models.

In Semantic Role Labeling (SRL), selecting the right argument frameset is crucial for accurate sentence understanding. We choose PASBio+ for its biomedical-specific argument roles, explicit predicate-argument relationships, and larger dataset size, making it more effective for SRL training than other biomedical PAS resources.

III. RELATED WORKS

Initial efforts in biomedical Semantic Role Labeling (SRL) focused on adapting general-domain SRL techniques to biomedical corpora. BIOSMILE was among the earliest systems tailored for biomedical text, employing PropBank-style annotation [12] and training a maximum entropy model to classify semantic roles of biomedical predicates [14]. While BIOSMILE achieved reasonable performance, it remained constrained by a general-domain argument frameset (BioProp), limiting its applicability in biomedical contexts. A later adaptation expanded the predicate set from 30 to 82 verbs and incorporated syntactic parsing-based features, improving role classification but still struggling to generalize to unseen biomedical text due to its continued reliance on general-domain argument frames [15]. To overcome this limitation, BelSmile extended BIOSMILE by replacing the maximum entropy model with Conditional Random Fields (CRF) and incorporating domain-specific linguistic rules to refine SRL outputs, particularly for biological expression language extraction [16]. This approach introduced biomedical-specific constraints, reducing reliance on general-domain framesets; however, it remained rule-based, requiring extensive domain expertise and lacking scalability across diverse biomedical subdomains.

To improve efficiency in biomedical SRL, a resource-efficient SRL approach was introduced, leveraging Markov Logic Networks (MLN) for collective learning and integrating tree-pruning and argument candidate identification to reduce memory usage and training time [17]. This method enhanced accuracy by preventing duplicate and overlapping arguments; however, it struggled with less common argument structures, particularly those involving discourse markers or mistakenly pruned arguments.

Beyond these efforts, several studies have highlighted the critical role of Named Entity Recognition (NER) in SRL, particularly in refining argument classification. While NER alone does not capture full predicate-argument structures, it serves as a foundation for anchoring key arguments and refining semantic interpretations [18]. A systematic review emphasized that NER aids SRL by enhancing entity identification, which in turn improves predicate-argument role classification [19]. Biomedical SRL resources such as GREC [3], PASBio [5], and PASBio+ [6] integrate Named Entity (NE) categorization within SRL annotations, explicitly labeling entities such as DNA, proteins, and biological processes within semantic arguments. This explicit entity annotation ensures that machine learning models can leverage domain-specific entity types to improve SRL. Studies also demonstrate that named entity categorization aids in determining the types of phrases that fill specific semantic roles, reinforcing its necessity in accurate and domain-adapted SRL applications [3]. A recent study further revealed that 63

Advancements in deep learning-based SRL have increasingly sought to integrate NER knowledge. A study by [20] explored the integration of Entity Recognition (ER) as an auxiliary task to enhance Semantic Role Labeling (SRL) performance in low-resource, conversational domains. By combining multi-task learning with active learning, the proposed model leveraged shared semantic information between ER and SRL, where ER provides entity type cues that help regularize SRL predictions. Their experiments on Indonesian dialogue data demonstrated that multi-task learning not only improved SRL accuracy but also reduced the need for annotated data, showing the utility of ER in refining argument role classification through shared representation learning. A bidirectional LSTM-CRF model was designed to jointly perform NER and SRL in medical documents, leveraging biomedical NER embeddings to capture contextual relationships between predicates and arguments [8]. However, due to the absence of a large language model (LLM) architecture, the model struggled with ambiguities in role classification, as biomedical entities often exhibit complex interdependencies that classic deep learning architectures fail to capture.

While many recent models have explicitly incorporated syntax-based SRL approaches, they implicitly used NER knowledge. A biomedical SRL model trained on PASBio+ leveraged entity-centered SRL annotations to improve argument classification [21]. Additionally, the model integrated BioBERT as a contextual encoder, leveraging entity-aware embeddings to enhance semantic role representations in biomedical text. Further research has demonstrated that constituency parsing can serve as a structural foundation for SRL, with named entities often appearing as constituents within parse trees, thereby aiding predicate-argument extraction [22]. Moreover, its TreeCRF-based structured model followed NER span-detection methods, reinforcing the structural similarities between predicate-argument spans in SRL and entity spans in NER. These findings suggest that advancements in NER span-based modeling could benefit SRL, particularly in argument role assignment.

Given their close interdependence, [23] introduced a unified framework for low-resource sequence labeling that jointly learns NER and SRL, using a shared augmented language representation. By incorporating named entity information within the input format and adapting model parameters based on underperforming instances, the approach significantly improved SRL performance. These results highlight the value of entity cues in guiding role assignment, demonstrating that improvements in NER can

contribute meaningfully to more accurate SRL, particularly in low-resource settings.

Despite widespread recognition of NER's benefits in SRL, current methodologies have limitations (Table II). Specifically, they primarily incorporate entity knowledge indirectly, either via entity-centered SRL corpora or through structural similarities between NER and SRL spans. While some neural models integrate NER and SRL, they remain limited to conventional neural network architectures, lacking the representational power of modern LLMs. Recent work in 2025 has increasingly embraced model merging as a scalable, modular alternative to full fine-tuning, especially in sensitive domains like healthcare. For example, the PatientDx framework shows that merging pre-trained models can preserve patient privacy while improving performance on clinical prediction tasks [24]. Likewise, a domain-adaptive strategy demonstrates how merging general and biomedical LLMs enhances downstream performance without retraining [25]. The Modular Clinical SLM framework illustrates how pre-instruction tuning and task-specific alignment through merging can streamline adaptation for clinical use cases [26]. Complementing these, STAR method introduces spectral truncation and rescaling to mitigate instability during merging and optimize transferability [27]. Motivated by these advances, our work proposes a novel approach that leverages model merging to directly integrate NER with SRL in an LLM-based architecture. This enables more precise argument extraction in biomedical texts by leveraging entity-aware contextual representations, positioning our method at the forefront of emerging model integration techniques.

IV. Method

1. Foundation model selection

Our approach involves merging a baseline SRL model with a pre-existing NER model via confidence-based model merging. This merging strategy operates at the level of task vectors, which represent layer-wise weight differences from a shared base model. As a result, both constituent models (SRL and NER) must be derived from the same underlying architecture and initialization to ensure compatibility during the merging process. The publicly available NER model we leverage, achieving an F1 score of up to 93.04

While alternative biomedical language models such as PubMedBERT [29] and ClinicalBERT [30] have been introduced, they are not suitable for our merging mechanism in this context. While PubMedBERT has demonstrated strong performance across various biomedical NLP tasks, its direct application to SRL tasks within the biomedical domain hasn't been featured in the literature. ClinicalBERT, on the other hand, is fine-tuned on MIMIC-III clinical notes and is therefore optimized for processing clinical narratives, rather than scientific biomedical literature.

In contrast, BioBERT has been adopted and validated on biomedical SRL [21]. Moreover, the availability of high-quality, fine-tuned BioBERT-based NER models made it the suitable choice for our model merging setup without retraining an auxiliary model from scratch. We therefore selected BioBERT as the foundation model for our SRL approach.

2. Constituent models for merging

We use the NER model provided by BioBERT [7]. As for the SRL model, we fine-tune the BioBERT pre-trained model to identify semantic roles in biomedical text by leveraging contextual embeddings. A token-level classification layer with a softmax activation function is added to predict predicates and argument roles based on the model's final hidden states. The model is optimized using cross-entropy loss, which aligns predicted probabilities with true labels.

3. Model merging method

Given these two constituent models, the goal is to create a merged SRL model that leverages beneficial knowledge from the NER model to enhance SRL performance. Researchers typically develop models only for their main task and reuse pre-trained auxiliary models from other studies, which are usually shared without their original training data. Therefore, our study focuses on a merging solution that does not require training data for the auxiliary task (in this case, the NER task). The process must avoid retraining on the original NER datasets and instead rely on unlabeled samples for optimization.

The merging process is based on the concept of task vectors [31], which represent task-specific adaptations at both the model and layer levels. Each task vector is obtained by subtracting the weights of a pre-trained foundation model from those of a fine-tuned model. Formally, let $\theta_{BioBERT}$ denote the weights of the foundation model before fine-tuning (i.e., BioBERT), and let θ_{SRL} and θ_{NER} represent the weights of the fine-tuned SRL and NER models, respectively. The overall task vectors representing SRL and NER are specified in Equations (1):

$$\begin{aligned} T_{SRL} &= \theta_{SRL} - \theta_{BioBERT} \\ T_{NER} &= \theta_{NER} - \theta_{BioBERT} \end{aligned} \tag{1}$$

At the layer level, the task vector for the l-th layer, denoted as $T_{task}^{(l)}$, is computed similarly by subtracting the

Table II
BIOMEDICAL AND GENERAL ARGUMENT FRAMES OF THE PREDICATE MUTATE.

Theme	Works	Key Contributions	Limitations
General SRL Adapted to Biomedical	BIOSMILE [14]	Adapted PropBank-style SRL to biomedical verbs using MaxEnt models.	Relied on general-domain frames; limited generalization to unseen biomedical text.
	Extended BIOSMILE [15]	Expanded verb coverage, added syntactic parsing features.	Still used general-domain PAS frames; poor performance on novel biomedical inputs.
Biomedical SRL	BelSmile [16]	Introduced domain-specific linguistic rules; used CRFs.	Rule-based; lacked scalability; required domain expertise.
	MLN-based SRL [17]	Efficient inference using collective reasoning; avoided duplicate and overlapping arguments.	Struggled with less frequent structures (e.g., discourse markers).
NE in SRL corpora	GREC, PASBio, PASBio+ [3, 5, 6]	Provided SRL corpora with explicit entity annotations (e.g., DNA, protein).	May suffer from label granularity or corpus size constraints.
	GENEVA [28]	Showed 63% of argument roles are entity-based, confirming NER's importance.	Entity overlap still poses challenge.
Joint or Integrated NER-SRL Models	Multi-task ER-SRL [20]	Used ER as auxiliary task to regularize SRL in low-resource conversational data.	Low-resource setting; not LLM-based.
	BiLSTM-CRF (Joint) [8]	Joint model for NER and SRL using medical NER embeddings.	Lacked LLM backbone; struggled with complex biomedical interdependencies.
	TreeCRF model [22], Syntax-aware SRL [21]	Incorporated syntactic and span-based NER insights into SRL.	Indirect NER use; relies on parse tree quality.
	Unified ER+SRL framework [23]	Joint low-resource learning using shared representations and input augmentation.	Not optimized for high-resource biomedical texts.

weight matrix of the foundation model at layer l, $\theta^{(l)}_{BioBERT}$, from that of the fine-tuned model ($\theta^{(l)}_{SRL}$ or $\theta^{(l)}_{NER}$).

$$\begin{aligned} T^{(l)}_{SRL} &= \theta^{(l)}_{SRL} - \theta^{(l)}_{BioBERT} \\ T^{(l)}_{NER} &= \theta^{(l)}_{NER} - \theta^{(l)}_{BioBERT} \end{aligned} \tag{2}$$

The layer-wise task vectors in Equations (2) capture the localized layer-specific modifications introduced during fine-tuning, enabling a more granular integration of task-specific knowledge when merging models. The merging of the SRL model with the NER model is achieved by combining their respective task vectors $T^{(l)}_{SRL}$ and $T^{(l)}_{NER}$ at each layer using a weighted sum. This process forms a unified task vector $T^{(l)}_{Merged}$ for the l-th layer of the merged model, which captures both semantic role labeling and named entity recognition knowledge:

$$T^{(l)}_{Merged} = T^{(l)}_{SRL} + \lambda^{(l)} T^{(l)}_{NER} \tag{3}$$

In Equation (3), $\lambda^{(l)}$ is the layer-specific coefficient for the NER task vector at the l-th layer, determining its contribution to each layer of the merged SRL model ($0 \leq \lambda^{(l)} \leq 1$). The SRL task vector's coefficient is 1, as it is the primary task in our study. Lower layers, capturing general features, tend to have lower coefficients, while higher layers, capturing task-specific features, receive higher coefficients [32].

The merged model's weights for each layer are computed by Equation (4):

$$\theta^{(l)}_{Merged} = \theta^{(l)}_{BioBERT} + T^{(l)}_{Merged} \tag{4}$$

Figure 1 illustrates an example of this merging process for layers 1 and 2. This layer-wise approach ensures that each layer incorporates the right amount of NER information while preserving the overall structure of the SRL model. A critical challenge is determining the optimal layer-wise coefficients $\lambda^{(l)}$ to effectively leverage NER knowledge for improving SRL performance. This is particularly difficult since the auxiliary model (in our case, the NER model) is typically available through other studies, but its original training data is often inaccessible. To overcome this, we propose Maximum-Confidence optimization.

4. Maximum-Confidence optimization

Entropy has been shown to correlate strongly with prediction loss. Grouping test samples by entropy levels reveals a consistent trend where higher entropy aligns with increased loss, further supported by a Spearman correlation of 0.87 between entropy and loss, confirming entropy as a reliable proxy for loss minimization [32]. These findings suggest that entropy minimization can effectively guide optimization in multi-task learning when access to original training data is unavailable.

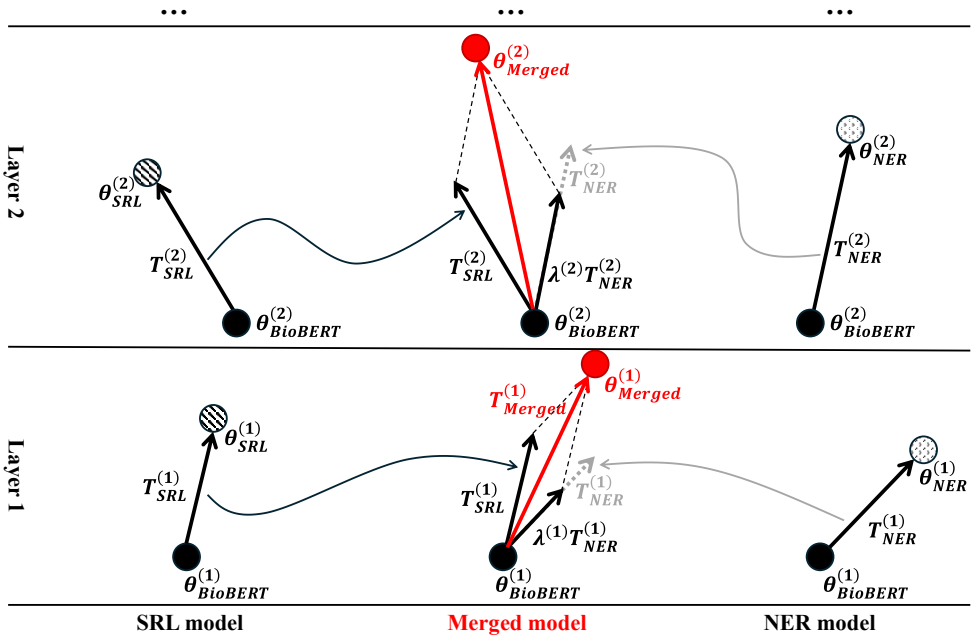

Figure 1. An example of layer-wise merging, showing how the NER task vector contributes to layers 1 and 2 of the merged SRL model.

However, our study finds that prediction confidence exhibits an even stronger correlation with loss. Specifically, when evaluating the NER model on the JNLPBA corpus [33], we observe a Spearman correlation of -0.91 between confidence and loss, surpassing the 0.87 correlation between entropy and loss reported by [32]. These results suggest that higher confidence consistently corresponds to lower prediction loss, reinforcing confidence as a more reliable indicator of model performance.

Given this stronger correlation, confidence maximization presents a more effective optimization objective than entropy minimization. Maximizing confidence ensures that the model produces more certain predictions, which typically correlate with higher accuracy. By directly encouraging high-confidence predictions, optimization can better reduce loss, leading to improved multi-task learning outcomes. Formally, the optimization function is defined as:

$$\underset{\{\lambda^{(l)}\}_{l=1}^{L}}{\operatorname{argmin}} \left(\sum_{x_i \in D_{SRL}} \mathcal{L}_{SRL}\left(x_i; \{\lambda^{(l)}\}_{l=1}^{L}\right) + \beta \sum_{x_k \in D_{NER}} (1 - C_{NER}(x_k)) \right) \quad (5)$$

In Equation (5):

- L is the total number of model layers.
- D_{SRL} is the training data for the main task (SRL).
- D_{NER} is the unlabeled test dataset for the auxiliary task (NER).
- $C_{NER}(x_k)$ represents the NER prediction confidence for the test sample x_k.

The layer-wise coefficients $\lambda^{(l)}$ are optimized for the main task based on loss, as SRL-annotated training data is available, while Maximum-Confidence optimization is used for the auxiliary task to compensate for the absence of original NER training data. The hyperparameter β, empirically set to 0.5, regulates the relative importance of the NER task in the optimization.

Figure 2 summarizes our overall framework. The workflow begins by computing task vectors that capture how the NER and SRL models differ from the shared BioBERT foundation. Each model's adaptations are extracted layer by layer, forming task-specific representations. These vectors are then passed to the *Layer-wise merging* component, where they are combined using adjustable coefficients that control the influence of NER on the SRL model. Next, in the *Merged weight computation* component, the merged task vectors are added back to the BioBERT base to generate the parameters of a new, integrated model. This merged model is evaluated using the SRL loss function, which compares its predictions against gold-standard SRL annotations. Simultaneously, the model's NER output confidence is measured. Finally, the *Coefficient tuning* component adjusts the merging coefficients based on both the SRL loss and NER confidence. These updated coefficients

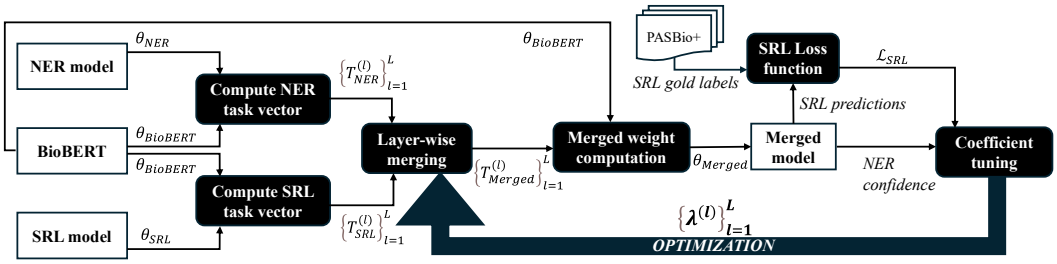

Figure 2. Overview of the confidence-based model merging framework.

are fed back into the merging process, completing the loop and iteratively refining the merged model's performance.

V. Experiments

1. Experimental data and setting

We assess NER models using the JNLPBA corpus [33]. For SRL evaluation, multiple datasets were reviewed, though most were found lacking. We ultimately chose PASBio+ [6], a specialized biomedical corpus featuring 35 core predicates. PASBio+ was selected based on several strengths. It is built specifically for the biomedical domain, ensuring task relevance. It offers clearly annotated argument structures grounded in domain-specific verb semantics. It provides gold-standard predicate–argument structures, enabling rigorous evaluation of SRL outputs. Additionally, it has been recently expanded to include named entity annotations, supporting both SRL and NER assessments.

- Biomedical Focus: Drawn exclusively from domain-specific literature, PASBio+ aligns closely with our study's subject matter.
- Adequate Scale: With more than 15,000 annotated sentences, the dataset provides sufficient volume for training. In contrast, smaller sets like BioVerbNet [13] and GREC [3], each under 1,500 sentences, risk data sparsity and performance degradation.
- Specialized Argument Structures: The argument structures, designed by domain experts, reflect biomedical semantics in detail. This is a key advantage over resources such as [1] or BioVerbNet [13], which adapt general linguistic frames to biomedical texts.

SRL evaluation: We evaluate the performance of our entity-aware SRL model, which is optimized via confidence-based model merging (Section IV), on biomedical text. The model is compared against the following models, all evaluated on the same dataset:

- A baseline SRL model without NER integration.
- An entity-aware SRL model using traditional multitask learning.
- Another entity-aware SRL model that uses entropy-based Adamerging to integrate NER knowledge [32].
- SRL models that incorporate syntactic knowledge instead of NER, achieving state-of-the-art results on PASBio+ [21].

This setup enables a comparison not only between different methods of integrating NER knowledge but also between NER-based and syntax-based enhancements to SRL performance on the same biomedical dataset.

2. Experimental result and analysis

We evaluate our entity-aware SRL model, which integrates NER knowledge via confidence-based model merging. We compare our model against four baselines introduced in Section III. Experimental results show that our method outperforms the baselines in both SRL performance and training speed.

Table V presents the SRL and NER F1 scores across models. Our entity-aware SRL model achieves the highest SRL F1 and accuracy, indicating superior overall performance and predictive correctness across all sentences. In addition, it attains strong precision (93.86%) and recall (94.31%), reflecting both accurate role identification and comprehensive argument span coverage.

Although the NER F1 score slightly drops compared to multitask learning, the performance remains adequate. This suggests that while the merged model does not preserve the full standalone NER performance, it retains sufficient entity-related information to improve SRL accuracy, as evidenced by: (i) the model achieving the improved SRL F1 score, and (ii) maintaining a moderate NER F1 score (76.00%), indicating that a substantial amount of NER knowledge remains usable after merging.

The models were tested on a server with an Intel(R) Xeon(R) CPU @ 2.00 GHz, a Tesla T4 GPU with 15 GB VRAM, and 12.7 GB of system RAM. Table III compares training time. Our method is also more efficient, with training time nearly halved compared to multitask learning and lower than entropy-based merging. This improvement

is attributed to the computational simplicity of confidence score calculation over entropy estimation. Additionally, in terms of average processing time per sentence, our model achieves the lowest latency at 0.685 seconds per sentence, compared to 0.768 for multitask learning and 0.826 for entropy-based merging. This demonstrates that our confidence-based merging method is not only efficient in total training time but also scalable for real-time or large-scale applications where per-sentence inference speed matters.

Table III
TRAINING TIME AND INFERENCE SPEED COMPARISON (IN SECONDS).

Metric	Multitask SRL+NER	Entropy Merging	Confidence Merging
Training time	7136	3539	2940
AVG time per sentence	0.768	0.826	0.685

We examined the effect of NER knowledge across individual predicates. For a small subset of predicates, performance declined after merging, as shown in Table IV.

Table IV
PREDICATES WITH F1 DECLINE POST-MERGING.

Predicate	F1 after	F1 before	after - before
result	0.8608	0.9898	-0.1290
lead	0.4822	0.7727	-0.2905
abolish	0.7702	0.7806	-0.0104
begin_2	0.9657	0.9787	-0.0130
translate_3	0.9806	0.9832	-0.0026
begin_1	0.9893	0.9904	-0.0011

Importantly, according to the PASBio predicate–argument frameset [5], each predicate is associated with two to five arguments, including both entity-based and non-entity-based arguments. Figure 3 illustrates an example of this structure. All the predicates in Table IV lack entity-based arguments according to the GENIA taxonomy. Figure 4 illustrates this using the predicate *result*, whose arguments are not categorized as biomedical entities, leading NER information to contribute noise rather than guidance.

To better understand how entity knowledge contributes to SRL improvement, we grouped predicates (excluding those in Table IV) into four tiers based on their baseline SRL F1 scores (i.e., the F1 score of the SRL model before merging with the NER model): (a) below 50%, (b) 50% to below 80%, (c) 80% to below 90%, and (d) 90% and above (Fig 5). This partitioning reflects the assumption that lower-performing predicates generally have more room for improvement and are thus more likely to benefit from the integration of NER knowledge. We then calculated the proportion of entity-based arguments. For example, the predicate *alter* in Figure 3 has five arguments, three of which are entity-based, resulting in a proportion of $3/5 = 0.6$.

Figure 5 visualize the relationship between F1 gain and the proportion of entity-based arguments across these four tiers. A consistent pattern emerges: predicates with higher entity-based argument proportions, indicated by taller yellow columns, yield greater SRL performance improvement, as reflected by a larger gap between the blue lines (SRL F1 after merging with the NER model) and red lines (SRL F1 before merging with the NER model).

Experimental results also show that predicates with high ambiguity in entity class assignments tend to exhibit smaller F1 improvements, even compared to predicates with a lower proportion of entity-based arguments. This is likely because such ambiguity weakens the ability of entity types to signal semantic roles or distinguish between different arguments.

One type of ambiguity occurs when a single entity-based argument is annotated with multiple entity classes. For example, predicate *delete*'s `arg1` (the entity being removed) spans five different entity classes, resulting in lower F1 gains than predicates *disrupt*, *eliminate*, and *decrease_2*, which have the same proportion of entity-based arguments (Fig 5a). Similarly, predicate *skip*'s `arg2` is linked to both DNA and RNA classes and improves less than predicates *translate_2* and *decrease_1* (Fig 5d).

Another form of ambiguity arises when multiple arguments of the same predicate share the same entity class, making it difficult to leverage NER knowledge to distinguish between them. In predicate *recognize*, both `arg0` and `arg1` are associated with the protein class, leading to limited improvement (Fig 5a). In predicate *encode*, `arg0` (gene) and `arg1` (protein) fall under the same GENIA class and are often referred to using the same term in biomedical texts, reducing the benefit of NER (Fig 5b). Similarly, predicates *splice_2* and *truncate* involve multiple arguments tied to the RNA class, which hampers disambiguation and yields modest F1 gains (Fig 5c and d).

A distinct case involves entity-based arguments that only partially include an entity mention. For instance, predicate *block*'s `arg1` may describe a biological process that references a gene or protein, but the argument as a whole is not an entity. This limits the contribution of NER to SRL in such instances (Fig 5d).

In some cases, F1 gains are modest despite a high percentage of entity-based arguments due to ceiling effects, where performance is already near optimal. Predicate *tran-*

Table V
SRL AND NER PERFORMANCE METRICS ACROSS MODELS.

Metric	Baseline SRL	Multitask SRL+NER	Syntax-aware SRL	Entropy-based SRL	Confidence-based SRL
SRL Precision	82.70	92.41	89.52	44.38	93.86
SRL Recall	86.80	93.01	91.24	66.62	94.31
SRL F1	83.80	92.71	90.37	53.27	94.08
SRL Accuracy	73.46	86.41	82.43	66.62	88.83
NER F1	–	80.91	–	76.03	76.00

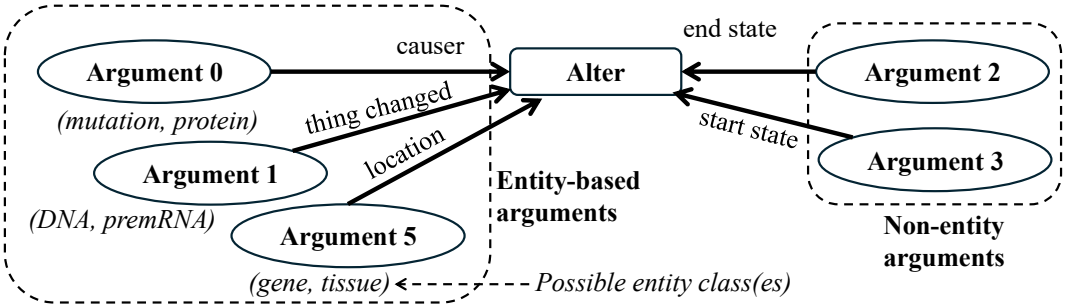

Figure 3. Example of PASBio predicate-argument structure with both entity-based and non-entity-based arguments.

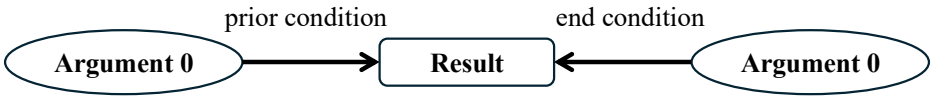

Figure 4. An example predicate (result) with no entity-based arguments.

scribe (Fig 5b) illustrates this: all arguments are entity-based, yet the baseline F1 is already very high.

Conversely, predicates like *proliferate* achieve greater F1 gains than predicates with a higher proportion of entity-based arguments. Only one of *proliferate*'s three arguments is entity-based. However, it appears as a predicate in 173 sentences, 49 of which contain only its `arg0` (i.e., 100% of the mentioned arguments in those cases are entity-based), and another 71 contain only `arg0` and `arg1` (i.e., 50% of the mentioned arguments are entity-based). This elevated frequency increases the practical influence of the entity-based argument beyond what the raw $1/3$ proportion suggests (Fig 5c).

Taken together, these findings suggest that entity-aware model merging is particularly effective when predicates are structurally aligned with biomedical entity distinctions, and that confidence-based merging provides a computationally efficient and semantically sound mechanism for integration.

VI. CONCLUSION AND FUTURE DIRECTIONS

This paper introduced a novel approach to enhancing Semantic Role Labeling (SRL) for biomedical texts by directly merging SRL and NER models using confidence-based merging optimization. The approach is specifically tailored for biomedical SRL, where domain-specific entities are prevalent and tend to align well with predicate-argument structures, making the integration of NER knowledge particularly effective . Our method significantly improves SRL performance on the PASBio+ dataset, outperforming state-of-the-art multitask and syntax-aware methods while requiring less training time. Empirical results confirm that biomedical entity knowledge plays a critical role in improving SRL, especially for predicates with high proportions of entity-based arguments. However, F1 gains vary depending on argument ambiguity, entity class overlap, and the completeness of entity span coverage.

We acknowledge that when predicates exhibit argument overlap in entity classes or span multiple semantic types, entity-based guidance may be less effective. Addressing such ambiguity is an important direction for future enhancement. For example, future works could explore fine-grained entity typing to replace coarse biomedical classes with more expressive representations that capture nuanced functional roles. Exploring adaptive entity representations

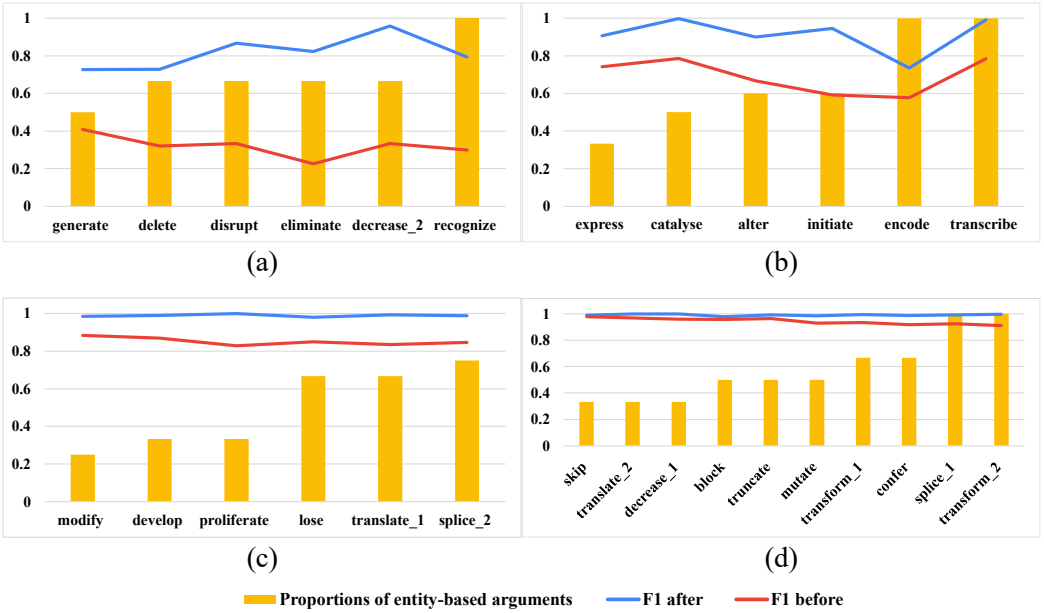

Figure 5. F1 improvements by predicate, grouped by baseline SRL F1 tiers: (a) below 50%, (b) 50% to below 80%, (c) 80% to below 90%, and (d) 90% and above. Within each tier, predicates are sorted by the proportion of entity-based arguments (represented by yellow columns).

or hierarchical typing strategies is also a promissing future approach to mitigate the limitations of argument ambiguity . Leveraging graph-based representations, such as AMR (Abstract Meaning Representation) or UMLS knowledge graphs, could further contextualize arguments and support better predicate-argument alignment.

Moreover, our merging strategy could be extended to integrate additional linguistic features, such as coreference chains or discourse relations, to enable deeper semantic understanding.

Lastly, as although our current focus is biomedical, future efforts may be to improve generalizability beyond domain-specific predicates. This involves applying our model merging method in cross-domain SRL settings to enhance SRL performance in domains with limited annotation.

REFERENCES

[1] W.-C. Chou, R. T.-H. Tsai, Y.-S. Su, W. K., T.-Y. Sung, and W.-L. Hsu, "A semi-automatic method for annotating a biomedical proposition bank," in *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, 2006, pp. 5–12.

[2] M. Miwa, P. Thompson, J. McNaught, D. B. Kell, and S. Ananiadou, "Extracting semantically enriched events from biomedical literature," *BMC Bioinformatics*, vol. 13, no. 1, p. 108, 2012.

[3] P. Thompson, P. Cotter, S. Ananiadou, J. Mcnaught, S. Montemagni, A. Trabucco, and G. Venturi, "Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[4] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, "Deep semantic role labeling: What works and what's next," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 473–483.

[5] T. Wattarujeekrit, P. K. Shah, and N. Collier, "Pasbio: Predicate-argument structures for event extraction in molecular biology," *BMC Bioinformatics*, vol. 5, no. 1, p. 155, 2004.

[6] H.-D. Tuan-Nguyen, H.-S. Pham, and V.-T. Hoang, "A semi-automatic approach to biomedical semantic role corpus construction," *Science and Technology Development Journal - Natural Sciences*, vol. 6, no. 2, pp. 2083–2094, 2022.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[8] H.-D. Tuan-Nguyen, L.-T. Tran-Tien-Loi, and V.-H. Le-Dinh, "A deep-learning model for semantic role labelling in medical documents," *Science and Technology Development Journal - Natural Sciences*, vol. 5, no. 2, pp. 1032–1039, 2021.

[9] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics*, 1998, pp. 86–90.

[10] M. S. Palmer and K. K. Schuler, "Verbnet: A broad-coverage, comprehensive verb lexicon," University of Pennsylvania, Computer and Information Science Dept., 2000 South 33rd St., Philadelphia, PA, United States, Tech. Rep., 2005.

[11] S. Pradhan, J. Bonn, S. Myers, K. Conger, T. O'gorman, J. Gung, K. Wright-Bettner, and M. Palmer, "Propbank comes of age—larger, smarter, and more diverse," in *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, 2022, pp. 278–288.

[12] P. Kingsbury and M. Palmer, "From treebank to propbank,"

in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, M. G. Rodríguez and C. P. S. Araujo, Eds. European Language Resources Association (ELRA), 2002.

[13] O. Majewska, C. Collins, S. Baker, J. Björne, S. W. Brown, A. Korhonen, and M. Palmer, "Bioverbnet: A large semantic-syntactic classification of verbs in biomedicine," *Journal of Biomedical Semantics*, vol. 12, no. 1, 2021.

[14] R. T.-H. Tsai, W.-C. Chou, Y.-C. Lin, C.-L. Sung, W. Ku, T.-Y. Sung, and W.-L. Hsu, "Biosmile: Adapting semantic role labeling for biomedical verbs," in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, 2006, pp. 57–64.

[15] R. T.-H. Tsai, W.-C. Chou, Y.-S. Su, Y.-C. Lin, C.-L. Sung, H.-J. Dai, I. T.-H. Yeh, W. Ku, T.-Y. Sung, and W.-L. Hsu, "Biosmile: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features," *BMC Bioinformatics*, vol. 8, no. 1, p. 325, 2007.

[16] P.-T. Lai, Y.-Y. Lo, M.-S. Huang, Y.-C. Hsiao, and R. T.-H. Tsai, "Belsmile: A biomedical semantic role labeling approach for extracting biological expression language from text," *Database*, 2016.

[17] R. T.-H. Tsai and P.-T. Lai, "A resource-saving collective approach to biomedical semantic role labeling," *BMC Bioinformatics*, vol. 15, no. 1, pp. 160–171, 2014.

[18] F. Eckert and M. Neves, "Semantic role labeling tools for biomedical question answering: a study of selected tools on the bioasq datasets," in *Proceedings of the 6th BioASQ Workshop A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering*, 2018, pp. 11–21.

[19] A. D. P. Ariyanto, D. Purwitasari, and C. Fatichah, "A systematic review on semantic role labeling for information extraction in low-resource data," *IEEE Access*, vol. 12, pp. 57 917–57 946, 2024.

[20] F. Ikhwantri, S. Louvan, K. Kurniawan, B. Abisena, V. Rachman, A. F. Wicaksono, and R. Mahendra, "Multi-task active learning for neural semantic role labeling on low resource conversational corpus," in *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 2018, pp. 43–50.

[21] H.-D. Tuan-Nguyen, T. Duong Luu, and Q. Duy Huynh, "A syntax-aware deep-learning model for biomedical semantic role labelling," *Science and Technology Development Journal - Natural Sciences*, vol. 7, no. 4, pp. 2750–2762, 2023.

[22] W. Liu, S. Yang, and K. Tu, "Structured mean-field variational inference for higher-order span-based semantic role," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 918–931.

[23] S. S. S. Das, H. Zhang, P. Shi, W. Yin, and R. Zhang, "Unified low-resource sequence labeling by sample-aware dynamic sparse finetuning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6998–7010.

[24] J. G. Moreno, J. Lovon-Melgarejo, M. Rick, Robin-Charlet, C. Damase-Michel, and L. Tamine, "Patientdx: Merging large language models for protecting data-privacy in healthcare," in *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*. Albuquerque, New Mexico: Association for Computational Linguistics, May 2025, pp. 1–11, retrieved June 1, 2025. [Online]. Available: https://aclanthology.org/2025.cl4health-1.1/

[25] T. Rousset, T. Kakibuchi, Y. Sasaki, and Y. Nomura, "Merging language and domain specific models: The impact on technical vocabulary acquisition," February 2025, preprint on arXiv.

[26] J.-P. Corbeil, A. Dada, J.-M. Attendu, A. B. Abacha, A. Sordoni, L. Caccia, F. Beaulieu, T. Lin, J. Kleesiek, and P. Vozila, "A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment," May 2025, preprint on arXiv.

[27] Y.-A. Lee, C.-Y. Ko, T. Pedapati, I.-H. Chung, M.-Y. Yeh, and P.-Y. Chen, "Star: Spectral truncation and rescale for model merging," in *Proceedings of the 2025 Conference of the North, Central and South American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Albuquerque, New Mexico: Association for Computational Linguistics, April 2025, pp. 496–505, retrieved June 1, 2025. [Online]. Available: https://aclanthology.org/2025.naacl-short.42/

[28] T. Parekh, I.-H. Hsu, K.-H. Huang, K.-W. Chang, and N. Peng, "Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 3664–3686.

[29] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, January 2022. [Online]. Available: https://doi.org/10.1145/3458754

[30] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," April 2019, preprint.

[31] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," in *Proceedings of the Eleventh International Conference on Learning Representations*, February 2023.

[32] E. Yang, Z. Wang, L. Shen, S. Liu, G. Guo, X. Wang, and D. Tao, "Adamerging: Adaptive model merging for multi-task learning," in *Proceedings of the International Conference on Learning Representations (ICLR 2024)*, October 2024. [Online]. Available: https://arxiv.org/abs/2310.02575

[33] M.-S. Huang, P.-T. Lai, R. T.-H. Tsai, and W.-L. Hsu, "Revised jnlpba corpus: A revised version of biomedical ner corpus for relation extraction task," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, 2024, pp. 73–78.

Hoai-Duc Tuan-Nguyen earned his B.Sc. in Information Systems in 2005 and his M.Sc. in Computer Science in 2009 from the same university, where he is currently pursuing a Ph.D. Now he is a lecturer in the Faculty of Information Technology at the University of Science, Vietnam National University, Ho Chi Minh City. His research interests include deep learning, natural language processing, and explainable artificial intelligence.

Email: tnhduc@fit.hcmus.edu.vn