Iterative Silver-Label Refinement for Biomedical Temporal Information Extraction

Chan Lee and Shaikh Arifuzzaman
Department of Computer Science
University of Nevada, Las Vegas
Las Vegas, NV, USA
Emails: leec56@unlv.nevada.edu, shaikh.arifuzzaman@unlv.edu

***Abstract*—Temporal information extraction is essential for biomedical text mining, where identifying events and their temporal relationships can support the interpretation of experimental findings, disease progression, treatment effects, and scientific evidence across the literature. Despite its importance, biomedical temporal information extraction remains challenging because high-quality annotated corpora are scarce, expensive to construct, and often insufficient for training modern transformer-based models. Moreover, temporal extraction models developed for other domains, such as news or clinical narratives, often exhibit limited transferability to biomedical literature due to differences in terminology, discourse structure, and label distributions.**

This paper presents an iterative silver-label refinement framework for adapting temporal information extraction models to biomedical literature without requiring comprehensive manual annotation. Starting from a temporal extraction model trained on news-domain data, the proposed approach applies cycles of automatic silver labeling, targeted human correction, and model retraining to progressively improve both event recognition and temporal relation classification. The framework supports multi-class temporal relation classification across six relation types and includes a practical corpus-conversion workflow for adapting existing temporal annotations to transformer-based architectures.

Experimental results show that both named entity recognition and temporal relation extraction improve meaningfully across refinement iterations. The study further demonstrates that the ratio of corrected instances to silver-labeled instances plays a critical role in successful domain adaptation, particularly under substantial source-target label distribution shifts. These findings highlight the value of data-centric refinement strategies for low-resource biomedical NLP and suggest that iterative silver-label correction can provide a practical pathway for adapting temporal information extraction systems to specialized scientific domains.

***Index Terms*—temporal information extraction, biomedical NLP, domain adaptation, silver labeling, transformer models, relation extraction**

I. Introduction

Temporal information extraction is a critical capability for biomedical text mining because scientific findings are often meaningful only when events are placed in temporal context. In biomedical literature, temporal information can help identify drug interaction timelines in pharmacology, reconstruct disease progression patterns in epidemiology, and accelerate protocol synthesis in systematic reviews. These applications require more than detecting isolated biomedical entities. They require identifying events, temporal expressions, and the relations that determine how events unfold, overlap, precede, or contain one another.

Despite its importance, temporal information extraction remains difficult to deploy in biomedical research literature. The primary obstacle is not only model architecture, but also the scarcity and quality of annotated data. Modern Transformer-based models [1], [2] have become the dominant architecture for natural language processing (NLP), yet many foundational temporal extraction datasets were created before these architectures became standard. As a result, existing resources often require conversion, normalization, or redesign before they can be used effectively with current modeling pipelines. More importantly, the major temporal corpora are concentrated in domains other than biomedical research literature, leaving a substantial gap between available training data and the target application domain.

Temporal Corpora and the Biomedical Gap. The foundational resources for temporal information extraction are primarily concentrated in the news and clinical domains. In the news domain, the TimeML standard [5] established a widely used annotation framework by defining TIMEX3 annotations for temporal expressions, EVENT annotations for event mentions, and TLINK annotations for temporal relationships based on Allen's interval algebra [6]. The TimeBank corpus [7] provided 183 annotated newswire documents and became a central benchmark for temporal extraction research. Subsequent efforts extended this foundation by increasing annotation density [8], simplifying temporal relation schemes [9], and providing standardized shared-task benchmarks through the TempEval evaluations [10], [11].

In the clinical domain, the THYME project [12] adapted TimeML-style temporal annotation to clinical narratives, while the i2b2 2012 Temporal Relations Challenge [13] provided another major corpus for clinical temporal relation extraction. These resources have shaped much of the clinical temporal extraction literature. Alfattni et al. [14] systematically examined 51 studies on clinical temporal relation extraction and found that the field relies almost exclusively on these two resources. However, both corpora require institutional data use agreements, which restricts accessibility, reproducibility, and broader experimentation.

For biomedical research literature, the resource gap is more severe. No annotated corpus currently exists for temporal relation extraction in biomedical research articles. Although the BioNLP shared tasks [15], [16] advanced biomedical

event extraction, they focused on event arguments rather than temporal ordering. This distinction is important: extracting that a biomedical event occurred is not equivalent to determining when it occurred, how long it persisted, whether it preceded another event, or whether it temporally contained another process. Consequently, existing biomedical event resources do not directly solve the temporal relation extraction problem.

The gap cannot be closed simply by transferring models trained on clinical or news-domain corpora. Biomedical research articles differ substantially from both newswire and clinical notes in vocabulary, temporal expressions, discourse structure, and event semantics. News articles often describe externally observable events in chronological narrative form, while clinical notes frequently describe patient histories, diagnoses, procedures, and care timelines. Biomedical research articles, by contrast, describe experimental designs, biological mechanisms, interventions, measurements, observations, and conclusions, often using dense scientific language and nonlinear discourse. These differences create substantial source-target mismatch and make direct cross-domain transfer unreliable.

The Annotation Bottleneck. Constructing a fully manually annotated biomedical temporal corpus would be expensive, time-consuming, and difficult to scale. Temporal annotation requires not only identifying events and temporal expressions, but also resolving complex semantic relations such as precedence, overlap, containment, and duration. In biomedical text, these relations may be implicit, distributed across clauses, or embedded in experimental descriptions, making annotation challenging even for trained annotators.

This work addresses the annotation bottleneck through iterative silver-label refinement. Starting from a temporal extraction model trained on a source-domain corpus, we automatically generate silver labels for biomedical text, selectively correct them, and retrain the model. Through repeated cycles of automatic labeling, targeted correction, and retraining, the model progressively adapts to biomedical literature while requiring substantially less manual effort than full corpus annotation. This data-centric strategy reframes domain adaptation as an iterative refinement problem, which is especially important when source and target domains exhibit different label distributions and temporal semantics.

Contributions. The contributions of this study are summarized as follows:

1) **First transformer-compatible temporal dataset for biomedical research literature.** We develop a temporal extraction dataset for biomedical research articles covering both temporal entity recognition and relation extraction, addressing a domain where no such resource previously existed to the best of our knowledge.
2) **Iterative silver-label refinement for cross-domain adaptation.** We introduce and evaluate an iterative refinement methodology that adapts temporal extraction from the news domain to biomedical text through cycles of automatic labeling, targeted correction, and retraining. The results demonstrate that the ratio of corrections to silver labels is a critical factor governing adaptation success.
3) **TimeML-to-transformer conversion workflow.** We provide a workflow for translating existing temporal corpora from XML-based TimeML-style formats into transformer-compatible training formats, enabling modern neural architectures to leverage legacy temporal annotation resources.
4) **Multi-class biomedical temporal relation classification.** We formulate temporal relation extraction as a six-class classification problem in biomedical text and identify a systematic DURING/INCLUDES distribution shift between source and target domains, highlighting broader challenges for cross-domain temporal annotation and model transfer.

II. Related Work

A. Transformer Models for Temporal Extraction

The application of transformer models to temporal relation extraction has yielded performance improvements but adoption remains limited. Domain-specific variants such as BioBERT [17], SciBERT [18], and ClinicalBERT [19], pretrained on biomedical and clinical corpora respectively, have shown improvements over general-domain models on biomedical NLP tasks. However, the major temporal corpora are annotated in formats that are not directly compatible with the token-level and sequence-level inputs that transformer models require, and no established conversion workflow exists for translating these annotations into transformer-compatible formats.

Lin et al. [20] applied BERT to clinical temporal relation extraction on the THYME corpus but focused exclusively on binary CONTAINS classification with gold entity annotations. Han et al. [21] combined deep neural networks with structured SVMs for temporal relation extraction. Neither work addressed format conversion from legacy temporal annotations to transformer-compatible training data, nor did they target biomedical research literature.

B. Silver Labeling and Domain Adaptation

Silver labeling refers to the automatic generation of training labels, as opposed to gold labels produced by human annotators. The TempEval-3 shared task [22] introduced the use of silver data for temporal annotation in the news domain, creating approximately 500,000 tokens of data by merging outputs from multiple temporal annotation systems. Self-training provides a framework for improving models using their own predictions. Lin et al. [23] applied this approach with recurrent neural networks to clinical temporal relation extraction, achieving notable results for both in-domain and cross-domain evaluation prior to the widespread adoption of transformers.

Zhao et al. [24] demonstrated that distant supervision can generate silver labels for temporal relation extraction, enabling zero-shot and few-shot transfer to new domains. In the biomedical domain, Fries et al. [25] applied weak

supervision with medical ontologies for clinical entity classification, demonstrating that automatically generated labels can approach human annotation quality. The viability of iterative self-labeling for NLP was established by foundational bootstrapping methods [3], [28], [31], though these approaches face semantic drift challenges [29], [30]. While these prior approaches have demonstrated the viability of automatic annotation for temporal tasks, none has addressed cross-domain adaptation for multi-class temporal relation classification in biomedical research literature.

III. Problem Formulation and Annotation Framework

A. Task Definition

Temporal information extraction consists of two interconnected tasks. First, named entity recognition (NER) takes input text and identifies event and time expression spans at the token level. Events include occurrences, actions, states, and processes that can be temporally located. Time expressions consist of phrases denoting points, duration, or intervals on a timeline [5]. The output is a labeled sequence indicating entity boundaries and types.

Second, relation extraction (RE) classifies the temporal relationship between pairs of entities as identified by the NER model. The model considers candidate pairs, particularly event-time or event-event pairs, and predicts the temporal relation between them. The output is a set of labeled pairs, each consisting of two entity spans and a relation type indicating their temporal connection.

These tasks are executed sequentially where errors in NER propagate down the pipeline to RE. Any missed entities result in missing relations, and incorrect boundaries lead to misclassified relations. To isolate each task's performance from these cascading errors, both tasks are evaluated independently using gold labels.

B. Temporal Annotation Framework

The annotation framework follows TimeML annotation scheme [5] adapted for practical implementation with transformer-based models. The two entity types recognized are events and time expressions, corresponding to TimeML's EVENT and TIMEX3 entities, respectively. The temporal relation is represented as TLINK connecting entity pairs within the sentence.

For the NER training dataset, annotations take the form of a BIO tagging scheme, which stands for Begin, Inside, and Outside tags [32] assigned to each token. The B tag marks the first token of an entity span, the I tag marks continuation tokens, and the O tag designates tokens outside any entity. Each tag is combined with an entity type, yielding labels such as B-EVENT, I-EVENT, B-TIMEX3, I-TIMEX3, and O. This scheme allows multi-token entities to be captured while distinguishing adjacent entities of the same type.

For RE training, annotations consist of labeled entity pairs. Given a sentence with identified entity spans, each event-time pair serves as a classification instance. Special tokens are inserted at entity boundaries to mark the spans. The model predicts one of six relation types: BEFORE, AFTER, BEGINS, ENDS, INCLUDES, and DURING. INCLUDES and DURING both represent temporal containment but from opposite perspectives. As a convention, the entity order is fixed with the event as the first entity and the time expression as the second. As discussed in the results, annotators from different domains applied INCLUDES and DURING inconsistently for similar patterns. Figure 1 illustrates the NER and RE pipeline with a biomedical example sentence.

The six relation types were derived from TimeBank's original labels by merging types with subtle distinctions that were difficult to annotate consistently or had too few training examples. Specifically, IAFTER/IBEFORE were merged into AFTER/BEFORE, BEGUN_BY/ENDED_BY into BEGINS/ENDS, and SIMULTANEOUS/DURING_INV/OVERLAP/IDENTITY into DURING, with IS_INCLUDED mapped to INCLUDES. The resulting six types form three directional pairs (BEFORE/AFTER, BEGINS/ENDS, INCLUDES/DURING) with broad temporal coverage and sufficient training examples.

C. Strict Interpretation and Dataset Density

This work adopts strict TIMEX3 interpretation [5] and dense event annotation. Strict TIMEX3 considers only time expressions that can be normalized to a standard value (e.g., "January 2020," "30 minutes," "daily"), excluding vague expressions like "recently" or "soon." This trades fewer time expressions per sentence for higher-confidence anchors whose predictable patterns enable systematic identification through regular expression matching. Dense event annotation marks all actions, states, and occurrences rather than only those deemed significant, capturing verbs like "incubated," "observed," and "measured" to reduce annotator subjectivity and provide more candidates for relation extraction.

IV. Methods

A. Overall Pipeline Overview

1) Stage 1: Temporal Named Entity Recognition (NER): During the NER stage, preprocessed biomedical sentences are given as input, and BIO-tagged labels for events and time expressions are produced as output. Input sentences are tokenized using the BioBERT tokenizer, producing subword tokens that may subdivide words. The encoder processes the token sequence through transformer layers [2], producing contextualized representations for each position. The classification head applies a linear transformation to these representations, producing logits over the BIO label vocabulary, specifically O, B-EVENT, I-EVENT, B-TIMEX3, and I-TIMEX3. The Softmax formula is used to calculate the label probabilities, and the highest probability label is predicted for each token.

BioBERT was chosen as the encoder backbone because it is pretrained on large-scale biomedical corpora, including PubMed abstracts and PubMed Central articles [17], [26], [27], giving the model familiarity with biomedical terminology and

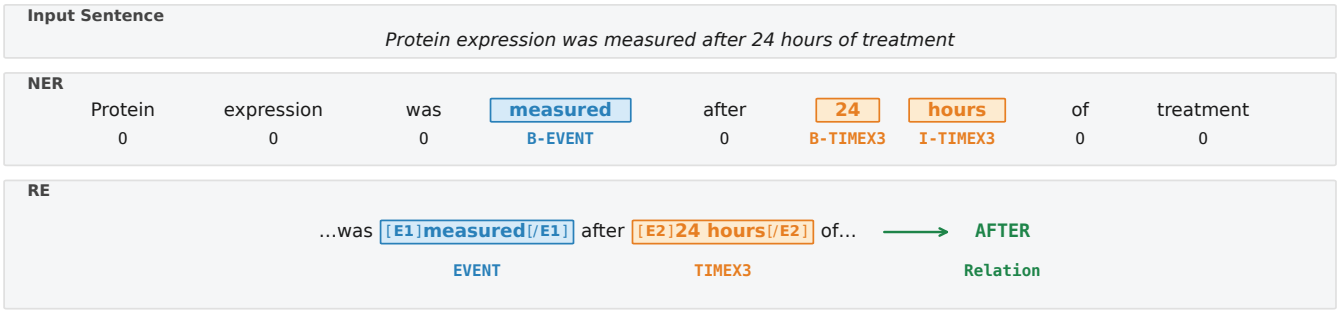

Fig. 1. Example of the NER and RE pipeline, with BIO tagging followed by relation classification.

reducing the adaptation problem to temporal extraction rather than basic language understanding.

2) Stage 2: Temporal Relation Extraction (RE): The RE stage takes sentences with marked entity spans as input and predicts temporal relation labels for event-time pairs as output. Only event-timex pairs are considered because time expressions provide objective temporal anchors, whereas event-event relations are more ambiguous. Entity spans identified by NER are paired to form candidate relation instances. Entity spans are marked by inserting a pair of special tokens at their boundaries. For consistency, event is marked as the first entity and time expression is marked as the second entity as [E1]/[/E1] and [E2]/[/E2], respectively. These markers allow the model to identify participating entities during encoding. The CLS token representation from the final transformer layer is passed through a linear classifier to predict one of the six relation types.

During silver-label generation, predictions below a confidence threshold of 0.70 are filtered. This threshold value was selected empirically to balance retention of data against noise. Higher thresholds would exclude too many potentially useful predictions, while lower thresholds would introduce noisy data that would degrade model performance.

3) Shared Training Procedures and Design Choices: To ensure consistency, NER and RE follow the same iterative refinement approach. Both use the same BioBERT encoder checkpoint [17], tokenizer, and preprocessing steps. Hyperparameter settings such as number of epochs, learning rate, and batch size are held constant across iterations for each task to isolate the effect of data composition changes. Evaluation uses consistent metrics computed on the same gold set for both pipelines. This ensures that performance changes reflect the iterative refinement process and not the training variability.

Given the limited corpus size and the availability of a separate gold set for evaluation, the training dataset used 100% of the available data without a validation split. For NER, models were trained for 3 epochs with a batch size of 16, learning rate of 2×10^{-5}, 1000 warmup steps, and weight decay of 0.01. For RE, models were trained for 7 epochs with a batch size of 8 and learning rate of 2×10^{-5}. Both tasks used mixed precision training on a GPU and a fixed random seed for reproducibility.

B. Data Preparation and Annotation Flow

1) Source Data and Corpus Overview: The training pipeline relies on two primary data sources. First, the TimeML source corpus [5], [7] consists of news articles expertly annotated with events, time expressions, and temporal links, providing the initial training signal. Second, open-access PMC articles provide biomedical texts across topics ranging from molecular biology to clinical research. Articles were selected at random and filtered to contain at least 12 unique temporal expressions across 3 distinct temporal categories, identified through regular expression matching under strict TIMEX3 interpretation.

2) Corrections Dataset: The corrections dataset consisted of PMC articles reserved for target domain supervision. Articles were selected at random and filtered using the same criteria as the training dataset to ensure sufficient number and variety of temporal examples. Twenty articles were added at each iteration, resulting in sixty total articles by iteration 3.

For the RE experiments that varied silver volume, corrections were drawn from a separate, disjoint article partition. The same correction schedule was applied in every configuration, adding 356 corrected pairs per iteration and reaching 1,068 by iteration 3. Larger silver volumes consequently yielded a smaller proportion of corrected examples.

Initial annotations were generated by consensus between GPT-4.5 and Claude Opus 4.1, accessed through their respective web interfaces. Both models were instructed to follow strict TIMEX3 interpretation and dense event annotation guidelines, and each was prompted separately with one article per prompt using default interface settings. For NER, a regex-based matcher assisted in identifying temporal expressions since the strict interpretation scheme made most TIMEX3 entities identifiable through predictive patterns. Token-level agreement across annotators determined final labels, with manual review applied to TIMEX3 entities. Events identified by either language model were accepted given the dense annotation policy. For RE, the same two models labeled temporal relations for event-time pairs. When the pair of models disagreed on the label, they were excluded to ensure annotation consistency.

3) Gold Evaluation Set: The gold evaluation set was constructed using the same filtering and annotation procedures as

the corrections but remained entirely separate from training. For NER evaluation, 10 PMC articles were annotated. For RE evaluation, 50 PMC articles yielded 916 candidate sentences with entity pairs. Of these, 124 (13.5%) were discarded because NER had produced incorrect entity spans. Among the pairs with valid spans, the two models agreed on 564 (71.2%) and disagreed on 228 (28.8%) which were excluded, leaving 564 labeled sentences for evaluation. This agreement level is consistent with the low inter-annotator agreement long reported for temporal relation annotation, and excluding pairs without a clear relation follows established practice in this literature [7], [8]. A human annotator verified the remaining labels to establish a gold standard.

4) TimeML Conversion Workflow: The TimeBank corpus uses annotations in TimeML's XML format, which is not directly compatible with transformer model inputs. Data conversion requires addressing several challenges. First, the interspersed inline XML tags for EVENT and TIMEX3 entities along with the raw text had to be extracted while preserving entity span information. Second, temporal relations are stored separately as TLINK elements, requiring resolution through MAKEINSTANCE elements to map relations back to referenced entities. Third, line breaks in the source files reflect newspaper column formatting rather than sentence boundaries, requiring sentence-level segmentation before tokenization.

Two separate conversions produce the BIO-tagged token sequences for NER and entity-marked sentence pairs for RE, as illustrated in Figure 1. For NER, the conversion parses inline EVENT and TIMEX3 XML tags to identify entity spans, tokenizes the extracted text, and assigns BIO labels to each token. The conversion handles the XML tree structure carefully, as text can appear both inside and between entity tags.

For RE, the conversion resolves MAKEINSTANCE indirection to map event instance identifiers back to event identifiers, locates the sentence containing each entity referenced in a TLINK, and replaces the original XML tags with entity boundary markers. When both entities appear in the same sentence, the sentence is used directly. When they span multiple sentences, the sentences are concatenated. After conversion, the label set is reduced from TimeBank's original nine relation types to six.

C. Training and Iterative Refinement

The iterative approach follows established self-training methodology [4], [23]. Each iteration begins with the current model generating predictions on unlabeled PMC articles. Predictions above the confidence threshold become silver labels, which are combined with the TimeML source corpus and corrections to form the next training set. The model is then retrained and evaluated on the gold set. Figure 2 summarizes this process.

NER and RE were trained sequentially, as RE requires identified entity spans. NER training began with a baseline on the TimeML source corpus [7], followed by three refinement iterations. The final NER model was then used to annotate entity spans in the RE dataset, which followed the same iterative scheme.

TABLE I
NER PERFORMANCE ACROSS ITERATIONS

Iter.	Prec.	Rec.	F1	EVT F1	TMX F1
0 (Baseline)	0.673	0.536	0.597	0.62	0.45
1	0.686	0.583	0.630	0.65	0.52
2	0.692	0.648	0.670	0.67	0.66
3	0.689	0.667	0.678	0.67	0.72

D. Evaluation Metrics and Procedure

Evaluation was performed using a held-out gold set that was never used for training or silver-label generation. The same set was used across all iterations to enable tracking of progress. NER was evaluated using entity-level precision, recall, and F1 with strict span matching requiring exact boundary correspondence. Performance on events and time expressions was measured separately. For RE evaluation, the model operates on gold entity spans so that relation performance is isolated from upstream NER errors. RE metrics include accuracy, macro F1 (averaging equally across relation types), and weighted F1 (weighting by class frequency).

V. RESULTS

A. Overview

NER training consisted of a baseline iteration followed by three refinement iterations. For RE, extensive experiments with varying silver-label volumes (200, 500, and 1000 articles) were analyzed to determine the trade-off between data quantity and correction influence. The final RE configuration used 500 silver articles with a 0.70 confidence threshold, which provided a balance between silver-label coverage and correction ratio.

B. Temporal NER Performance

Table I presents the NER performance metrics across all iterations. The baseline TimeBank-trained model achieved an F1 score of 0.597, with notably lower performance on TIMEX3 entities ($F1 = 0.45$) compared to EVENT entities ($F1 = 0.62$). This disparity reflected the domain shift between news and biomedical text, particularly in how the temporal expressions were articulated.

The iterative refinement process yielded consistent improvements. Overall F1 increased from 0.597 to 0.678 (+13.6%), driven primarily by recall improvements (0.536 to 0.667). Most notably, TIMEX3 recognition improved by 60% (F1: 0.45 to 0.72), indicating successful adaptation to biomedical temporal expression patterns.

C. Temporal Relation Extraction Performance

The RE task presented greater challenges due to fundamental label distribution differences between TimeBank and PMC. TimeBank's event-time pairs are predominantly labeled INCLUDES (47%), while PMC articles favor DURING (84%).

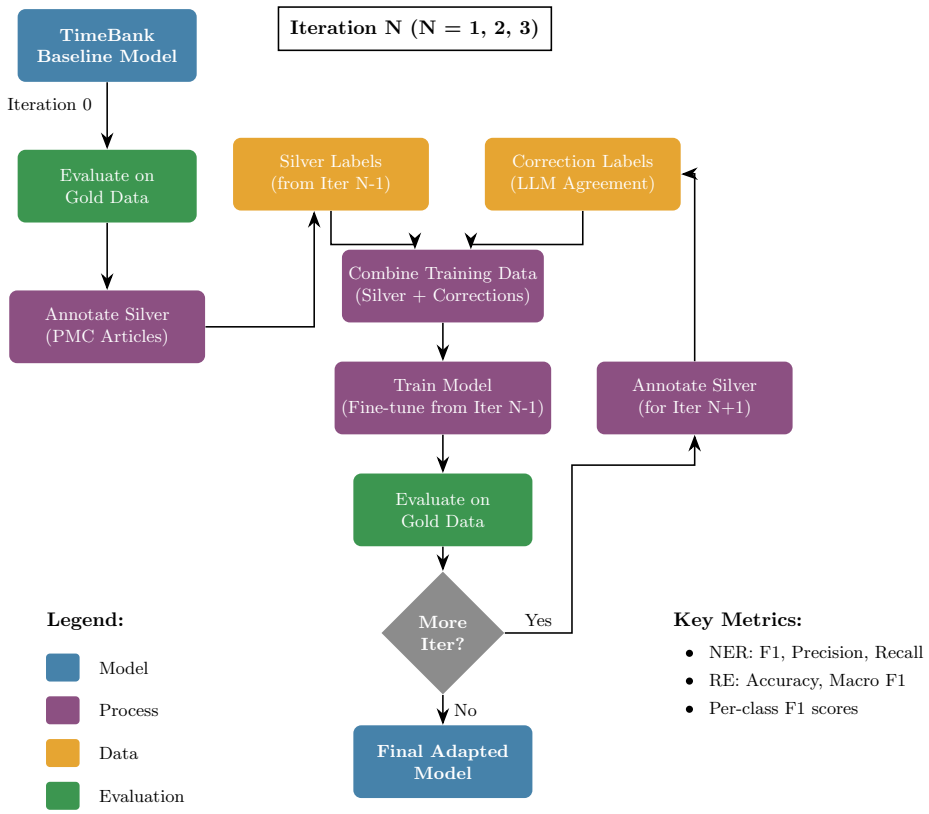

Fig. 2. The iterative refinement training process used in our methodology.

TABLE II
RE PERFORMANCE ACROSS ITERATIONS (500 SILVER ARTICLES)

Iter.	Acc.	Mac. F1	Wt. F1	Corr. %
0 (Baseline)	21.1%	0.22	0.32	0%
1	34.9%	0.42	0.49	24%
2	50.9%	0.43	0.64	38%
3	60.5%	0.49	0.71	48%

Table II shows the RE performance across iterations using 500 silver articles.

The baseline model achieved only 21.1% accuracy, reflecting severe distribution mismatch. The model predicted INCLUDES for 63.8% of examples where the gold distribution contained 84.2% DURING. Through iterative refinement, accuracy improved to 60.5% (+39.4%), demonstrating successful domain adaptation.

The distribution shift in model predictions illustrates the adaptation process (Figure 3). INCLUDES predictions decreased from 360 to 165, while DURING predictions increased from 113 to 331, approaching the gold distribution of 475 DURING examples.

1) Per-Class Performance: Table III presents per-class F1 scores comparing baseline and final iteration performance. Most classes showed marked improvement, with BEGINS achieving the highest F1 (0.80) and DURING reaching 0.75.

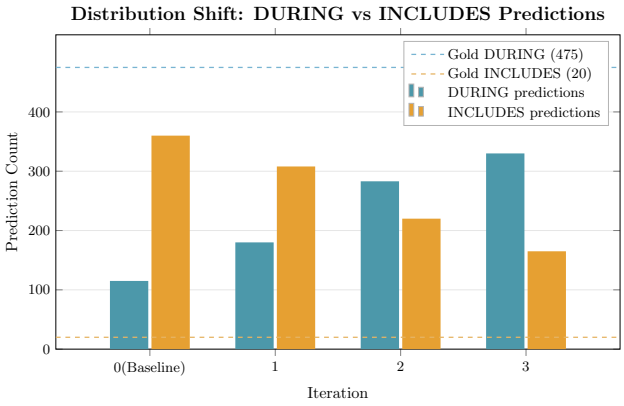

Fig. 3. Distribution Shift: DURING vs INCLUDES Predictions.

D. Effect of Correction to Silver Ratio

The impact of silver-label volume on model performance was investigated by varying the number of silver articles (200, 500, and 1000). The correction set was identical across all three configurations, so only silver volume varied. Table IV summarizes the final iteration results for each configuration.

A clear inverse relationship emerged between silver-label volume and final performance. With 200 articles (71% corrections), the model achieved 76.6% accuracy. When increased

TABLE III
PER-CLASS F1 SCORES

Label	Supp.	Base. F1	Final F1	Chg.	Impr.
BEFORE	7	0.40	0.60	+0.20	+50%
AFTER	51	0.28	0.66	+0.38	+136%
BEGINS	8	0.24	0.80	+0.56	+233%
ENDS	3	0.07	0.12	+0.05	+71%
DURING	475	0.33	0.75	+0.42	+127%
INCLUDES	20	0.03	0.01	−0.02	−67%

TABLE IV
EFFECT OF SILVER ARTICLE COUNT ON FINAL PERFORMANCE

Articles	Examples	Corr. %	Accuracy	Mac. F1
200	∼435	71%	76.6%	0.54
500	∼1,148	48%	60.5%	0.49
1000	∼2,296	32%	42.6%	0.38

to 1000 articles (32% corrections), its accuracy reduced to 42.6%. This demonstrates that silver labels, which inherit the source domain's distribution bias, can dilute the effect of target domain annotation corrections.

E. Error Analysis

The primary sources of errors for the final RE model are illustrated in the confusion matrix (Figure 4). The dominant pattern was DURING-INCLUDES confusion, where the model predicted 151 DURING examples as INCLUDES and 18 INCLUDES examples as DURING, stemming from silver labels reflecting TimeBank's INCLUDES preference. A smaller category involves AFTER to DURING errors (9 instances), where sequential events were misclassified as overlapping. ENDS ($F1 = 0.12$) and INCLUDES ($F1 = 0.01$) underperformed due to limited gold examples and overcorrection toward DURING, respectively.

VI. DISCUSSION

A. Interpretation of Results

1) Domain Distribution Mismatch: The vast contrast between baseline performance for NER (59.7% F1) and RE (21.1% accuracy) revealed fundamentally different types of domain shift. For NER, the TimeBank model retained reasonable entity recognition capability because EVENT and TIMEX3 definitions remained relatively consistent across domains. The challenge was in recognizing entity boundaries and expressions unique to biomedical text, which iterative refinement successfully addressed.

For RE, there was a severe distribution mismatch in label frequencies between the source and target domains. TimeBank labeled 47% of event-time pairs as INCLUDES, while the biomedical corrections favored DURING at 84%. Since these labels represent inverse perspectives on the same bounded relationship, this distribution shift likely reflects different annotator conventions across domains, further compounded by text that may lack sufficient context for annotators to determine which perspective applies.

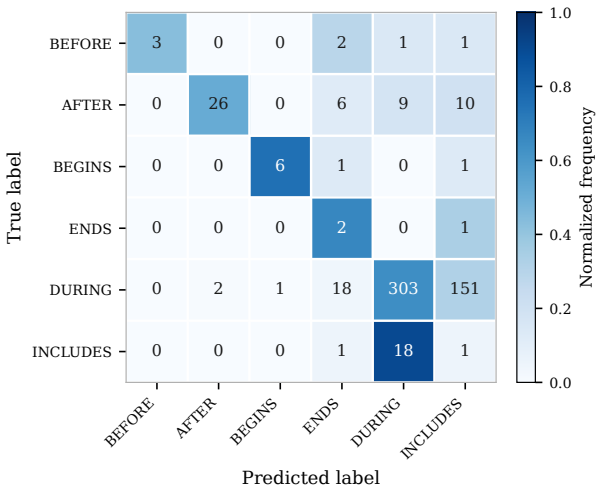

Fig. 4. RE Confusion Matrix (Iteration 3).

2) Effectiveness of Iterative Refinement: Both tasks benefited from iterative refinement, although the improvement patterns differed. NER showed consistent incremental gains across iterations, suggesting stable learning where each round of corrections built on the previous one. RE showed more dramatic adaptation, with accuracy nearly tripling from baseline, indicating that the model successfully overcame the initial distribution bias despite the severe mismatch between source and target domains.

The ratio of corrections to silver labels proved critical to adaptation success. As shown in Table IV, performance declined sharply as silver-label volume increased, suggesting that the silver labels' source-domain bias overwhelmed the corrective signal at lower correction ratios.

3) The INCLUDES-DURING Problem: The persistent INCLUDES-DURING confusion reflects differing annotation conventions across domains rather than model deficiency. TimeBank annotators preferred INCLUDES for temporal containment patterns, while biomedical annotators consistently labeled similar patterns as DURING. Since TimeML's definitions leave room for interpretation when text lacks explicit temporal boundaries, different conventions produced conflicting training signals that the model could not fully resolve.

B. Limitations

1) Class Imbalance: The gold data exhibited severe class imbalance where DURING comprised 84.2% of examples, while ENDS had only 3 examples (0.5%). This imbalance reflects the natural distribution of temporal relations in biomedical text rather than a limitation of the gold set. However, the limited samples for minority classes made reliable evaluation of their performance difficult and may have biased the model toward majority class predictions. The low F1 scores for ENDS (0.12) and BEFORE (0.60) should be interpreted with caution given their limited sample sizes.

2) Event-Timex Pair Restriction: The RE task focused on event-time pairs only, excluding event-event temporal relations. TIMEX3 entities follow strict TimeML definitions [5] and can be identified objectively, whereas biomedically significant events are more subjective to annotate. This restriction limited the model's applicability to complete temporal relation extraction, particularly for BEFORE and AFTER relations that occur more naturally between events.

C. Implications

Despite limitations, the success of data-centric refinement [23], [24] demonstrates that practical extraction capabilities can be achieved in domains lacking annotated resources, with potential generalization to other relation extraction tasks facing similar constraints. The annotation policies also illustrate how definitional choices shape model training, as strict temporal anchoring provided precise time expressions while dense event annotation ensured comprehensive coverage.

D. Future Directions

Three extensions would strengthen this work. First, a larger corrections dataset prioritizing texts with frequent model errors would improve evaluation evidence. Second, cross-sentence relation extraction would capture temporal links spanning sentence boundaries for full timeline reconstruction. Third, cross-domain experiments applying the iterative approach to historical texts, legal documents, or other specialized domains would test the methodology's generalizability.

VII. Conclusion

This work demonstrated that iterative silver-label refinement can adapt temporal information extraction to a new domain without requiring comprehensive manual annotation. Starting from a model trained on TimeBank [7], the approach progressively adapted extraction capabilities to biomedical research articles through cycles of silver-label generation, targeted correction, and retraining. NER and relation extraction performance improved substantially across iterations, with gains stabilizing after three refinement rounds.

The key methodological insight was that silver-label volume must be carefully controlled. Excessive silver labels carry source-domain bias and dilute the corrective signal. The methodology's reliance on data composition rather than architectural novelty suggests broad applicability to other information extraction tasks facing similar resource constraints. The conversion workflow from TimeML to transformer-compatible formats and the multi-class classification across six relation types provide a foundation for future work.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in NeurIPS*, vol. 30, pp. 5998–6008, 2017.

[3] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," in *Proc. 33rd Annual Meeting of the ACL*, pp. 189–196, 1995.

[4] D. McClosky, E. Charniak, and M. Johnson, "Effective Self-Training for Parsing," in *Proc. HLT-NAACL*, pp. 152–159, 2006.

[5] J. Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz, "TimeML: Robust Specification of Event and Temporal Expressions in Text," in *Proc. IWCS-5*, 2003.

[6] J. F. Allen, "Maintaining Knowledge about Temporal Intervals," *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, 1983.

[7] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo, "The TimeBank Corpus," in *Proc. Corpus Linguistics*, pp. 647–656, 2003.

[8] T. Cassidy, B. McDowell, N. Chambers, and S. Bethard, "An Annotation Framework for Dense Event Ordering," in *Proc. 52nd Annual Meeting of the ACL*, pp. 501–506, 2014.

[9] Q. Ning, H. Wu, and D. Roth, "A Multi-Axis Annotation Scheme for Event Temporal Relations," in *Proc. 56th Annual Meeting of the ACL*, pp. 1318–1328, 2018.

[10] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky, "SemEval-2007 Task 15: TempEval Temporal Relation Identification," in *Proc. SemEval-2007*, pp. 75–80, 2007.

[11] M. Verhagen, R. Saurí, T. Caselli, and J. Pustejovsky, "SemEval-2010 Task 13: TempEval-2," in *Proc. 5th International Workshop on Semantic Evaluation*, pp. 57–62, 2010.

[12] W. F. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, and J. Pustejovsky, "Temporal Annotation in the Clinical Domain," *Transactions of the ACL*, vol. 2, pp. 143–154, 2014.

[13] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge," *JAMIA*, vol. 20, no. 5, pp. 806–813, 2013.

[14] G. Alfattni, N. Peek, and G. Nenadic, "Extraction of Temporal Relations from Clinical Free Text: A Systematic Review," *Journal of Biomedical Informatics*, vol. 108, 103488, 2020.

[15] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 Shared Task on Event Extraction," in *Proc. BioNLP 2009 Workshop*, pp. 1–9, 2009.

[16] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, "Overview of BioNLP Shared Task 2011," in *Proc. BioNLP Shared Task 2011 Workshop*, pp. 1–6, 2011.

[17] 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.

[18] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proc. EMNLP*, pp. 3615–3620, 2019.

[19] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *arXiv:1904.05342*, 2019.

[20] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova, "A BERT-based Universal Model for Both Within- and Cross-sentence Clinical Temporal Relation Extraction," in *Proc. 2nd Clinical NLP Workshop*, pp. 65–71, 2019.

[21] R. Han, I.-H. Hsu, M. Yang, A. Galstyan, R. M. Weischedel, and N. Peng, "Deep Structured Neural Network for Event Temporal Relation Extraction," in *Proc. CoNLL*, pp. 666–676, 2019.

[22] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky, "SemEval-2013 Task 1: TempEval-3," in *Proc. SemEval 2013*, pp. 1–9, 2013.

[23] C. Lin, T. Miller, D. Dligach, S. Bethard, and G. Savova, "Self-training Improves Recurrent Neural Networks Performance for Temporal Relation Extraction," in *Proc. LOUHI*, pp. 165–176, 2018.

[24] X. Zhao, S.-T. Lin, and G. Durrett, "Effective Distant Supervision for Temporal Relation Extraction," in *Proc. Second Workshop on Domain Adaptation for NLP*, pp. 195–203, 2021.

[25] J. A. Fries, P. Steinberg, S. Khattar, S. L. Fleming, J. Posada, A. Callahan, and N. H. Shah, "Ontology-driven Weak Supervision for Clinical Entity Classification in Electronic Health Records," *Nature Communications*, vol. 12, no. 1, pp. 2017, 2021.

[26] Y. Peng, S. Yan, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," in *Proc. 18th BioNLP Workshop*, pp. 58–65, 2019.

[27] 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, 2021.

[28] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr., and T. M. Mitchell, "Toward an Architecture for Never-Ending Language Learning," in *Proc. AAAI*, pp. 1306–1313, 2010.

[29] J. R. Curran, T. Murphy, and B. Scholz, "Minimising Semantic Drift with Mutual Exclusion Bootstrapping," in *Proc. PACLING*, pp. 172–180, 2007.

[30] T. McIntosh and J. R. Curran, "Reducing Semantic Drift with Bagging and Distributional Similarity," in *Proc. ACL-IJCNLP*, pp. 396–404, 2009.

[31] E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," in *Proc. AAAI*, pp. 474–479, 1999.

[32] L. A. Ramshaw and M. P. Marcus, "Text Chunking Using Transformation-Based Learning," in *Proc. Third ACL Workshop on Very Large Corpora*, pp. 82–94, 1995.