# A Method for Dynamic Expansion of Domain Dictionaries and Weakly Labeled Named Entity Recognition for Low-Resource Specialized Corpora

**Yanjun Ma**[1,*]
[1]School of Cybersecurity, Liaoning Police College, Dalian 116036, China
Corresponding author: Yanjun Ma (e-mail: mayanjun0183@163.com).

**ABSTRACT** In order to solve the problems of domain entity lack and high noise, little human supervision in low resource specialized corpora, this work adopts the method of multi-feature fusion of dynamic domain dictionary expansion. The method includes dictionary matching, semantic confidence filtering and confidence-weighted named entity recognition, and realizes the mutual iteration between the dictionary and the recognition model by feeding back the entity. The results demonstrate that, the dynamic expansion raised the entity coverage from 38.10% to 77.22% while the accuracy of adding new entities is 94.67% for the dynamic expansion; the F1 score of the confidence based filtering for weakly labelled data is 84.68%, and the Micro-F1 score of the complete model is 89.17%, which is 13.28 percentage points better than baseline model under 1% human annotation. It shows that the approach proposed in this study is able to complement the long tail specific entities while keeping noise from weak labels under control and the entity recognition performance still good under low level manual annotation.

**INDEX TERMS** domain-specific corpus; dynamic domain dictionary expansion; weak annotation; named entity recognition; confidence-weighted

## I. INTRODUCTION

NAMED entity recognition (NER) refers to the extraction of the boundaries of entities and the recognition of entity semantic categories in the domain of specific text, and the recognition effect will directly influence the credibility of knowledge extraction, information retrieval and organization of domain knowledge. Current supervised NER generally requires large manually annotated corpora, but specialized texts, such as in medicine, law, and industrial technology, have dense terminology, a high rate of long tail entities, and new entities are continually emerging, which means there is a trade-off between the cost of manually annotating a large corpus and the coverage of entities. In low-resource scenarios, data augmentation might mitigate the lack of samples, but the ideal level of augmentation will vary depending on the size of the training set, and increasing data by making more synthetic instances doesn't always solve the problem of inadequate coverage of specialized entities (Torres et al. , 2026 [1]).

For specialty corpora, domain dictionaries are frequently applied to entity discovery and normalization and can directly give the boundaries of an entity and the priors of categories. But fixed dictionaries are influenced by the development of new terminology, aliases and changes in the expression of a domain and their coverage capability will decrease as the corpus is updated. Simultaneously, dictionary matching can also be faced with the problem of homonymy, boundary overlapping and context ambiguity. Deep models have been used to prune the set of contextually inconsistent matches provided by existing dictionary-enhanced methods of NER, which has shown that dictionary information can only be used reliably in combination with contextual discrimination (Nastou et al. , 2024 [2]). Thus the main problem for the resource limited case has changed from the question of “do we use a dictionary?” to “how can we grow the dictionary continuously while keeping a control of the expansion noise?”

Weakly labelled data also gives supervised data that is usable for other, unlabeled domain-specific corpora, but labels generated automatically will certainly have category misclassifications, entity boundary changes and even error propagation. Although distant-supervised NER has proven to benefit from the use of external knowledge for label construction, the quality of weak labels still directly affects the stability of model training (Bali and Anandaraj, 2023 [3]). If these three tasks are performed independently, newly identified entities will not be in the dictionary for quick labels, and low-confidence labels will likely get into the

training data again if they were not found in the dictionary. Even existing NER systems are still facing significant constraints of generalization under low resource, new entity and cross domain conditions (Seow et al. , 2025 [4]). Hence, under limited manual annotation conditions, multi-feature candidate entity screening, dynamic domain dictionary updating, weak-label confidence controlling and entity recognition should be combined together into an iterative process to jointly benefit the enhancement of entity coverage and the reliability of entity labels.

## II. METHODS AND MATERIALS

### A. CONSTRUCTION OF A LOW-RESOURCE PROFESSIONAL CORPUS AND ENTITY ANNOTATION SYSTEM

The technical documents included in the specialized corpus are 2, 436. The corpus contained 39, 860 sentences and 1, 196, 430 tokens after the removal of duplicated texts, characters normalization, the segmentation of texts into sentences and segments and the cleaning of the anomalous symbols. The corpus was split into manually annotated and unannotated parts, 5, 200 sentences were manually annotated, and 34, 660 sentences were kept as unannotated. Of the manually annotated data, 3, 600 of the sentences were used for training, while 800 sentences were used for validation and 800 sentences were used for independent test; note that the latter did not participate in the dictionary construction or in the weak label generation. To mitigate the issues arising from boundary ambiguity introduced by categories, entities are boundary-encoded using the BIOES framework, which classifies them into five classes: equipment, components, processes, parameters, and fault states based on the conceptual hierarchy of the technical text (Zanoli et al. , 2024 [5]).

The manually annotated corpus includes all together 5900 different entity types and 16800 instances. It can be seen from Table 1 that, there is a disparity in the distribution of different entities in the different categories, and equipment and component entities occupy 25.67% and 22.89% respectively, taking up nearly 50% of all the entities. The initial domain dictionary has 2, 248 entries that cover 38.10% of the 5, 900 unique entities, where fault status entities are the least covered at 35.81% initially. Such category variation in category coverage maintains the realistic long tail qualities of low resource specialized corpora and offers enough "uncovered" space for new entity identification. Multi-category annotation schemes can lead to recognition biases which can be mitigated by unified annotation boundaries and category rules (Belhajem, 2024 [6]).

The percentages of manually annotated sentences used in training for the low-resource setting with respect to the above participation model are controlled as follows:

$$\rho_r = \frac{N_r}{N_t} \times 100\% \tag{1}$$

where $\rho_r$ represents the proportion of human-annotated resources; $N_r$ represents the number of human-annotated training sentences actually used under the current conditions; and $N_t$ represents the size of the complete human-annotated training set, fixed at 3,600 sentences. Five ratios of annotations (1%, 5%, 10%, 20%, and 50%) were tested, which equated to 36, 180, 360, 720, and 1, 800 training sentences respectively. The only difference between each ratio was the number of manually labelled sentences used for the training phase and the number of manually labelled sentences remained the same for the validation set and test set as well as for the unlabeled corpus. This allowed for comparing changes in entity coverage as well as differences in entity recognition performance within the same data boundaries.

### B. METHOD FOR DYNAMIC EXPANSION OF DOMAIN-SPECIFIC DICTIONARIES BASED ON MULTI-FEATURE FUSION

Only 2, 248 of 5, 900 unique entities are included in the initial domain dictionary, which covers 38.10% of the entities; as a result, many entities with a low frequency, specialized abbreviations as well as composite entities are not covered by the dictionary. Candidate entity extraction uses occurrence frequency, internal cohesion, contextual freedom (to the left, to the right, above, and below) as well as semantic proximity to the category to reduce the negative effect of any single statistical feature on the low-frequency entities. Specialized concept extraction can be used to complement the existing domain knowledge by using the associations between terms in the corpus (Behr et al., 2023 [7]), whereas the use of semantic representations can lower the false positive rate that is generated by ambiguous terms generated by string co-occurrence (Morazzoni et al., 2024 [8]). The comprehensive score for candidate entities $S(e)$ is defined as:

$$\begin{aligned} S(e) = \alpha \frac{\ln(1+f_e)}{\ln(1+f_{\max})} + \beta \frac{\mathrm{PMI}(e) - p_{\min}}{p_{\max} - p_{\min}} \\ + \gamma \frac{2H_e^L H_e^R}{H_e^L + H_e^R + \varepsilon} + \delta \max_{c \in C} \cos(\mathbf{z}_e, \boldsymbol{\mu}_c) - \lambda A_e. \end{aligned} \tag{2}$$

where $S(e)$ denotes the comprehensive score of the candidate entity; $f_e$ and $f_{\max}$ denote the frequency of the candidate entity and the maximum frequency in the candidate set, respectively; $\mathrm{PMI}(e)$ denotes the intra-word point mutual information; $p_{\min}$ and $p_{\max}$ denote the minimum and maximum values of the point mutual information, respectively; $H_e^L$ and $H_e^R$ denote the left and right context information entropy, respectively; $\varepsilon$ denotes the smoothing term; $\mathbf{z}_e$ denotes the context vector of the candidate entity; $\boldsymbol{\mu}_c$ represents the semantic prototype of category $c$ ; $C$ represents the set of entity categories; $A_e$ represents the ambiguity penalty term; $\alpha, \beta, \gamma, \delta, \lambda$ represents the feature weights.

**TABLE 1.** Scale of the Professional Corpus and Distribution of Entity Categories

| Data Set/Entity Category | Number of Documents | Number of Sentences | Number of Tokens | Number of Entity Instances | Number of Unique Entities | Number of entries in the initial dictionary | Percentage of Entity Instances | Initial coverage /% |
|---|---|---|---|---|---|---|---|---|
| Total professional corpus | 2 436 | 39 860 | 1 196 430 | - | - | - | - | - |
| Manually Annotated Dataset | - | 5 200 | 156 820 | 16 800 | 5 900 | 2 248 | 100.00 | 38.10 |
| Unlabeled corpus | - | 34 660 | 1 039 610 | - | - | - | - | - |
| Device Entity | - | - | - | 4 312 | 1 286 | 512 | 25.67 | 39.81 |
| Part Entity | - | - | - | 3 846 | 1 442 | 568 | 22.89 | 39.39 |
| Process Entities | - | - | - | 2 915 | 1 037 | 396 | 17.35 | 38.19 |
| Parameter Entity | - | - | - | 2 604 | 962 | 352 | 15.50 | 36.59 |
| Fault State Entity | - | - | - | 3 123 | 1 173 | 420 | 18.59 | 35.81 |
| Total | - | - | - | 16 800 | 5 900 | 2 248 | 100.00 | 38.10 |

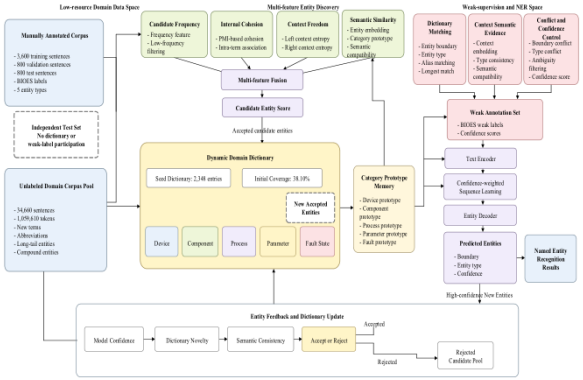


**Fig. 1.** Overall Framework for Dynamic Expansion of Domain Dictionaries and Weakly Annotated Named Entity Recognition

In the candidate entity evaluation, it is required to preserve the two kinds of information: corpus statistical information and category semantic information. Together, the specialized corpus and the seed dictionary with 2, 248 entities form a basis for entity discovery as illustrated in Figure 1. On the one hand, the corpus side generates the term frequency, PMI, and left-right information entropy of the technical terms and, on the other hand, the dictionary side constructs category semantic prototypes of five categories of technical terms, namely equipment, components, processes, parameters and fault states, and these two types of information are integrated at the candidate entity scoring stage to suppress the statistically significant terms with semantic deviation as much as possible, while the low-frequency technical terms that are closer to the category semantic prototype are retained as much as possible. An entity that has more than threshold value of composite score is added to the entity's corresponding category dictionary, and the ones that have less than threshold value of the composite score are not directly added into the dictionary but stay in the unlabeled corpus.

The text-matching module is fed with entity boundaries, category information and source confidence scores, which are obtained from the expanded classification dictionary. The weak labels thus generated are then added to the manually annotated samples and passed to the named entity recognition model, and then the high confidence named entities generated by the model on unannotated text undergo testing for frequency, semantic prototype and ambiguity constraints, the ones that accept the criteria are added to the next iteration of dictionary. Semantic consistency limits the size of the dictionary (as opposed to building it without restriction due to the number of possible candidate terms) while mitigating category shifts as a result of changes in professional context (Marjan and Amagasa, 2025 [9]).

The joint threshold that regulates the determination of categories for candidate entities and their inclusion in the dictionary is:

$$\hat{c}_e = \arg\max_{c \in C} \cos(\mathbf{z}_e, \boldsymbol{\mu}_c),$$
$$D_{\hat{c}_e}^{(k+1)} = D_{\hat{c}_e}^{(k)} \cup \{e \mid S(e) \geq \tau_s,\ \cos(\mathbf{z}_e, \boldsymbol{\mu}_c) \geq \tau_c,\ f_e \geq 3\}. \tag{3}$$

where $\hat{c}_e$ denotes the predicted category of a candidate entity; $D_{\hat{c}_e}^{(k)}$ and $D_{\hat{c}_e}^{(k+1)}$ denote the domain dictionaries for the corresponding categories in two adjacent rounds, respectively; $\tau_s$ denotes the composite score threshold; and $\tau_c$ denotes the category semantic similarity threshold. A minimum frequency of 3 is used to filter out random character combinations, and the composite score and category similarity control the quality of the entity, so that the number of dictionary extensions, entry precision, and entity coverage can be assessed as uniform measures in performance evaluation.

### C. WEAK LABEL GENERATION METHOD BASED ON DICTIONARY MATCHING AND SEMANTIC CONFIDENCE

Naming in a dynamic dictionary cannot be equated to reliable labelling since abbreviations, nested entities and homographs are easily used in specialized texts, which can cause incorrect matches. In the case of unlabeled corpora, the system first finds the candidates for the entities by longest-prefix matching starting from the dictionary type, and carries out the same process as described above when there are multiple candidates in the same character range. External knowledge can be used for distant supervision to generate entity labels for unlabeled text, yet, noise in automatically generated labels still needs to be explicitly controlled (Etudo and Yoon, 2024 [10]). In order to make the cut to the weakly labeled set, as illustrated in Figure 2, the candidate entities undergo a series of filtering steps which also include boundary conflict resolution, contextual semantic consistency checks, and confidence-based filtering.

The weak label confidence for each candidate entity fragment $\omega_i$ is defined as:

$$\omega_i = \sigma(\theta_1 d_i + \theta_2 s_i + \theta_3 b_i + \theta_4 r_i - \theta_5 u_i) \tag{4}$$

where,

longest-prefix matching starting from the dictionary type, and carries out the same process as described above when there are multiple candidates in the same character range. External knowledge can be used for distant supervision to generate entity labels for unlabeled text, yet, noise in automatically generated labels still needs to be explicitly controlled (Etudo and Yoon, 2024[10]). In order to make the cut to the weakly labeled set, as illustrated in Figure 2, the candidate entities undergo a series of filtering steps which also include boundary conflict resolution, contextual semantic consistency checks, and confidence-based filtering.

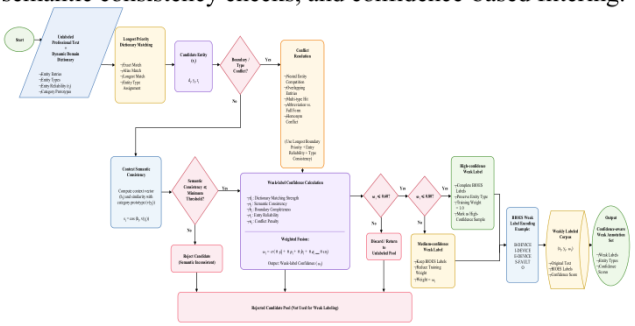

**Fig. 2.** Process for Weak Label Generation and Confidence Filtering

$$s_i = \cos(\mathbf{h}_i, \mathbf{v}_{y_i}), \qquad \sigma(z) = \frac{1}{1+\exp(-z)} \tag{5}$$

In the formula, $\omega_i$ represents the confidence of the weak label for the $i$ th candidate entity; $d_i$ represents the dictionary match strength; $s_i$ represents the cosine similarity between the candidate entity's contextual representation $\mathbf{h}_i$ and the semantic prototype $\mathbf{v}_{y_i}$ of its category $y_i$ ; $b_i$ represents the completeness of the entity boundaries; $r_i$ represents the reliability of the corresponding dictionary entry; $u_i$ represents the penalty for boundary or category conflicts; $\theta_j$ represents the weights of each evidence item; and $\sigma(\cdot)$ represents the sigmoid normalization function.

Confidence-based filtering employs a three-tier control mechanism: $\omega_i \geq 0.80$ entities generate complete BIOES weak labels; $0.60 \leq \omega_i < 0.80$ entities retain their labels but have their training weights reduced; $\omega_i < 0.60$ candidates are directly discarded. Prioritization of latent entities reduces the consumption of limited annotation budgets by low information text (Şapcı et al., 2024 [11]). This hierarchical method keeps the number of weak tags and their reliability, and makes it possible to quantify and compare the accuracy, recall, F1 score, boundary precision, and classification precision.

### *D. CONFIDENCE-WEIGHTED NAMED ENTITY RECOGNITION MODEL*

There are two types of annotation corpora: manual and filtered weakly annotated, and the training errors computed by the model need to use different weights to account for the different supervision levels. A pre-trained language encoder is used to provide the sequence inputs with contextual representations and a BiLSTM network to add in entity boundary dependency before and after the entity. The 5 categories of professional entities are further expanded to 20 entity labels and in combination with the non-entity label O, the output space consists of 21 states. BiLSTM-CRF can also use both of the additional properties of the context and neighbouring label constraints and has relatively stable sequence modelling abilities in terms of entity recognition in unstructured specialised texts (Arslan, 2024 [12]).

The main body of the model is unfolded vertically from "token representation" to "CRF decoding", with weak label confidence being an independent path that goes straight to the loss layer without changing the supervision strength of the manually annotated samples, as illustrated in Figure 3. The pre-trained encoder outputs the token representation at position $t$ , $g_t$ , and the bidirectional sequence representation is defined as:

$$\mathbf{q}_t = \left[\overrightarrow{\mathrm{LSTM}}(\mathbf{g}_t, \overrightarrow{\mathbf{q}}_{t-1}); \overleftarrow{\mathrm{LSTM}}(\mathbf{g}_t, \overleftarrow{\mathbf{q}}_{t-1})\right] \tag{6}$$

The label emission score is calculated as:

$$\mathbf{\Psi}_t = \mathbf{W}_q \mathbf{q}_t + \mathbf{c}_q \tag{7}$$

where $\mathbf{g}_t$ represents the context encoding vector of the $t$ th token; $\mathbf{q}_t$ represents the sequence features obtained by concatenating the forward and backward hidden states; $\mathbf{W}_q$ represents the label mapping matrix; $\mathbf{c}_q$ represents the bias vector; and $\mathbf{\Psi}_t$ represents the emission scores corresponding to the 21 label states.

The CRF layer simultaneously considers the label scores at word positions and the transition constraints between adjacent labels. Given the text sequence $X^{(m)}$ at position $m$ and the label sequence $Y^{(m)}$ , the conditional probability is defined as:

$$P(\mathbf{Y}^{(m)} \mid \mathbf{X}^{(m)}) = \frac{\exp\left(\sum_{t=1}^{T_m} \psi_{t,\mathbf{Y}_t^{(m)}} + \sum_{t=1}^{T_m-1} \bar{\gamma}_{\mathbf{Y}_t^{(m)},\mathbf{Y}_{t+1}^{(m)}}\right)}{\sum_{\tilde{\mathbf{Y}}} \exp\left(\sum_{t=1}^{T_m} \psi_{t,\tilde{\mathbf{Y}}_t} + \sum_{t=1}^{T_m-1} \bar{\gamma}_{\tilde{\mathbf{Y}}_t,\tilde{\mathbf{Y}}_{t+1}}\right)} \tag{8}$$

where $T_m$ denotes the length of the $m$ th sequence; $\psi_{t,\mathbf{Y}_t^{(m)}}$ denotes the emission score of the true label at position $t$ ; $\bar{\gamma}$ denotes the label transition matrix; and $\tilde{\mathbf{Y}}$ denotes the set of all possible label sequences.

The training loss is differentially weighted based on the source of supervision:

$$L_{cw} = \frac{1}{M}\sum_{m=1}^{M} \eta_m \log P(\mathbf{Y}^{(m)} \mid \mathbf{X}^{(m)}),$$
$$\eta_m = \begin{cases} 1, & \mathbf{X}^{(m)} \in G, \\ \omega_m, & \mathbf{X}^{(m)} \in W. \end{cases} \tag{9}$$

where $L_{cw}$ denotes the confidence-weighted sequence loss; $M$ denotes the number of samples in a training

where $g_t$ represents the context encoding vector of the $t$ th token; $ч_t$ represents the sequence features obtained by concatenating the forward and backward hidden states; $▾▾_q$ represents the label mapping matrix; $◡_q$ represents the bias vector; and $Ψ_t$ represents the emission scores corresponding to the 21 label states.

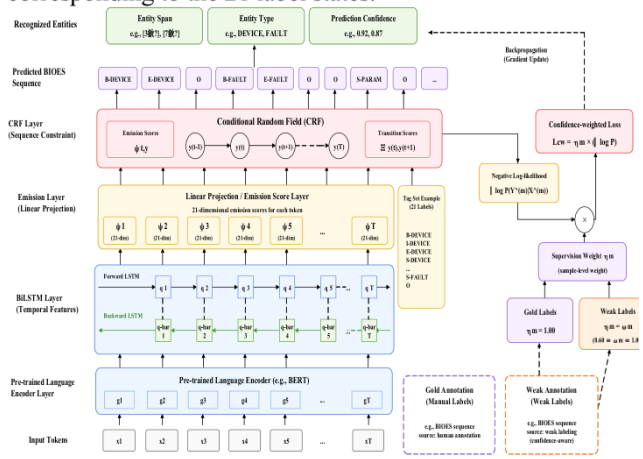


**Fig. 3.** Structure of the Confidence-Weighted Named Entity Recognition Model

batch; $\eta_m$ denotes the supervision weight for a sample; $G$ denotes the set of manually labeled gold standard samples; $W$ denotes the set of weakly labeled samples; and $\omega_m$ uses the confidence scores obtained during the weak label generation stage. Samples that are manually labelled always have a weight of 1 and samples that are weakly labelled with high confidence have a greater training contribution than those that are weakly labelled with low confidence, as they are in the range 0.60-0.80 and contribute to gradient updates according to their actual confidence scores. This mechanism helps avoid that noisy labels affect all of the model parameters equally, and also preserves the entity information from unlabeled domain specific corpora, which thus provides a uniform computational basis for assessing the influence of such changes on Precision, Recall, F1 and performance at different confidence levels.

### *E. MECHANISM FOR COLLABORATIVE ITERATIVE UPDATES BETWEEN THE DOMAIN DICTIONARY AND THE RECOGNITION MODEL*

However, after the expansion of the dictionary, the uncovered terms can be added, but if there are some wrong terms involved in the building of weak labels, these errors can be propagated when the model is trained. Hence, in every round, the model only fetches the entities with high predicted probability which have not been included in the dictionary and subsequently put them under a thorough feature and category semantic constraint. The fusion of knowledge resources and neural models improves the coherence of the professional entity boundaries and their classification into categories (García-Barragán et al., 2025 [13]). The set of entities eligible for the th $k$ round is defined as:

$$A^{(k)} = \{e_j \notin D^{(k)} \mid \hat{y}_j^{(k)} = c,\ p_j^{(c)} \geq \tau_p,\ S(e_j) \geq \tau_s,\ \cos(\mathbf{z}_{e_j}, \boldsymbol{\mu}_c) \geq \tau_c,\ f_{e_j} \geq 3\},$$
$$D^{(k+1)} = D^{(k)} \cup A^{(k)}. \tag{10}$$

where $A^{(k)}$ denotes the set of new entities admitted in the $k$ th round; $p_j^{(c)}$ denotes the prediction confidence of the recognition model for entity $e_j$ ; $\tau_p$ denotes the model's re-entry threshold, set to 0.90; $D^{(k)}$ denotes the current domain dictionary; $S(e_j)$ retains the previously defined meaning; and $f_{e_j}$ denotes the frequency of occurrence of the entity in the unlabeled corpus.

As illustrated in Figure 4, in order to minimize the propagation of weakly supervised mistakes in the cycles, the process starts by re-aligning 34, 660 unlabeled sentences against the current dictionary and assigning weak labels with confidence score. Once the parameter of the recognition model is updated, it will be used to make new predictions on the unlabeled corpus. Only predicted entities which meet the above three criteria (a model confidence of 0.90, multi-feature eligibility criteria, and category semantic consistency) are added to the new entry set and don't participate in the current round of dictionary updates, while those which don't meet any of the above 3 criteria are returned into the candidate pool. Instead of directly writing into the dictionary the output of the model, the dictionary is updated again, and the entities are again sent back to the model for re-estimation, creating a cycle of "dictionary update—label reconstruction—model re-estimation—entity re-entry, ". In the case of weakly supervised learning with different sources of labels, the contribution of noisy labels to the training process needs to be consistently controlled (Alotaibi et al. , 2025 [14]).

Iteration termination is determined by simultaneously evaluating both the dictionary increment and the change in validation set recognition performance:

$$\Delta_D^{(k)} = \frac{|D^{(k+1)} \setminus D^{(k)}|}{|D^{(k)}|}, \qquad \Delta_F^{(k)} = |F_1^{(k+1)} - F_1^{(k)}|,$$
$$\mathrm{Stop}^{(k)} = I\left(\Delta_D^{(k)} < 0.01 \wedge \Delta_F^{(k)} < 0.002\right). \tag{11}$$

where $\Delta_D^{(k)}$ represents the relative increment of the dictionary between two consecutive rounds; $\Delta_F^{(k)}$ represents the change in the validation set F1 score; and $I(\cdot)$ represents the conditional indicator function. Updates are stopped when both conditions are met in two successive rounds, and the maximum number of iterations is 6. At the same time, this criterion limits the expansion of invalid entries and fluctuations in model performance, which allows for a convergent comparison of the size of the dictionary, the coverage of the entity, and the F1 score at the same iteration scale.

consistency) are added to the new entry set and don't participate in the current round of dictionary updates, while those which don't meet any of the above 3 criteria are returned into the candidate pool. Instead of directly writing into the dictionary the output of the model, the dictionary is updated again, and the entities are again sent back to the model for re-estimation, creating a cycle of "dictionary update—label reconstruction—model re-estimation—entity re-entry, ". In the case of weakly supervised learning with different sources of labels, the contribution of noisy labels to the training process needs to be consistently controlled (Alotaibi et al. , 2025[14]).

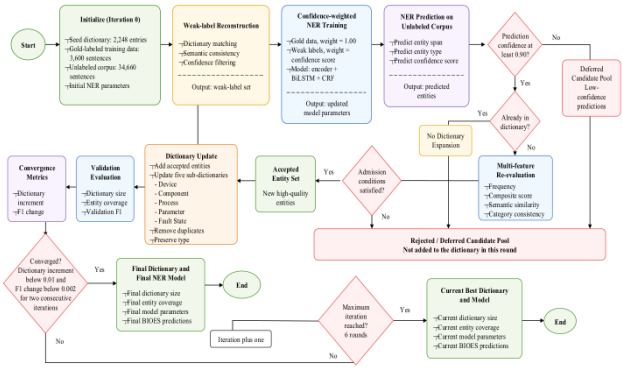


**Fig. 4.** Collaborative Iterative Update Process of Domain Dictionary—Weak Labels—Entity Recognition

## III. EXPERIMENTAL RESULTS AND ANALYSIS

### A. EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

The experiments were performed on an Ubuntu 22.04 system and the computing platform is an Intel Core i9-13900K processor with 64 GB RAM and a 24 GB NVIDIA RTX 4090 GPU. The model was coded using python 3.10 and pytorch 2.2. The maximum sequence length of the pre-trained encoder was 128, the hidden dimension of BiLSTM was 256, the dropout rate was 0.10 and the initial learning rate of AdamW was $2 \times 10^{-5}$. In such a resource-constrained scenario, it is beneficial to prevent data boundary changes from influencing the findings of the model comparison, so the validation set, test set and unlabeled corpus are kept the same (Lancheros et al. , 2025 [15]).

The manually annotated data was set to 1%, 5%, 10%, 20% and 50% as shown in Table 2. The thresholds for weak labels were set to 0.60 for acceptance and 0.80 for high confidence, and the threshold for entity re-entry is set to 0.90. The collaborative updating was carried out at most 6 times, with the condition that the relative dictionary increment is less than 0.01, and the change of F1 score on the validation set is less than 0.002 after two times. Random fluctuations from model initialization were controlled by using 5 random seeds for each experimental group, and averaging the results.

### B. PERFORMANCE ANALYSIS OF DYNAMIC DOMAIN DICTIONARY EXPANSION

The accuracy of newly added entries, entity coverage on the independent test set and the error expansion rate are used to measure the performance of dictionary expansion, where newly added entries are decided according to the manually annotated entities and the manual review results. Specialized domain entities are very context-dependent and relying only on strings or word frequencies can easily result in missing low-frequency entities and introducing semantic ambiguity (Çetindağ et al. , 2023 [16]). As shown in Table 3, frequency-based expansion expanded the set by 1,532 entries, with low accuracy (81.66%) and entity coverage (59.31%), but with the addition of PMI and information entropy the coverage rose to 64.27%, suggesting that some coincidental phrases could be filtered out by incorporating the internal cohesion and contextual freedom.

With the addition of semantic prototypes, the accuracy of the new entries was raised to 89.24%, showing that adding semantic constraints for the categories is effective in suppressing homographic entities. Dynamic expansion by multi-feature eventually expanded the dictionary by 2, 438 entries, with an accuracy rate of 94.67%, and an increase in coverage by 39.12% from the original dictionary and a false expansion rate of 5.33%. There is still a gap of 6.17 percentage points between the results of single-round expansion and dynamic expansion, which means that the feedback of models can still be used to find more new entities with weak statistical features but stable semantics, while multiple admission conditions can suppress the accumulation of noises caused by dictionary expansion.

### C. ANALYSIS OF WEAK LABEL GENERATION QUALITY

Weak labels were assessed with 800 independent test sentences annotated with BIOES labels (as reference). The manual labels were concealed and weak labels were created only using the dictionary and unlabeled text during evaluation. Direct impacts of consistency handling on multiple annotation sources relate to the quality of entity boundaries and category classification (Belhajem, 2024 [17]). The results of recall for direct matches for the fixed dictionary as shown in Table 4 were 58.77%, whereas the results for dynamic dictionary were 71.46%, showing that the incorporation of new entities, primarily helped in the detection capability of the previously unaccounted entities but the precision remained at 86.72% suggesting that the problem of ambiguous matches cannot be solved with dictionary expansion.

Boundary conflict resolution improved boundary accuracy by 3.67 percentage points and the accuracy of the semantic constraints improved category accuracy to 93.05%. With the minimum value of admission threshold was 0.60, the F1 score of comprehensive screening method was 84.68%, and the boundary and category accuracy rates were 95.36% and 94.71% respectively. These results show that dynamic expansion is mostly used to complete the coverage of the

TABLE 2. Experimental Environment and Key Model Parameter Settings

| Parameter Category | Key Parameters and Settings |
|---|---|
| Hardware Environment | Intel Core i9-13900K; 64 GB RAM; NVIDIA RTX 4090 24 GB |
| Software Environment | Ubuntu 22.04; Python 3.10; PyTorch 2.2 |
| Encoding and Network Parameters | Maximum sequence length: 128; BiLSTM hidden dimension: 256; Dropout: 0.10 |
| Model Training Parameters | AdamW; learning rate $2 \times 10^{-5}$; batch size 32; maximum 30 epochs |
| Resource-Constrained Settings | Manual annotation ratio: 1%, 5%, 10%, 20%, 50% |
| Weak label threshold | Minimum threshold: 0.60; High-confidence threshold: 0.80 |
| Dynamic parameter updates | Entity re-inclusion threshold: 0.90; maximum iterations: 6 |
| Convergence and repeated experiments | Dictionary increment < 0.01; F1 change < 0.002; 2 consecutive rounds; 5 random seeds |

TABLE 3. Performance Comparison of Dictionary Expansion Methods Across Different Domains

| Dictionary Expansion Method | Number of New Entries/Entries | Accuracy of New Entries / % | Entity Coverage /% | Error Expansion Rate (%) |
|---|---|---|---|---|
| Fixed Seed Dictionary | 0 | - | 38.10 | - |
| Term Frequency Threshold Expansion | 1 532 | 81.66 | 59.31 | 18.34 |
| PMI + Information Entropy Extension | 1 786 | 86.45 | 64.27 | 13.55 |
| Semantic Prototype Extension | 1 924 | 89.24 | 67.20 | 10.76 |
| Multi-feature single-round expansion | 2 106 | 92.31 | 71.05 | 7.69 |
| Multi-feature dynamic scaling | 2 438 | 94.67 | 77.22 | 5.33 |

entities, and semantic and confidence control are mostly used to filter out the noise created by new additions; these two mechanisms are intertwined in deciding whether or not weak labels are reliable to be added to weighted training.

### D. ANALYSIS OF NAMED ENTITY RECOGNITION PERFORMANCE

All methods were applied to the same training, validation and test sets and a 10% human annotated dataset was used to compare the performance of the model. For low-frequency entity boundaries, the incorporation of contextual information through domain-specific texts can boost identification performance, along with class classification (Abdullahi et al. 2025 [18]). The Micro-F1 score of the CRF is 68.09% as shown in Table 5, but the score is improved to 79.84% with BERT-CRF, suggesting that the pre-trained semantic representations can greatly mitigate the feature sparsity problem under limited annotation.

In comparison with the baseline model BERT-BiLSTM-CRF, the dynamic dictionary with weak labels improved Micro-F1 by 5.21 percentage points, and Recall by 6.80 percentage points, confirming that the newly introduced entities had an impact on uncovering the domain-specific ones. Confidence weighting further increased the Micro-F1 score to 89.17% and also the Macro-F1 score to 87.32% which shows that the low frequency categories were not hidden by the general improvement in accuracy. The advantages of BERT based models over their counterparts are comparable for named entity recognition tasks in low-resource languages (Gopalakrishnan et al. , 2025 [19]).

### E. ANALYSIS OF MODEL ABLATION EXPERIMENTS

The ablation experiments all used the same 360 manually annotated training sentences, 800 test sentences, and other training parameters and reduced the features one by one: the dynamic dictionary expansion, the semantic confidence constraints, the confidence weighting and the collaborative iteration. However, evaluating the contribution of each information extraction feature independently can hide the failure of the information extraction system locally, especially due to the domain knowledge coverage and the quality of supervision that affects the performance of domain-specific information extraction (Premasiri et al. , 2025 [19]). As presented in Table 6, the results of Micro-F1 score for BERT-BiLSTM-CRF and BERT-BiLSTM-CRF with all features are 81.81% and 89.17% respectively, with a gap of 7.36 percentage points.

The absence of dynamic dictionary resulted in Micro-F1 loss of 4.11 percentage points and Recall loss of 4.80 percentage points, suggesting that not having enough coverage on the entities was the key reason for incorrect recognitions. Once semantic constraints have been removed and confidence weighting, Micro-F1 dropped 2.41 and 2.15 percentage points respectively, showing that both wrong categories and low-confidence labels hurt the model's parameter estimation. Eliminating collaborative iteration, Macro-F1 fell by 1.69 percentage point, which shows that entity feedback not only enlarges the scale of the dictionary, but also enhances the recognition stability of low-frequency categories. The all configuration shows that the gap between Precision and Recall is not very large, indicating that the effect of each component is complementary to each other, and it is not possible to use a single component to improve the results.

### F. ANALYSIS OF MODEL PERFORMANCE UNDER DIFFERENT LOW-RESOURCE ANNOTATION RATIOS

Performance of the full model and the baseline recognition model both steadily improved as the percentage of manually annotated data rose from 1% to 50% but the gap between the two models slowly decreased as this percentage increased.

TABLE 4. Comparison of Annotation Quality Among Different Weak Label Generation Methods

| Weak Label Generation Method | Precision/% | Recall/% | F1/% | Boundary Accuracy/% | Category Accuracy/% |
|---|---|---|---|---|---|
| Exact Match from Fixed Dictionary | 82.14 | 58.77 | 68.52 | 86.05 | 84.73 |
| Dynamic Dictionary Exact Match | 86.72 | 71.46 | 78.35 | 89.21 | 87.64 |
| +Resolution of Boundary Conflicts | 89.63 | 72.18 | 79.96 | 92.88 | 89.15 |
| + Semantic Consistency Constraints | 92.41 | 73.86 | 82.10 | 93.42 | 93.05 |
| + Confidence Interval Stratified Filtering | 94.18 | 76.92 | 84.68 | 95.36 | 94.71 |

TABLE 5. Comparison of Performance Among Different Named Entity Recognition Methods

| Method | Precision/% | Recall/% | Micro-F1/% | Macro-F1/% |
|---|---|---|---|---|
| CRF | 72.84 | 63.91 | 68.09 | 64.87 |
| BiLSTM-CRF | 78.16 | 70.82 | 74.31 | 71.56 |
| BERT-CRF | 83.74 | 76.29 | 79.84 | 77.12 |
| BERT-BiLSTM-CRF | 85.22 | 78.67 | 81.81 | 79.24 |
| + Fixed Dictionary with Weak Labels | 86.73 | 82.16 | 84.38 | 81.97 |
| +Dynamic Dictionary with Weak Labels | 88.62 | 85.47 | 87.02 | 84.90 |
| +Confidence-Weighted | 90.41 | 87.96 | 89.17 | 87.32 |

TABLE 6. Ablation experiment results for different combinations of functional modules

| Model Configuration | Dynamic Dictionary | Semantic Confidence Constraints | Confidence Weighting | Cooperative Iteration | Precision/% | Recall/% | Micro-F1/% | Macro-F1/% |
|---|---|---|---|---|---|---|---|---|
| Base Recognition Model | × | × | × | × | 85.22 | 78.67 | 81.81 | 79.24 |
| Remove dynamic dictionary | × | √ | √ | × | 87.04 | 83.16 | 85.06 | 82.61 |
| Removing semantic belief constraints | √ | × | √ | √ | 88.21 | 85.36 | 86.76 | 84.38 |
| Remove confidence-weighted values | √ | √ | × | √ | 88.62 | 85.47 | 87.02 | 84.90 |
| Exclude collaborative iteration | √ | √ | √ | × | 89.03 | 86.49 | 87.74 | 85.63 |
| Full Model | √ | √ | √ | √ | 90.41 | 87.96 | 89.17 | 87.32 |

To prevent the self-reliance bias in evaluation which can be caused by the presence of weak labels in the test data, 800 manually annotated sentences were consistently used throughout the testing stage, where they served as an independent gold standard corpus for entity boundary and category performance evaluation (Murphy et al. 2025 [20]). As shown in Figure 5, the Micro-F1 scores of the complete model at annotation ratios of 1%, 5%, 10%, 20%, and 50% reached 77.46%, 84.02%, 89.17%, 91.63%, and 93.08%, respectively, compared to 64.18%, 74.96%, 81.81%, 86.24%, and 89.41% for the baseline model, respectively.

The full model obtained a 13.28 percentage point improvement of the Micro-F1 score at 1% annotation, which dropped to 3.67 percentage points when annotated at 50%. The performance gap decreases with the increase of human supervision, and human supervision mainly serves to compensate for the phase of insufficient entity coverage, the more human supervision is provided, the more stable boundaries and categories features the model will learn.

### G. ANALYSIS OF DYNAMIC DICTIONARY ITERATION AND MODEL CONVERGENCE PERFORMANCE

They ran collaborative iteration under 10% human annotation, while maintaining the numbers of unlabeled sentences (34, 660) and validation sentences (800) as fixed. After the first round, the dictionary size grew to 4, 354 with the entity coverage rising from 38.10% to 71.05% and Micro-F1 from 81.81% to 87.74% as shown in Figure 6. The dictionary size kept increasing to 4, 665 in rounds 2-4, the coverage reached 76.58% and Micro-F1 rose to 89.01%. The first two phases were largely aimed at supplementing entities, with the intermediate phases largely focusing on adding low-frequency entities and entities with unclear semantic boundaries.

In rounds 5 and 6, the dictionary sizes were 4, 682 and 4, 686 entries, respectively; coverage increased slightly from 77.12% to 77.22%, and Micro-F1 rose from 89.15% to 89.17%. The size of the dictionaries did not increase by more than 1% in both rounds, and F1 change was less than 0.002, which meets the requirement for consecutive rounds of joint convergence: $\Delta_D^{(k)} < 0.01$ and $\Delta_F^{(k)} < 0.002$. The experiments show that the positive side effects of the high confidence recirculation can quickly fill gaps in the dictionary in early stages and the multi-feature admission constraints can effectively reduce the amount of new entities added to the dictionary when it is close to completion, avoiding the accumulation of low quality entities.

## IV. DISCUSSION

### A. IMPACT OF DYNAMIC DOMAIN DICTIONARIES ON ENTITY COVERAGE IN RESOURCE-CONSTRAINED SETTINGS

A dynamic dictionary in a low resource situation has not just the role to increase the number of entries, but also to

throughout the testing stage, where they served as an independent gold standard corpus for entity boundary and category performance evaluation (Murphy et al. 2025[20]). As shown in Figure 5, the Micro-F1 scores of the complete model at annotation ratios of 1%, 5%, 10%, 20%, and 50% reached 77.46%, 84.02%, 89.17%, 91.63%, and 93.08%, respectively, compared to 64.18%, 74.96%, 81.81%, 86.24% and 89.41% for the baseline model, respectively.

The full model obtained a 13.28 percentage point improvement of the Micro-F1 score at 1% annotation, which dropped to 3.67 percentage points when annotated at 50%. The performance gap decreases with the increase of human supervision, and human supervision mainly serves to compensate for the phase of insufficient entity coverage, the more human supervision is provided, the more stable boundaries and categories features the model will learn.

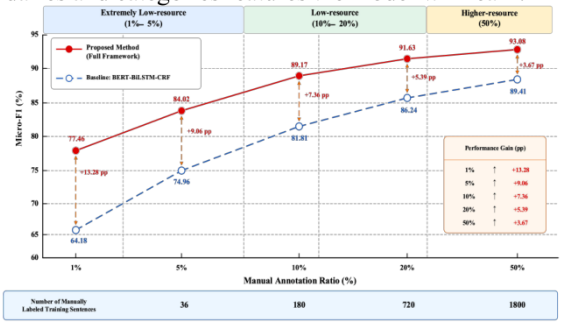


**Fig. 5.** Changes in Named Entity Recognition Performance Under Different Manual Annotation Ratios

completion, avoiding the accumulation of low quality entities.

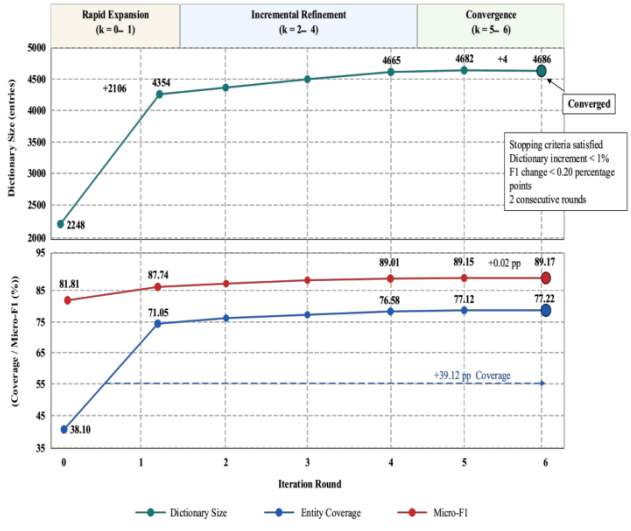


**Fig. 6.** Changes in dictionary size, entity coverage, and recognition performance across different iteration rounds

change the way by which domain specific entities come into the supervision system. The first dictionary of 2482 entries only included 38.10% of entities. 2, 438 new entries were added to the dictionary, after multi-feature screening and model feedback, the coverage rate of the dictionary is 77.22%, the accuracy of the newly added entries is 94.67%. These results show that the four constraints, term frequency, internal cohesion, contextual freedom and category semantic prototypes, can be combined to increase the likelihood of long-tail terms with stable domain semantics being included in the expansion while decreasing the likelihood of coverage inflation from invalid expansion.

The greater the information gain from dictionary expansion, the lower the level of resources. The Micro-F1 score of the complete model achieved a 13.28 percentage point improvement over the baseline model when 1% of the data was human-annotated, and a 3.67 percentage point improvement when 50% of the data was human-annotated. This difference reflects the fact that, when supervised samples are limited, the dynamic dictionary can be used to fill gaps in the external domain knowledge, while the more the number of samples manually given labels, the more the category features can be directly learned from them, and the less additional contribution the dynamic dictionary can bring. The scale of dictionary coverage and the reliability of labels need to be constrained jointly so that the recognition performance can be steadily improved when the new added entities can be turned into reliable weak labels.

### *B. THE ROLE OF THE WEAK LABEL CONFIDENCE MECHANISM IN NOISE SUPPRESSION*

Label noise is the result of uncertain dictionary matches, changes in entity boundaries and misclassification of entities into categories, and increasing the coverage of the dictionary does not reduce the noise. The hierarchical confidence mechanism takes into account dictionary match strength, semantics of the context, boundaries integrity, and entry reliability, and punishes conflicting information. With full filtering, the F1 value for weak labels has improved from 78.35% (obtained by direct dynamic dictionary matching) to 84.68%, the boundary accuracy is 95.36% and the category accuracy is 94.71%, which means that noise control is both entity scope-based and category-based.

Confidence weighting is used to further regulate the effect of low confidence labels on parameter changes. As soon as this mechanism had been removed, Micro-F1 declined from 89.17% to 87.02% showing that uniform weighting worsens the effect of incorrect labels. This is because the 0.60 admission threshold filters out obviously poor samples, and the 0.80 hierarchical threshold indicates a different level of supervision strength and is a soft constraint, meaning that some valuable uncertain samples are not filtered. This mechanism also decreases the chance that wrong predictions get fed back into the dynamic dictionary, which means noise can't build up in iterative updates.

### C. METHOD APPLICABILITY AND LIMITATIONS

It is especially suitable for entity recognition problems with limited or nonexistent human-annotated data, and large amounts of domain-specific unannotated data, and with some initial set of terminology resources. In our experiments, we used seed dictionary consisting of 2, 248 terms and a set of 34, 660 unannotated sentences, which facilitated candidate discovery and the generation of weak-labels. Even when human annotation is limited to only 1%, we achieved 13.28 percentage points improvement on Micro-F1, proving that this mechanism can be effective in alleviating information deficiency during the supervision-scarce phase. The number of rounds to reach a stable state of the model is limited to 6, and the size of the dictionaries and the parameters to be updated is not too large. But there are still some obvious boundaries of applicability to this method. So far, only 5 types of specialized entities exist in the current entity schema, and the semantic prototypes and word frequency thresholds as well as the confidence threshold (0.60, 0.80 and 0.90) are all based on the distribution of the training set, and need to be calibrated when they are transferred to other domains. The final Entity Coverage rate is 77.22%, meaning that there is still a possibility of filtering extremely low-frequency terms, new acronyms and high context-ambiguous entities. Increasing the manually annotated data to 50% resulted in a more marginal benefit of only 3.67 percentage points, supporting the observation that as the amount of supervisory data becomes relatively rich, the benefits of elaborate weak supervision loops is minimal.

## V. CONCLUSIONS

The approach that updates a dynamic domain dictionary together with weakly labelled confidence scores to effectively reduce the problems of insufficient entity coverage and supervised noise in low-resource specialised corpora in a collaborative way. After six rounds of iteration, under 1% human annotation, the performance of the complete model was Micro-F1=89.17%, F1 for weakly labeled data was 84.68%, the dictionary coverage increased from 38.10% to 77.22%. These results are consistent with the common sense of a domain knowledge compensation in existing low-resource entity recognition research and a weak-supervision denoising, and suggest that they should be restricted in a same iterative mechanism. The current validation is limited to five entity classes and only a single specialized corpus and the fixed threshold is also corpus-dependent. The transfer of entities across domains and dynamic confidence thresholds as well as the detection of open category entities should be further explored to make the method more stable to changes in terminology and domain.

## FUNDING

This work was supported by the Liaoning Provincial Science and Technology Program Joint Project (Natural Science Foundation—General Program) (2024-MSLH-203).

## REFERENCES

[1] A. E. Torres, E. S. de Moura, A. S. da Silva, et al., "An experimental study on data augmentation techniques for named entity recognition on low-resource domains," Natural Language Processing, vol. 32, no. 2, pp. 184–203, 2026.
[2] K. Nastou, M. Koutrouli, S. Pyysalo, et al., "Improving dictionary-based named entity recognition with deep learning," Bioinformatics, 40(Supplement_2): ii45-ii52, 2024.
[3] M. Bali and S. P. Anandaraj, "Reinforcement learning based distantly supervised biomedical named entity recognition," Intelligent Decision Technologies, vol. 17, no. 2, pp. 317–330, 2023.
[4] W. L. Seow, I. Chaturvedi, A. Hogarth, et al., "A review of named entity recognition: from learning methods to modelling paradigms and tasks," Artificial Intelligence Review, vol. 58, no. 10, p. 315, 2025.
[5] R. Zanoli, A. Lavelli, D. V. do Amarante, et al., "Assessment of the E3C corpus for the recognition of disorders in clinical texts," Natural Language Engineering, vol. 30, no. 4, pp. 851–869, 2024.
[6] I. Belhajem, "Effects of Multiple Annotation Schemes on Arabic Named Entity Recognition," Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 17060–17067, 2024.
[7] A. S. Behr, M. Völkenrath, and N. Kockmann, "Ontology extension with NLP-based concept extraction for domain experts in catalytic sciences," Knowledge and Information Systems, vol. 65, no. 12, pp. 5503–5522, 2023.
[8] I. Morazzoni, V. Scotti, and R. Tedesco, "Def2Vec: you shall know a word by its definition," International Journal of Speech Technology, vol. 27, no. 4, pp. 887–899, 2024.
[9] M. A. Marjan and T. Amagasa, "Domain-adaptive entity recognition: unveiling the potential of CSER in cybersecurity and beyond," International Journal of Machine Learning and Cybernetics, vol. 16, no. 5, pp. 2849–2867, 2025.
[10] U. Etudo and V. Y. Yoon, "Ontology-based information extraction for labeling radical online content using distant supervision," Information Systems Research, vol. 35, no. 1, pp. 203–225, 2024.
[11] A. O. B. Şapcı, H. Kemik, R. Yeniterzi, et al., "Focusing on potential named entities during active label acquisition," Natural Language Engineering, vol. 30, no. 3, pp. 602–624, 2024.
[12] S. Arslan, "Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text," Neural Computing and Applications, vol. 36, no. 15, pp. 8371–8382, 2024.
[13] Á. García-Barragán, A. Sakor, M. E. Vidal, et al., "NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes," Medical & biological engineering & computing, vol. 63, no. 3, pp. 749–772, 2025.
[14] A. Alotaibi, F. Nadeem, and M. Hamdy, "Weakly supervised deep learning for arabic tweet sentiment analysis on education reforms: Leveraging pretrained models and llms with snorkel," IEEE Access, vol. 13, pp. 30523–30542, 2025.
[15] B. S. Lancheros, G. Corpas Pastor, and R. Mitkov, "Data augmentation and transfer learning for cross-lingual Named Entity Recognition in the biomedical domain," Language Resources and Evaluation, vol. 59, no. 2, pp. 665–684, 2025.
[16] C. Çetindağ, B. Yazıcıoğlu, and A. Koç, "Named-entity recognition in Turkish legal texts," Natural Language Engineering, vol. 29, no. 3, pp. 615–642, 2023.
[17] I. Belhajem, "Voting Strategies for Arabic Named Entity Recognition using Annotation Schemes," Engineering, Technology & Applied Science Research, vol. 14, no. 6, pp. 17690–17695, 2024.
[18] A. A. Abdullahi, M. C. Ganiz, U. Koç, et al., "Deep learning for named entity recognition in Turkish radiology reports," Diagnostic and Interventional Radiology, vol. 31, no. 5, p. 430, 2025.
[19] A. Gopalakrishnan, K. P. Soman, S. Rajendran, et al., "Efficient text analysis: a BERT-based approach to named entity recognition (NER) and classification for Malayalam language," International Journal of Information Technology, vol. 17, no. 7, pp. 4021–4027, 2025.
[20] D. Premasiri, T. Ranasinghe, R. Mitkov, et al., "Survey on legal information extraction: current status and open challenges: D. Premasiri et al," Knowledge and Information Systems, vol. 67, no. 12, pp. 11287–11358, 2025.
[21] R. M. Murphy, D. A. Dongelmans, N. F. de Keizer, et al., "Creation of a gold standard Dutch corpus of clinical notes for adverse drug event detection: the Dutch ADE corpus," Language Resources and Evaluation, vol. 59, no. 3, pp. 2763–2779, 2025.