Automatic Generation of Entity Class Names via Iterative Semantic Refinement and Divide-and-Transfer Mechanisms

1. Introduction

1.1 The Annotation Bottleneck in Named Entity Recognition

The field of Natural Language Processing (NLP) has witnessed a paradigm shift in recent years, moving from rule-based systems to statistical models, and subsequently to deep learning architectures that dominate the current landscape. Within this evolution, Named Entity Recognition (NER) remains a foundational task, serving as the bedrock for downstream applications ranging from Knowledge Graph (KG) construction and relation extraction to question answering and semantic search.[1] Traditionally, NER has been formulated as a supervised sequence labeling problem, relying heavily on large-scale, manually annotated corpora such as CoNLL-2003, OntoNotes, or ACE2004. These datasets define rigid taxonomies of entity types (e.g., PERSON, ORGANIZATION, LOCATION) and require annotators to mark specific spans of text that correspond to these categories.[2]

However, this supervised paradigm faces a critical bottleneck: the scarcity of labeled data in specialized domains and the inherent rigidity of predefined schemas. As industries digitize, the volume of unstructured text in fields like biomedicine, material science, legal contracting, and technical manufacturing is exploding. These domains are characterized by highly specific entity types—such as "chemical compound," "legal plaintiff," "machine part," or "process parameter"—that are absent from general-purpose datasets.[4] The cost of creating new annotated datasets for every new domain is prohibitive, both in terms of financial resources and the requisite subject matter expertise.[1] Furthermore, the dynamic nature of language means that new entity types emerge continuously (e.g., "cryptocurrency," "Covid-19 variant"), rendering static models obsolete without frequent, labor-intensive retraining.

1.2 The Shift Toward Unsupervised and Open-World NER

In response to these challenges, the research community has increasingly pivoted toward low-resource approaches, including few-shot, zero-shot, and unsupervised learning. Zero-shot methods, particularly those leveraging Large Language Models (LLMs) via prompting, have demonstrated remarkable ability to identify entities based on descriptive labels rather than training examples.[7] However, even zero-shot models typically operate under a "closed-world" assumption: they require the user to pre-specify the set of target entity

types. This limits their utility in truly exploratory scenarios where the ontology of the target domain is unknown *a priori*.[1]

True "Open-World" NER—where a system discovers both the entities and their categories from raw text without predefined labels—remains an elusive goal. Existing unsupervised attempts often struggle with two fundamental issues:

1. **Imprecise Mention Detection:** Without supervision, identifying where an entity begins and ends is difficult. Heuristic methods fail on complex noun phrases, while unsupervised grammar induction is often too noisy.[10]
2. **Semantic incoherence in Typing:** Clustering entity embeddings often yields groups that are geometrically distinct but semantically incoherent or misaligned with human intuition. For instance, a clustering algorithm might group entities based on syntactic role rather than semantic type, or fail to distinguish between closely related concepts like "Protein" and "Gene".[12]

1.3 Proposed Framework: Iterative Cluster-Critic (ICC)

This report introduces a novel, comprehensive framework designed to address the unique challenges of automated entity class generation: the **Iterative Cluster-Critic (ICC)** architecture. This framework synthesizes three advanced methodologies into a coherent, self-correcting pipeline:

1. **Divide-and-Transfer Mechanism for Mention Detection:** We adopt the "Divide-and-Transfer" paradigm, which decouples entity span detection from type classification. By observing that span detection (identifying valid entity boundaries) is significantly more transferrable across domains than type classification, we utilize a robust, multi-view ensemble extractor trained on a high-resource source domain to identify candidate spans in the target domain without supervision.[7]
2. **Adaptive Semantic Clustering:** To discover entity types in an open-world setting, we employ adaptive clustering algorithms (specifically **Dip-means**) that do not require a pre-specified number of clusters (k). This allows the model to "discover" the natural granularity of the target domain's ontology based on the data distribution.[15]
3. **Critic-Based Refinement Loop:** The core innovation of the ICC framework is the introduction of a feedback loop where a provisional NER model acts as a "Critic." By training a classifier on the initial clusters and analyzing its **confusion matrix**, the system identifies clusters that are indistinguishable in context (candidates for merging) or internally inconsistent (candidates for splitting). This iterative process uses the model's own confusion as a training signal to refine the ontology dynamically.[17]

The primary objective of this research is to automate the generation of high-quality, semantically meaningful entity class names directly from corpus data. By eliminating the dependency on external knowledge bases or predefined schemas, the ICC framework aims to democratize access to advanced Information Extraction (IE) capabilities for

resource-constrained domains.

2. Theoretical Framework and Literature Synthesis

To establish the validity of the proposed ICC framework, it is necessary to deconstruct the theoretical underpinnings of its components: the Divide-and-Transfer paradigm, the mechanics of ensemble tagging schemes, and the statistical principles of adaptive clustering.

2.1 The Divide-and-Transfer Paradigm in Cross-Domain NER

The "Divide-and-Transfer" paradigm addresses the fundamental challenge of domain shift in NER. Standard supervised NER models learn to identify spans and assign types simultaneously (e.g., using a single BiLSTM-CRF or BERT-based token classifier). When applied to a new domain, these monolithic models fail catastrophically because the label distribution changes (e.g., "Bank" means a financial institution in News but a river feature in Geography).[7]

However, research suggests that the two sub-tasks of NER—**Entity Span Detection (ESD)** and **Entity Type Classification (ETC)**—exhibit discrepant transferability.

- **ESD Transferability:** The syntactic structure of named entities is largely robust across domains. Entities are typically noun phrases, often capitalized, and occupy specific syntactic roles (subject, object) within a sentence. A model trained to recognize entity boundaries in news text can often identify entity boundaries in biomedical text with reasonable accuracy, even if it cannot name them.[7]
- **ETC Specificity:** Entity types are highly domain-specific. The semantic class "Protein" does not exist in the news domain, and "Geopolitical Entity" may not be relevant in a manufacturing corpus.

The Divide-and-Transfer approach exploits this by training a dedicated ESD module on a high-resource Source Domain ($\mathcal{D}_S$) and transferring it to the Target Domain ($\mathcal{D}_T$) to generate candidate spans. This effectively bootstraps the unsupervised process by solving the "where is the entity?" problem using transfer learning, leaving the "what is the entity?" problem for unsupervised discovery.[7]

2.2 Advanced Tagging Schemes for Robust Detection

In the context of the Divide-and-Transfer mechanism, the choice of tagging scheme significantly impacts the robustness of the boundary detection. While the **BIO** (Begin, Inside, Outside) scheme is standard, it has limitations, particularly regarding discontinuous entities or boundary ambiguity in complex technical text.[1] The ICC framework employs an ensemble of three distinct schemes to maximize recall:

1. **BIO Scheme:** The traditional standard. Effective for simple, continuous entities but prone to error propagation if the "B" tag is missed.[21]

2. **Start-End (SE) Scheme:** This scheme treats boundary detection as two binary classification tasks: predicting the probability of a token being a Start index and the probability of it being an End index. This approach is more robust to long entities, as the prediction for the end token is not strictly dependent on the sequence of intermediate Inside tags.[19]
3. **Tie-or-Break (TB) Scheme:** This scheme explicitly models the relationship between *adjacent* tokens. A "Tie" label indicates that token t_i and t_{i+1} belong to the same entity, while a "Break" indicates a boundary. This captures local coherence and is particularly effective for multi-word entities (e.g., "sodium chloride") where the connection between words is strong.[19]

By ensembling these views, the ICC framework can recover entities that might be missed by a single scheme, providing a high-recall set of candidates for the clustering phase.[19]

2.3 Adaptive Clustering via Dip-means

A central challenge in unsupervised NER is determining the number of entity types (k). Standard clustering algorithms like K-means require k to be pre-specified, which is impossible in an open-world setting where the ontology is unknown. Heuristics like the Silhouette Score or the Elbow Method are often ambiguous and computationally expensive to compute for every possible k.[27]

The **Dip-means** algorithm offers a statistically grounded solution. It is an incremental clustering algorithm that starts with a single cluster and recursively attempts to split clusters based on **Hartigan's Dip Test** for unimodality.[15]

- **The Dip Test:** This is a non-parametric test that measures the departure of a distribution from unimodality.
- **The Mechanism:** For a given cluster, Dip-means computes the pairwise distances between all points. If the distribution of these distances is significantly multimodal (i.e., has multiple "humps"), it suggests the cluster actually contains two distinct sub-groups (e.g., "City" and "Country" lumped together). The algorithm then splits the cluster (e.g., using 2-means) and repeats the test on the sub-clusters.[16]

This approach allows the data distribution itself to dictate the granularity of the taxonomy, avoiding the arbitrary imposition of a fixed number of classes.[15]

2.4 The Critic Concept in Machine Learning

The "Critic" component of the ICC framework draws inspiration from **Actor-Critic** methods in Reinforcement Learning (RL) and **Generative Adversarial Networks (GANs)**. In RL, the Critic estimates the value function to guide the Actor's policy updates.[31] In the context of unsupervised classification, we can frame the clustering module as the "Actor" (proposing

labels) and a downstream classifier as the "Critic" (evaluating learnability).

Recent work in "Self-Critique" and feedback loops suggests that model performance on a downstream task can serve as a powerful signal for optimizing upstream data representation.[8] If a supervised model trained on the generated clusters fails to generalize (i.e., it confuses two clusters or has low confidence), this "confusion" is a signal that the upstream clustering is flawed.[17] The ICC framework formalizes this intuition into a rigorous refinement algorithm.

3. Methodology: The Iterative Cluster-Critic (ICC) Framework

The ICC Framework is a cyclic pipeline composed of four integrated modules. It moves from unsupervised mention extraction to embedding, followed by an iterative cycle of clustering and refinement, and finally, semantic naming.

3.1 Module I: Cross-Domain Ensemble Mention Detection

The first module is responsible for identifying entity mentions in the target corpus $\mathcal{D}_T$ without access to any target annotations. We utilize the Divide-and-Transfer strategy trained on a source domain $\mathcal{D}_S$ (e.g., CoNLL-2003).

Architecture:
We employ a shared encoder backbone (e.g., DeBERTa-v3-Large) with three distinct decoder heads, each corresponding to a different tagging scheme: BIO, SE, and TB.
Mathematical Formulation of the Tagging Schemes:

1. **BIO Decoder:** Standard CRF layer maximizing $P(\mathbf{y}|\mathbf{x})$.
2. Start-End (SE) Decoder: Two binary classifiers for each token x_i:

   $$P_{start}(x_i) = \sigma(W_s h_i + b_s)$$
   $$P_{end}(x_i) = \sigma(W_e h_i + b_e)$$

   An entity is formed by any pair (i, j) where $P_{start}(x_i) > \tau$ and $P_{end}(x_j) > \tau$ with $j \ge i$.19
3. Tie-or-Break (TB) Decoder: This explicitly models the transition between adjacent tokens. Let h_{i-1} and h_i be the contextual embeddings of tokens w_{i-1} and w_i. We concatenate these to form a "token interaction representation" 19:

   $$v_{inter} = [h_{i-1}; h_i]$$

   The probability of the relation $r_k \in \{ \text{Tie, Break} \}$ is computed as:
   $$ P(r_k | w_{i-1}, w_i) = \frac{\exp(\mathbf{w}k^T v{inter} + b_k)}{\sum_{j \in R} \exp(\mathbf{w}j^T v{inter} + b_j)} $$
   A "Tie" implies the tokens are part of the same entity; "Break" implies a boundary. This

formulation is notably robust for detecting multi-token entities in technical domains (e.g., "reactive oxygen species") where the semantic bond between words is strong.[19]

Ensemble Strategy:
To maximize recall, the final set of candidate mentions M_{cand} is the union of mentions detected by all three heads:

$$M_{cand} = M_{BIO} \cup M_{SE} \cup M_{TB}$$

Using the union rather than intersection is a deliberate design choice for unsupervised learning: false positives can be filtered out during clustering (where they will form outliers or incoherent clusters), but false negatives are lost forever.19

3.2 Module II: Contextual Embedding and Initial Clustering

Once mentions are extracted, we must group them into semantic categories.

Contextual Embedding:
For each extracted mention m, we derive a fixed-length vector representation. Simple averaging of token embeddings is often insufficient. We use Prompt-Based Encoding following the OWNER architecture.1 For a mention m in context X, we append a prompt:

$$\text{Input}: X m \text{ is a}$$

The embedding of the `` token is taken as the representation of the entity type. This leverages the pre-trained knowledge of the Language Model to project the entity into a semantic "type space" rather than just a contextual space.1
Initial Clustering with Dip-Means:
We initialize the clustering using Dip-means to avoid guessing k.

1. **Initialization:** Start with a single global cluster C_1 containing all entity embeddings.
2. Recursion: For each leaf cluster C_i:
   a. Compute the pairwise distance matrix D for all points in C_i.
   b. Project the data onto the vector defined by the two furthest points (rough principal component).
   c. Calculate the Dip Statistic D_n on this 1D projection to test for unimodality.
   d. If $P(D_n) < \alpha$ (reject unimodality), split C_i into two sub-clusters using 2-means and repeat.15
   e. Else, mark C_i as a final cluster.

This results in an initial set of clusters $\mathcal{C}_{init}$ that respects the geometric structure of the embedding space.[16]

3.3 Module III: The Critic-Based Refinement Loop

Geometric clustering alone is insufficient because "semantic distinctness" in vector space does not always map to "functional distinctness" for a classifier. The ICC framework introduces a **Refinement Loop** to align the clusters with the decision boundaries of a discriminative model.

Step 1: Train the Critic
We treat the current cluster assignments $\mathcal{C}_{t}$ as pseudo-labels. We split the corpus into a training set T_{train} and a hold-out validation set T_{val}. We train a lightweight classifier (e.g., DistilBERT), denoted as f_θ, to predict the cluster ID of an entity given its context.
Step 2: Compute Confusion Matrix
We evaluate f_θ on T_{val} to generate a confusion matrix M, where M_{ij} represents the number of entities belonging to cluster i (according to the current clustering) that were predicted as cluster j by the Critic.17
Step 3: Refinement Heuristics
We derive two scalar metrics from M to drive the split/merge decisions.

- Merge Heuristic (Addressing Over-Segmentation):
  If two clusters i and j are frequently confused, they likely represent the same semantic category (e.g., "US" assigned to Cluster A and "France" to Cluster B, but both are "Countries"). We define the Symmetric Confusion Score (SCS):

  $$SCS(i, j) = \frac{M_{ij} + M_{ji}}{|c_i| + |c_j|}$$

  If $SCS(i, j) > \tau_{merge}$, we merge clusters i and j.18 This uses the Critic's inability to distinguish classes as proof of their semantic identity.
- Split Heuristic (Addressing Under-Segmentation):
  If a cluster i has low diagonal agreement, it means the Critic cannot consistently learn a representation for it. This suggests the cluster is an incoherent amalgamation of different types (e.g., a "Misc" cluster containing dates and money). We define the Internal Confidence Score (ICS):

  $$ICS(i) = \frac{M_{ii}}{|c_i|}$$

  If $ICS(i) < \tau_{split}$, we force a split on cluster i. Unlike the initial Dip-means which uses geometry, this step splits based on learnability. We apply X-means or recursive K-means on just the points in c_i to find a sub-structure that minimizes internal variance.36

Step 4: Iteration
The clusters are updated based on the split/merge operations. The Critic is re-trained (or fine-tuned) on the new labels. This cycle repeats until the Delta AMI (change in Adjusted

Mutual Information between iterations) falls below a threshold ϵ, indicating stability.9

3.4 Module IV: Semantic Naming via Prototype Prompting

The final output of the refinement loop is a set of stable, unlabeled clusters. To assign human-readable names, we utilize an LLM (e.g., GPT-4) with a **Prototype-Based Prompting** strategy.

Instead of feeding random samples, we select:

1. **Centroids:** The 5 entities closest to the cluster center (representing the core concept).
2. **Boundary Cases:** The 5 entities furthest from the center (representing the scope/diversity).

Prompt Structure:

> "You are an expert taxonomist. I will provide a list of entities that belong to the same category.
> Core Examples: {centroids}
> Edge Examples: {boundaries}
> Identify the single most specific semantic category that encompasses all these examples. Return only the category name."

This approach ensures the generated name is robust enough to cover the variance within the cluster while remaining specific enough to be useful.[1]

4. Experimental Design

To rigorously evaluate the ICC framework, we propose a comprehensive experimental protocol that tests both the quality of the generated ontology and the accuracy of the entity extraction.

4.1 Datasets and Domains

We utilize a **Cross-Domain** setup to strictly enforce the unsupervised/open-world constraint. The Source Domain ($\mathcal{D}_S$) is used *only* to train the Mention Detector (Module I). The Target Domains ($\mathcal{D}_T$) are used for unsupervised discovery; no labels from $\mathcal{D}_T$ are used for training.

Source Domain ($\mathcal{D}_S$):

- **CoNLL-2003 (News):** The standard benchmark. Contains 4 broad types (PER, LOC, ORG, MISC). Used to train the "Divide-and-Transfer" mention detector.[1]

Target Domains ($\mathcal{D}_T$):
We select four diverse datasets to test generalization across different semantic landscapes:

1. **OntoNotes 5.0 (General):** Contains 18 fine-grained types (e.g., PRODUCT,

WORK_OF_ART, EVENT). High cardinality challenges the clustering algorithm.[39]

2. **BC5CDR (Biomedical):** Contains "Chemical" and "Disease" entities. Tests adaptability to scientific lexicon and complex chemical names (e.g., "1,3-dimethylamylamine").[40]
3. **WNUT-17 (Social Media):** Noisy, informal text with emerging entities. Tests robustness to non-standard syntax.[41]
4. **FabNER (Manufacturing):** A specialized technical dataset containing entity types like "Manufacturing Process," "Machine," and "Material." This represents the "long tail" of industrial applications where unsupervised NER is most needed.[1]

4.2 Baselines

We compare ICC against three categories of baselines to contextualize its performance:

Category	Models	Description
State-of-the-Art Zero-Shot	**UniNER** [7], **GliNER** [1], **ChatIE** [7]	These models use massive pre-training or LLM prompting. They operate in a "closed-world" setting (require label names as input) but represent the current upper bound.
Unsupervised Frameworks	**OWNER** [1], **UNER**	Direct competitors. OWNER uses static K-means without a Critic loop. Comparing against OWNER isolates the benefit of the *Critic Refinement* and *Dip-means*.
Ablated ICC Variants	**ICC-NoCritic**, **ICC-KMeans**, **ICC-BIO**	Internal baselines to quantify the contribution of specific modules (The Critic Loop, Adaptive Clustering, and Ensemble Tagging).

4.3 Evaluation Metrics

Evaluating unsupervised NER is non-trivial because the predicted cluster IDs ($0, 1, 2...$) do

not match ground truth labels ($PER, LOC...$). We employ a mapping-based evaluation strategy alongside cluster validity indices.

1. **Mapping-Based F1 Score:** We map each predicted cluster to the ground truth class with which it has the highest overlap (using the Hungarian algorithm for optimal bipartite matching). We then calculate standard Precision, Recall, and F1 based on this mapping. This measures the *utility* of the clusters for a downstream task.[1]
2. **Adjusted Mutual Information (AMI):** An information-theoretic metric that measures the agreement between the ground truth clustering and the predicted clustering, corrected for chance. AMI is crucial for assessing how well the model captured the true ontology structure, independent of the labels.[9]
3. **Silhouette Score:** An internal validation metric that measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Used to assess quality in the absence of any ground truth.[43]
4. **Semantic Consistency (Human Eval):** We conduct a blinded study where human annotators rate the LLM-generated names for **Descriptiveness** (Does the name describe the entities?) and **Coherence** (Do the entities in the cluster belong together?) on a 5-point Likert scale.

4.4 Hyperparameters and Implementation Details

To ensure reproducibility, we define the specific configuration for the experiments:

- **Mention Detector Backbone:** DeBERTa-v3-Large, chosen for its superior performance on span-level tasks.[45]
- **Critic Model:** DistilBERT-base-cased. We use a smaller model for the Critic to reduce the computational overhead of the iterative loop. The Critic is trained for 3 epochs per refinement iteration with a learning rate of 5e-5.[46]
- **Dip-means Parameters:** Significance level $\alpha = 0.05$ for the Dip Test.
- **Refinement Thresholds:**
  - Merge Threshold $\tau_{merge} = 0.35$ (Clusters confusing >35% of samples are merged).
  - Split Threshold $\tau_{split} = 0.60$ (Clusters with <60% internal consistency are split).[47]
- **Stopping Criterion:** The loop terminates when $\Delta AMI < 0.01$ or after a maximum of 10 iterations to prevent infinite oscillation.[48]

5. Detailed Analysis of Framework Mechanics

5.1 The "Singularity" of Cluster Stability

A key hypothesis of this research is that the Refinement Loop will exhibit a "singularity" or phase transition—a point where the clusters align with the latent semantic boundaries of the

data, leading to a rapid stabilization of the confusion matrix.

In early iterations, we expect high volatility. The initial Dip-means clusters might over-segment significantly (e.g., creating separate clusters for "French politicians" and "German politicians" due to surface-level distributional differences). The Critic, however, will likely confuse these clusters because they appear in identical syntactic contexts (e.g., "Mr. said..."). This confusion drives the **Merge Heuristic** to unify them into a broader "Politician" cluster.

Conversely, a "Misc" cluster formed by outliers in the embedding space will exhibit high internal variance in the confusion matrix (low diagonal), triggering the **Split Heuristic**. As these operations proceed, the system theoretically converges toward an equilibrium where the clusters are maximally separable by the Critic. Monitoring the trajectory of the **Delta AMI** across iterations will provide empirical evidence of this convergence and may serve as a proxy for the "learnability" of the discovered ontology.[46]

5.2 The Role of "Noise" Clusters

Standard K-means forces every data point into a cluster, often contaminating pure clusters with noise (false positive mentions). The ICC framework's **Split Heuristic** offers a unique mechanism for noise handling.

When "junk" spans (e.g., "the", "of", random punctuation) are included in the candidate set, they typically form loose, incoherent clusters. The Critic model will struggle to classify these consistently (M_{ii} will be low). The system will repeatedly split these clusters. Over time, we expect valid entities to merge into large, stable clusters, while noise remains fragmented in small, unstable micro-clusters.

- **Filtration Strategy:** We introduce a post-processing step where any cluster with a size below a threshold δ (e.g., $< 1\%$ of total mentions) after convergence is discarded as noise. This effectively turns the refinement loop into an unsupervised filter for false positives.[49]

5.3 Emergent Ontologies vs. Prescriptive Taxonomies

Traditional NER evaluates success based on adherence to a rigid schema (e.g., "Did you label 'iPhone' as PRODUCT?"). However, in many domains, the "correct" taxonomy is ambiguous. Is "iPhone" a PRODUCT, or an OBJECT, or a TECHNOLOGY?

The ICC framework allows for **Emergent Ontologies**. Instead of forcing data into predefined boxes, it allows the data to suggest the categories.

- **Implication:** In the **FabNER** experiments, we anticipate the model might discover distinctions not present in the original dataset. For example, it might separate "Manufacturing Tools" (drills, lathes) from "Manufacturing Processes" (welding, casting), even if the ground truth lumps them together. The Human Evaluation metric (Semantic Consistency) is crucial here, as it may reveal that the unsupervised model has discovered

a *more* useful taxonomy than the gold standard for certain applications.[1]

6. Challenges and Mitigation Strategies

Implementing the ICC framework involves navigating several technical risks. We outline specific mitigation strategies for each.

Challenge	Description	Mitigation Strategy
Catastrophic Forgetting in Critic	The Critic model might overfit to the noisy pseudo-labels in early iterations, reinforcing errors rather than correcting them.	**Re-initialization:** We re-initialize the Critic model weights at the start of each refinement iteration. This ensures the Critic provides an unbiased assessment of the *current* clusters, rather than carrying over bias from previous, flawed clusters.[50]
Semantic Drift	Iterative merging might cause clusters to drift into overly broad categories (e.g., merging everything into "Noun Phrase").	**Anchor Constraints:** If available, we can use a small set of "high-confidence" seed entities (identified via high LLM confidence or heuristic rules) as "Anchors" that cannot be merged. This pins the semantic space.[52]
Computation Cost	Retraining a BERT classifier at every iteration is computationally expensive ($O(N \times \text{TrainTime})$).	**Freeze Encoder:** During the intermediate loops, we freeze the transformer backbone of the Critic and only fine-tune the classification head. Full fine-tuning is performed only in the final iteration. This drastically reduces the backward pass cost.[53]

Oscillation	The system might enter a cycle of merging and splitting the same clusters (A+B -> C -> A+B).	**Damping Factor:** We introduce a "momentum" or damping term to the thresholds. If a merge operation is reversed in the next step, the threshold for that specific operation is increased to prevent infinite loops.[54]

7. Conclusion

The **Iterative Cluster-Critic (ICC)** framework represents a significant departure from static unsupervised NER approaches. By replacing the "train-once" clustering paradigm with a dynamic, feedback-driven refinement loop, ICC addresses the core weakness of low-resource Information Extraction: the inability to adaptively define the boundaries between entity concepts.

The integration of the **Divide-and-Transfer** ensemble ensures that the system begins with a high-recall set of entity candidates, leveraging the cross-domain robustness of syntactic structure. The **Adaptive Clustering** (Dip-means) provides a flexible starting point for the ontology, free from the constraints of a fixed k. Finally, the **Critic Loop** transforms the inherent "confusion" of a neural model into a constructive signal, refining the ontology until it aligns with the semantic learnability of the data.

If successful, this research will demonstrate that high-performance, domain-specific NER is achievable not just with *less* data, but with *no* target annotations at all. This has profound implications for the accessibility of AI technologies in specialized fields such as digital humanities, historical archiving, and rare disease research, where annotated data serves as a barrier to entry. The ICC framework moves us closer to systems that do not just extract knowledge based on what they were told, but actively *discover* the structure of the world they analyze.

Works cited

1. OWNER__Toward_Unsupervised_Open-World_Named_Entity_Recognition.pdf
2. (PDF) ContrastSkill: Task-Oriented Contrastive Pre-Training for Enhanced Skill Extraction in Job Data - ResearchGate, accessed on December 8, 2025, https://www.researchgate.net/publication/395993025_ContrastSkill_Task-Oriented_Contrastive_Pre-Training_for_Enhanced_Skill_Extraction_in_Job_Data
3. How to perform Named Entity Recognition (NER) - Foundry Tools | Microsoft Learn, accessed on December 8, 2025, https://learn.microsoft.com/en-us/azure/ai-services/language-service/named-entity-recognition/how-to-call

4. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition - PMC, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC1764467/
5. Assessing named entity recognition by using geoscience domain schemas: the case of mineral systems - Frontiers, accessed on December 8, 2025, https://www.frontiersin.org/journals/earth-science/articles/10.3389/feart.2025.1530004/full
6. Evolution and emerging trends of named entity recognition: Bibliometric analysis from 2000 to 2023 - PMC - PubMed Central, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11066397/
7. DOZEN: Cross-Domain Zero Shot Named Entity Recognition with Knowledge Graph, accessed on December 8, 2025, https://www.researchgate.net/publication/353185720_DOZEN_Cross-Domain_Zero_Shot_Named_Entity_Recognition_with_Knowledge_Graph
8. Reinforcement learning from human feedback - Wikipedia, accessed on December 8, 2025, https://en.wikipedia.org/wiki/Reinforcement_learning_from_human_feedback
9. OWNER — Toward Unsupervised Open-World Named Entity Recognition - IEEE Xplore, accessed on December 8, 2025, https://ieeexplore.ieee.org/iel8/6287639/10820123/10930473.pdf
10. The merge-and-split heuristic and the (k,l)-means - arXiv, accessed on December 8, 2025, https://arxiv.org/abs/1406.6314
11. NER with Unsupervised Learning? - Data Science Stack Exchange, accessed on December 8, 2025, https://datascience.stackexchange.com/questions/61009/ner-with-unsupervised-learning
12. Towards Efficient and Effective Deep Clustering with Dynamic Grouping and Prototype Aggregation - arXiv, accessed on December 8, 2025, https://arxiv.org/html/2401.13581v1
13. Clustering Algorithms: Their Application to Gene Expression Data - PMC - PubMed Central, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5135122/
14. Data Augmentation for Cross-Domain Named Entity Recognition - ResearchGate, accessed on December 8, 2025, https://www.researchgate.net/publication/357123047_Data_Augmentation_for_Cross-Domain_Named_Entity_Recognition
15. Dip-means: an incremental clustering method for estimating the number of clusters, accessed on December 8, 2025, https://proceedings.neurips.cc/paper/2012/file/a8240cb8235e9c493a0c30607586166c-Paper.pdf
16. Dip-means - Argyris Kalogeratos, accessed on December 8, 2025, https://kalogeratos.com/psite/material/dip-means/
17. Feature-Dependent Confusion Matrices for Low-Resource NER Labeling with Noisy Labels, accessed on December 8, 2025, https://aclanthology.org/D19-1362/
18. Feature-Dependent Confusion Matrices for Low-Resource NER Labeling with

Noisy Labels - ACL Anthology, accessed on December 8, 2025, https://aclanthology.org/D19-1362.pdf
19. Cross-Domain NER under a Divide-and-Transfer Paradigm - Tingwen Liu, accessed on December 8, 2025, http://liutingwen.ac.cn/papers/TOIS2024-Cross-domain%20NER%20under%20a%20Divide-and-Transfer%20Paradigm.pdf
20. Cross-Domain NER using Cross-Domain Language Modeling | Request PDF - ResearchGate, accessed on December 8, 2025, https://www.researchgate.net/publication/335784138_Cross-Domain_NER_using_Cross-Domain_Language_Modeling
21. Inside–outside–beginning (tagging) - Wikipedia, accessed on December 8, 2025, https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)
22. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text - PMC - NIH, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5925775/
23. Joint Entity and Relation Extraction Network with Enhanced Explicit and Implicit Semantic Information - MDPI, accessed on December 8, 2025, https://www.mdpi.com/2076-3417/12/12/6231
24. Automatic Named Entity Recognition - Master Computer Science, accessed on December 8, 2025, https://theses.liacs.nl/pdf/2020-2021-SunJiakun.pdf
25. Learning Named Entity Tagger using Domain-Specific Dictionary - Jiawei Han, accessed on December 8, 2025, http://hanj.cs.illinois.edu/pdf/emnlp18_jshang.pdf
26. Few-shot Named Entity Recognition: Definition, Taxonomy and Research Directions - IRIS Unina, accessed on December 8, 2025, https://iris.unina.it/retrieve/fa0573f8-0687-4bc9-b132-53f99de5565e/3609483.pdf
27. An upper bound of the silhouette validation metric for clustering - arXiv, accessed on December 8, 2025, https://arxiv.org/html/2509.08625v1
28. Stopping Criteria for Iterative Solution Methods - UFPR, accessed on December 8, 2025, http://servidor.demec.ufpr.br/CFD/velhos/Mach2D7/documentos/Recktenwald_2012.pdf
29. (PDF) Dip-means: An incremental clustering method for estimating the number of clusters, accessed on December 8, 2025, https://www.researchgate.net/publication/304462885_Dip-means_An_incremental_clustering_method_for_estimating_the_number_of_clusters
30. (PDF) Dip-means: An incremental clustering method for estimating the number of clusters, accessed on December 8, 2025, https://www.researchgate.net/publication/289245736_Dip-means_An_incremental_clustering_method_for_estimating_the_number_of_clusters
31. A Guide to Actor Critic Model in Reinforcement Learning - upGrad, accessed on December 8, 2025, https://www.upgrad.com/blog/actor-critic-model-in-reinforcement-learning/
32. Learning to Learn By Self-Critique, accessed on December 8, 2025,

http://papers.neurips.cc/paper/9185-learning-to-learn-by-self-critique.pdf
33. Flow chart of merge operation and split operation in the ISODATA clustering algorithm., accessed on December 8, 2025, https://www.researchgate.net/figure/Flow-chart-of-merge-operation-and-split-operation-in-the-ISODATA-clustering-algorithm_fig3_373152768
34. A Brief History of Named Entity Recognition - arXiv, accessed on December 8, 2025, https://arxiv.org/html/2411.05057v1
35. A mathematical and computational framework for quantitative comparison and integration of large-scale gene expression data - NIH, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC1092273/
36. X-Means Clustering Explained - by Alif Arshad Bakshi - Medium, accessed on December 8, 2025, https://medium.com/@arshadbakshi/x-means-clustering-explained-e4335a31cf8e
37. K-Medoids Clustering of Data Sequences With Composite Distributions - NSF PAR, accessed on December 8, 2025, https://par.nsf.gov/servlets/purl/10119169
38. Systematic Evaluation of Convergence Criteria in Iterative Training for NLP - The Association for the Advancement of Artificial Intelligence, accessed on December 8, 2025, https://cdn.aaai.org/ocs/45/45-2332-1-PB.pdf
39. Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training - ACL Anthology, accessed on December 8, 2025, https://aclanthology.org/2021.emnlp-main.810.pdf
40. Weakly Supervised Sequence Tagging from Noisy Rules, accessed on December 8, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/6009/5865
41. clusterMaker2: a major update to clusterMaker, a multi-algorithm clustering app for Cytoscape - PMC - NIH, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10074866/
42. Refining Filter Global Feature Weighting for Fully Unsupervised Clustering - MDPI, accessed on December 8, 2025, https://www.mdpi.com/2076-3417/15/16/9072
43. Unsupervised Learning Evaluation Metrics Explained - Insight7 - Call Analytics & AI Coaching for Customer Teams, accessed on December 8, 2025, https://insight7.io/unsupervised-learning-evaluation-metrics-explained/
44. Silhouette Score. I understand that learning data science... | by Amit Yadav | Biased-Algorithms | Medium, accessed on December 8, 2025, https://medium.com/biased-algorithms/silhouette-score-d85235e7638b
45. Do LLMs Surpass Encoders for Biomedical NER? - arXiv, accessed on December 8, 2025, https://arxiv.org/html/2504.00664v1
46. Re-Examine Distantly Supervised NER: A New Benchmark and a Simple Approach - ACL Anthology, accessed on December 8, 2025, https://aclanthology.org/2025.coling-main.727.pdf
47. Improving replicability in single-cell RNA-Seq cell type discovery, accessed on December 8, 2025, https://www.biorxiv.org/content/10.1101/2020.03.03.974220v1.full.pdf
48. Unsupervised Graph Neural Network Framework for Balanced Multipatterning in Advanced Electronic Design Automation Layouts - arXiv, accessed on December

8, 2025, https://arxiv.org/html/2511.16374v1
49. High-Dimensional Cluster Analysis with the Masked EM Algorithm - MIT Press Direct, accessed on December 8, 2025, https://direct.mit.edu/neco/article/26/11/2379/8010/High-Dimensional-Cluster-Analysis-with-the-Masked
50. An efficient Split-Merge re-start for the K-means algorithm | Request PDF - ResearchGate, accessed on December 8, 2025, https://www.researchgate.net/publication/342222141_An_efficient_Split-Merge_re-start_for_the_K-means_algorithm
51. A pre-training and self-training approach for biomedical named entity recognition - NIH, accessed on December 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7872256/
52. Unlearnable Examples Detection via Iterative Filtering - arXiv, accessed on December 8, 2025, https://arxiv.org/html/2408.08143v1
53. Efficient Machine Learning in Dynamic Environments - Carnegie Mellon University, accessed on December 8, 2025, https://ml.cmu.edu/research/phd-dissertation-pdfs/donkurid_phd_ml_2025.pdf
54. 2.3. Clustering — scikit-learn 1.7.2 documentation, accessed on December 8, 2025, https://scikit-learn.org/stable/modules/clustering.html