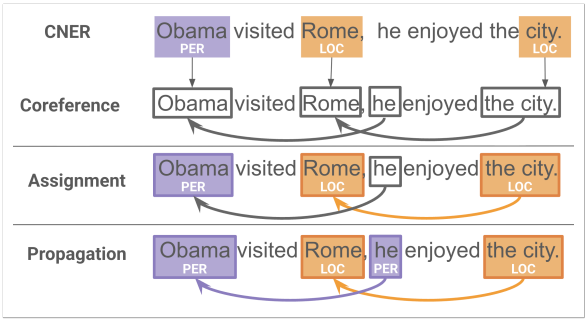
Interpretable Coreference Resolution Evaluation Using Explicit Semantics

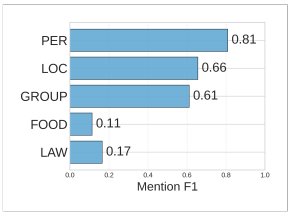
Bruno Gatti[1,†], **Giuliano Martinelli**[1,†], **Roberto Navigli**[1,2,†]
[1]Sapienza NLP Group, Sapienza University of Rome
[2]Babelscape
{gatti, martinelli, navigli}@diag.uniroma1.it

Abstract

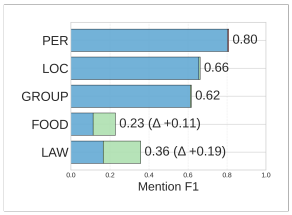
Coreference resolution is typically evaluated using aggregate statistical metrics such as CoNLL-F_1, which measure structural overlap between predicted and gold clusters. While widely used, these metrics offer limited diagnostic insights, penalizing errors without revealing whether a system struggles with specific semantic categories, such as people, locations, or events, and making it difficult to interpret model capabilities or derive actionable improvements. We address this gap by introducing a semantically-enhanced evaluation framework for coreference resolution. Our approach overlays Concept and Named Entity Recognition (CNER) onto coreference outputs, assigning semantic labels to nominal mentions and propagating them to entire coreference clusters. This enables the computation of typed scores aimed at evaluating mention extraction and linking capabilities stratified by semantic class. Across our experiments on OntoNotes, LitBank, and PreCo, we show that our framework uncovers systematic weaknesses that remain obscured by aggregate metrics. Furthermore, we demonstrate that these diagnostics can be used to design targeted, low-cost data augmentation strategies, achieving measurable out-of-domain improvements.



(a) Our labeling and propagation technique.



(b) Semantically-enhanced evaluation.



(c) Targeted data augmentation.

Figure 1: Our three main contributions: (a) Our labeling and propagation technique that overlays CNER onto coreference outputs; (b) Our interpretable per-class semantic analysis of coreference capabilities; (c) The improvements of targeted data augmentation on less represented classes.

1 Introduction

While coreference resolution models have improved substantially in recent years, evaluation methodologies have evolved far more slowly and continue to suffer from longstanding limitations.

Standard metrics such as MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_{\phi 4}$ (Luo, 2005), popularized by the CoNLL-2012 shared task (Pradhan et al., 2012), are based on statistical methods that focus on exact matches of mentions and coreference links. While effective as aggregate measures, they ignore semantic and contextual information. Therefore, minor span discrepancies or annotation mismatches can be penalized as full errors, complicating meaningful comparison across annotation schemes, genres, and domains.

A further limitation of these metrics is their lack of interpretability, as single aggregate scores often obscure systematic failure modes and conflate qualitatively different errors. For example, a model may perform well on person-centric narrative chains while failing on event- or object-related coreference, yet this degradation is not reflected in standard evaluation scores. This issue becomes especially pronounced under domain shift, where

† Authors contributed equally.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45854–45872

models may degrade unevenly across semantic categories without any clear diagnostic signal. Prior work has attempted to improve the interpretability of coreference evaluation by incorporating explicit semantic information (Agarwal et al., 2019). This approach is, however, constrained by the limited scope of the conventional Named Entity Recognition (NER) tagset, which only covers named entities and excludes common nouns, despite the fact that nominal concepts constitute a large fraction of coreference mentions. For this reason, NER-based evaluation can only sparsely annotate coreference clusters and is restricted to a small set of coarse-grained categories (typically PER, ORG, LOC, and MISC), limiting meaningful analysis across richer semantic distinctions.

To address these limitations, we propose a new evaluation framework that adopts a novel two-step labeling and propagation technique to tag coreference clusters using Concept and Named Entity Recognition (Martinelli et al., 2024b, CNER) as a semantic layer. Unlike conventional NER systems, CNER assigns semantic categories to both named entities (e.g., *Obama*, *Rome*, *Moby Dick*) and nominal concepts (e.g., *president*, *city*, *whale*) within a unified label inventory, ensuring that a meaningful semantic label can be assigned to the large majority of coreference mentions. As illustrated in Figure 1, we overlay CNER annotations onto coreference outputs using a simple overlap-based alignment, followed by cluster-level label propagation. This enables the computation of typed Mention and Link F_1 scores, turning coreference evaluation into a semantically-aware diagnostic interface. Beyond improved interpretability, we also show that these measures can be leveraged to design targeted and low-cost data augmentation strategies yielding measurable improvements in out-of-domain generalization. Specifically, our contributions are as follows:

1. **A labeling and propagation technique for semantically-enhanced coreference evaluation**, based on a simple two-step approach that merges CNER tags with coreference outputs, providing dense and granular semantic annotations.

2. **A fine-grained analysis of coreference models and datasets**, revealing systematic, semantically interpretable error patterns that remain hidden under standard metrics.

3. **An actionable evaluation framework**, demonstrating how our novel diagnostics can guide targeted and cost-effective data augmentation, leading to improved out-of-domain performance.

We release the code of our evaluation framework and the data used at `https://github.com/SapienzaNLP/cner-coref`.

2 Related Work

2.1 Coreference Resolution Evaluation

In recent years, coreference resolution has typically been evaluated using MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_{\phi 4}$ metrics (Luo, 2005), later combined into the CoNLL score (Pradhan et al., 2012).

While these metrics provide stable aggregate performance estimates, they offer little diagnostic insight: they cannot reveal whether a model systematically fails on particular semantic categories or mention types, and they often obscure qualitatively distinct error patterns.

To address structural biases, Moosavi and Strube (2016) propose the LEA metric to better weight entity importance, while Moosavi et al. (2019) introduce MINA algorithms to decouple the evaluation of coreference logic from strict boundary detection. Parallel efforts focus on diagnosing specific failures through algorithmic categorization; Kummerfeld and Klein (2013) develop a toolkit to automatically group errors into intuitive types based on entity transformations, revealing that many persistent failures stem from semantic or discourse mismatches invisible to link-based metrics. From a different methodological perspective, Martschat et al. (2015) introduce a framework that uses spanning trees to represent coreference structures, allowing for a systematic analysis of precision and recall errors.

Despite the availability of these analytical tools, recent literature suggests that evaluation challenges are deeply tied to the underlying datasets' annotation guidelines. Porada et al. (2024) demonstrate that perceived failures in out-of-domain generalization are frequently artifacts of inconsistent mention definitions rather than genuine linguistic shortcomings. This echoes arguments by Zeldes (2021), who contends that the field's heavy reliance on the OntoNotes schema has artificially constrained the task's definition. A prominent example of this constraint is the assumption of strict coreference; while psycholinguistic evidence demonstrates that

coreference often involves a more nuanced "near-identity" rather than a strict binary classification (Recasens et al., 2013), standard datasets enforce strict-identity annotations.

However, coreference evaluation continues to rely on aggregate statistical scores that lack fine-grained, semantically informed diagnostics, which is the gap our work aims to address.

2.2 Semantics and Coreference Models

A substantial body of prior work has explored incorporating semantic information within coreference models. Early feature-based systems leverage semantic class information and lexical relations to constrain antecedent selection (Ng, 2007). Subsequent joint models integrate typing, linking, and coreference into unified architectures (Durrett and Klein, 2014; Chen et al., 2017; Agarwal et al., 2022), demonstrating that shared semantic representations can improve cluster coherence. Neural approaches further refine this paradigm by injecting type information directly into span- or entity-level representations (Clark and Manning, 2015; Khosla and Rose, 2020). More recently, Mtumbuka and Schockaert (2024) show that coreference structure itself can act as supervision for fine-grained entity typing. Differently, Agarwal et al. (2019) focus on using semantics as an evaluation lens for coreference behavior, analyzing coreference performance by semantic type using a standard NER-style tagset. However, this approach comes with two limits: first, only a small set of coarse-grained categories is considered (typically PER, ORG, LOC and MISC); second, the majority of nominal, abstract, and conceptual mentions are left untyped. The resulting evaluation offers a shallow diagnostic view that is unable to reveal category-specific failure modes beyond named entities. Our work addresses this limitation by enabling semantic evaluation at a much finer granularity, covering both entities and concepts, without altering the coreference model itself. To this end, we employ Concept and Named Entity Recognition (Martinelli et al., 2024b, CNER), which is introduced specifically to address the narrow scope of traditional NER. CNER brings together named entities and nominal concepts under a unified tagset with 29 categories, ranging from PERSON and LOCATION to RELATION, EVENT, PLANT, and SUPERNATURAL, among others. We detail the full inventory in Appendix E. This yields not only a much denser annotation layer – since nearly every mention in a coreference chain can be mapped to a precise CNER category – but also far broader semantic granularity than standard NER.

3 Labeling and Propagation Technique

In this Section, we introduce our two-step labeling and propagation technique that overlays coreference outputs with the semantic annotations produced by CNER. This simple-yet-effective method assigns a semantic category to each coreference cluster by first aligning mentions to labeled spans using an overlap-based criterion, and then propagating these labels at the cluster level via majority voting.

3.1 Formal Overview

Let D be a document for which a coreference model has predicted a set of coreference mentions $M = \{m_1, m_2, \ldots, m_n\}$ and clusters $\mathcal{G}$, where each cluster $G \in \mathcal{G}$ contains a set of coreferential mentions $m_1^G, m_2^G, \ldots, m_l^G$. For D we compute a set of CNER-annotated spans $C = \{c_1, c_2, \ldots, c_k\}$ with semantic labels $\mathcal{L}(c_j) \in \mathcal{T}$, where $\mathcal{T}$ is the inventory of CNER categories presented in Appendix E, Table 12. The goal of our technique is to assign to each coreference cluster G a semantic label $\mathcal{S}(G) \in \mathcal{T}$. To do so, we base our approach on two steps: i) Mention Assignment, in which we adopt a span overlap-based technique to assign a CNER label to all nominal mentions, and ii) Category Propagation, in which we assign a label to each coreference cluster through majority voting and propagate the labels to all the mentions in the cluster, including pronouns.

3.2 Mention Assignment

In the first step, we attempt to directly assign a semantic label to each mention. To quantify the alignment between a mention m_i and a CNER span c_j, we define an overlap function Ω based on token-level Jaccard similarity:

$$\Omega(m_i, c_j) = \frac{|\text{span}(m_i) \cap \text{span}(c_j)|}{|\text{span}(m_i) \cup \text{span}(c_j)|}$$

where $\text{span}(\cdot)$ denotes the set of indices of the words encompassed by mention m_i or CNER tag c_j. For a given mention m_i, we select the CNER span $\hat{c}_j$ with the highest overlap score. If this maximum score exceeds a threshold τ (set to 0.5 in all experiments), the mention m_i is assigned the corresponding label $l_i = \mathcal{L}(\hat{c}_j)$. Mentions that do

not sufficiently overlap with any CNER span are left unlabeled.

3.3 Category Propagation

In the second step, we assign a semantic label to each coreference cluster based on the labels of its mentions, and then propagate this label to all unlabeled mentions in the cluster. For each cluster $G \in \mathcal{G}$ that contains at least one labeled mention, we compute its cluster-level label via majority voting. Specifically, for each semantic category $t \in \mathcal{T}$, we count the number of mentions in G labeled as t, and select the most frequent category:

$$\mathcal{S}(G) = \arg\max_{t \in \mathcal{T}} \left| \{ m^G \in G \mid \mathcal{L}(m^G) = t \} \right|.$$

If two or more labels occur with equal frequency, we break ties by selecting the label whose mentions have, on average, the highest overlap Ω with their corresponding CNER spans. Once the dominant label $\mathcal{S}(G)$ is determined, it is propagated to all mentions in $m^G \in G$, also including pronominal and other unlabeled mentions. In the particular edge case where coreference clusters contain no nominal mentions, our technique cannot assign tags; however, as shown in Section 6, this occurs infrequently due to the high density of CNER annotations.

4 Semantic Evaluation Framework

We now present our semantic framework for interpretable and actionable evaluation of coreference resolution. After propagation, each cluster (thus, each mention) is associated with a CNER type. We use these types to report coreference performance stratified by semantic class via two typed F_1 scores, i.e., Mention F_1 and Link F_1.

Mention F_1 Score The Mention F_1 score quantifies, for a given class $t \in \mathcal{T}$, the quality of mention extraction independently of clustering. A predicted mention m with label $\mathcal{L}(m) = t$ is treated as a true positive if m is a gold-standard mention. It is defined as a traditional F_1 score, i.e., the harmonic mean of mention precision and recall, where precision measures the proportion of predicted mentions that exactly match gold-standard mentions, and recall measures the proportion of gold mentions that are correctly predicted.

Link F_1 Score The Link F_1 score evaluates the quality of predicted coreference links between mentions, i.e., pairs of coreference mentions that belong to the same cluster. A predicted link $p^G = \{m_1^G, m_2^G\}$ between two mentions in the same cluster G with label $\mathcal{S}(G) = t$ is treated as a true positive for class t if p^G is present among the gold links. Unlike Mention F_1, which isolates span detection, Link F_1 captures the structural quality of clustering. In our experiments, we compute Link F_1 using gold mentions to control for mention detection effects and to focus exclusively on the model's linking performance.

Metrics Interpretation and Downstream Usage Together, Mention F_1 and Link F_1 provide a coherent and interpretable decomposition of coreference performance. Mention F_1 captures *what* the model recognizes as entity mentions among those annotated in the dataset, while Link F_1 captures *how* these mentions are connected into referential structures. When computed per semantic category, these metrics provide fine-grained diagnostics; we use them in Section 6.3 to guide targeted data augmentation toward underperforming categories.

5 Experimental Setup

5.1 Datasets

We conduct experiments on three English coreference resolution datasets that differ substantially in domain, annotation scope, and semantic coverage.
OntoNotes (Pradhan et al., 2012) is a large, multi-genre corpus released as part of the CoNLL-2012 shared task on multilingual coreference resolution. Its diverse genre and variety of noun types (proper nouns, common-noun mentions, pronominal mentions) cover a broad and balanced distribution of entity types.
LitBank (Bamman et al., 2020) is a coreference corpus derived from literary texts, annotated under restrictive guidelines that include only six entity types: PER (persons), FAC (facilities), LOC (locations), GPE (geopolitical entities), ORG (organizations), and VEH (vehicles). According to the original report, PER mentions account for 83.1% of all mentions, followed by FAC (8.0%), LOC (4.4%), GPE (3.3%), VEH (0.7%), and ORG (0.5%).
PreCo (Chen et al., 2018) is a large-scale English coreference dataset comprising terminology mostly coming from preschoolers' vocabulary, which emphasizes entities in everyday contexts.

5.2 Comparison Systems

5.2.1 Coreference Models

To isolate the effects of domain and semantic distribution, we evaluate models with the same under-

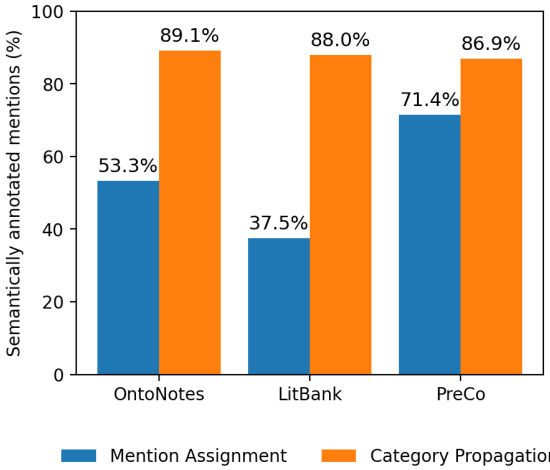


Figure 2: Percentage of CNER-annotated coreference mentions after Mention Assignment (in blue) and after Category Propagation (in orange).

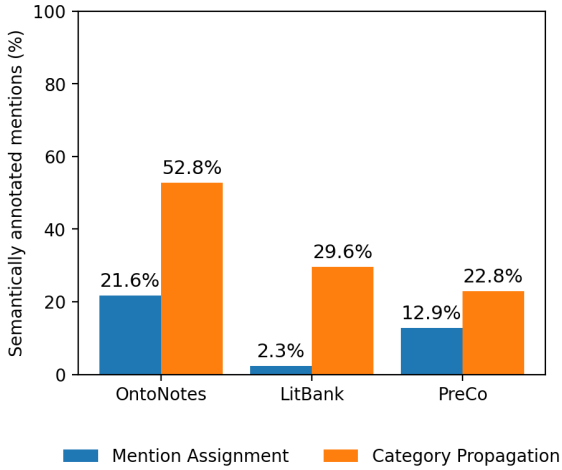


Figure 3: Percentage of NER-annotated coreference mentions after Mention Assignment (in blue) and after Category Propagation (in orange).

WikiNEuRal (NER)		**CNER**	
Type	Support	Type	Support
PER	46,147	PER	58,153
LOC	12,988	LOC	12,458
ORG	9,756	ORG	8,645
MISC	7,096	GROUP	5,222
		MEDIA	4,034
		SUPER	3,782
		ARTF	2,291
		EVENT	1,091
		(...)	
		REL	15
		CULT	15
		DIS	10
		PLANT	5

Table 1: Distribution comparison of semantic types and supports between NER and CNER on OntoNotes.

lying coreference architecture trained on different datasets. This checks for architectural confounds and allows us to focus on out-of-domain behavior. As the underlying coreference model, we use Maverick (Martinelli et al., 2024a) in its multi-expert scorers version, a recent encoder-only neural architecture that jointly models mention extraction and clustering, achieving state-of-the-art performance while remaining computationally efficient. We consider three Maverick variants from the official repository[1]: **maverick-mes-ontonotes**, **maverick-mes-litbank**, **maverick-mes-preco**, trained respectively on OntoNotes, LitBank and PreCo.

5.2.2 Semantic Annotation Layer

Our framework relies on CNER as the semantic annotation layer, and all our experiments are conducted using the official CNER checkpoints[2].

To compare against prior approaches that rely on standard NER-style semantic categories, we experiment with replacing CNER with a standard NER, i.e., WikiNEuRal (Tedeschi et al., 2021), a widely used NER system that predicts four traditional coarse-grained NER classes (PER, LOC, ORG, MISC).

6 Results

In Section 6.1, we measure the effectiveness of our labeling and propagation technique in terms of mention coverage and include a comparison between CNER and standard NER. Subsequently, in Section 6.2, through our evaluation framework, we analyze semantic gaps in current coreference datasets and models. Finally, in Section 6.3, we explore a possible downstream application of our evaluation to improve model results.

6.1 Effectiveness of our technique

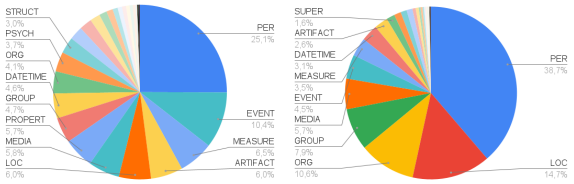
We evaluate the effectiveness of our labeling and propagation technique by i) measuring mention coverage with CNER, ii) comparing CNER and standard NER as semantic annotation layers, and iii) manually measuring the performance of our labeling and propagation technique on 30% of the entire LitBank test set, thereby checking the correctness of CNER labels assigned to gold clusters.

[1] Maverick official github repository

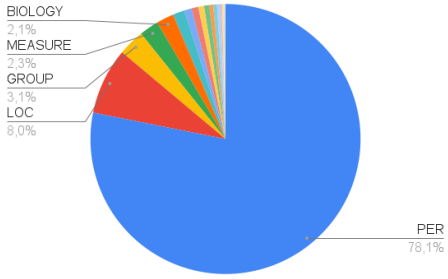
[2] Official CNER checkpoints

Model	Mention F_1			Link F_1		
	OntoNotes	LitBank	PreCo	OntoNotes	LitBank	PreCo
maverick-mes-ontonotes	**0.85**	0.48	0.40	**0.77**	**0.53**	0.57
maverick-mes-litbank	0.40	**0.78**	0.31	0.43	**0.53**	0.47
maverick-mes-preco	0.53	0.35	**0.93**	0.47	0.46	**0.82**

Table 2: Macro Mention F_1 and Link F_1 results for Maverick models across OntoNotes, LitBank, and PreCo.



(a) PreCo. (b) OntoNotes.



(c) LitBank.

Figure 4: Distribution of propagated CNER semantic categories for gold coreference mentions.

Mention Coverage. Figure 2 reports the mention coverage, i.e., the proportion of coreference mentions assigned to a CNER category after Mention Assignment and after Category Propagation. Mention Assignment labels only nominal mentions aligned via span overlap (Section 3.2), covering between 37.5% of mentions on LitBank and 71.4% on PreCo. After Category Propagation, coverage rises to ∼90% on all datasets. The remaining unlabeled mentions, after inspection, are found to be mostly clusters containing only pronominal mentions, which cannot be labeled during the Mention Assignment step. Further analysis on pronominal mentions coverage is provided in Appendix A in Table 5 and 6, and in Appendix C.1, where we present qualitative examples of our labeling and propagation technique.

Investigating NER vs CNER. Since our framework is agnostic to the choice of the semantic annotation layer, we test whether standard NER can provide an alternative to CNER. We compare CNER and NER along two dimensions: i) mention coverage, as the percentage of CNER annotated mentions, and ii) semantic granularity, i.e., the number and diversity of semantic categories available for analysis. Regarding mention coverage, Figure 3 shows the percentage of annotated mentions by NER during our two-step labeling and propagation technique. Compared to CNER, NER assigns semantic labels to substantially fewer mentions across all datasets. Even after propagation, NER coverage remains limited (52.8% on OntoNotes, 29.6% on LitBank, and 22.8% on PreCo), making CNER more suited as an underlying semantic layer.

Regarding semantic granularity, Table 1 shows that NER-based evaluation collapses model behavior into only four coarse categories (PER, LOC, ORG, MISC), each with high support and limited differentiation, providing low diagnostic resolution. In contrast, CNER offers a substantially richer label space (29 categories), with diverse supports and performance patterns, enabling a finer-grained and more interpretable analysis. A full comparison of downstream effects, including the impact of splitting the MISC class into its CNER counterparts, is reported in Appendix A.3.

Manual validation The reliability of our evaluation framework critically depends on the accuracy of the semantic labels. We therefore conduct a quantitative analysis, manually annotating 30% of the entire LitBank test set, and verifying the label assigned by CNER on gold clusters. This yields an overall 90% precision, 87% recall and 88% F_1 score, confirming the high accuracy and reliability of our semantic label propagation observed in the qualitative evaluation presented in Appendix C.1.

6.2 Interpretable Evaluation

We apply our semantic evaluation framework in order to identify semantic gaps in datasets that result in systematic model errors.

Category Distribution of Coreference Datasets. In Figure 4, we show the distribution of CNER semantic categories in OntoNotes, LitBank, and PreCo. In LitBank, the category distribution produced by our labeling and propagation technique

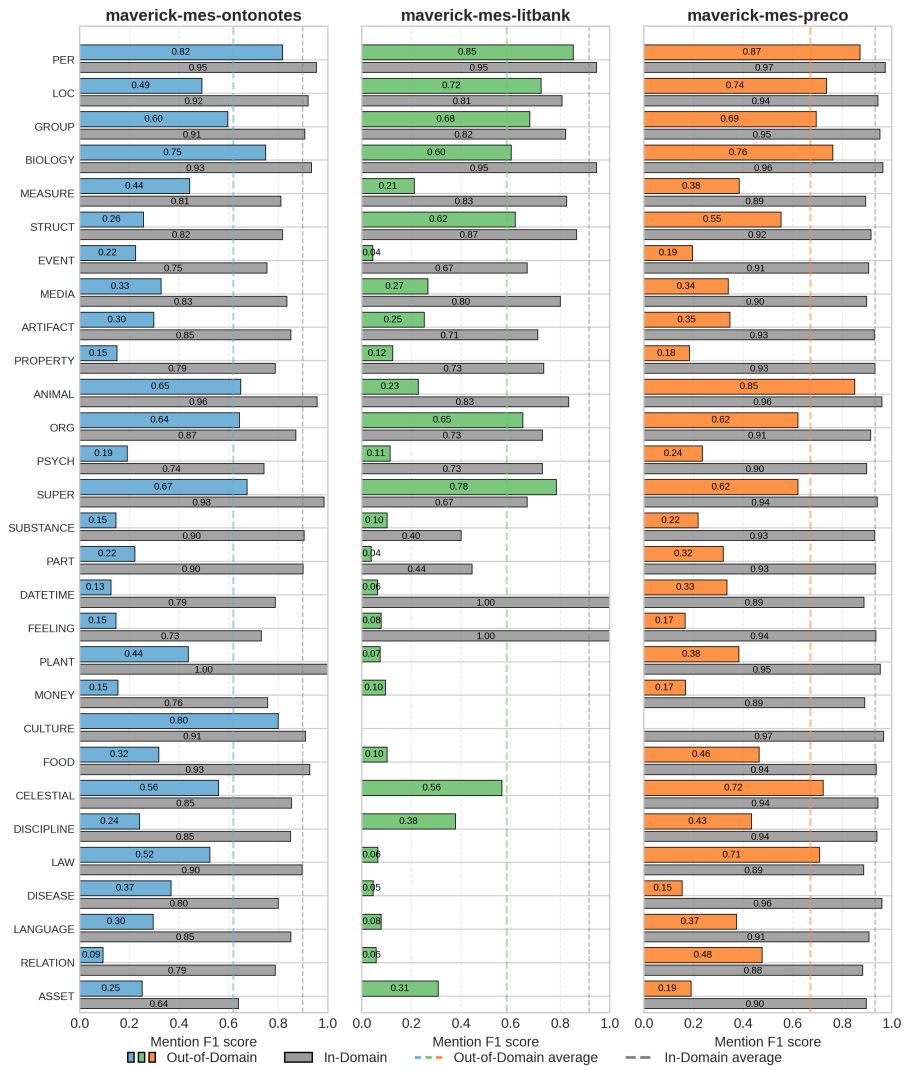


Figure 5: Per-class Mention F_1 scores for each model. In-domain results are shown in grey, while out-of-domain results are shown in different colors and computed as the average performance on the two datasets not used for training. Dashed vertical lines indicate the mean in-domain and out-of-domain scores. Classes are ordered by decreasing category frequency in LitBank so as to highlight out-of-domain performance.

closely resembles the statistics reported by the dataset authors (cf. Section 5), providing an external validation for the quality of our method. At the same time, the distribution reveals the strong skew of LitBank annotations, a direct consequence of its restrictive annotation guidelines: PER mentions are dominant, reflecting character-centered narratives, while categories such as LOC and ORG occur rarely. Several semantic categories are entirely absent, including MONEY, PLANT, ASSET, FOOD, DISCIPLINE, CELESTIAL, LANGUAGE, DISEASE, RELATION, LAW, and CULTURE. In contrast, OntoNotes and PreCo exhibit a more balanced distribution, with substantial coverage of PER, ORG, LOC, and EVENT, reflecting their multi-genre nature and less restrictive design choices. These patterns illustrate a core challenge in out-of-domain evaluation that we will inspect in the following section: models trained on highly person-centric corpora such as LitBank may overfit to human-referential semantics, obtaining low performance compared to models trained more evenly across diverse entity types, as in OntoNotes and PreCo.

Model results. We evaluate the three variants of Maverick fine-tuned on OntoNotes, LitBank, and PreCo, and assess their performance both in-domain and out-of-domain, i.e., on the test set of the same dataset used for training, or the test set of a different one. Table 2, left column, reports the macro-averaged Mention F_1 scores across all

configurations, while we refer the reader to Appendix B Table 8 for a complete micro-averaged Mention F_1 score table. We report that all models achieve strong in-domain performance. However, when evaluated out-of-domain, Mention F_1 drops substantially, highlighting the difficulty of transferring mention extraction across datasets that differ in genre and annotation conventions.

A finer-grained view is reported in Figure 5, with per-class Mention F_1 scores for each model. All models maintain strong in-domain performance across most semantic categories. In contrast, out-of-domain results reveal pronounced discrepancies across both models and categories. The LitBank-trained model, in particular, exhibits the lowest performance across multiple categories when evaluated on OntoNotes and PreCo.

To disentangle whether this result is due to the model predicting different span boundaries or wrongly predicting mentions of underrepresented semantic classes, we analyze Link F_1 scores using gold mentions as input (Table 2, right column). We report that the LitBank-trained model has lower performance in the out-of-domain setting, also when starting from gold mentions, and refer the reader to Appendix B for a per-class analysis of Link F_1, where the performance drop on underrepresented classes is even more evident, confirming the concerns on person-centric bias of LitBank we raised in the previous section. In Appendix C.2, we also report a qualitative error analysis.

6.3 Downstream Usage

We investigate whether our semantic diagnostics can be leveraged to improve out-of-domain performance through targeted, low-cost data augmentation for coreference resolution focused on underrepresented semantic categories. Our goal is not to scale data augmentation, but to test whether semantically targeted diagnostics enable measurable improvements even with a very small amount of additional annotated data.

Specifically, as shown in the previous section, the LitBank-trained model exhibits a pronounced drop in out-of-domain coreference performance, particularly for semantic categories that were not annotated by design choice. Motivated by this finding, we focus on the LitBank-trained model and evaluate whether adding a small number of documents in which those semantic categories appear can overcome these limitations.

We generate three synthetic narrative documents using GPT 5.1[3], each approximately 2,000 words long, written to match the LitBank narrative style while including mentions from semantic categories that are underrepresented or missing in the original training data (the prompt is provided in Appendix D). The documents are manually annotated under two alternative guidelines:

- **Restricted annotation**, which restricts the annotation to entity types PER, FAC, LOC, GPE, ORG, and VEH, following the original LitBank guidelines and the directly associated classes that are present in the original CNER paper.
- **Unrestricted annotation**, which covers nominal and pronominal mentions to entities and concepts, without type-based restrictions.

In Table 3, we also report the distribution and support of semantic classes in our synthetic documents for both restricted and unrestricted annotation guidelines.

We train two augmented models on the LitBank training set by adding the three

Unrestricted		**Restricted**	
Type	Count	Type	Count
PER	82	PER	69
EVENT	65	EVENT	1
PSYCH	54	PSYCH	1
PROPERTY	52	PROPERTY	0
ARTIFACT	47	ARTIFACT	9
MEDIA	43	MEDIA	0
PLANT	27	PLANT	0
MEASURE	26	MEASURE	5
FEELING	17	FEELING	0
LOC	17	LOC	12
SUBSTANCE	17	SUBSTANCE	1
DISEASE	16	DISEASE	0
GROUP	16	GROUP	5
ASSET	15	ASSET	0
ORG	15	ORG	9
DATETIME	14	DATETIME	0
LANGUAGE	14	LANGUAGE	1
STRUCT	13	STRUCT	8
PART	12	PART	0
FOOD	11	FOOD	0
DISCIPLINE	7	DISCIPLINE	1
MONEY	7	MONEY	0
CULTURE	5	CULTURE	0
RELATION	5	RELATION	0
CELESTIAL	4	CELESTIAL	3
LAW	1	LAW	0

Table 3: Comparison of mention counts between Unrestricted and Restricted settings across semantic types for our synthetic data.

[3] ChatGPT 5.1 System Card

Model	CoNLL-F_1			Link F_1			Mention F_1		
	PreCo	Onto	Avg	PreCo	Onto	Avg	PreCo	Onto	Avg
maverick-mes-litbank	45.5	51.7	48.6	30.53	29.25	29.89	34.40	26.75	30.58
maverick-mes-litbank-augmented	44.7	51.9	48.3	**34.96**	26.37	30.67	26.52	**29.49**	28.01
maverick-mes-litbank-augmented-NR	**49.7**	**52.5**	**51.1**	**34.77**	**29.26**	**32.02**	**45.10**	**29.88**	**37.49**
Diff. between augmented & augmented-NR			**+2.8**			**+1.35**			**+9.49**

Table 4: Comparative out-of-domain results of our three models used to test the benefit of targeted data augmentation, in terms of CoNLL-F1 and Macro Mention F_1 and Link F_1 scores on OntoNotes and PreCo.

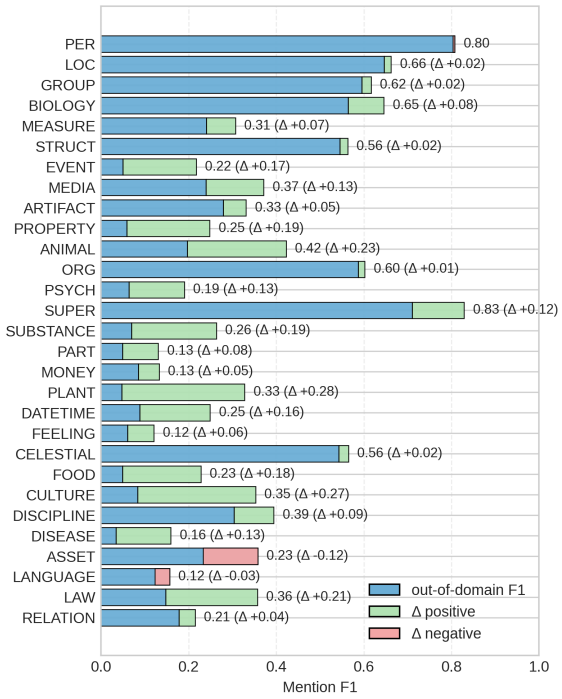


Figure 6: Performance difference between maverick-mes-litbank-NR and maverick-mes-litbank in terms of Mention F_1 in the out-of-domain setting. Positive values indicate improvements over the baseline.

synthetic documents annotated under each guideline. This yields two additional models: (i) `maverick-mes-litbank-augmented`, trained with LitBank-like annotation; and (ii) `maverick-mes-litbank-augmented-NR`, trained with unrestricted annotation (NR).[4]

In Table 4 we show that the model trained with unrestricted annotation consistently outperforms both the LitBank baseline and the model augmented with LitBank-like annotations, improving +2.5 and +2.8 CoNLL-F_1 points, respectively.

These trends are reflected in the semantic evaluation metrics: Link F_1 shows a modest but consistent improvement, while Mention F_1 exhibits a substantially larger improvement of roughly 9.5 points. This suggests that unrestricted annotation primarily benefits mention extraction, improving the model's ability to detect and represent semantic categories that were previously underrepresented or absent. A per-class analysis of out-of-domain Mention F_1 performance is presented in Figure 8, and this further highlights per-class improvements. Below, in Appendix B, we provide the Link F_1-based figures along with further analysis of their comparison. We believe these results highlight the novel downstream capabilities of our semantic evaluation framework, which, unlike traditional aggregate statistical metrics, can be used directly to improve model performance. Not only that, we hope that future annotation campaigns will be able to leverage our evaluation framework to provide semantically diverse annotations, with the objective of training models that perform better out of domain.

7 Conclusions

In this work, we present a semantically-enhanced evaluation framework for coreference resolution that improves the interpretability of standard aggregate metrics. By overlaying Concept and Named Entity Recognition (CNER) onto coreference outputs, our two-step labeling and propagation technique densely annotates clusters with a fine-grained semantic categorization, enabling an analysis of mention extraction and coreference linking beyond what metrics such as CoNLL-F_1 can capture. Using semantically typed Mention F_1 and Link F_1 scores, we exposed systematic domain-dependent weaknesses in current benchmarks and models, and showed how taking direct action to improve data quality leads to measurable gains in scores.

Overall, we demonstrate that semantic grounding of mentions enhances evaluation by addressing the interpretability gap and improving data quality, ultimately leading to improved model performance.

[4]Training details are provided in Appendix D.1.

Limitations

Despite the strengths of the proposed framework, several limitations should be acknowledged. First, while we show that Concept and Named Entity Recognition (CNER) provides substantially higher coverage and finer-grained semantic categories than standard NER on coreference data, we do not perform a full intrinsic evaluation of the semantic propagation procedure used to assign cluster-level labels. In particular, due to the lack of manually annotated cluster-level semantic labels, we do not directly measure the accuracy of propagated semantic tags. As a result, the effect of semantic labeling errors on the proposed typed evaluation metrics is not explicitly quantified. We emphasize, however, that our goal is not to propose a state-of-the-art method for semantic annotation of coreference clusters, but rather to employ a simple and scalable baseline that enables semantic analysis of model behavior. We view this work as a step toward semantic coreference evaluation, and we hope it will motivate future efforts to construct high-quality, manually annotated benchmarks for this task.

Second, approximately 10% of mentions remain unlabeled in each dataset. Qualitative inspection suggests that these predominantly correspond to clusters composed exclusively of pronouns, which cannot be directly tagged by CNER. While future work could potentially focus on handling these cases by training a lightweight classifier weakly supervised by the propagated labels, in this work, we restrict our analysis to the labeled subset of mentions.

Finally, the proposed technique relies on English-language CNER resources and has only been evaluated on English datasets. Extending the framework to other languages would require multilingual CNER models, which we leave for future work.

Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

We also gratefully acknowledge the support of the AI Factory IT4LIA project.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658, Seattle, United States. Association for Computational Linguistics.

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. 2019. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, Minneapolis, USA. Association for Computational Linguistics.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225, Vancouver, Canada. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Sopan Khosla and Carolyn Rose. 2020. Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31, Online. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*,

pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024a. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024b. CNER: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico. Association for Computational Linguistics.

Sebastian Martschat, Thierry Göckel, and Michael Strube. 2015. Analyzing and visualizing coreference resolution errors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 6–10, Denver, Colorado. Association for Computational Linguistics.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Frank Mtumbuka and Steven Schockaert. 2024. EnCore: Fine-grained entity typing by pre-training entity encoders on coreference chains. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1768–1781, St. Julian's, Malta. Association for Computational Linguistics.

Vincent Ng. 2007. Semantic class induction and coreference resolution. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 536–543, Prague, Czech Republic. Association for Computational Linguistics.

Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Cheung. 2024. Challenges to evaluating the generalization of coreference resolution models: A measurement modeling perspective. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15380–15395, Bangkok, Thailand. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Amir Zeldes. 2021. Can we fix the scope for coreference? problems and solutions for benchmarks beyond ontonotes. *CoRR*, abs/2112.09742.

A Additional Details on Effectiveness of our Labeling and Propagation Technique

In this Section we provide additional details on the process of evaluating our labeling and propagation technique. We assess its effectiveness by i) measuring CNER mention coverage and ii) comparing CNER and standard NER as semantic annotation layers.

A.1 Mention Coverage Additional Details

In Table 5 we see the details of CNER mention coverage, with an additional focus on pronominal mentions. As expected the percentage is lower than labeled mentions in general, which hints that, as

Dataset	System	Overall Mentions			Pronoun Mentions		
		%Dir	%Prop	%Any	%Dir	%Prop	%All
OntoNotes	CNER	53.27	35.80	89.07	0.00	83.47	83.47
LitBank	CNER	37.50	50.53	88.03	0.00	84.09	84.09
PreCo	CNER	71.36	15.57	86.94	0.00	62.46	62.46

Table 5: Mention coverage percentages for CNER, reported for overall mentions and pronominal mentions. Direct coverage (%Dir) corresponds to mentions labeled directly by the CNER tagger, propagated coverage (%Prop) via coreference links, and %Any/%All indicates total labeled coverage.

Dataset	System	Overall Mentions			Pronoun Mentions		
		%Dir	%Prop	%Any	%Dir	%Prop	%All
OntoNotes	WikiNEuRal	21.65	31.13	52.77	0.00	47.71	47.71
LitBank	WikiNEuRal	2.26	27.37	29.63	0.00	36.44	36.44
PreCo	WikiNEuRal	12.86	9.98	22.84	0.00	28.90	28.90

Table 6: Mention coverage percentages for the NER-based baseline (WikiNEuRal), reported for overall mentions and pronominal mentions. Direct coverage (%Dir) corresponds to mentions labeled via mention assignment, propagated coverage (%Prop) via category propagation, and %Any/%All indicates total labeled coverage.

WikiNEuRal (NER)			CNER		
Type	Link F_1	Support	Type	Link F_1	Support
PER	0.896	46,147	PER	0.890	58,153
LOC	0.849	12,988	LOC	0.828	12,458
ORG	0.772	9,756	ORG	0.761	8,645
MISC	0.823	7,096	GROUP	0.854	5,222
			MEDIA	0.838	4,034
			SUPER	0.888	3,782
			ARTIFACT	0.723	2,291
			EVENT	0.678	1,091
			DATETIME	0.714	890
			MEASURE	0.655	866
			BIOLOGY	0.921	658
			PROPERTY	0.894	839
			PART	0.886	76
			CELESTIAL	0.881	53
			LAW	0.806	29
			ASSET	0.542	29
			DISCIPLINE	0.473	29
			LANGUAGE	0.698	23
			FOOD	0.722	18
			CULTURE	0.800	15
			FEELING	0.813	15
			RELATION	0.621	15
			DISEASE	0.588	10
			PLANT	1.000	5

Table 7: Distribution of semantic types with Link F_1 and support, comparing standard NER (WikiNEuRal) and fine-grained CNER. While NER aggregates diverse concepts under MISC, CNER redistributes them across semantically specific categories, providing both broader coverage and improved interpretability.

we hypothesized, the gap of roughly 10% of unlabeled mentions is mostly composed of pronominal clusters.

A.2 NER Mention Coverage and Pronoun Propagation

Table 6 reports detailed statistics on mention labeling when using a NER-based baseline (WikiNEuRal). The table distinguishes between labels assigned directly by the tagger and those obtained via coreference-based propagation, and reports results separately for all mentions and for pronominal mentions.

From Table 6, we observe the following trends. First, the proportion of unlabeled pronominal mentions is substantially higher than that of non-pronominal mentions, mirroring the overall limited coverage of the NER-based baseline. Second, while label propagation partially mitigates this issue, the percentage of pronominal mentions receiving a semantic label remain relatively low across all datasets.

A.3 Semantic Granularity of NER vs. CNER

Beyond raw mention coverage, we also analyze the semantic granularity of NER when used as a semantic layer for evaluation. In this subsection, we report a complete comparison between semantic evaluation conducted with a standard NER-based layer and with CNER as the semantic layer.

Table 7 shows that using CNER enables a substantially richer semantic label space, comprising 29 entity classes, compared to the more limited tagset typically provided by NER, also showing Link F_1 as a performance measure. We show that more diverse class supports reveal finer-grained performance patterns that are not observable while using NER, which brings scores to be aggregated under the MISC class.

B Complete Evaluation Results

In this Section, we report additional evaluation results, omitted from the main paper for clarity and space constraints.

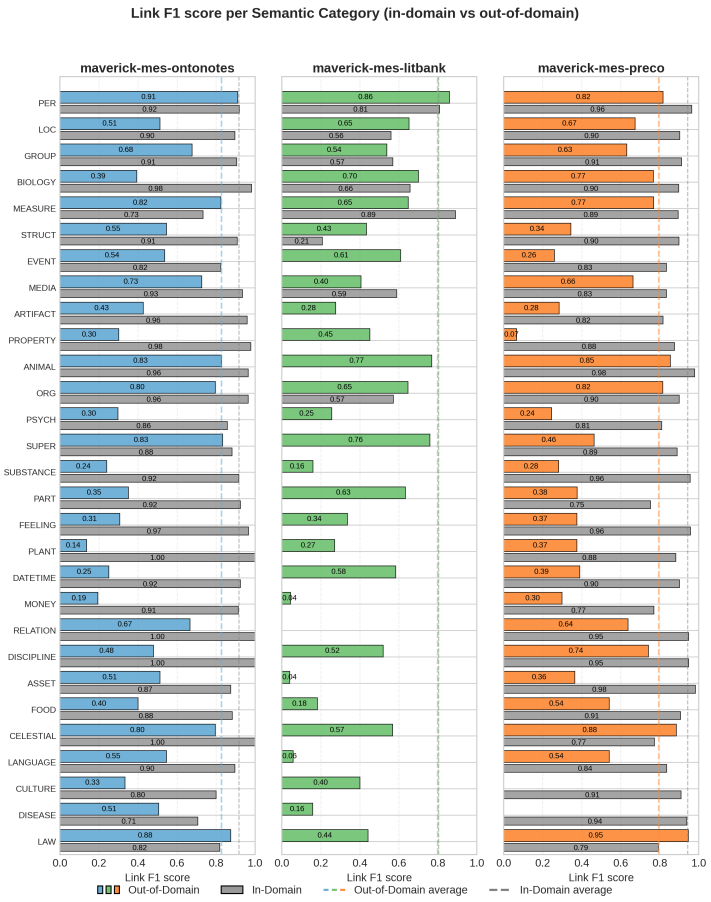


Figure 7: Per-class Link F_1 scores for maverick-mes-ontonotes, maverick-mes-litbank, and maverick-mes-preco. Grey bars indicate in-domain performance, while colored bars indicate out-of-domain performance. Classes are ordered by decreasing support in the LitBank dataset.

We first present the Micro-averaged results of our evaluation on maverick-mes-ontonotes, maverick-mes-litbank and maverick-mes-preco, then complement the per-class Mention F_1 evaluation presented in Section 6.2 with Link F_1 scores.

B.1 Micro-averaged Results

Table 8 reports the Micro F_1 scores for the analyzed models. We observe a decline in out-of-domain performance on Mention F_1 scores for all models, with a steep decrease in maverick-mes-litbank as it reaches the lowest score on Ontonotes and ties with maverick-mes-ontonotes on the PreCo dataset. In Link-F_1, we observe maverick-mes-ontonotes outperforming maverick-mes-litbank on its own domain.

B.2 Link F_1

Figure 7 reports per-class Link F_1 scores for our comparison systems. For each model we show in-domain performance in grey and out-of-domain performance in color. Semantic classes are ordered by decreasing support in the LitBank dataset.

The results highlight several consistent patterns. In line with the quantitative results reported in the main corpus, the LitBank-trained model exhibits poor out-of-domain performance, particularly on semantic categories with low or zero support in the LitBank training set due to its restrictive annotation guidelines. Notably, several classes are entirely absent from LitBank, leading to near-zero performance when evaluated out of domain. Moreover, even in-domain, maverick-mes-litbank underperforms compared to maverick-mes-ontonotes. As reported in Table 8, the LitBank model reaches an in-domain Link F_1 of 80, which is lower than the corresponding score achieved by the OntoNotes-trained model. This further suggests that limited semantic coverage negatively affects both generalization and in-domain linking quality. A clear performance decrease is also observable on the OntoNotes and PreCo trained models from in-

Model	Mention F_1			Link F_1		
	OntoNotes	**LitBank**	**PreCo**	**OntoNotes**	**LitBank**	**PreCo**
maverick-mes-ontonotes	**0.90**	0.74	0.49	**0.92**	**0.89**	0.77
maverick-mes-litbank	0.67	**0.92**	0.50	0.80	0.80	0.82
maverick-mes-preco	0.70	0.64	**0.93**	0.82	0.77	**0.94**

Table 8: Micro Mention F1 and Link F1 results for each of the comparison systems on all our datasets.

domain to out-of-domain performance, especially in classes such as SUBSTANCE, PART, FEELING, PLANT, DATETIME and MONEY, where we have a decrease that is often more than 60 Link F_1 points.

C Qualitative Error Analysis

This section provides qualitative examples illustrating the behavior of the proposed semantic propagation framework and the types of semantic and linking coreference errors revealed by our evaluation. The examples complement the quantitative results by offering concrete illustrations of domain-specific patterns discussed in the main corpus.

C.1 CNER labeling and propagation examples.

In Table 9, we show two qualitative examples of CNER labeling and propagation onto coreference clusters.

In the first example, from a news-domain text in OntoNotes, CNER labeling and propagation ensure type consistency across all mentions within the same entity cluster. In the first row, we show the high coverage of named entities and concepts as CNER annotated spans, while in the second row, we report the different span boundaries found in the gold annotated coreferences of OntoNotes. During the Mention Assignment stage, we can assign to every non-pronoun mention a CNER label, and with cluster propagation, we can obtain complete coverage of coreference mentions.

The same happens in the second example, in which we can propagate our CNER labels for plural references. In fact, CNER annotation assigns PER to the named individuals Ehud Barak and Yasser Arafat, and the Mention Assignment step tags them correctly as PER, leaving out plural references, which are inherited in the following Cluster Propagation step, demonstrating the technique's ability to extend semantic categories across implicitly linked mentions.

C.2 Interpretable Error Examples.

Table 10 presents qualitative examples illustrating cross-domain discrepancies among our Maverick-mes models trained on LitBank, OntoNotes, and PreCo. These examples highlight systematic differences in mention boundary detection, semantic category coverage, and cross-type linking behavior. In the first two examples, we observe mention extraction failures in the LitBank-trained model, particularly for underrepresented categories such as CULTURE and DISEASE. In Example 1, the LitBank model identifies only generic *Group* mentions (e.g., "people," "some," "others"), whereas conceptual entities like *the movement* remain unlabeled. In contrast, the OntoNotes model correctly extracts these and assigns CULTURE and MEDIA categories, reflecting broader semantic coverage.

In Example 2, the LitBank model omits all disease-related mentions, while OntoNotes successfully labels *malaria* and related expressions, as well as biological entities such as *the insects*. This underscores OntoNotes' stronger lexical diversity and its inclusion of scientific and factual content, compared to LitBank's narrative focus.

Examples 3 and 4 highlight linking errors involving financial entities. Even with gold mentions, the LitBank model fails to link money-related expressions, treating them as singletons, whereas the PreCo model clusters them accurately and correctly handles possessive references (e.g., *my money*).

Finally, Example 5 demonstrates that the LitBank-trained model fails to link disease mentions (*colon cancer*, *the cancer*), due to the scarcity of such categories in LitBank. The PreCo model, by contrast, clusters them consistently.

Overall, these qualitative examples reveal that LitBank-trained models struggle with non-human and abstract referents, such as cultural, financial, and biomedical entities, due to LitBank's narrative bias and limited exposure to factual or technical domains. This domain imbalance results in reduced mention recall and weaker cross-type linking generalization.

Stage	Annotation
Example 1	
CNER	While all this is going on, Mr. [Clinton]$_{\text{PER}}$ is overseas. [President Clinton]$_{\text{PER}}$ was in [Northern Ireland]$_{\text{LOC}}$ when he heard the [Supreme Court]$_{\text{ORG}}$ [decision]$_{\text{EVENT}}$. He talked to [Al Gore]$_{\text{PER}}$ on the [phone]$_{\text{ARTIFACT}}$ from [Belfast]$_{\text{LOC}}$. This is Mr. [Clinton]$_{\text{PER}}$'s [third]$_{\text{MEASURE}}$ [visit]$_{\text{EVENT}}$ to [Northern Ireland]$_{\text{LOC}}$"
Coreference	"While all this is going on, [Mr. Clinton]$_1$ is overseas. [President Clinton]$_1$ was in [Northern Ireland]$_2$ when [he]$_1$ heard the Supreme Court decision. [He]$_1$ talked to Al Gore on the phone from Belfast. This is [Mr. Clinton]$_1$'s third visit to [Northern Ireland]$_2$"
(1) Mention Assignment	"While all this is going on, [Mr. Clinton]$_{\text{1-PER}}$ is overseas. was in [Northern Ireland]$_{\text{2-LOC}}$ when [he]$_{\text{1-Unassigned}}$ heard the Supreme Court decision . [He]$_{\text{1-Unassigned}}$ talked to Al Gore on the phone from Belfast. This is [Mr. Clinton]$_{\text{1-PER}}$'s third visit to [Northern Ireland]$_{\text{2-LOC}}$"
(2) Cluster Propagation	" While all this is going on, [Mr. Clinton]$_{\text{1-PER}}$ is overseas. [President Clinton]$_{\text{1-PER}}$ was in [Northern Ireland]$_{\text{2-LOC}}$ when [he]$_{\text{1-PER}}$ heard the Supreme Court decision . [He]$_{\text{1-PER}}$ talked to Al Gore on the phone from Belfast. This is [Mr. Clinton]$_{\text{1-PER}}$'s third visit to [Northern Ireland]$_{\text{2-LOC}}$"
Example 2	
CNER	"[Tomorrow]$_{\text{DATETIME}}$'s [summit]$_{\text{EVENT}}$ [meeting]$_{\text{EVENT}}$ will bring [Ehud Barak]$_{\text{PER}}$ and [Yasser Arafat]$_{\text{PER}}$ to the [resort city]$_{\text{STRUCT}}$ of [Sharm El - Sheikh]$_{\text{LOC}}$. Getting both to attend was not an easy task."
Coreference	"Tomorrow's summit meeting will bring [[Ehud Barak]$_1$ and [Yasser Arafat]$_2$]$_3$ to the resort city of Sharm El-Sheikh. Getting [both]$_3$ to attend was not an easy task."
(1) Mention Assignment	"Tomorrow's summit meeting will bring [[Ehud Barak]$_{\text{1-PER}}$ and [Yasser Arafat]$_{\text{2-PER}}$]$_{\text{3-PER}}$ to the resort city of Sharm El-Sheikh. Getting [both]$_{\text{3-Unassigned}}$ to attend was not an easy task."
(2) Cluster Propagation	"Tomorrow's summit meeting will bring [[Ehud Barak]$_{\text{1-PER}}$ and [Yasser Arafat]$_{\text{2-PER}}$]$_{\text{3-PER}}$ to the resort city of Sharm El-Sheikh. Getting [both]$_{\text{3-PER}}$ to attend was not an easy task."

Table 9: Propagation examples annotation.

Model	Output
Ex. 1	*maverick-mes-litbank does not recognize CULTURE mentions.*
maverick-litbank	"... And [people]$_{\text{singleton-GROUP}}$ have different opinions about the movement. [Some]$_{\text{singleton-GROUP}}$ think street art is a crime and destroys property. But [others]$_{\text{singleton-GROUP}}$ see this art as a rich form of non-traditional cultural expression. [Many experts]$_{\text{singleton-GROUP}}$ say the movement began in [New York City]$_{\text{1-LOC}}$ in the nineteen sixties..."
maverick-ontonotes	... And people have different opinions about [the movement]$_{\text{2-CULTURE}}$. Some think [street art]$_{\text{3-MEDIA}}$ is a crime and destroys property. But others see [this art]$_{\text{3-MEDIA}}$ as a rich form of non-traditional cultural expression. Many experts say [the movement]$_{\text{2-CULTURE}}$ began in [New York City]$_{\text{4-LOC}}$ in the nineteen sixties..."
Ex. 2	*maverick-mes-litbank does not recognize* DISEASE *mentions.*
maverick-litbank	"... The insects carry serious diseases like malaria. It is estimated that almost 630,000 people died from malaria and malaria-related causes in 2012, and most of these cases were in African countries..."
maverick-ontonotes	"... [The insects]$_{\text{0-BIOLOGY}}$ carry serious diseases like [malaria]$_{\text{1-DISEASE}}$. It is estimated that almost 630,000 people died from [malaria]$_{\text{1-DISEASE}}$ and [malariarelated]$_{\text{1-DISEASE}}$ causes in 2012, and most of these cases were in African countries..."
Ex. 3	*Using gold mentions, maverick-mes-litbank is unable to cluster* MONEY.
maverick-litbank	"... [With the previous 800 yuan income tax threshold]$_{\text{singleton-MONEY}}$, if [it]$_{\text{singleton-Unassigned}}$ were strictly enforced, 80% of beggars would have to pay..."
maverick-preco	" ... With [the previous 800 yuan income tax threshold]$_{\text{6-MONEY}}$, if [it]$_{\text{6-MONEY}}$ were strictly enforced, 80% of beggars would have to pay..."
Ex. 4	*Using gold mentions, maverick-mes-litbank is unable to cluster* MONEY.
maverick-litbank	"...When [he]$_{\text{7-PER}}$ came home, [he]$_{\text{7-PER}}$ said, 'Call [those servants who have [my]$_{\text{7-PER}}$ money]$_{\text{10-PER}}$...'"
maverick-preco	"...When [he]$_{\text{8-PER}}$ came home, [he]$_{\text{8-PER}}$ said, 'Call [those servants who have [[my]$_{\text{8-PER}}$ money]$_{\text{15-MONEY}}$]$_{\text{10-PER}}$...'"
Ex. 5	*Using gold mentions, maverick-mes-litbank is unable to cluster* DISEASE.
maverick-litbank	"...[Sharon Osbourne, [Ozzy's]$_{\text{0-PER}}$ long-time manager, wife and best friend]$_{\text{1-PER}}$, announced to the world that [she]$_{\text{1-PER}}$ 'd been diagnosed with colon cancer. ... [Sharon]$_{\text{1-PER}}$ announced in April that the cancer was in remission, but just weeks after that announcement, [son Jack]$_{\text{5-PER}}$ entered drug and alcohol rehabilitation in California..."
maverick-preco	"...[Sharon Osbourne, [Ozzy's]$_{\text{0-PER}}$ long-time manager, wife and best friend]$_{\text{1-PER}}$, announced to the world that [she]$_{\text{1-PER}}$ 'd been diagnosed with [colon cancer]$_{\text{3-DISEASE}}$. ... [Sharon]$_{\text{1-PER}}$ announced in April that [the cancer]$_{\text{3-DISEASE}}$ was in remission, but just weeks after that announcement, [son Jack]$_{\text{4-PER}}$ entered drug and alcohol rehabilitation in California..."

Table 10: Representative cross-domain annotation errors. Each example contrasts the LitBank-trained model with OntoNotes or PreCo models, showing domain-specific gaps in semantic coverage and mention linking.

D Data Augmentation Experiment Details

In this Section, we provide a more detailed analysis of the data augmentation experiments. Specifically,

we report: i) the training setup and hyperparameters of the models, and ii) details about the newly generated training data used in our experiments.

D.1 Training Details

All models were trained using the official Maverick codebase[5]. Training was performed on a single NVIDIA RTX 4090 GPU with 24GB of VRAM, using CUDA 12.4.

We focus on the maverick-mes architecture, employing a DEBERTA-V3-LARGE encoder as the underlying language model. The encoder was fine-tuned during training (i.e., not frozen), and span representations were constructed by concatenating the start and end token embeddings.

The most relevant hyperparameters are reported below:

- Optimizer: Adafactor
- Learning rate: 3×10^{-5}
- Weight decay: 0.01
- Warm-up steps: 6,000
- Total training steps: 8,000
- Gradient accumulation: 4 steps
- Gradient clipping: 1.0
- Precision: FP16

Training was conducted using PyTorch Lightning with a deterministic setup (random seed set to 30). Validation was performed twice per epoch. Early stopping was enabled based on the CoNLL-2012 F_1 validation score, with a patience of 120 validation checks. Model checkpointing was performed using the same metric.

D.2 Training Data Details

In this subsection, we describe the prompt used to generate the additional training data employed in the data augmentation experiments.

The data was generated using GPT 5.1[6]. The model was prompted to produce synthetic training examples consistent with the target task formulation. The exact prompt used for data generation is reported in Table 11.

[5] https://github.com/SapienzaNLP/maverick-coref

[6] https://cdn.openai.com/gpt-5-system-card.pdf

Annotation Details The generated texts were annotated according to two different annotation schemes: i) *unrestricted annotation*, and ii) *restricted annotation*, inspired by the LitBank annotation guidelines.

In the restricted setting, annotations were designed to mirror the LitBank scheme, which does not annotate all semantic classes but is limited to the following entity types: PERSON (PER), FACILITY (FAC), LOCATION (LOC), GEO-POLITICAL ENTITY (GPE), ORGANIZATION (ORG), and VEHICLE (VEH). These entity types have a direct map with CNER classes, as described in the original Concept and Named Entity Recognition paper (Martinelli et al., 2024b).

D.3 Evaluation Results on LitBank Models

In this subsection, we report the per-class results of our evaluation framework applied to Maverick-mes models trained on data augmented with generated texts. We focus on out-of-domain performance, and we analyze the impact of unrestricted versus restricted annotation schemes, to complement the results obtained with a fine grained per-class evaluation.

Figure 8 reports the delta in Mention F_1 scores, between a maverick-mes-litbank-augmented-NR and maverick-mes-litbank. We observe a clear improvement in Mention F_1 for nearly all entity classes, with particularly strong gains for classes that are underrepresented or entirely absent in the original LitBank training data.

Figure 9 reports the corresponding results for out-of-domain Link F_1, comparing the same augmented model against the baseline.

In contrast to the mention-level results, improvements in Link F_1 are more limited. This mirrors the behavior already observed in the main paper. A possible explanation is a mismatch in the distribution of coreference links between the generated data and the test sets, for instance, due to an over-representation of singleton mentions in the augmented data. A more detailed analysis of this phenomenon is left for future work.

E CNER Categories

In this Section, we provide a detailed description of the CNER categories used in our experiments. We include this information for completeness, since CNER is employed as the semantic layer of our model. A clear understanding of the entity and

Synthetic Document Generation Prompt.
We aim to generate from scratch a long-form document written in the style of texts found in the LitBank coreference benchmark, a well-known dataset for long-document coreference resolution.
Document requirements:

- **Length:** approximately 2,000 words.
- **Style:** narrative, cohesive, and information-dense, resembling LitBank benchmark documents.
- **Structure:** suitable for coreference resolution, with recurring entities, implicit references, and long-range dependencies.

The text must naturally and meaningfully include all of the following CNER semantic classes, with multiple occurrences distributed across the document.
Required semantic classes:

[...]

All semantic classes must be embedded organically within the narrative rather than enumerated explicitly in the generated text, and the overall tone may be historical, academic, or narrative, provided it remains consistent with the LitBank document style and supports complex, long-distance coreference phenomena.

Table 11: Prompt used for synthetic document generation; the list of required semantic classes includes class name, description, and examples.

concept types covered by each category is essential both for interpreting our results and for enabling potential users of our evaluation framework to conduct more informed analyses.

Table 12 reports the full list of CNER categories, together with their definitions and representative examples.

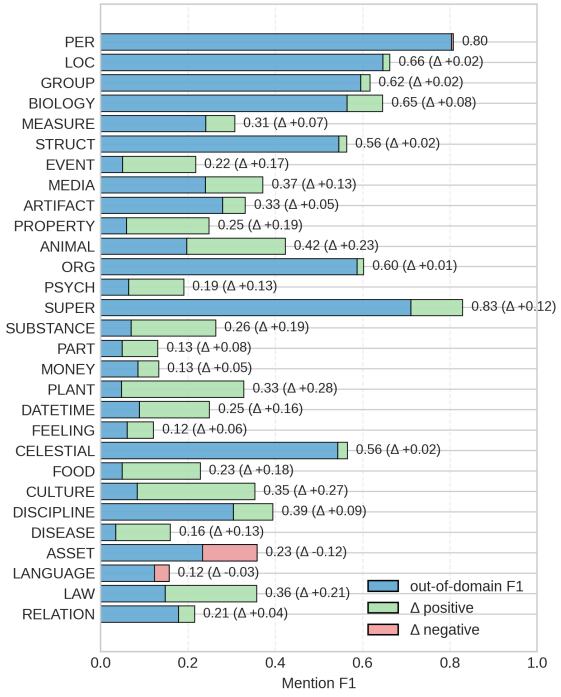


Figure 8: Delta of out-of-domain Mention F_1 for maverick-mes-litbank-NR, compared to maverick-mes-litbank. Positive values indicate improvements over the baseline.

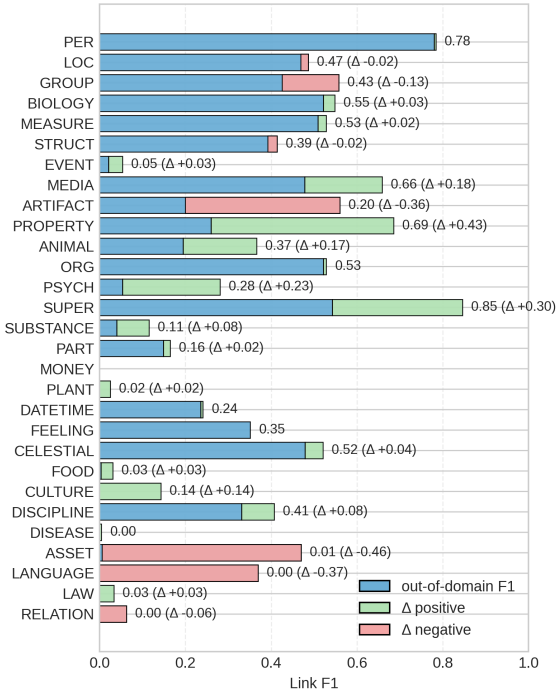


Figure 9: Delta of out-of-domain Link F_1 for maverick-mes-litbank-NR, compared to maverick-mes-litbank. Positive values indicate improvements over the baseline.

Category	Description	Examples
ANIMAL	Living beings (excluding humans) with the ability to move and perceive their surroundings.	dog, cat, mammal, carnivore, brown bear, African Wild Dog, Great White Shark
ARTIFACT	All the objects, artifacts, tools, products and items	vehicle, software, mouse, data stream, Windows XP, Fiat Panda
ASSET	Assets, resources, or possessions with economic or intrinsic value.	capital, stock, wealth, resource, phone bill, Federal Perkins Loan, Investment in Russia
BIOLOGY	Biological entities, including living organisms, cells, or biological components	protein, cell, living organism, lipid, Herpes Simplex Virus, Escherichia Coli
CELESTIAL	celestial bodies as Planets, stars, asteroids, galaxies and other astronomical objects.	comet, nebulae, Sun, Neptune, Asteroid 187 Lamberta, Proxima Centauri
CULTURE	Cultural aspects, traditions, customs, and practices associated with specific groups or societies.	religion, feminism, socialism, capitalism, anarchism, doctrine, cult, Islam, Buddhism
DATETIME	Dates and times	18 March, Saturday, 1979, the evening of 19 November, 15:30 am
DISEASE	medical conditions, illnesses, disorders, and health-related issues affecting living organisms.	infection, allergy, metastasis, complication, acne, Alzheimer's Disease, Cystic Fibrosis
DISCIPLINE	specific fields of study, knowledge, or expertise. It includes academic disciplines, areas of research, and professional domains.	discipline, sport, football, computer science, anatomy, long jump.
EVENT	Events, phenomenon or activities that occur at specific times or places. It includes both significant and everyday occurrences	crime, professorship, temperature change, 2003 Wimbledon Championships, Cannes Film Festival.
FEELING	Emotions, sensations, and subjective experiences related to human or animal consciousness.	affection, attachment, agitation, craving, urge, temptation.
FOOD	edible items, dishes, beverages, and culinary products that are consumed for nourishment or enjoyment	beverage, dish, pork, lasagna, Carbonara, Sangiovese, Cheddar Beer Fondue, Pizza Margherita.
GROUP	group of people or animals	staff, social group, panel, militia, community, trio, duo, family, genealogy, alliance, nationality, peoples
LANGUAGE	individual language-related items, such as words, phrases, or idiomatic expressions	discourse, context, lexeme, morpheme, appellation, eponym, nickname, vowel, syllable, headword
LAW	legal principles, regulations, and rules governing society and various aspects of life	law, civil law, administrative law, martial law, shariah, ordinance, civil right, Magna Carta, Islamic Law
LOC	geographical locations, such as villages, towns, cities, regions, countries, continents, landmarks, or natural features	space, surface, street, road, town, Rome, Lake Paiku, Mississippi River.
MEASURE	units of measurement and quantification used to determine the size, quantity, or quality of various objects or phenomena.	day, microsecond, millisecond, two, 35, 45%, first, temperature, length
MEDIA	various forms of communication and entertainment media, such as newspapers, television shows, movies, social media or digital content.	soundtrack, report, publication, language, English, Forbes, American Psycho
MONEY	monetary units, currencies, and financial values used in different contexts	monetary unit, dollar, 15 euros, 1116 CHF
ORG	organizations, institutions, and companies involved in diverse sectors or activities	Industry, commercial enterprise, San Francisco Giants, Google, Democratic Party.
PART	individual components or sections of larger entities or objects	finger, chin, head, tail, femur, airplane wing, airplane's wings, flower's stem
PER	individuals or persons, including real people and historical figures	doctor, historian, professor, musician, Ray Charles, Jessica Alba
PLANT	Types of trees, flowers, and other plants, including their scientific names.	grass, peach tree, Forsythia, Artemisia Maritima.
PROPERTY	properties or attributes of objects, entities, or concepts	thickness, height, dimension, shape, age
PSYCH	psychological concepts, mental states, and phenomena related to the human mind and behavior	psychological feature, cognition, attention, necessity
RELATION	relationships, connections, and associations between entities or concepts	apport, competition, comparison, bridge, relatedness, parentage, function, parity, transitivity
STRUCT	physical structures, including buildings, architectural designs, and engineered constructions made by humankind	shelter, gravestone, refuge, tent, loft, San Peter's Church, Golden Bridge
SUBSTANCE	chemical substances	acid, bactericide, carbonyl, explosive, fertilizer, Zyclon B
SUPER	Mythological and religious entities.	Apollo, Persephone, Aphrodite, Saint Peter, Pope Gregory I, Hercules.

Table 12: Label, description, and instance examples of each of our CNER categories.