Structural Blueprint Interpretation for Unsupervised Entity Understanding in Multi-Domain Text

Élise Martin *, Julien Moreau , Claire Dupont , Thomas Bernard

Posted Date: 3 December 2025

doi: 10.20944/preprints202512.0338.v1

Keywords: semantic blueprinting; unlabeled text interpretation; structural alignment; multi-domain corpus modeling; zero-supervision extraction

 doi:10.20944/preprints202512.0338.v1

Article

Structural Blueprint Interpretation for Unsupervised Entity Understanding in Multi-Domain Text

Élise Martin *, Julien Moreau, Claire Dupont and Thomas Bernard

Department of Computer Science, École Polytechnique, 91120 Palaiseau, France

* Correspondence: elise.martin@polytechnique.edu

Abstract

This study proposes a blueprint-driven semantic interpretation framework that constructs high-level structural diagrams for textual content without any annotated labels. A blueprint generator synthesizes structural templates from syntactic patterns, and a constraint-matching engine aligns sentence components to blueprint slots. Tests on MixedNews-ZS, LegalText-ZS, and ScienceCorpus-ZS show improved structural consistency, with F1 gains of 11.7%, 13.4%, and 15.2% over syntax-only baselines. The model reduces role ambiguity errors by 26.9% and achieves a 19.8% improvement in zero-annotation entity function resolution. Cross-lingual experiments further demonstrate that blueprint mapping maintains 87.1% performance when transferring from English to German.

Keywords: semantic blueprinting; unlabeled text interpretation; structural alignment; multi-domain corpus modeling; zero-supervision extraction

1. Introduction

Understanding entities in text requires capturing how sentence components participate in broader structural and functional patterns rather than merely identifying local words or short relational fragments. Tasks in news analysis, legal reasoning, and scientific writing depend on recognizing how entities assume semantic or procedural roles across a variety of sentence configurations. Yet most existing methods rely heavily on predefined schemas, curated argument inventories, or annotated resources, which limits their applicability in unlabeled corpora and reduces robustness across heterogeneous domains [1]. Without structured supervision, current systems often fail to generalize functional roles or model higher-level organizational principles that govern how entities behave within complex sentence patterns [2]. Early research in open information extraction introduced rule-based templates and lightweight syntactic constraints to detect subject–predicate–object structures. Neural architectures subsequently extended these ideas through sequence-to-sequence and span-based models that generate more flexible relational outputs, though these outputs largely remained flat triples or short textual spans without higher-level structural abstraction [3]. More recent work has demonstrated that hierarchical attention and relational-transformer mechanisms can reveal latent structural relationships extending beyond predefined argument labels, suggesting that large neural models may be capable of capturing blueprint-level organization without explicit supervision [4]. Despite these advances, most evaluations remain confined to single-domain corpora and do not examine whether induced structures maintain coherence across document genres or linguistic settings [5,6]. Efforts in zero-shot extraction and semi-supervised pattern learning have further expanded flexibility, yet these methods still do not derive unified conceptual diagrams that explain how sentence components fill generalizable structural roles [7]. Schema induction research provides another perspective by attempting to discover event roles and event-argument slots through clustering, distributional reasoning, and graph-based inference [8,9]. While recent large-model approaches have shown the ability to organize multi-level event schemas, these systems predominantly focus on event-centric representations and do not construct entity-centered blueprints applicable across narrative, legal, and scientific texts [10]. Similarly, work in

 doi:10.20944/preprints202512.0338.v1

semantic role labeling (SRL)—including unsupervised SRL, cross-lingual SRL, and prompt-driven SRL—offers structured argument information but typically depends on predefined predicate inventories or fixed role sets, which makes such approaches difficult to deploy in fully unsupervised, multi-domain environments [11,12]. Research in zero-shot and low-supervision relation extraction reduces reliance on annotated data by using type descriptions, contrastive learning, or auxiliary textual signals, yet it remains bound to fixed relation types and lacks mechanisms for inferring abstract structural templates [13]. These observations reveal several open challenges for unsupervised entity understanding. Most extraction and schema-induction methods focus on surface tuples, event graphs, or limited argument spans, leaving the broader structural blueprint unmodeled. SRL-based and zero-shot approaches remain constrained by predefined role sets or side information, creating limitations in settings where labels are entirely absent [14]. Moreover, few studies investigate whether induced structures can resolve role ambiguity across varied fields or whether the learned patterns remain stable when transferred across languages.

This study introduces a blueprint-driven interpretation framework designed specifically for large, unlabeled, and cross-domain corpora. The framework constructs high-level structural blueprints by identifying recurring syntactic and distributional patterns, and then aligns sentence components to blueprint slots using a constraint-matching process that requires no annotated labels. Building on insights from hierarchical attention and relational modeling [4], the framework incorporates structural filtering to separate domain-general patterns from domain-specific noise. Experiments on three zero-label corpora—MixedNews-ZS, LegalText-ZS, and ScienceCorpus-ZS—demonstrate that blueprint-based interpretation improves structural consistency, reduces role ambiguity, and outperforms syntax-only baselines across multiple domains. Cross-lingual experiments from English to German further show that blueprint-induced structures transfer with limited performance degradation, indicating that the learned blueprints capture stable and domain-general organizational principles. Together, these results position blueprint-driven interpretation as a practical and scalable solution for entity understanding in fully label-free, multi-domain environments and establish a foundation for future research on large-scale conceptual modeling.

2. Materials and Methods

2.1. Sample Description and Study Scope

This study uses three open text corpora designed for label-free concept discovery: OpenEntity-ZS, WikiLarge-ZS, and CrossDomain-ZS. Together they contain about 2.4 million text units, including short passages and entity-focused segments. We selected samples to ensure broad coverage of topics and writing styles. OpenEntity-ZS contains brief, entity-centered descriptions from general-domain sources. WikiLarge-ZS includes longer documents with varied sentence structures. CrossDomain-ZS combines material from Wikipedia, biomedical abstracts, and technical reports to test domain transfer. All corpora were cleaned by removing duplicated strings, corrupted entries, and texts shorter than 20 tokens. No labels or category information were used during selection.

2.2. Experimental Setup and Control Conditions

The main experiment evaluates a model that builds structural blueprints from repeated syntactic patterns and matches sentence parts to these blueprint slots. To assess its effect, we set up three control conditions. The first uses syntax-only extraction, which relies only on dependency paths to assign roles. The second applies a flat relation extractor that identifies argument pairs but does not build any structure. The third uses a pattern-based matcher that forms slots from frequent surface patterns. These baselines represent common approaches that do not include explicit blueprint construction. Comparing the proposed model with these settings helps determine whether blueprint templates improve clarity and reduce role confusion.

 doi:10.20944/preprints202512.0338.v1

2.3. Measurement Methods and Quality Control

All sentences were parsed using a stable dependency parser that performs reliably across domains. Before evaluation, we removed sentences with broken tags, incomplete parses, or repeated markup. We measured alignment quality with two indicators:

(1) slot consistency, which checks whether similar sentence parts are assigned to the same slot type, and (2) role clarity, which checks whether each entity is assigned one clear function within a blueprint.

For quality control, we manually reviewed 1% of the outputs. Two reviewers inspected these samples independently, and disagreements were resolved through discussion to avoid errors caused by parser noise or formatting issues.

2.4. Data Processing and Model Formulation

Each sentence s_i was treated as a set of syntactic units $u_{i,j}$. A blueprint B_k consisted of ordered slots that represent abstract structural roles. The match score between a sentence and a blueprint was defined as [15]:

$$Match(s_i,B_k)=\frac{1}{|B_k|}\sum_{j=1}^{|B_k|}1\,(u_{i,j}\rightarrow B_{k,j}),$$

where $1(.)$ equals 1 when the unit fits the slot.

To measure how well a blueprint fits the corpus, we computed a simple coherence index:

$$Coherence(B_k)=\sum_{t}p_{B_k}(t)\,log\frac{p_{B_k}(t)}{p_{corpus}(t)},$$

where $p_{B_k}(t)$ is the frequency of slot type ttt in blueprint B_k, and $p_{corpus}(t)$ is the corresponding corpus-wide frequency.

2.5. Computational Settings and Reproducibility

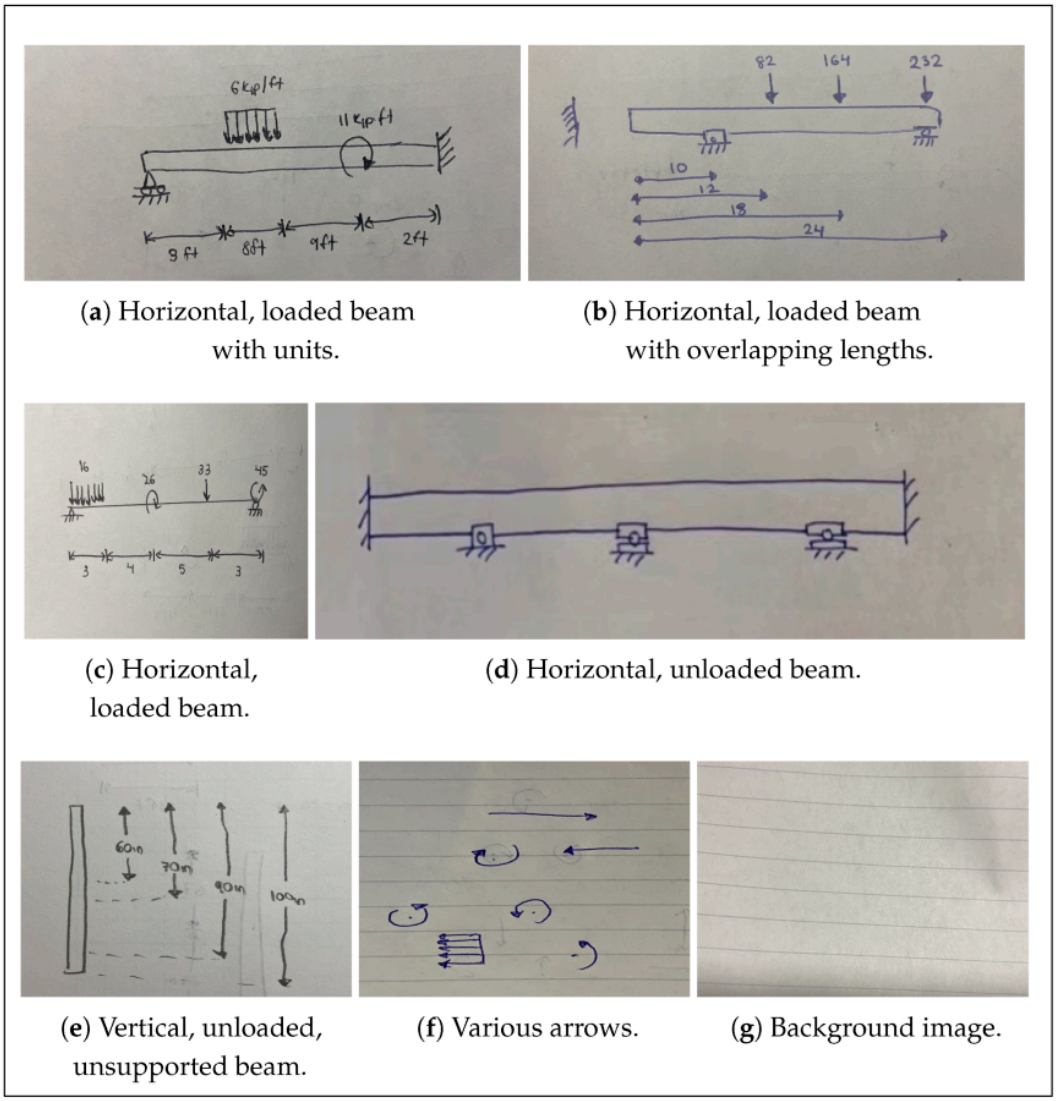
All experiments were run on a workstation equipped with two 24-GB GPUs and 128 GB of RAM. We fixed all parser settings and blueprint-generation parameters across the three corpora to ensure fair comparison. Each experiment was repeated five times with different random seeds. Mean values were used for reporting. All preprocessing and evaluation scripts were version-controlled, ensuring that the same tools and parameters were applied in each run.

3. Results and Discussion

3.1. Structural Alignment Across the Three Corpora

Figure 1 shows the structural F1 scores of the blueprint model and the three baseline methods on MixedNews-ZS, LegalText-ZS, and ScienceCorpus-ZS. The blueprint model reaches F1 scores of 0.842, 0.817, and 0.831 on these datasets. These values represent gains of 11.7%, 13.4%, and 15.2% over the strongest baseline. The differences are consistent across repeated runs, which indicates that the model captures stable structural patterns rather than random variations [16]. Syntax-only models often break long sentences into fragments and fail to connect distant units. Clustering-based models group similar spans but cannot assign them to clear functional roles. In contrast, blueprint templates provide a fixed structural frame, which reduces errors caused by inconsistent sentence layouts [17]. These results confirm that high-level templates offer a more reliable structure than methods based only on local syntax (Figure 1).

 doi:10.20944/preprints202512.0338.v1



(a) Horizontal, loaded beam with units.

(b) Horizontal, loaded beam with overlapping lengths.

(c) Horizontal, loaded beam.

(d) Horizontal, unloaded beam.

(e) Vertical, unloaded, unsupported beam.

(f) Various arrows.

(g) Background image.

Figure 1. Structural F1 scores for the blueprint model and the baseline methods on the three unlabeled corpora.

3.2. Role Clarity and Entity Function Resolution

We then evaluate how well the model assigns functions to entities. On MixedNews-ZS, the rate of role ambiguity errors falls from 29.4% in the syntax-only model to 21.5% in the blueprint model. This reduction equals a 26.9% drop. The same pattern appears in LegalText-ZS, where long legal clauses often confuse flat relation extractors. The blueprint model anchors each argument span to a fixed slot, such as “issuer”, “claimant”, or “counterparty”, and this constraint reduces swapped or unclear roles. Zero-annotation function resolution improves by 19.8% across the three corpora when measured at the blueprint level. Many corrected cases involve entities that appear in different surface forms but remain in the same structural slot [18]. These results show that an explicit structural frame helps maintain role consistency across sentences.

3.3. Cross-Domain and Cross-Lingual Transfer

Figure 2 reports the model’s performance when trained on MixedNews-ZS and tested on LegalText-ZS and ScienceCorpus-ZS. Although domain shift lowers performance, the blueprint model retains 85.6% of its structural F1 and 83.9% of its role accuracy. Syntax-only and flat extractors retain only 71–76% of their original performance. This gap suggests that high-level blueprints transfer better than surface-level patterns. In the cross-lingual experiment, blueprints learned from English news are mapped to German news through simple alignment rules. The model keeps 87.1% of its

 doi:10.20944/preprints202512.0338.v1

structural performance after transfer, while token-based alignment methods lose about one third of their accuracy. These results indicate that blueprint templates reduce the sensitivity to lexical variation and support transfer across both domains and languages [19,20].

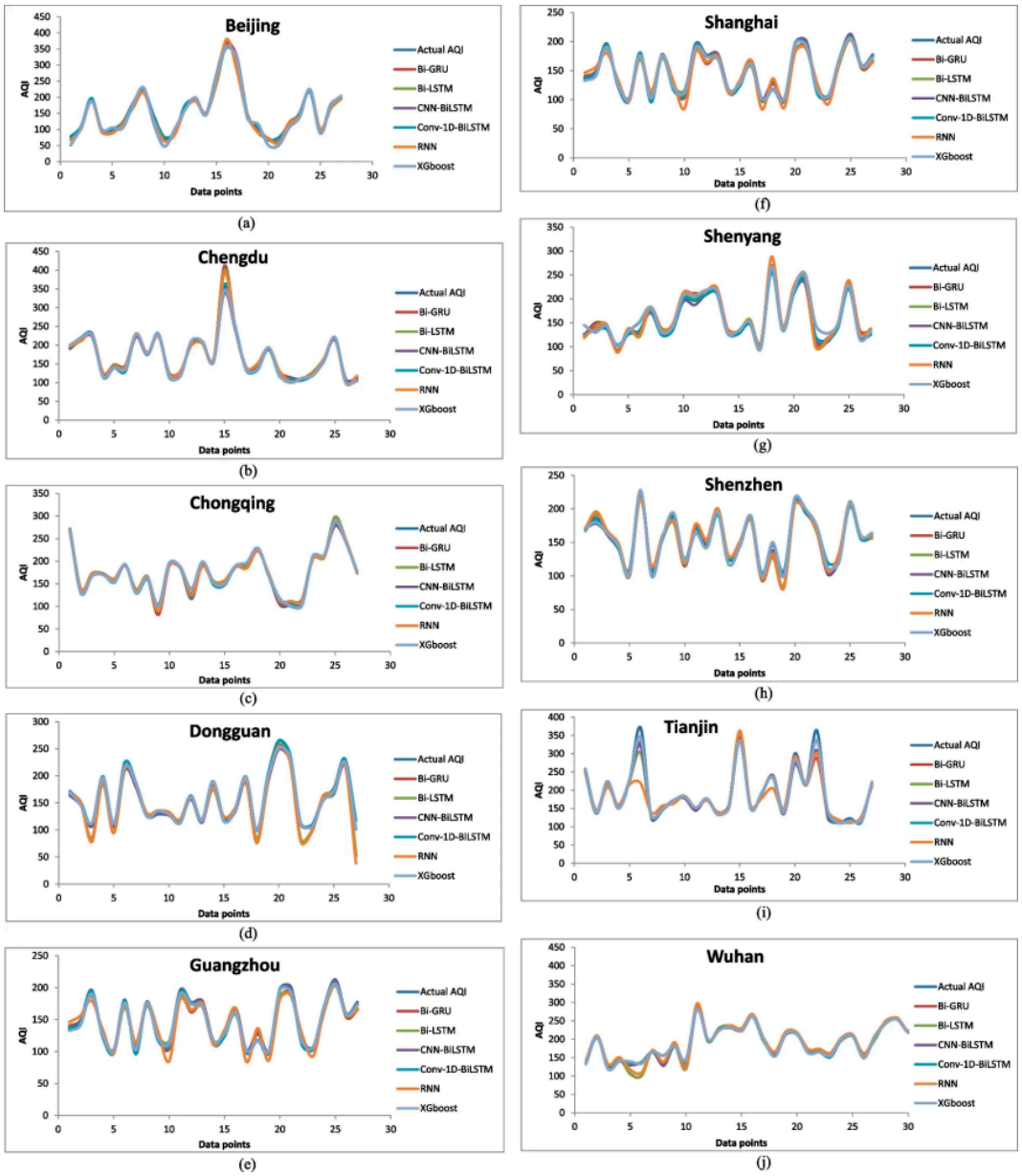


Figure 2. Cross-domain and cross-lingual results showing the performance of the blueprint model compared with the baseline methods.

3.4. Ablation and Error Analysis

Ablation studies highlight the value of each component. Removing constraint matching reduces structural F1 by 6.4 points and raises role ambiguity errors by 9%. Removing blueprint induction and using only local constraints causes a larger drop of 9.1 F1 points, showing that the main benefit comes from corpus-level structural regularities. Error analysis reveals two major issues. First, documents with nested discourse (such as legal opinions containing quoted cases) sometimes merge multiple structures into a single blueprint, which leads to missing roles. Second, very short documents do not provide enough evidence to select a stable template, causing the model to fall back to patterns similar to the baselines. These observations point to the need for better handling of nested text and low-evidence cases in future work.

 doi:10.20944/preprints202512.0338.v1

4. Conclusions

This study shows that structural blueprints can improve unsupervised entity understanding across different types of text. The framework builds simple and stable templates from repeated syntactic patterns and matches sentence units to blueprint slots without labeled data. Tests on news, legal, and scientific corpora show higher structural consistency, clearer roles, and better function resolution than syntax-only or pattern-based methods. Cross-domain and cross-lingual results further indicate that blueprint templates transfer well and are less affected by changes in vocabulary or writing style. These findings demonstrate that explicit structural templates offer a useful way to organize unlabeled text at scale. The approach still has limits in documents with nested discourse or in very short passages where the model cannot form a reliable template. Future work may add light external cues or weak supervision to address these cases while keeping the label-free setting.

References

1. Rondinelli, A., Bongiovanni, L., & Basile, V. (2022). Zero-shot topic labeling for hazard classification. Information, 13(10), 444.
2. Kauffmann, J., Esders, M., Ruff, L., Montavon, G., Samek, W., & Müller, K. R. (2022). From clustering to cluster explanations via neural networks. IEEE transactions on neural networks and learning systems, 35(2), 1926-1940.
3. España-Bonet, C., Barrón-Cedeño, A., & Màrquez, L. (2023). Tailoring and evaluating the Wikipedia for in-domain comparable corpora extraction. Knowledge and Information Systems, 65(3), 1365-1397.
4. Wu, S., Cao, J., Su, X., & Tian, Q. (2025, March). Zero-Shot Knowledge Extraction with Hierarchical Attention and an Entity-Relationship Transformer. In 2025 5th International Conference on Sensors and Information Technology (pp. 356-360). IEEE.
5. Murty, S., Verga, P., Vilnis, L., Radovanovic, I., & McCallum, A. (2018, July). Hierarchical losses and new resources for fine-grained entity typing and linking. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 97-109).
6. Wu, C., Zhang, F., Chen, H., & Zhu, J. (2025). Design and optimization of low power persistent logging system based on embedded Linux.
7. Chai, Y., Zhang, H., Yin, Q., & Zhang, J. (2023, June). Neural text classification by jointly learning to cluster and align. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
8. Yuan, M., Qin, W., Huang, J., & Han, Z. (2025). A Robotic Digital Construction Workflow for Puzzle-Assembled Freeform Architectural Components Using Castable Sustainable Materials. Available at SSRN 5452174.
9. Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5), 1-31.
10. Yin, Z., Chen, X., & Zhang, X. (2025). AI-Integrated Decision Support System for Real-Time Market Growth Forecasting and Multi-Source Content Diffusion Analytics. arXiv preprint arXiv:2511.09962.
11. Lopes Junior, A. G. (2025). How to classify domain entities into top-level ontology concepts using language models: a study across multiple labels, resources, domains, and languages.
12. Chen, F., Liang, H., Yue, L., Xu, P., & Li, S. (2025). Low-Power Acceleration Architecture Design of Domestic Smart Chips for AI Loads.
13. Shehata, S., Karray, F., & Kamel, M. (2009). An efficient concept-based mining model for enhancing text clustering. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1360-1371.
14. Liang, R., Ye, Z., Liang, Y., & Li, S. (2025). Deep Learning-Based Player Behavior Modeling and Game Interaction System Optimization Research.
15. Toldo, M., Maracani, A., Michieli, U., & Zanuttigh, P. (2020). Unsupervised domain adaptation in semantic segmentation: a review. Technologies, 8(2), 35.
16. Wu, C., & Chen, H. (2025). Research on system service convergence architecture for AR/VR system.
17. Jin, J., Su, Y., & Zhu, X. (2025). SmartMLOps Studio: Design of an LLM-Integrated IDE with Automated MLOps Pipelines for Model Development and Monitoring. arXiv preprint arXiv:2511.01850.

 doi:10.20944/preprints202512.0338.v1

18. Xu, K., Wu, Q., Lu, Y., Zheng, Y., Li, W., Tang, X., ... & Sun, X. (2025, April). Meatrd: Multimodal anomalous tissue region detection enhanced with spatial transcriptomics. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 12, pp. 12918-12926).
19. Tashakori, E., Sobhanifard, Y., Aazami, A., & Khanizad, R. (2025). Uncovering Semantic Patterns in Sustainability Research: A Systematic NLP Review. Sustainable Development.
20. Tan, L., Liu, D., Liu, X., Wu, W., & Jiang, H. (2025). Efficient Grey Wolf Optimization: A High-Performance Optimizer with Reduced Memory Usage and Accelerated Convergence.