Cross-Domain Zero-Shot Performance of Small Language Models for Knowledge Extraction Tasks

Adrian Brasoveanu
Modul Technology
Vienna, Austria
brasoveanu@modul.ac.at

Albert Weichselbraun
webLyzard technology
Vienna, Austria
weichselbraun@weblyzard.com

Lyndon J.B. Nixon
Storypact
Vienna, Austria
Modul Technology
Vienna, Austria
nixon@storypact.com

Arno Scharl
webLyzard technology
Vienna, Austria
arno.scharl@modul.ac.at

Abstract

This paper presents a comprehensive evaluation of small language models (up to 500M parameters) fine-tuned for Named Entity Recognition (NER) on the CoNLL dataset. We examine the performance of publicly available models obtained from Hugging Face in a cross-domain, zero-shot setting across three tasks: NER, Named Entity Linking (NEL), and Relation Extraction (RE). Our experiments enable an assessment of their suitability as ready-to-use components in natural language processing (NLP) pipelines and provide insights into their robustness and generalization capabilities across tasks and domains.

All models were integrated into a standardized entity linking and extraction pipeline that employs a consistent evaluation algorithm. Our experiments reveal substantial variations in cross-domain NER performance. For NEL, linking accuracy was highly sensitive to domain shifts, while for RE, the choice of integration algorithm significantly affected overall performance, resulting in comparable outcomes across models. These findings highlight the continued usefulness and relevance of smaller Transformer models for specialized knowledge extraction tasks and emphasize the importance of advances in representation learning to enhance their generalization and robustness.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Computing methodologies** → **Information extraction**; • **Computing methodologies** → **Transfer learning**; • **Computing methodologies** → **Relation extraction**; • **Information systems** → **Information extraction**; • **Information systems** → **Language models**; • **Information systems** → **Entity resolution**;

Keywords

Small Language Models, Transformers, Named Entity Recognition, Named Entity Linking, Relation Extraction, Zero-Shot Learning.

ACM Reference Format:
Adrian Brasoveanu, Albert Weichselbraun, Lyndon J.B. Nixon, and Arno Scharl. 2026. Cross-Domain Zero-Shot Performance of Small Language Models for Knowledge Extraction Tasks. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

This paper addresses a critical gap in understanding how well contemporary Transformer-based models, fine-tuned specifically for NER, can serve as foundational components for broader knowledge extraction (KE) pipelines in real-world applications. Although most evaluations focus on performance within the same domain as the training data, practical deployment scenarios often require models to extract structured information from diverse text sources that may differ significantly from their original training corpora. Our systematic assessment provides essential information on the robustness and generalization capabilities of these models in different domains, revealing how architectural innovations and training strategies translate to performance when faced with the linguistic diversity found in the scientific literature, political discourse, and entertainment content. Taken together, these domains provide broad cross-domain coverage, capturing distinct linguistic registers, communication intents, and stylistic conventions.

The broader implications of this work extend beyond the NER task to encompass two other critical knowledge extraction tasks: NEL for connecting entities to knowledge base entries, and RE for identifying semantic relationships between entities. Understanding the cross-domain transfer capabilities of different Transformer architectures across all three tasks helps practitioners make informed decisions about model selection for multi-domain applications, while our systematic evaluation provides useful insights for developing more robust evaluation methodologies. Our goal is to understand how foundational NER capabilities propagate through

Preprint. This manuscript has not undergone peer review.

Make sure to enter the correct conference title from your rights confirmation email,

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

the knowledge extraction pipeline, highlighting both the potential and limitations of repurposing models trained on standard NER benchmarks for complex real-world information extraction scenarios that demand consistent performance across diverse textual domains.

Some recent small language models performance evaluations such as [Golde et al.(2025)] and [Teglia et al.(2025)] reveal that even well-trained encoder models exhibit variable zero-shot generalization across closely related knowledge extraction tasks. This highlights a persistent gap between strong in-domain performance and cross-task transferability. Studying zero-shot transfer from fine-tuned NER models to downstream tasks like NEL or RE can expose how much structured knowledge about entities and their semantics these models actually encode, beyond surface recognition. Such insights are crucial for designing efficient, modular KE pipelines that minimize retraining costs while maintaining performance across tasks. Moreover, they offer a principled way to evaluate the representational depth of NER models, bridging classic KE and modern foundation model paradigms.

Besides bridging the gap, an evaluation of the zero-shot performance of small language models like Transformers fine-tuned on CoNLL-2003 [Tjong Kim Sang and De Meulder(2003)] for KE tasks, like this one, highlights accessible baselines that require significantly fewer computational resources than Large Language Models (LLMs). This makes such models viable for deployment in resource-constrained environments, edge devices, and applications where cost efficiency and reduced environmental impact are priorities. Their smaller size offers advantages in interpretability and faster inference times, crucial factors for many real-world applications that prioritize consistent performance and low latency over state-of-the-art accuracy. Furthermore, understanding how these small models transfer across diverse KE tasks provides valuable insights into fundamental knowledge representation in Transformer architectures. Such evaluations help democratize NLP research and application development, allowing researchers and organizations with limited computational resources to participate meaningfully in the field.

The paper is organized as follows: Section 2 presents important milestones and recent developments from the history of the three tasks; Section 3 describes the approach we took for selecting models and the basic algorithms we used for evaluations; Section 4 describes and discusses the experiments; and Section 5 concludes the paper with an overview of the results and some ideas for further improvements.

2 Related Work

The evolution of NER, NEL, and RE can be traced back to the Message Understanding Conferences (MUC) of the 1990s, particularly MUC-6 [Sundheim(1995)] and MUC-7 [Chinchor et al.(1998)], which formalized NER and early template-based relation extraction as key components of information extraction systems. However, it was not until the mid-2000s that Named Entity Linking emerged as a distinct task, with early systems such as Bunescu and Pasca [Bunescu and Pasca(2006)] proposing methods to disambiguate named entities against Wikipedia entries. A significant milestone came with SemEval 2013 Task 9 [Segura-Bedmar et al.(2013)], which introduced a more nuanced formulation of entity linking, including partial matching, NIL prediction (e.g., predicting which entities do not have a link in the knowledge graph), and clustering of entities based on types — aspects that directly influenced later developments. The most impactful phase in defining the modern landscape of NEL and RE occurred during the TAC-KBP (Text Analysis Conference – Knowledge Base Population) evaluations, especially in the late years of the previous decade through the development of large-scale benchmarks for multilingual NEL [Ji et al.(2017)] and fine-grained NEL [Ji et al.(2019)]. TAC-KBP established large-scale benchmarks for both entity linking (including NIL and coreference resolution) and slot filling (a form of fine-grained relation extraction), emphasizing realistic, open-domain scenarios grounded in knowledge base construction.

Automated evaluation tools such as neleval [Hachey et al.(2014)], GERBIL [Röder et al.(2018)] and Orbis [Odoni et al.(2018)] also appeared during the same period. Neleval and GERBIL provided standardized benchmarking, while Orbis focused on visual benchmarking. By offering visual cues and contextual information, Orbis supported researchers in better understanding individual results and improving their NER and NEL systems.

The development of Transformers and LLMs [Zhao et al.(2023)] has democratized open knowledge extraction. To follow the developments of the last few years, it is best to consult surveys dedicated to each of the tasks. Seow et al. [Seow et al.(2025)] explore NER, showcasing various evolving modeling paradigms beyond traditional sequence labeling, including span-based, machine reading comprehension, and prompt-based learning. They underscore that Large Language Models (LLMs) are crucial for adapting to downstream NER tasks through fine-tuning or prompting, despite challenges such as high computational costs and potential biases. Guellil et al. [Guellil et al.(2024)] observe that Transformers and contextual embeddings are primarily used for named entity disambiguation (NED), although most research efforts continue to focus on NER rather than comprehensive end-to-end entity linking systems. Zhao et al. [Zhao et al.(2024)] provide a comprehensive survey on RE, noting the dominance of Large Pre-trained Language Models (PLMs), which have elevated state-of-the-art RE to new levels, with LLMs like Claude or GPT further complementing their capabilities in handling complex texts. Across all three surveys, there is a clear emphasis on continued challenges and advancements in addressing low-resource, cross-domain, cross-lingual, multimodal and fine-grained scenarios within all three tasks.

3 Method

3.1 Model Selection

To identify suitable Transformer models for our knowledge extraction tasks, we developed a systematic testing framework that validated model accessibility and performance quality on sample data. Our selection criteria focused primarily on well-known Transformer model architectures tagged with "ner", "conll", and "token-classification" labels on the Hugging Face repository, ensuring they were specifically designed or fine-tuned for NER tasks rather than general language modeling. Our testing pipeline loaded each candidate model using the Transformers library and evaluated them on a small set of representative sentences containing various entity types. For each model, we verified that the tokenizer and model

weights could be loaded successfully, confirmed that the model produced reasonable named entity predictions with valid entity types (PER, ORG, LOC, MISC), and measured the F1 score and accuracy as basic quality indicators. We also looked at the processing speed and methods to quickly improve it (e.g. moving the computation of similarities to GPUs to achieve 10x speed improvements compared to sequential computation on CPUs, etc.), as our goal was to run most of these models on single A100 GPUs. This preliminary screening process revealed that many advanced models were either unavailable in properly fine-tuned versions for NER tasks or produced poor quality outputs, highlighting the importance of careful model selection.

From our initial selection of over 25 candidate models, we selected eight models (see Table 1) that demonstrated both technical compatibility and reasonable performance in our test suite. Our selection specifically excluded models that were only available as base pre-trained versions without NER fine-tuning, as evidenced by poor F1 scores during preliminary testing. Notably, we encountered significant challenges with accessing properly fine-tuned versions of advanced models like GLiNER, LUKE and KnowBERT, which either required separate evaluation pipelines, authentication gateways or had no additional fine-tuning, ultimately leading to their exclusion from our comparative study despite their theoretical advantages for entity-related tasks.

To narrow down from the initial 25 candidates to our final 8 models, we performed preliminary testing on a subset of 100 CrossNER sentences [Liu et al.(2021)], evaluating both performance and technical compatibility with our evaluation framework. This pilot test revealed significant variations in model outputs, tokenization schemes, and aggregation strategies that required careful handling to ensure fair comparison. Models that did not produce reasonable entity predictions, had incompatible output formats, or demonstrated obvious fine-tuning issues were eliminated at this stage.

The evaluation was deliberately constrained by practical usability and methodological consistency to ensure a fair comparison rather than exhaustive optimization. We limited our technical adaptations of the selected models to changes that could be implemented and validated in a single day of development work, focusing on standardization rather than optimization of individual models. This constraint meant that we avoided complex post-processing techniques, custom fine-tuning, or architectural modifications that might have improved performance but would have compromised the comparative nature of our evaluation. The DeBERTa model's extensive changes nearly exceeded this constraint, highlighting how architectural innovations can introduce unexpected complexity in practical deployment scenarios. Instead of pursuing model-specific optimizations, we concentrated on ensuring that each model could operate under equivalent conditions, using their default hyperparameters and inference settings while applying only the minimal adaptations necessary for consistent evaluation across the diverse architectural landscape.

DeBERTa required special handling due to its unique tokenization and entity boundary detection requirements. Initial testing revealed that the standard entity post-processing pipeline produced severely degraded performance, despite the model card showing promising results. To address this, we implemented a pipeline that properly handles DeBERTa's tokenization artifacts, improves entity boundary detection that accounts for the model's disentangled attention mechanism, and adds an entity merging algorithm that combines fragmented entities of the same type when they appear within close proximity.

Our selected models (cf. Table 1) represent a balance between architectural diversity, demonstrated competency in standard benchmarks based on model cards and early evaluation, and technical feasibility within our evaluation constraints. The inclusion of both base and large variants of popular architectures like BERT allowed us to assess the impact of model scale, ModernBERT allowed us to study the recent architectures, while the presence of efficiency-focused models like DistilRoBERTa provided insights into the performance-computational trade-offs relevant for practical deployment scenarios. The near-exclusion of DeBERTa due to its technical complexity served as a valuable reminder that structural innovation does not always translate to practical usability, though the model's eventual strong performance vindicated the effort required to integrate it properly.

3.2 NER Strategy

Since the models we selected have all been fine-tuned on CoNLL [Tjong Kim Sang and De Meulder(2003)], a dataset focused on the main entity types (PER, ORG, LOC, MISC), we started directly with an evaluation on CrossNER [Liu et al.(2021)], a cross-domain dataset. To adjust to CrossNER, we implemented type mapping from the diverse CrossNER entity categories to the standard four-class CoNLL schema (PER / ORG / LOC / MISC). For example, entities such as politician, scientist, musical artist, writer, researcher, and person corresponding to the general interest and the respective domains were all mapped to person (PER). This allows for meaningful comparison despite the different annotation schemes used in the source training data versus the evaluation dataset.

Entity boundary reconstruction proved to be one of the most challenging technical hurdles, as Transformer models often produce fragmented predictions due to subword tokenization breaking entities across multiple tokens. Our solution involved implementing a sophisticated entity merging algorithm that could reconnect fragmented predictions while maintaining type consistency and avoiding false mergers across different entity classes. This process included handling common tokenization artifacts, adjusting character offsets for models that included or excluded whitespace in their predictions, and implementing fuzzy boundary matching to account for minor discrepancies in entity span detection that do not reflect genuine semantic errors.

The evaluation framework itself incorporated lessons learned from the NER evaluation literature (e.g., SemEval 2013 Task 9.1) and the nervaluate package, implementing both strict boundary matching and relaxed evaluation with 0.5 partial credit for entities with correct types but imperfect boundaries. This evaluation approach proved essential given the boundary detection challenges inherent in cross-domain transfer, where models trained on news text might struggle with the varied linguistic patterns found in scientific literature, political discourse, or music reviews.

Table 1: Transformer architectures for knowledge extraction

geckos/deberta-base-fine-tuned-ner (139M)
Innovation: Disentangled attention separating content and position representations Attention: Content-to-content, content-to-position, position-to-content matrices Training: Fine-tuned on CoNLL-2003 (F1: 96.08%)
Jean-Baptiste/roberta-large-ner-english (355M)
Innovation: Optimized BERT training with dynamic masking, specialized for NER Attention: Standard multi-head self-attention mechanism (24 layers, 1024 hidden) Training: Fine-tuned on CoNLL-2003, optimized for informal text/entities
philschmid/distilroberta-base-ner-conll2003 (82M)
Innovation: Knowledge distillation of RoBERTa for efficient NER Attention: Distilled multi-head self-attention (6 layers, 768 hidden) Training: Student model, fine-tuned on CoNLL-2003 (F1: 90.74%)
dbmdz/bert-large-cased-finetuned-conll03-english (340M)
Innovation: Case-sensitive BERT-large fine-tuned for multilingual NER Attention: Standard multi-head self-attention mechanism (24 layers, 1024 hidden) Training: Fine-tuned bert-large-cased on CoNLL-2003
dslim/bert-large-NER (340M)
Innovation: BERT-large optimized specifically for Named Entity Recognition Attention: Standard multi-head self-attention mechanism (24 layers, 1024 hidden) Training: Fine-tuned on CoNLL-2003 (F1: 91.7%)
dbmdz/electra-large-discriminator-finetuned-conll03-english (335M)
Innovation: Generator-discriminator approach with replaced token detection for NER Attention: Standard multi-head self-attention mechanism (24 layers, 1024 hidden) Training: Fine-tuned on CoNLL-2003 for entity classification
dslim/bert-base-NER (110M)
Innovation: BERT-base fine-tuned specifically for Named Entity Recognition Attention: Standard multi-head self-attention mechanism (12 layers, 768 hidden) Training: Fine-tuned on CoNLL-2003 (F1: 91.3%)
IsmaelMousa/modernbert-ner-conll2003 (149M)
Innovation: ModernBERT with RoPE, alternating attention, and 8K context length Attention: Local-Global alternating attention with Rotary Positional Embeddings (22 layers, 768 hidden) Training: Fine-tuned answerdotai/ModernBERT-base on CoNLL-2003 (F1: 0.84%)

3.3 NEL Strategy

Building upon the successful NER evaluation framework, we extended our analysis to include Named Entity Linking and Relation Extraction tasks, leveraging the same foundational infrastructure while implementing task-specific adaptations. The NEL evaluation required developing custom linking strategies that could connect the entities identified by our NER models to knowledge base entries, implementing both classic NEL approaches that relied on surface form matching and context similarity, as well as more sophisticated embedding-based linking methods. We adapted the entity extraction pipeline to preserve not only entity boundaries and types but also the contextual information necessary for disambiguation, allowing us to evaluate how well models could not only identify entities but also resolve them to specific knowledge base concepts. This extension revealed additional challenges in cross-domain transfer, as linking accuracy proved even more sensitive to domain shift than basic entity recognition, particularly when entities from specialized domains like AI or science required disambiguation against technical knowledge bases.

The Named Entity Linking algorithm operates in two main phases: entity recognition and entity linking. First, a pre-trained

Algorithm 1: Named Entity Linking algorithm using multi-strategy candidate selection and ranking.

Input: Text document `text`, Knowledge graph `knowledge_graph`
Output: List of linked entities

```
# Step 1: Entity Recognition;
entities ← NER_MODEL(text);
cleaned_entities ← MERGE_FRAGMENTS(entities, text);
# Step 2: Knowledge Graph Lookup;
linked_entities ← [ ];
foreach entity in cleaned_entities do
    candidates ← [ ];
    candidates += EXACT_MATCH(entity.text, kg.surface_forms);
    candidates += FUZZY_MATCH(entity.text, kg.surface_forms, threshold=0.85);
    candidates += SEMANTIC_MATCH(entity.text, kg.tfidf_vectors);
    candidates += PARTIAL_MATCH(entity.text, kg.surface_forms);
    ranked ← RANK_CANDIDATES(candidates, entity.type, context);
    best_match ← ranked[0] if ranked else FALLBACK_URI(entity.text);
    linked_entities.append(entity + best_match);
end
```

Named Entity Recognition model processes the input text to identify entity mentions and their boundaries. The raw NER output often contains fragmented entities due to subword tokenization, so an advanced merging step consolidates fragments using contextual rules specific to entity types like persons, organizations, and locations. This merging considers gaps between fragments, entity type consistency, and common linguistic patterns to reconstruct complete entity mentions. The second phase performs knowledge graph lookup using a multi-strategy ensemble approach. For each recognized entity, the system attempts four different matching strategies: exact surface form matching against pre-built alias tables, fuzzy string matching using Levenshtein distance for handling spelling variations, semantic similarity matching using TF-IDF vectors to capture meaning-based relationships, and partial matching for incomplete mentions. All candidate matches are then ranked using a composite scoring function that considers string similarity, entity type compatibility, confidence scores, and local context. The highest-ranked candidate becomes the final link, with a fallback mechanism generating DBpedia-style URIs [1] for entities not found in the knowledge graph. This ensemble approach significantly improves linking accuracy by combining multiple complementary matching strategies.

[1] https://dbpedia.org/resource/Nirvana_(band) for the Nirvana rock band

Algorithm 2: Prototype-based few-shot relation extraction using weighted prototype+example embeddings and multi-view matching across Transformer models.

Input: Evaluation dataset `eval_data`, relation set `relations`, model list `MODELS`, evaluation size `EVAL_SIZE`
Output: Predicted relation labels and consistency report

```
MODELS ← [DistilRoBERTa, RoBERTa-Large, BERT variants, DeBERTa, ELECTRA, ModernBERT];
EVAL_SIZE ← user_choice(1k, 5k, 10k, full);
# Step 1: Hand-crafted Prototype Creation (e.g., "X crosses Y");
prototypes ← create_hand_crafted_prototypes()
# Step 2: Few-shot Example Extraction;
few_shot_examples ← extract_from_training_data();
// e.g., "London [RELATION] England"
relation_embeddings ← weighted_combine(prototypes=40%, few_shot_examples=60%);
foreach model in MODELS do
    sampled_data ← stratified_sample(eval_data, EVAL_SIZE);
    foreach example in sampled_data do
        head, tail, context ← extract_entities(example);
        representations ← create_multiple_views(head, tail, context);
        input_embeddings ← encode(representations);
        similarities ← cosine_similarity(input_embeddings, relation_embeddings);
        prediction ← best_matching_relation(similarities);
    end
end
analyze_model_consistency();          // Evaluate generalization across model types
```

3.4 Relation Extraction Strategy

We implemented prototyping strategies that could generalize relation patterns across domains. The relation extraction pipeline built upon the NER scaffolding by using identified entities as anchor points for relation detection, but required sophisticated span-pair analysis and relation classification components that operated at a higher semantic level than token-level entity recognition. This hierarchical approach allowed us to assess how well models trained primarily for entity recognition could be adapted for more complex information extraction tasks, though the results highlighted the limitations of repurposing NER models for relation-level understanding without task-specific fine-tuning.

Our approach develops and evaluates a prototype-based few-shot learning method for relation extraction that combines hand-crafted semantic prototypes with examples extracted from the train set. For each relation type, we create rich prototype representations using both expert-designed templates and examples extracted from training data, then combine these through weighted embedding averaging to create comprehensive relation signatures. The method generates multiple linguistic representations of each input example, computes similarity scores against all relation prototypes, and selects the best matching relation through cosine similarity with additional disambiguation rules for challenging relation pairs.

The key methodological contribution is demonstrating that this prototype-based approach works consistently across diverse model architectures without requiring model-specific adaptations or fine-tuning. By applying the identical few-shot prototype method to eight different pre-trained language models ranging from DistilRoBERTa to ModernBERT, we can isolate the effectiveness of the algorithmic approach from model architecture choices. This evaluation strategy allows us to determine whether advances in relation extraction should focus on developing better prototype-based learning methods or continue exploring different Transformer architectures, while establishing whether strong performance on named entity recognition tasks translates to effective relation extraction capabilities through our unified prototype framework.

4 Experiments

This section presents evaluation results for the individual tasks, and concludes with a discussion around these results.

All the models used in these evaluations were already fine-tuned on CoNLL 2003 dataset. Due to this aspect, the hyperparameters for the models were primarily taken from their model cards (if available), or inferred from their respective architecture papers. For Jean-Baptiste, dbmdz, and dslim models, values are inferred from RoBERTa [Liu et al.(2019)], Electra [Clark et al.(2020)], and BERT [Devlin et al.(2019)] paper defaults, as they were not included in model cards. These parameters represent standard practices for fine-tuning these architectures on sequence labeling tasks, though specific implementations may vary slightly between model developers. Table 2 presents the likely values of these parameters.

A fixed seed was introduced for all relevant libraries (random, numpy, and torch), and deterministic settings were enabled to minimize nondeterministic behaviour. This substantially increases reproducibility across runs by controlling most sources of randomness, though small differences may still occur due to hardware-level operations (e.g., differences between GPUs). The highest variation between runs was observed at the relation extraction evaluation, despite these changes.

4.1 NER Evaluation

Instead of evaluating on CoNLL, we decided to evaluate on a more recent dataset, namely CrossNER [Liu et al.(2021)], as this is a freely available, open corpora which has a more varied set of entities from multiple domains, e.g. politics, artificial intelligence, literature. We have mapped the entities associated with these 5 domains (Politics, Natural Science, Music, Literature, Artificial Intelligence) on the 4 entity types from CoNLL (PER, ORG, LOC, MISC) to get a feeling for what kind of scores one can expect. We computed both direct scores and relaxed scores. We prefer relaxed metrics for this cross-domain analysis as they better capture semantic understanding while accounting for boundary detection challenges inherent in domain transfer. They proved also to be more in line with the scores declared on the HuggingFace model cards.

Table 3 presents the results of the NER evaluation, which reveal significant performance variations. Following established practices from NER evaluation literature and the nervaluate [2] framework, we employ a 0.5 partial match scoring system that credits predictions when entity types are correct and boundaries have at least 50% overlap with gold annotations. Under this relaxed evaluation approach, geckos/deberta-base-fine-tuned-ner and Jean-Baptiste/roberta-large-ner-english models achieve the best performance, demonstrating superior cross-domain transfer capabilities through disentangled attention mechanisms that separates content and positional representations.

The domain-specific analysis reveals that literature and politics domains generally yielded the highest relaxed F1 scores across all models, while AI and science domains proved to be the most challenging, often showing performance drops of 10-15 percentage points compared to the other domains. This pattern suggests that technical terminology and the types of specialized entities in scientific domains present persistent challenges even under relaxed evaluation conditions. The consistent performance ordering across domains reinforces the reliability of the relaxed evaluation approach while highlighting the robustness differences between architectural approaches when faced with diverse linguistic patterns and entity distributions found in cross-domain transfer scenarios.

4.2 NEL Evaluation

We began this NEL evaluation by adapting the NER evaluation script to work with entity linking, using the TweekiGold dataset [Harandizadeh and Singh(2020)] which contains tweets with specialized entities (e.g., creative works, political entities, etc). The approach involved running top-performing NER models on the tweet text to detect entity mentions, then generating DBpedia links using an algorithmic system that mapped entity surface forms to knowledge base entries. Our first results showed a significant gap between mention detection performance, essentially the scores for NEL being less than 50% from the scores obtained for NER. To close this gap, we built a small script that created a small knowledge graph with Wikidata or Dbpedia links. Only the important fields like name, type or abstract were collected, as these were needed for disambiguation. The final NEL evaluation results are presented in Table 3.

This allowed us to develop a classic NEL algorithm that is presented in Figure 1. This algorithm also contained a smart entity decomposition system that could recognize merged entity patterns and intelligently split them into their constituent components leading to better boundary positions and individual links. For example, when a model predicted "Nirvanas Nevermind" as a single entity, our decomposer would split it into "Nirvanas" linked to the Nirvana band page and "Nevermind" linked to the album page, adjusting

[2] https://github.com/MantisAI/nervaluate

Table 2: Training hyperparameters reported or inferred for the NER models evaluated.

Model (HF id)	BS (train)	BS (eval)	LR	Epochs	Optimizer	LRS
Jean-Baptiste/roberta-large-ner-english	16	16	2e-5	4	Adam	linear
geckos/deberta-base-fine-tuned-ner	16	16	5e-5	5	Adam	linear
dbmdz/electra-large-discriminator-finetuned-conll03-english	32	16	3e-5	4	Adam	linear
philschmid/distilroberta-base-ner-conll2003	32	16	4.99e-5	6	Adam	linear
dbmdz/bert-large-cased-finetuned-conll03-english	16	16	2e-5	3	Adam	linear
dslim/bert-large-NER	16–32	16	2e-5–5e-5	3	Adam	linear
dslim/bert-base-NER	16–32	16	2e-5–5e-5	3	Adam	linear
IsmaelMousa/modernbert-ner-conll2003	8	8	1e-6	10	AdamW	linear

Table 3: Comparison of F1 scores across three evaluation tasks: CrossNER (NER), Tweegy (NEL), and FewRel (RE). Best score in each column is in bold.

Model	CrossNER (NER)	Tweegy (NEL)	FewRel (RE)
Jean-Baptiste/roberta-large-ner-english	**0.869**	**0.731**	0.633
geckos/deberta-base-fine-tuned-ner	0.860	0.706	0.624
dbmdz/electra-large-discriminator-finetuned-conll03-english	0.857	0.693	0.619
philschmid/distilroberta-base-ner-conll2003	0.849	0.727	0.637
dbmdz/bert-large-cased-finetuned-conll03-english	0.835	0.674	0.616
dslim/bert-large-NER	0.830	0.687	0.647
dslim/bert-base-NER	0.819	0.668	**0.651**
IsmaelMousa/modernbert-ner-conll2003	0.707	0.558	0.610

the boundary coordinates to match the expected fine-grained annotations. This approach addressed the fundamental issue that NER models naturally group related entities together while the evaluation dataset maintained strict separation between different entity types, allowing us to bridge the gap between coarse-grained model predictions and fine-grained evaluation requirements. Although this solution is not perfect, the fact that the models were all within a small range suggests that the linking algorithms are more important than the models themselves.

4.3 Relation Extraction Evaluation

We developed a systematic approach to evaluate relation extraction methods by testing the selected Transformers on the classic FewRel dataset [Han et al.(2018)]. The number of random samples was set to 4000 to keep the dataset in line with the datasets we used for the other tasks. Our initial attempts followed common practices in the literature by trying to leverage NER capabilities of these models to enhance relation classification, under the assumption that entity-type information would provide useful signals for determining semantic relationships. However, this hybrid approach consistently underperformed (e.g., F1 scores of 0.3).

The evaluation revealed that current approaches to relation extraction may benefit more from advances in representation learning methodology than from improved language models. Specifically, our prototype-based approach with weighted few-shot augmentation outperformed more complex hybrid methods (scores up to 0.65), suggesting that simpler, more interpretable algorithms may be more effective than approaches that try to combine multiple model capabilities. This finding aligns with recent trends in the literature that emphasize the importance of training, prompting and data representation over the pure model scale [Gao et al.(2021)].

4.4 Discussion

For Named Entity Recognition (NER), the geckos/deberta-base-fine-tuned-ner and Jean-Baptiste/roberta-large-ner-english models emerged as top performers in cross-domain scenarios. RoBERTa is typically used in classification tasks. DeBERTa's disentangled attention mechanism separating content and positional representations likely help a lot in the NER tasks. While these models demonstrated superior robustness in literature and politics, they faced greater challenges in AI and science domains. Critically, for Named Entity Linking (NEL) and Relation Extraction (RE), the specific Transformer architecture proved less influential; algorithmic strategies for linking and relation extraction consistently outranked model architectural innovations in importance, yielding results within a narrow range across different models. This indicates that the method of task adaptation significantly impacts performance for these complex downstream tasks.

Challenges. A key challenge lies in the pronounced sensitivity of model performance to domain shifts across all knowledge extraction tasks, particularly for NEL where linking accuracy suffered. Models often struggled with fragmented entity predictions and the nuanced distinction required by fine-grained annotations, necessitating complex entity decomposition and merging algorithms. Moreover, directly repurposing NER-tuned models for tasks like relation extraction revealed inherent limitations in achieving deep semantic understanding without task-specific approaches. What is likely to bring good results, even when using prototype-based few-shot learning is the prompting strategy for selecting high quality

examples. The architectural improvements alone do not overcome the practical usability gaps or the complexities of real-world linguistic diversity.

Potential Improvements. To achieve substantial score improvements, future work should prioritize advancements in representation learning methodology and the development of sophisticated algorithmic strategies over continuous reliance on new Transformer architectures. This includes creating more effective prototype generation methods that automatically discover optimal semantic representations for relation types. Furthermore, exploring ensemble approaches that strategically combine multiple algorithms, rather than just different models, holds significant promise. Finally, leveraging large volumes of unlabeled text through improved few-shot learning techniques and establishing more systematic frameworks to isolate algorithmic contributions from model-specific effects will be crucial for progress in information extraction.

Limitations. A well-known limitation of our current evaluation is the lack of inclusion of models that are specifically focused on NEL like BLINK, CLINK or GENRE (see [Zhu et al.(2023)]). Similarly, for relation extraction, more recent models focused on it like REXEL [Bouziani et al.(2024)] or ReLiK[Orlando et al.(2024)] are missing. This is mainly because we were interested in the performance of NER models for other tasks. Future work will address these limitations.

5 Conclusion

When bigger and better LLMs appear, people often rush to use them for all tasks, even if the previous technology (e.g., small Transformers) has not yet reached its peak. What we observed during these evaluations only confirms that smaller Transformer models are still worth improving and developing, especially for specialized knowledge extraction tasks.

Future work should focus on several key areas to advance beyond our current results. First, we plan to develop more sophisticated prototype generation methods that can automatically discover optimal semantic representations for each type of relation. Second, explore ensemble approaches that combine multiple algorithmic strategies rather than multiple models. Third, investigate how to better leverage the large amounts of unlabeled text available for relation extraction through improved few-shot learning techniques. Finally, developing more systematic evaluation frameworks that can better isolate algorithmic contributions from model-specific effects, helping the field make more principled progress on this fundamental natural language understanding task.

References

[Bouziani et al.(2024)] Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. REXEL: An End-to-end Model for Document-Level Relation Extraction and Entity Linking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track)*. Association for Computational Linguistics, Toronto, Canada, 119–130. doi:10.18653/v1/2024.naacl-industry.11

[Bunescu and Pasca(2006)] Razvan C. Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*, Diana McCarthy and Shuly Wintner (Eds.). The Association for Computer Linguistics. https://aclanthology.org/E06-1002/

[Chinchor et al.(1998)] Nancy Chinchor, Lynette Hirschman, and David Lewis. 1998. *Message Understanding Conference (MUC-7): Proceedings of the Seventh Message Understanding Conference*. Technical Report MUC-7. NIST, Gaithersburg, MD, USA. Available via Linguistic Data Consortium and NIST archives.

[Clark et al.(2020)] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=r1xMH1BtvB

[Devlin et al.(2019)] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[Gao et al.(2021)] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3816–3830. doi:10.18653/V1/2021.ACL-LONG.295

[Golde et al.(2025)] Jonas Golde, Patrick Haller, Max Ploner, Fabio Barth, Nicolaas Jedema, and Alan Akbik. 2025. FAMILIARITY: Better Evaluation of Zero-Shot Named Entity Recognition by Quantifying Label Shifts in Synthetic Training Data. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Albuquerque, New Mexico, 820–834. doi:10.18653/v1/2025.naacl-long.37

[Guellil et al.(2024)] Imane Guellil, Antonio García-Domínguez, Peter R. Lewis, Shakeel Hussain, and Geoffrey Smith. 2024. Entity Linking for English and Other Languages: A Survey. *Knowl. Inf. Syst.* 66, 7 (2024), 3773–3824. doi:10.1007/S10115-023-02059-2

[Hachey et al.(2014)] Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and Easy Entity Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*. The Association for Computer Linguistics, 464–469. doi:10.3115/V1/P14-2076

[Han et al.(2018)] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 4803–4809. doi:10.18653/V1/D18-1514

[Harandizadeh and Singh(2020)] Bahareh Harandizadeh and Sameer Singh. 2020. Tweeki: Linking Named Entities on Twitter to a Knowledge Graph. In *Proceedings of the Sixth Workshop on Noisy User-generated Text, W-NUT@EMNLP 2020 Online, November 19, 2020*, Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (Eds.). Association for Computational Linguistics, 222–231. doi:10.18653/V1/2020.WNUT-1.29

[Ji et al.(2017)] Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, and Cash Costello. 2017. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In *Proceedings of the 2017 Text Analysis Conference, TAC 2017, Gaithersburg, Maryland, USA, November 13-14, 2017*. NIST. https://tac.nist.gov/publications/2017/additional.papers/TAC2017.KBP_Entity_Discovery_and_Linking_overview.proceedings.pdf

[Ji et al.(2019)] Heng Ji, Avirup Sil, Hoa Trang Dang, Ian Soboroff, and Joel Nothman. 2019. Overview of TAC-KBP 2019 Fine-grained Entity Extraction. In *Proceedings of the 2019 Text Analysis Conference, TAC 2019, Gaithersburg, Maryland, USA, November 12-13, 2019*. NIST. https://tac.nist.gov/publications/2019/additional.papers/TAC2019.EDL.overview.proceedings.pdf

[Liu et al.(2019)] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019). https://arxiv.org/abs/1907.11692

[Liu et al.(2021)] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13452–13460. doi:10.1609/AAAI.V35I15.17587

[Odoni et al.(2018)] Fabian Odoni, Philipp Kuntschik, Adrian M. P. Brasoveanu, and Albert Weichselbraun. 2018. On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance. In *Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018, Vienna, Austria, September 10-13, 2018 (Procedia Computer Science, Vol. 137)*, Anna Fensel, Victor de Boer, Tassilo Pellegrini, Elmar Kiesling, Bernhard Haslhofer, Laura Hollink, and Alexander Schindler (Eds.). Elsevier, 33–42. doi:10.1016/J.PROCS.2018.09.004

[Orlando et al.(2024)] Riccardo Orlando, Pere-Lluis Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Toronto, Canada, 14114–14132. doi:10.18653/v1/2024.findings-acl.839

[Röder et al.(2018)] Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL - Benchmarking Named Entity Recognition and Linking consistently. *Semantic Web* 9, 5 (2018), 605–625. doi:10.3233/SW-170286

[Segura-Bedmar et al.(2013)] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, Mona T. Diab, Timothy Baldwin, and Marco Baroni (Eds.). The Association for Computer Linguistics, 341–350. https://aclanthology.org/S13-2056/

[Seow et al.(2025)] Wei Liang Seow, Iti Chaturvedi, Amber Hogarth, Rui Mao, and Erik Cambria. 2025. A Review of Named Entity Recognition: From Learning Methods to Modelling Paradigms and Tasks. *Artificial Intelligence Review* 58, 10 (2025), 315. doi:10.1007/s10462-025-11321-8

[Sundheim(1995)] Beth Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*. ACL, 13–31. doi:10.3115/1072399.1072402

[Teglia et al.(2025)] Simone Teglia, Simone Tedeschi, and Roberto Navigli. 2025. How Much Do Encoder Models Know About Word Senses?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 2266–2277. doi:10.18653/v1/2025.acl-long.113

[Tjong Kim Sang and De Meulder(2003)] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. doi:10.3115/1119176.1119195

[Zhao et al.(2023)] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *CoRR* abs/2303.18223 (2023). arXiv:2303.18223 doi:10.48550/ARXIV.2303.18223

[Zhao et al.(2024)] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers. *ACM Comput. Surv.* 56, 11 (2024), 293:1–293:39. doi:10.1145/3674501

[Zhu et al.(2023)] Fangwei Zhu, Jifan Yu, Hailong Jin, Juanzi Li, Lei Hou, and Zhifang Sui. 2023. Learn to Not Link: Exploring NIL Prediction in Entity Linking. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 10846–10860. doi:10.18653/v1/2023.findings-acl.690

Received 17 October 2025; revised -; accepted -