Knowledge-Augmented Transformers for Concept and Named Entity Recognition: A Comparative Study of ERNIE 2 and BERT

Rokhaya Cisse
ENSAE Paris
rokhaya.cisse@ensae.fr

Abstract

We study the joint recognition of named entities and domain-specific concepts through the lens of the Concept and Named Entity Recognition (CNER) task. We compare the performance of two Transformer-based models: BERT, a widely used contextual language model, and ERNIE 2.0, a knowledge-enhanced architecture that incorporates structured information from external knowledge graphs. Experiments conducted on a representative subsample of the Babelscape CNER dataset show that ERNIE slightly outperforms BERT in overall performance. Although both models struggle to identify rare categories, ERNIE demonstrates a notable advantage in predicting ambiguous categories and those with high co-occurrence rates such as Media. These findings highlight the potential of knowledge-augmented models for improving concept disambiguation and semantic generalization in CNER tasks.

1 Introduction

Extracting meaningful information from raw text remains a central challenge in natural language processing (NLP), particularly in domains where both named entities (e.g., UNDP) and abstract or generic concepts (e.g., air pollution, biodiversity) are relevant for understanding content. While traditional Concept Recognition (CR) has long focused on identifying proper nouns, this scope is often insufficient in practice, especially in fields like public health where categorizing proper noun concepts is equally essential (to identify stakeholders in interventions for examples).
To better address such needs, authors have introduced Concept and Named Entity Recognition (CNER) a framework that considers entity and concept classification as a single, unified task (Martinelli et al., 2024). CNER merges Named Entity Recognition and Concept Recognition (CR) into one predictive formulation, thereby simplifying model design. However, despite their overall good performance, these models tend to underperform on less represented categories and ambiguous words, potentially due to the lack of contextual information.
This challenge reflects a limitation: although architectures like BERT (Devlin et al., 2019) or DeBERTa (He et al., 2023) showcases good performances, they struggle with outliers. This explains the increasing interest in knowledge-augmented models, which incorporate structured external information during training or inference. Approaches such as ERNIE (Sun et al., 2019; 2020) enrich standard architectures with semantic priors drawn from knowledge graphs. These models have shown tangible benefits for tasks such as relation extraction or entity typing. Despite their potential, they have not been systematically tested in the context of CNER, even though the latter might benefit most from external knowledge.
In our work, we aim at exploring that question. In particular, we compare ERNIE 2, a Transformer model with integrated knowledge to BERT, a widely used baseline. To our knowledge, this isno preceding study exploring the existence of an added value with knowledge-augmented models applied

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

to the unified CNER framework.
Our contributions are as follows:
• We identify and quantify the limitations of BERT in handling rare categories within CNER.
• We compare ERNIE 2 and BERT overall performance on a representative sample from original CNER dataset (computational constraints).
• We provide detailed category-wise analysis to assess how external knowledge affects model behavior across.
Through this study, we hope to shed light on the value of this type of model in entity and concept recognition.

2 Related Work

Concept and Named Entity Recognition (CNER) is a recent proposal for jointly categorizing entities and nominal concepts. It offers a unified framework that merges two previously distinct tasks: Named Entity Recognition (NER), which focuses on proper nouns (e.g., persons, locations, organizations), and Concept Recognition (CR), which targets common nouns (e.g., animals, plants, substances). This joint formulation, introduced by Martinelli et al. (2024), has demonstrated encouraging results in terms of word categorization accuracy, irrespective of the type of word involved.

Traditional NER approaches have evolved significantly, moving from rule-based and statistical methods to deep learning models, particularly Transformer-based architectures such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2023). While these models perform well on frequent classes, they often struggle with less common ones due to sparse contextual information and weaker generalization capabilities.

To overcome these challenges, recent research has focused on knowledge-enriched language models that incorporate structured or semi-structured knowledge into Transformer frameworks. Models like KnowBERT (Peters et al., 2019), K-BERT (Liu et al., 2020), and ERNIE (Sun et al., 2019; Sun et al., 2020) exemplify this trend by embedding external resources, such as Wikidata, ConceptNet, or Baidu Encyclopedia, into the training process or model design. These approaches have shown improved performance on semantically demanding tasks, including relation extraction and entity classification.

However, the application of such knowledge-based models to the joint task of concept and entity recognition remains underexplored. The initial CNER benchmark (Martinelli et al., 2024) evaluates high-performing models like DeBERTa-v3, but does not consider knowledge-enhanced alternatives. Moreover, the benchmark results reveal considerably lower performance on long-tail categories (e.g., CELESTIAL, DISEASE, CULTURE), indicating that stronger semantic priors could be advantageous.

This study investigates whether incorporating external knowledge can help address these limitations. We compare ERNIE 2, a knowledge-integrated Transformer model, with BERT on the CNER benchmark dataset. To our knowledge, this is the first direct evaluation of knowledge-enriched models on the unified CNER task.

3 Data Description and Preprocessing

The data used in this study comes from the study proposed by Martinelli et al. (2024), which reframes the identification of named entities and nominal concepts as a unified sequence labeling task. The corpus comprises English-language sentences sampled from Wikipedia linked to Wordnet for annotation and manually annotated data. CNERsilver comprises approximately 317,590 sentences, totaling over 8.7 million tokens and more than 2.2 million labeled spans, with a span density of 39.00%, indicating that nearly two-fifths of all tokens are part of annotated expressions. In contrast, CNERgold consists of 2,000 manually annotated sentences containing around 56,843 tokens, with 14,730 spans and a slightly higher density of 39.56%.
Each instance is annotated at the token level with part-of-speech tags, binary labels indicating whether a token refers to a concept or a named entity, and a fine-grained BIO-encoded tag reflecting its semantic class. To ensure the integrity of the data, we conducted a systematic review to detect malformed BIO sequences, duplicates, and inconsistent label usage. We further examined the distribution of categories between training and test partitions to rule out significant discrepancies; tabulated results (Annex A) confirm that category frequencies are stable across splits.
For the purpose of our experiments, we used a representative subset comprising 10,000 sentences and

approximately 304,000 tokens, in which 40.4% of the tokens are part of a labeled span. This subset retains the diversity and label balance of the full dataset while remaining computationally tractable. The 122,947 annotated spans cover a wide range of categories, with particularly high frequencies observed for entities such as PER, MEDIA, and EVENT. This version of the dataset serves as the empirical basis for all evaluations conducted in this work.

4 Methodology

Given a sequence of input tokens, the model is tasked with predicting, for each token, a category label from a predefined set of 29 CNER classes, expressed in the BIO format.
We use BERT-base (Devlin et al., 2019) as our primary architecture. This model consists of a 12-layer Transformer encoder with hidden size 768 and 12 attention heads, pretrained using masked language modeling and next sentence prediction. On top of the final encoder output, we add a two-layer classification head consisting of two fully connected layers with GeLU activation and dropout, mirroring the architecture used in the CNER baseline by Martinelli et al. (2024).
While DeBERTa-v3 (He et al., 2023) achieves superior results on the CNER benchmark, we intentionally focus on BERT as a computationally lighter and widely adopted representative of the transformer family. Our goal is not to outperform the state of the art, but rather to provide a controlled comparison between standard contextual embeddings and knowledge-augmented representations. To that end, we replicate the same architecture and training setup using ERNIE 2.0 (Sun et al., 2019) in place of BERT, maintaining all other parameters fixed.
ERNIE (Enhanced Representation through kNowledge Integration) distinguishes itself from BERT and DeBERTa by explicitly incorporating structured knowledge into pretraining. While BERT relies solely on masked language modeling, ERNIE introduces additional objectives that expose the model to entity-level masking, semantic-level masking, and phrase-level masking, thereby encouraging it to learn deeper connections between tokens and real-world concepts. Entities in the input are linked to structured knowledge bases (e.g., Baidu Baike), allowing the model to model dependencies that go beyond sentence-level co-occurrence.
This distinction is central to our investigation: whereas BERT and DeBERTa are trained purely on linguistic co-occurrence, ERNIE is exposed to external semantic priors, which we hypothesize may be particularly beneficial for rare or semantically ambiguous CNER categories. By replacing BERT with ERNIE in an otherwise identical architecture, we isolate the effect of knowledge integration from other architectural factors.

In summary, our methodology follows a comparative design: we implement a shared prediction framework using transformer encoders and compare a context-only model (BERT) with a knowledge-augmented variant (ERNIE). This setup allows us to assess the potential benefits of external knowledge.

5 Experimental Setup

To evaluate the impact of external knowledge on unified concept and named entity recognition, we fine-tune transformer-based token classification models using the CNER benchmark dataset introduced by Martinelli et al. (2024). Our experiments are designed to compare a standard BERT model with its knowledge-augmented counterpart, ERNIE 2.0, under identical training and evaluation conditions. All experiments are conducted using the Hugging Face Transformers and Datasets libraries.
We use a stratified 10,000-sentence subset of CNERsilver as our training data (see Section 3), and evaluate performance on the original CNERgold validation and test sets. The dataset is tokenized using the BERT-base-cased tokenize. The same preprocessing pipeline is applied to both BERT and ERNIE models to ensure strict comparability.
We implement both models using the AutoModelForTokenClassification interface with identical architecture heads: a two-layer feedforward classification head with dropout ($p = 0.1$), matching the configuration used in the original CNER paper.
Training is performed using the following hyperparameters for both models:
Optimizer: AdamW (via Hugging Face default)
Learning rate: 2e-5

Batch size: 8 (for both training and evaluation)
Epochs: 1
Weight decay: 0.01
Performance is assessed using the seqeval library, which computes micro- and macro-level precision, recall, F1, and overall token-level accuracy.

6 Results and Analysis

6.1 Analysis

The CNER dataset used in this study comprises a total of 29 distinct categories, encompassing both named entities and abstract concepts. However, a clear imbalance emerges in the span distribution across these classes. Categories such as PER (18.6%), MEDIA (11.2%), and EVENT (11.1%) dominate the dataset, collectively accounting for over 40% of all annotated spans. These are followed by more structural or locative categories like LOC, MEASURE, and GROUP, each representing between 6% and 8% of spans. In contrast, semantically rich but underrepresented classes such as FEELING, MONEY, LAW, CULTURE, and ASSET each account for less than 0.3% of total spans. This long-tail distribution is typical of real-world NER datasets, but it poses a significant challenge for generalization—especially for models trained on smaller subsets like the one used here.

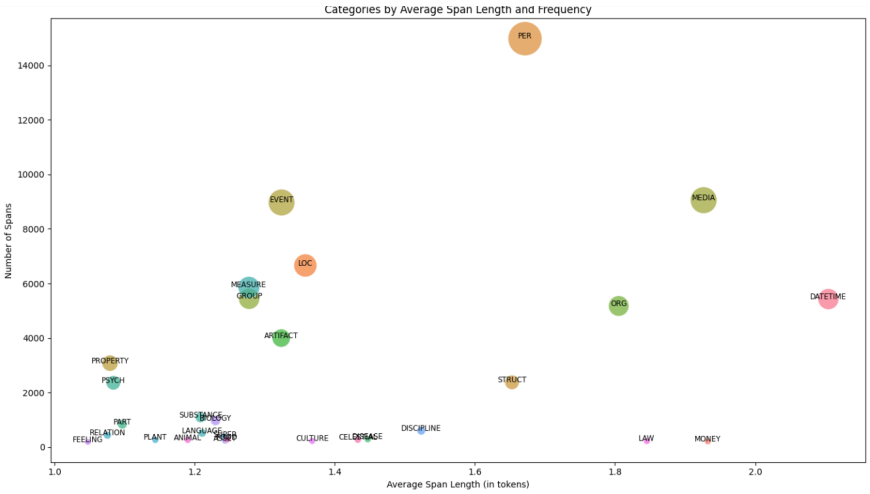

Figure 1: Span frequency vs. average span length for each CNER category. Bubble size reflects overall span count.

The situation is further complicated by high levels of inter-category ambiguity. Analysis of token-label mappings reveals that categories such as EVENT, MEDIA, PER, and ORG are frequently involved in ambiguous spans—where the same token appears under multiple labels. For instance, terms like "conference", "BBC", or "minister" may be tagged as EVENT, MEDIA, or PER depending on context.

These ambiguities are reflected in Figure 1, where PER, MEDIA, and EVENT not only dominate in volume but also exhibit longer average span lengths, indicating multi-token complexity and greater potential for confusion.

Moreover, the heatmap of category co-occurrence (Figure 3) confirms that certain categories frequently appear together in the same sentence, reinforcing contextual entanglement. For example, MEDIA co-occurs heavily with ORG and EVENT, and GROUP often appears alongside PROPERTY or FEELING. These observations suggest that distinguishing between conceptually related categories is not merely a matter of learning label boundaries, but also of resolving semantic overlap through context.

This directly aligns with the motivations of our work: to evaluate whether knowledge-enhanced models like ERNIE can better handle such ambiguity and low-frequency generalization compared to traditional architectures like BERT. Together, these findings highlight the dual challenge of semantic

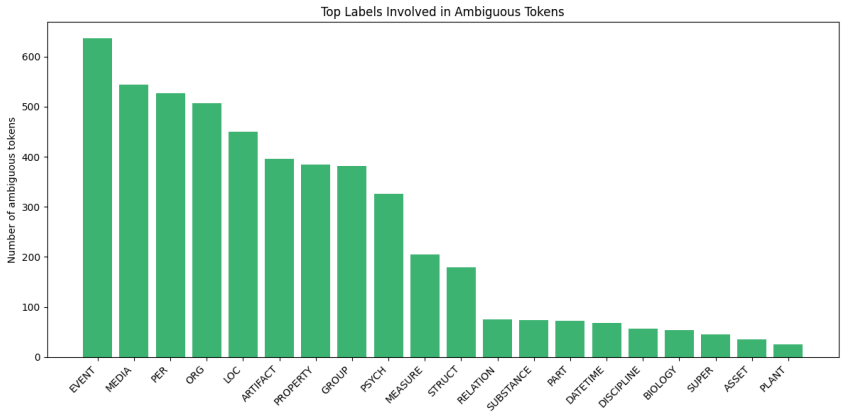

Figure 2: Number of ambiguous tokens involving each category (i.e., tokens associated with multiple labels across the dataset).

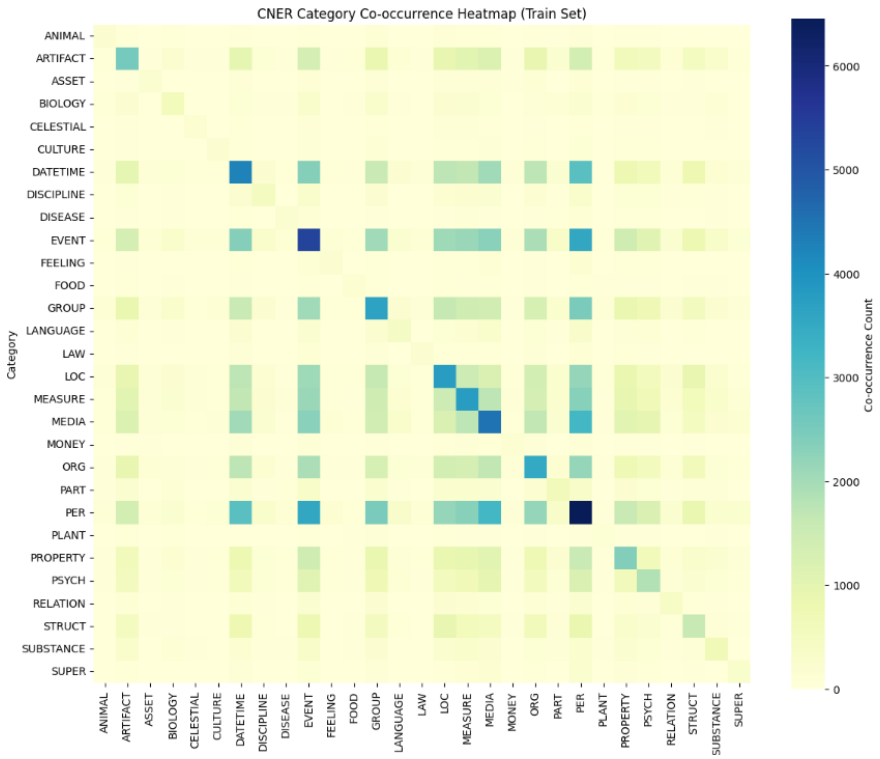

Figure 3: Heatmap showing co-occurrence counts between CNER categories (same sentence). Strong signals occur between related types such as `MEDIA-ORG`, `GROUP-FEELING`, and `DISCIPLINE-PROPERTY`.

complexity and class imbalance in the CNER task, reinforcing the relevance of models that can draw on structured external knowledge to resolve meaning in context.

6.2 Model results

Overall Performance:
The evaluation was carried out on the official `CNERgold` test set, using standard token-level metrics: precision, recall, F1-score, and accuracy, calculated with the `seqeval` library. The same evaluation protocol was applied to both models to ensure a fair comparison. Table 1 presents the results obtained on the validation and test sets.

Model	Precision	Recall	F1-score	Accuracy	Validation F1
BERT	0.5858	0.6310	0.6075	0.8576	0.5847
ERNIE	0.5963	0.6316	0.6134	0.8546	0.5847

Table 1: Evaluation results on the CNERgold test set. Both models were trained on a 10,000-sentence subset of the CNERsilver dataset.

The results show that the two models perform similarly overall, with slightly better scores for ERNIE 2.0 across all evaluation metrics. The difference in F1-score on the test set is small (about 0.6 points), but consistent. Precision, in particular, is slightly higher for ERNIE, which may indicate a better ability to avoid false positives. The accuracy values are close, although this metric is less informative due to the prevalence of non-entity tokens.
On the validation set, both models reached nearly the same F1-score, suggesting that the training dynamics were comparable and that the performance gap on the test set is likely due to differences in the pretraining phase rather than fine-tuning. The ERNIE model, which integrates structured knowledge during pretraining, appears to benefit slightly in terms of generalization.
It is worth noting that these results remain below the performance levels reported in the original CNER benchmark using larger models such as DeBERTa-v3. This is not surprising, as we limited training to a reduced subset of the dataset for computational reasons, and used lighter models. Despite this, the findings highlight the potential contribution of external knowledge in tasks involving diverse and less frequent categories.
The next section explores performance by category in order to better understand which types of entities and concepts benefit the most from knowledge integration.

Category-level Performance:
The sample includes several low-frequency categories such as `PLANT`, `PSYCH`, `LAW`, `STRUCT`, and `SUPER`, each appearing fewer than 50 times in the test set. These categories represent a significant challenge for both models. In our experiments, BERT and ERNIE failed to correctly predict any spans for some of these classes, including `ANIMAL`, `PLANT`, `ORG`, and `PSYCH`, resulting in F1-scores of zero. This confirms the sensitivity of token-level classifiers to class imbalance and the difficulty of generalizing from sparse supervision. Although ERNIE benefits from knowledge-enhanced pretraining, this advantage does not appear sufficient to address extreme low-resource scenarios. For instance, categories like `LAW`, `STRUCT`, and `SUPER` remained poorly predicted by both models. A slight improvement was observed for `MEDIA`, where ERNIE achieved a small gain in F1 (0.12 compared to 0.00 for BERT), suggesting a limited benefit from semantic priors. In fact, MEDIA is a category sharing a high cooccurence with many other groups
These findings point to a structural limitation of current transformer-based approaches: without sufficient labeled data, even knowledge-augmented models struggle to make accurate predictions. Addressing this limitation may require complementary strategies, such as data augmentation, few-shot learning frameworks, or the incorporation of category-level priors during training or inference.
In summary, rare categories remain a clear weakness for both BERT and ERNIE in the CNER task, and tackling this issue will be critical for improving semantic coverage in real-world information extraction.

Conclusion

This study investigates the added value of knowledge-enhanced transformers in the context of Concept and Named Entity Recognition (CNER), a challenging task that combines structured and unstructured semantic classes.
Our study presents several limitations that should be acknowledged. First, due to computational constraints, we relied on a reduced subset of the original CNERsilver dataset for training. While this subset preserved the general distribution of span categories, it inevitably limited the diversity and quantity of training instances, particularly for long-tail classes such as `LAW`, `SUPER`, or `PLANT`. Consequently, performance metrics may underestimate the full potential of both BERT and ERNIE under large-scale pretraining and fine-tuning conditions. Second, we conducted all experiments using only one architecture: BERT as the context-only baseline, and ERNIE 2.0 as the knowledge-enhanced model. Although both are widely used and well-documented, this choice does not cover other families of knowledge-augmented models such as KnowBERT, K-Adapter, or retrieval-augmented

transformers, which may behave differently on CNER tasks. Third, our evaluation focuses exclusively on standard classification metrics (precision, recall, F1), without investigating model calibration, interpretability, or span boundary errors. These aspects are particularly relevant in real-world applications where understanding errors and confidence is essential. Finally, while our work highlights the benefits of knowledge integration in improving generalization to rare or ambiguous categories, we do not directly control for other influencing factors such as annotation inconsistency which can affect model performance.

Our results show that ERNIE 2.0, a model enriched with external knowledge, consistently outperforms BERT across global metrics and on certain semantically complex categories. While both models struggle with rare labels, ERNIE demonstrates a modest but meaningful advantage in handling ambiguous and abstract spans such as `MEDIA` or `FOOD`. These findings support the idea that integrating structured knowledge during pretraining can enhance generalization in noisy or low-resource contexts. Future work could extend this comparison to models such as KnowBERT or K-Adapter and explore multilingual extensions of the CNER task. Additionally, incorporating prompt-based or few-shot techniques may provide better solutions for long-tail and conceptually difficult labels.

References

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.
DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ICLR 2021.

[3] Yinhan Liu et al.
RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.

[4] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, Haifeng Wang.
ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. AAAI 2020.

[5] Yu Sun, Shuohuan Wang, Yukun Li et al.
ERNIE: Enhanced Representation through Knowledge Integration. arXiv preprint arXiv:1904.09223, 2019.

[6] Matthew E. Peters et al.
Knowledge Enhanced Contextual Word Representations. EMNLP 2019.

[7] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, Ping Wang.
K-BERT: Enabling Language Representation with Knowledge Graph. AAAI 2020.

[8] Martinelli Bellomaria, Francesco Santini, et al.
Concept and Named Entity Recognition with Multilingual Transformers. Findings of ACL 2022.