Meta-Pretraining for Zero-Shot Cross-Lingual Named Entity Recognition

David Demitri Africa*, Suchir Salhan, Yuval Weiss, Paula Buttery, Richard Diehl Martinez
University of Cambridge

Introduction

Named-entity recognition (NER) identifies people, organisations, and locations in text. While high-resource languages dominate current benchmarks, Philippine languages like **Tagalog** and **Cebuano** remain under-served.

We test whether **meta-pretraining** (specifically Model-Agnostic Meta-Learning (MAML)) can improve zero-shot transfer to these languages using small decoder-only Transformers.

- **Goal:** Adapt small models efficiently to unseen languages.
- **Approach:** Replace part of standard pretraining with episodic MAML.
- **Hypothesis:** Meta-pretraining induces faster, more generalizable linguistic representations.

Typological Contrast

Feature	Tagalog	Cebuano
Voice System	4-way	2-way
Case Marking	Obligatory (*si*, *ni*)	Often dropped
Borrowing	High English mix	Conservative
Word Order	Flexible	Freer
Reduplication	Common	Widespread
Orthography	Mixed	More uniform

Contributions

- Introduce a **meta-pretrained small decoder LM** for low-resource NER.
- Empirically compare MAML vs. vanilla pretraining across four model sizes (11–570M).
- Show that **zero-shot F1 gains (1–6 pp)** are strongest for single-token person entities.
- Analyze representational effects and convergence behavior.

Method

Model: Pico decoder (LLaMA-style) with 12 layers, grouped-query attention, SwiGLU feed-forwards.

Training objective:

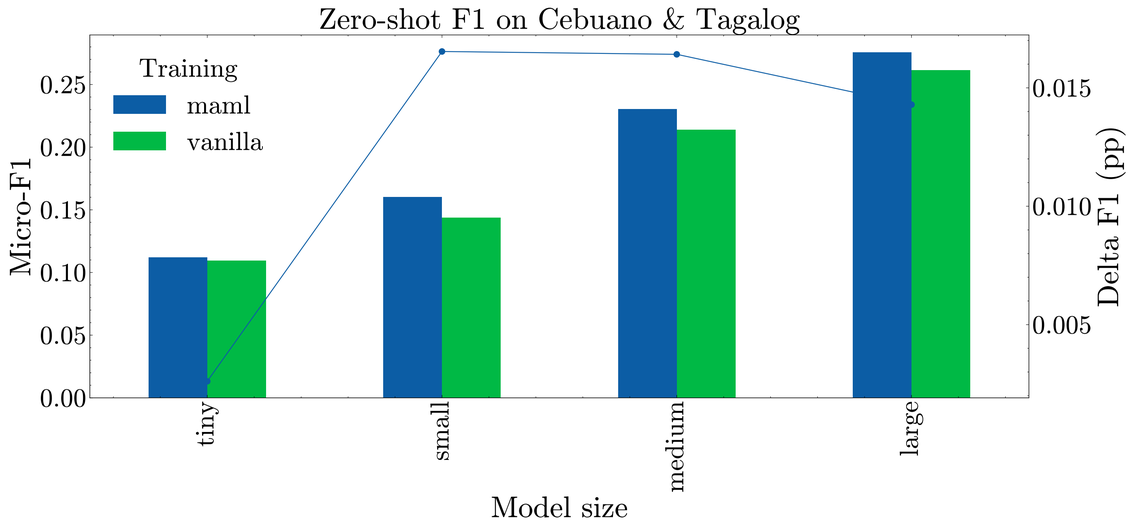
1. **Autoregressive LM step:** Next-token prediction on Dolma corpus.
2. **MAML episode:** 32-way, 4-shot masked-token prediction with 10 inner SGD steps.

Optimizer: AdamW (3×10^{-4}), 6k steps, $\rho = 0.5$ mix of MAML vs. LM updates.

Finetuning:

- CRF head for NER.
- Finetuned on 9 high-resource languages (EN, HR, SK, PT, etc.).
- Evaluated zero-shot on Tagalog and Cebuano (Universal NER).

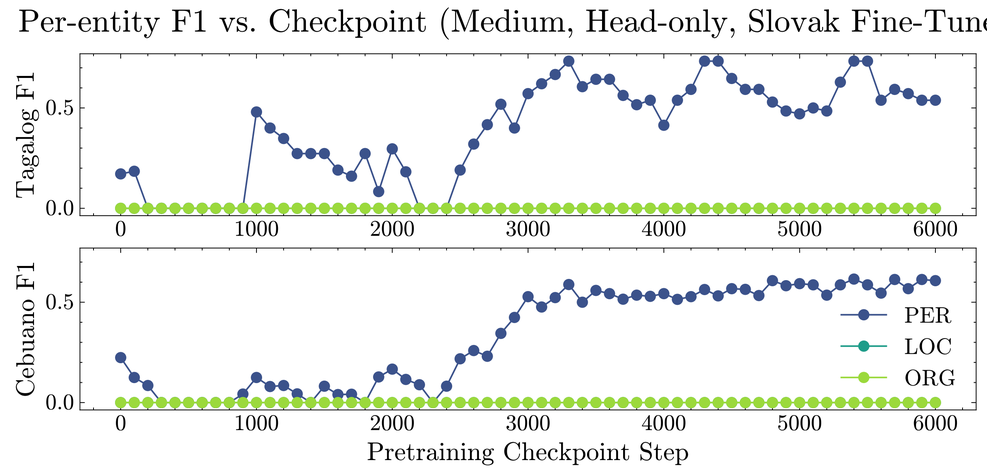
Results: Zero-Shot Transfer



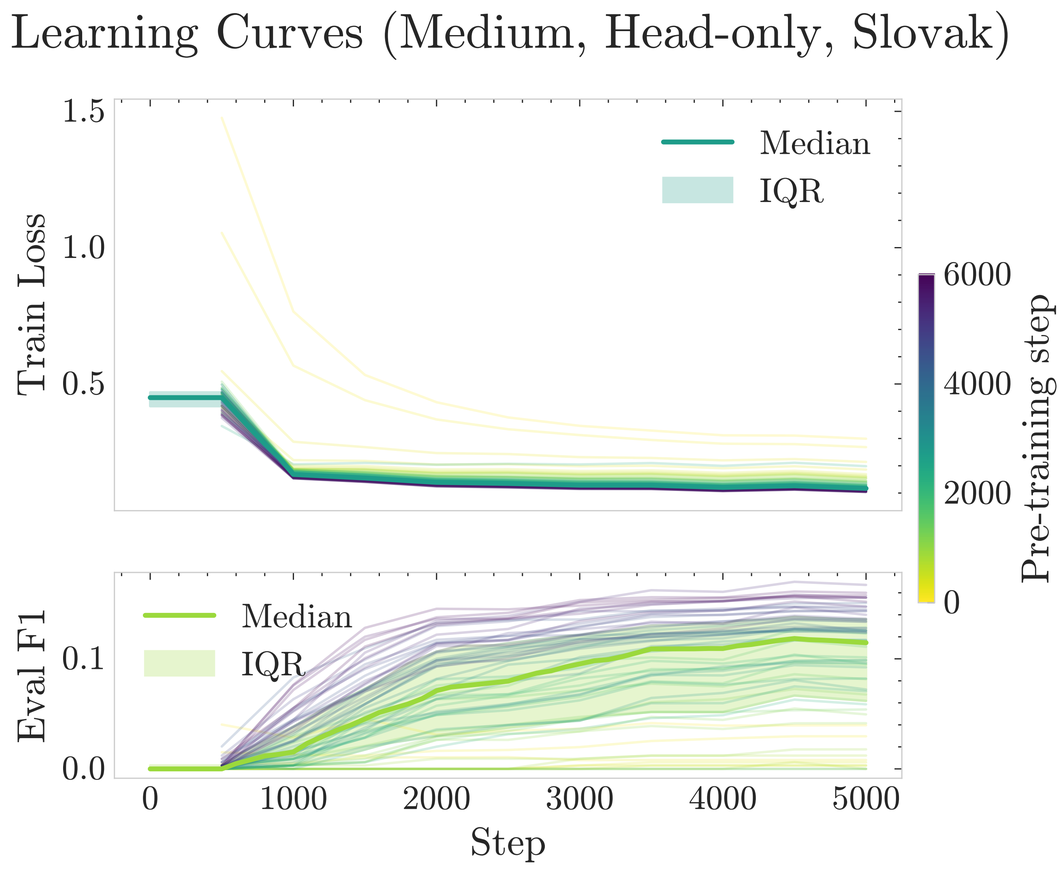
Meta-pretraining improves NER across all model scales.
Gains shrink with model size, suggesting diminishing returns beyond 500M params.

Analysis: What Transfers?

Entity-specific gains: MAML improves most on PER tags— (ingle-token person entities) while LOC and ORG remain flat.
Interpretation: Meta-learning sharpens lexical prototypes, especially for names anchored by Tagalog case particles (*si*, *ni*).



Convergence Speed



MAML accelerates convergence by up to 8 %, reorganizing the loss landscape for faster adaptation.

Conclusions

Meta-pretraining enables efficient transfer in small models. By shaping the initialization for rapid adaptation, even modest architectures can generalize across typologically diverse languages.

Key Takeaways:

- Zero-shot F1 ↑ 1–6 pp.
- Largest gains in low-capacity and head-only regimes.
- Improves transfer to Austronesian languages unseen in pretraining.

Limitations: Effects diminish with larger models, weak on multi-token entities.

Future Work: Explore alternative meta-objectives, expand to more low-resource languages, and analyze prototype dynamics.

Acknowledgments

Supported by the Accelerate Programme for Scientific Discovery, Cambridge Trust, Jardine Foundation, CUPA, and Gates Cambridge. Computing resources provided by CSD3 and DiRAC (EPSRC EP/T022159/1).

david.demitri.africa@gmail.com
github.com/DavidDemitriAfrica/pico-maml-train
huggingface.co/davidafrica

UNIVERSITY OF CAMBRIDGE