Tagalog now has a **dedicated NLP toolkit** that provides **out-of-the-box** and **high-performance** support for dependency parsing, parts-of-speech tagging, and named entity recognition.

calamanCy: A Tagalog Natural Language Processing Toolkit

Lester James V. Miranda

Introduction

Tagalog is a **low-resource** language from the Austronesian family with over 28 million speakers in the Philippines. However, it still lacks adequate NLP resources.

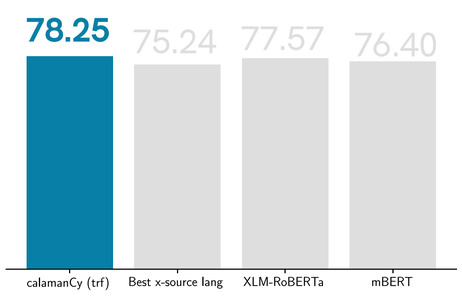
Methodology

We trained calamanCy using **gold-standard data** such as Tagalog Universal Dependencies (UD) and TLUnified-NER. We built **three pipelines** based on word embeddings (2x) and transformers (1x).

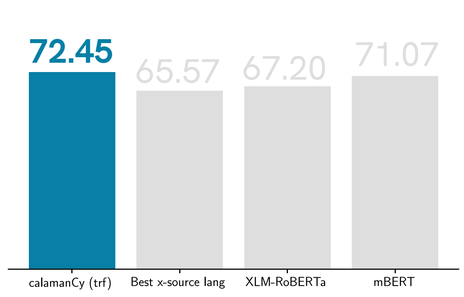
Results

Our transformer-based calamanCy pipeline **outperformed cross-lingual and multilingual transfer learning techniques** on different tasks:

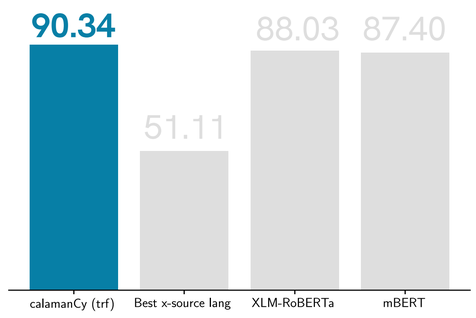
Hatespeech, *binary textcat*
(Cabasag et al., 2019)



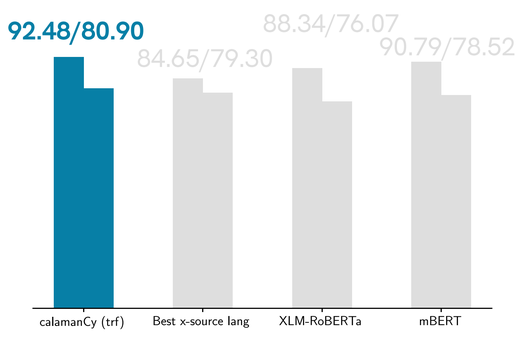
Dengue, *multilabel textcat*
(Livelo and Cheng, 2018)



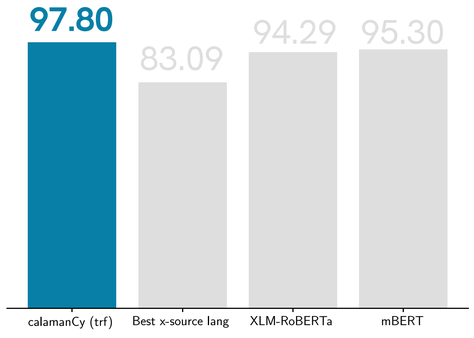
TLUnified-NER, *NER*
(Miranda, 2023)



Merged UD, *Dep. pars. (UAS/LAS)*
(Samson, 2018; Aquino and de Leon, 2020)



Merged UD, *POS tagging*
(Samson, 2018; Aquino and de Leon, 2020)



Notes

- Unless otherwise stated, reported results are F1-scores on the test set (avg. of three trials).
- "Best x-source lang" is the score of the best-performing source language for cross-lingual transfer (full results in the paper).

⚙️ Installation

```
pip install calamancy
```

👨‍💻 Sample usage

```
import calamancy as cl
nlp = cl.load("tl_calamancy_md")
# John went to Japan
doc = nlp("Pumunta si Juan sa Japan")
```

📦 Available pipelines

tl_calamancy_md

- Medium-sized pipeline
- Uses floret vectors trained on the TLUnified corpora
- 50k unique vectors (200 dims), 77 MB

tl_calamancy_lg

- Large-sized pipeline
- Uses fastText vectors trained on the CommonCrawl corpora
- 714k unique vectors (300 dims), 455 MB

tl_calamancy_trf

- Transformer-based pipeline
- Context-sensitive vectors from a transformer network.
- Finetuned on top of RoBERTa tagalog, 813 MB

EXPLOSION

ljvmiranda@gmail.com

github.com/ljvmiranda921

Portions of this work were done while the author, now affiliated with AI2, was employed at ExplosionAI GmbH

Take a picture to **download the full paper**