Developing a Named Entity Recognition Dataset for Tagalog

Lester James V. Miranda
ljvmiranda@gmail.com

Abstract

We present the development of a Named Entity Recognition (NER) dataset for Tagalog. This corpus helps fill the resource gap present in Philippine languages today, where NER resources are scarce. The texts were obtained from a pretraining corpora containing news reports, and were labeled by native speakers in an iterative fashion. The resulting dataset contains ~7.8k documents across three entity types: Person, Organization, and Location. The inter-annotator agreement, as measured by Cohen's κ, is 0.81. We also conducted extensive empirical evaluation of state-of-the-art methods across supervised and transfer learning settings. Finally, we released the data and processing code publicly to inspire future work on Tagalog NLP.

1 Introduction

Tagalog (`tl`) is one of the major languages in the Philippines with over 28 million speakers in the country (Lewis, 2009). It constitutes the bulk of Filipino, the country's official language, by sharing its lexical items and grammatical structure. Despite this fact, there are little to no resources for Tagalog (Cruz and Cheng, 2022), hampering the development of reliable language technologies.

In this paper, we present TLUNIFIED-NER,[1] a Tagalog dataset for Named Entity Recognition (NER). The texts were obtained from TLUnified (Cruz and Cheng, 2022), a pretraining corpora containing news reports and other types of text. We focused on NER because of its foundational role in several NLP tasks (Tjong Kim Sang and De Meulder, 2003; Lample et al., 2016), especially in problems that require the extraction of structured information. TLUNIFIED-NER consists of ~7.8k documents across three entity types (*Person*, *Organization*, *Location*), modeled closely to

the CoNLL Shared Tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). Three native speakers conducted the annotation process, resulting to an inter-annotator agreement (IAA) score of 0.81.

We hope that TLUNIFIED-NER will allow researchers to build better NER classifiers for Tagalog, and thereby inspire future research on Tagalog NLP through the following contributions:

1. We curated and annotated texts from a large pretraining corpora to represent the modern usage of Tagalog in the news domain.

2. We provided performance baselines across a variety of supervised and transfer learning settings.

2 Related Work

Tagalog language Tagalog is an agglutinative language within the Austronesian family (Kroeger, 1992). It uses the Latin script for its writing system with 28 letters in its alphabet. Twenty-six letters are the same as in English, with the addition of Ñ/ñ and Ng/ng. Tagalog typically follows the VSO word order, but VOS and SVO are also accepted (Schachter and Otanes, 1973). Although Filipino is the country's official language, it has little to no linguistic differences with Tagalog.

Tagalog NER datasets Unfortunately, resources for Tagalog NER are meager. One major resource is WikiANN (Pan et al., 2017), a silver-standard corpora based on a framework designed for 282 other languages. However, the Tagalog portion of WikiANN is full of annotation errors, often misconstruing one entity type as another. Another NER dataset is the Filipino Storytelling corpora (Costiniano et al., 2022). Although gold-standard, its entity labels (e.g., *Humans & Body*, *Natural Environment*, etc.) are too domain-specific for general use. Finally, the LORELEI project also provides

---

[1]The dataset is accessible at https://huggingface.co/datasets/ljvmiranda921/tlunified-ner

Entity	Short Description	Examples
Person (PER)	Person entities limited to humans. It may be a single individual or group.	Juan de la Cruz, Jose Rizal, Quijano de Manila
Organization (ORG)	Organization entities limited to corporations, agencies, and other groups of people defined by an organizational structure.	Meralco, DPWH, United Nations
Location (LOC)	Location entities are geographical regions, areas, and landmasses. Geo-political entities are also included within this group.	Pilipinas, Manila, CALABARZON, Ilog Pasig

Table 1: Entity types used for annotating TLUNIFIED-NER (derived from the TLUnified pretraining corpus of Cruz and Cheng, 2022).

language packs for Tagalog (Strassel and Tracey, 2016), but they're not publicly-accessible.

TLUNIFIED-NER aims to fill this resource gap by providing a publicly-assessible gold standard resource for Tagalog NER.

3 Dataset Collection

The texts were obtained from Cruz and Cheng (2022)'s TLUnified pretraining corpora. It combines news reports (Cruz et al., 2020), a preprocessed version of CommonCrawl (Suarez et al., 2019), and several other datasets. We manually filtered this dataset to contain news reports so as to resemble the CoNLL Shared Tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

The texts are diverse. It contains articles from different news sites online that ran a published print media or news channel in Metro Manila from 2009 to early 2020. The topics range from politics, weather, and popular science among others.

4 Annotation Setup

We used Prodigy as our annotation tool.[2] We set up a web server on the Google Cloud Platform and routed the examples through Prodigy's built-in task router. Figure 1 shows the labeling interface as seen by the annotator. Finally, we used the `ner.manual` recipe to highlight spans during the annotation process. We used three entity labels for TLUNIFIED-NER as shown in Table 1. Unlike CoNLL, we decided to exclude the *Miscellaneous (MISC)* tag to reduce confusion.

Annotation Process The annotation process was done iteratively with three annotators (including the author) who are native Tagalog speakers. Given
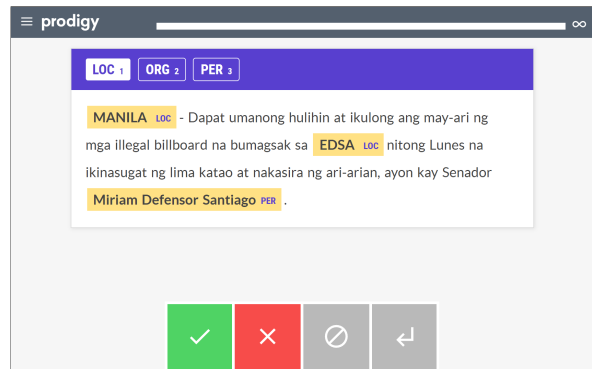

Figure 1: Prodigy's annotation interface for a given text. (Translation: *MANILA - The owner of the illegal billboards that fell on EDSA this Monday, injuring five people and damaging property, should be caught and imprisoned according to Senator Miriam Defensor Santiago.*)

a set annotation budget, we paid the annotators above the country's minimum daily wage. Each annotation round spans for two to three weeks, for a total of six rounds (18 weeks). The annotators labeled the same batch of examples to ensure high overlap.

After each round, the annotators hold a retrospective meeting and discussed examples they found confusing, inconsistent with the annotation guidelines, and noteworthy. This process continued until we reached ∼10k examples or if we exhausted our annotation budget. In addition, we also tracked the training curve to determine the quality of the collected annotations. If the F1-score improved within the last 25% of the training data, then it is a good sign that obtaining more labels will result to better accuracy.

Annotation Guidelines We developed the annotation guidelines in an iterative fashion.

---

[2]https://prodigy.ai

14

Dataset	Examples	Tokens	PER	ORG	LOC	Length	SD	BD
Training	6252	198579	6418	3121	3296	1.49	2.66	1.26
Development	782	25069	793	392	409	1.51	2.77	1.37
Test	782	25100	818	423	438	1.48	2.77	1.34

Table 2: Dataset statistics for TLUNIFIED-NER. It shows the number of examples, number of tokens, and span-level statistics. SD stands for span distinctiveness whereas BD is boundary distinctiveness (Papay et al., 2020).

Metric	IAA
Cohen's κ on all tokens	0.81
Cohen's κ on annotated tokens only	0.65
F1 score	0.91

Table 3: Inter-annotator agreement (IAA) measurements. We obtained these values by computing for the pairwise comparisons on all annotator-pairs and averaging the results.

The Automatic Content Extraction (ACE 2004/05) annotation document (Doddington et al., 2004) heavily inspired our initial draft. We co-developed the guidelines after each annotation round to improve clarity and reduce disagreements. These guidelines are accessible on GitHub: `https://github.com/ljvmiranda921/calamanCy/tree/master/datasets/tl_calamancy_gold_corpus/guidelines`

5 Corpus Statistics and Evaluation

Table 2 shows the final dataset statistics for TLUNIFIED-NER. We also included span- (SD) and boundary-distinctiveness (BD) metrics (Papay et al., 2020). They measure the KL-divergence of the unigram word distributions between the span (or its boundaries) and the rest of the corpora. These metrics can be used to gauge the difficulty of the span labeling task, (e.g., more distinct spans means it's "easier" to detect them in the text).

5.1 Inter-annotator Agreement (IAA)

Similar to Brandsen et al. (2020), we measured two types of Cohen's κ. The first metric calculates κ for tokens where at least one annotator has made an annotation. The second metric computes for all tokens while ignoring the 'O' label. In addition, we had a third measure: the F1-score using one set of annotations as reference (Deleger et al., 2012). We did these computations for each annotator-pair and averaged the results as shown in Table 3.
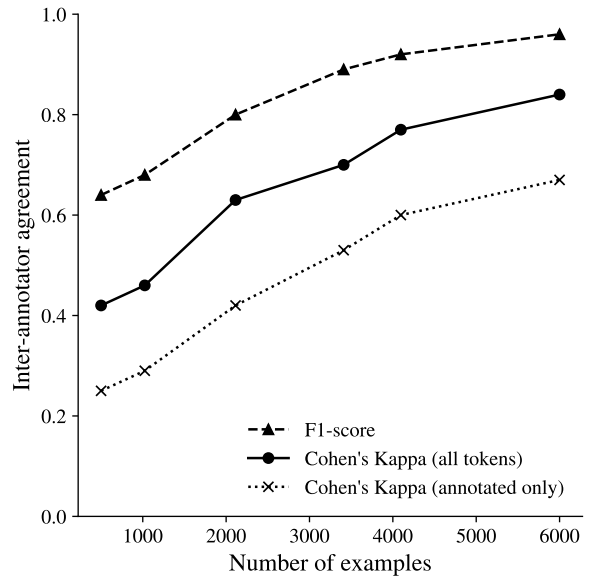
Finally, Figure 2 shows the growth of IAA for

Figure 2: Growth of IAA for each annotation round.

each annotation round. Because of our annotation process, we were able to label the same batch of documents and track the agreement every round.

5.2 Benchmark results

We trained several NER models using spaCy's transition-based parser (Honnibal et al., 2020). The state transitions are based on the BILUO sequence encoding scheme and the actions are decided by a convolutional neural network with a maxout (Goodfellow et al., 2013) activation function.

While keeping the NER classifier constant, we experimented with various word embeddings that led to the following configurations:

- **Baseline:** we trained the transition-based parser "from scratch" without additional information from static or context-sensitive vectors.

- **Static vectors:** we used Tagalog fastText vectors (Bojanowski et al., 2017) and included a simple pretraining process to initialize the

Word Embeddings	Person	Organization	Location	Overall
Baseline (no additional embeddings)	87.85±0.01	74.80±0.02	81.03±0.01	84.57±0.02
fastText (Bojanowski et al., 2017)	91.20±0.02	85.39±0.03	88.38±0.01	88.90±0.01
RoBERTa Tagalog (Cruz and Cheng, 2022)	92.18±0.01	87.30±0.00	90.01±0.02	90.34±0.02
XLM-RoBERTa (Conneau et al., 2020)	91.95±0.04	84.84±0.02	88.92±0.01	88.03±0.03
Multilingual BERT (Devlin et al., 2019)	90.78±0.03	85.08±0.01	88.45±0.03	87.40±0.02

Table 4: Benchmark results on TLUNIFIED-NER across different word embeddings using spaCy's transition-based parser (Honnibal et al., 2020). Reported results are F1-scores on the test set across three trials.

Figure 3: Development set confusion matrix of the Baseline model predictions in the IOB format.

Embeddings set-up	Rel. error reduction	
	ORG	LOC
Shared	+5%	+3%
Context-sensitive	+12%	+18%

Table 5: Relative error reduction (with respect to the Baseline) for classifying ORG and LOC entities. Reported results are F1-scores on the development set.

5.3 Error analysis

From our benchmark results, we noticed that most models are having trouble predicting the *Location* or *Organization* tags. Figure 3 shows the confusion matrix of the Baseline model on the development set in the IOB format.

Most of the mistakes came from incorrectly tagging a token with the outside 'O' label. However, we also noticed instances where the model confuses between the lexical and semantic tag of an entity. For example, in the span, *"...panukala ng Ombudsman..."* ("...proposed by the Ombudsman..."), the token Ombudsman might be a Person or Organization depending on the context. We hypothesize that including context-sensitive training, which the baseline model lacks, can help mitigate this issue.

To test this hypothesis, we experimented on two training configurations. First, we trained a POS tagger together with our transition-based NER with shared weights. This process may help provide extra information to the transition-based parser so it can disambiguate between entities. Second, we finetuned context-sensitive vectors from RoBERTa Tagalog (Cruz and Cheng, 2022) for NER. Table 5 shows the relative error reduction between LOC and ORG entities. Given these results, we encourage researchers to utilize context-sensitive vectors such as RoBERTa Tagalog (or other BERT variants) when training models from this corpora.

weights of the model. The pretraining objective asks the model to predict some number of leading and trailing UTF-8 bytes for the words—a variant of the cloze task.

- **Transformer-based vectors (monolingual):** we used RoBERTa Tagalog (Cruz and Cheng, 2022), the only pretrained language model for Tagalog, and finetuned it with our annotations.

- **Transformer-based vectors (multilingual):** we tested on XLM-RoBERTa (Conneau et al., 2020) and multilingual BERT (Devlin et al., 2019) for transfer learning. These models include Tagalog in their training pool albeit underrepresented.

This experimental setup allows us to see the expected performance when training Tagalog NER classifiers using standard techniques. Table 4 reports the F1-score on the test set across three trials.

Model	Training dataset	
	WikiANN	TLUnified-NER
Baseline (no additional embeddings)	19.92±0.03	**30.24±0.02**
fastText (Bojanowski et al., 2017)	24.41±0.01	**45.09±0.02**
RoBERTa Tagalog (Cruz and Cheng, 2022)	23.38±0.02	**58.90±0.03**
XLM-RoBERTa (Conneau et al., 2020)	31.28±0.01	**57.67±0.01**
Multilingual BERT (Devlin et al., 2019)	29.20±0.03	**59.26±0.03**

Table 6: Cross-dataset comparison between WikiANN (Pan et al., 2017) and TLUnified-NER. We trained a model from WikiANN then applied it to TLUnified-NER (and vice-versa). Reported results are F1-scores on the test set across three trials.

Entity label	F1-score
Person (PER)	67.95
Organization (ORG)	00.59
Location (LOC)	35.17

Table 7: Comparing the overlap between the original (silver-standard) WikiANN annotations against our reannotated version.

Model	F1-score
GPT-4 (OpenAI, 2023)	65.89±0.44
GPT-3.5-turbo	53.05±0.42
Claude v1 (Anthropic, 2023)	58.88±0.03
Command (Cohere, 2023)	25.48±0.11
Dolly v2* (Conover et al., 2023)	13.07±0.14
Falcon* (Almazrouei et al., 2023)	8.65±0.04
StableLM v2* (Stability-AI, 2023)	0.25±0.03
OpenLLaMa* (Geng and Liu, 2023)	15.09±0.48

Table 8: Benchmark results on TLUnified-NER across a variety of open-source and commercial LLMs. We used the 7B-parameter variants for models denoted with an asterisk (∗) due to budget constraints.

5.4 Comparison to WikiANN

The WikiANN dataset (Pan et al., 2017) is another resource for Tagalog NER. However, we found many annotation errors in the dataset, from misclassifications to fragmented sentences. We investigated how TLUnified-NER fares against WikiANN's silver-standard annotations.

We finetuned several models similar to Section 5.2 on the Tagalog portion of WikiANN's training set and tested it on TLUnified-NER's test set (and vice-versa). In order to properly evaluate on WikiANN, we reannotated the test dataset using the same annotation guidelines described in Section 4.

Our results in Table 6 suggest that models built from the TLUnified-NER corpus are more performant than with WikiANN. Additionally, the gap between WikiANN's silver-standard annotations and our corrections is large, as shown in Table 7. We then posit that the gold-standard nature of TLUnified-NER led to better performance than WikiANN, which predominantly consists of text fragments and low-quality annotations.

5.5 Experiments on large language models

Large language models (LLMs) have been shown to exhibit multilingual capabilities—incidental or not (Briakou et al., 2023). We investigated this property by performing a zero-shot prompting approach on TLUnified-NER's test set across a variety of commercial and open-source LLMs. Table 8 reports the F1-score across three trials.

Our results suggest that supervised learning reliably outperforms zero-shot prompting for TLUnified-NER given our prompt (see Appendix A.1). However, we acknowledge that these results are not a definitive comparison between two methods as prompt engineering is unstable with high variance (Webson and Pavlick, 2022; Zhao et al., 2021). In the future, we plan to explore different prompting techniques such as PromptNER (Ashok and Lipton, 2023) and chain-of-thought (Wei et al., 2023) to uncover the language models' full capabilities.

6 Conclusion

In this paper, we introduced TLUnified-NER, a Named Entity Recognition dataset for Tagalog. Unlike other Tagalog NER datasets, TLUnified-NER is publicly-accessible and gold standard. Our iterative annotation process, together with our inter-

annotator agreement, shows that the corpus is of high quality. In addition, our benchmarking results suggest that the task is learnable even with a simple baseline method. We hope that TLUNIFIED-NER fills the resource gap present in Tagalog NLP today. In the future, we plan to create a more fine-grained (and perhaps, overlapping) NER tag set similar to the ACE project and expand on other major Philippine languages. Finally, the dataset is available online (https://huggingface.co/datasets/ljvmiranda921/tlunified-ner) and we encourage researchers to improve upon our benchmark results.

Limitations

The TLUNIFIED-NER corpora is comprised mostly by news reports. Although the texts demonstrate the standard usage of Tagalog, its domain is limited. In addition, we only trained a transition-based parser model for our NER classifier. In the future, we plan to extend these benchmarks and include CRFs or other tools such as Stanford Stanza.

Acknowledgements

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Anthropic. 2023. Model card and evaluations for claude models.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Cohere. 2023. Command Model: The AI-Powered Solution for the Enterprise.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Sherwyne Costiniano, Rose Ann Mae Santos, Julius Simon Mendoza, and Allen Jay Gale. 2022. Custom Coarse Grained Named Entity Recognition for Filipino Storytelling Data Using Uncased Transformer Models. *SSRN Electronic Journal*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving Large-scale Language Models and Resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Ko Cheng. 2020. Exploiting News Article Structure for Automatic Corpus Generation of Entailment Datasets. In *Pacific Rim International Conference on Artificial Intelligence*.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, pages 144–53.

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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinyang Geng and Hao Liu. 2023. OpenLLaMA: An Open Reproduction of LLaMA.

Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1319–1327, Atlanta, Georgia, USA. PMLR.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Paul R. Kroeger. 1992. Phrase Structure and Grammatical Relations in Tagalog.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Paul M. A. Lewis. 2009. Ethnologue: languages of the world. https://ethnologue.com/language/tgl. Accessed: June 2023.

OpenAI. 2023. GPT-4 Technical Report.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Sean Papay, Roman Klinger, and Sebastian Padó. 2020. Dissecting span identification tasks with performance prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online. Association for Computational Linguistics.

Paul Schachter and Fe T. Otanes. 1973. Tagalog reference grammar. *The Journal of Asian Studies*, 32:760 – 761.

Stability-AI. 2023. StableLM-Alpha v2.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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*, pages 142–147.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Appendix

A.1 Zero-shot prompt template

You are an expert Named Entity Recognition (NER) system. Your task is to accept Text as input and extract named entities for the set of predefined entity labels. From the Text input provided, extract named entities for each label in the following format:

- *PER: <comma delimited list of strings>*
- *ORG: <comma delimited list of strings>*
- *LOC: <comma delimited list of strings>*

Below are definitions of each label to help aid you in what kinds of named entities to extract for

each label. Assume these definitions are written by an expert and follow them closely.

- *PER: PERSON*
- *ORG: ORGANIZATION*
- *LOC: LOCATION OR GEOPOLITICAL EN-TITY*

Text: {{ text }}

A.2 Reproducibility

All the experiments and models in this paper are available publicly. Readers can head over to https://github.com/ljvmiranda921/calamanCy/reports/aacl for all related code and assets. Note that the XLM-RoBERTa and multilingual BERT experiments may at least require a T4 or V100 GPU.