Constructing Indonesian-English Travelogue Dataset

Eunike Andriani Kardinata[1], **Hiroki Ouchi**[1,2], **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology [2] RIKEN
{eunike.kardinata.ef9, hiroki.ouchi, taro}@is.naist.jp

Abstract

Research in low-resource language is often hampered due to the under-representation of how the language is being used in reality. This is particularly true for Indonesian language because there is a limited variety of textual datasets, and majority were acquired from official sources with formal writing style. All the more for the task of geoparsing, which could be implemented for navigation and travel planning applications, such datasets are rare, even in the high-resource languages, such as English. Being aware of the need for a new resource in both languages for this specific task, we constructed a new dataset comprising both Indonesian and English from personal travelogue articles. Our dataset consists of 88 articles, exactly half of them written in each language. We covered both named and nominal expressions of four entity types related to travel: location, facility, transportation, and line. We also conducted experiments by training classifiers to recognise named entities and their nominal expressions. The results of our experiments showed a promising future use of our dataset as we obtained F1-score above 0.9 for both languages.

Keywords: Corpus (Creation, Annotation, etc.), Less-Resourced/Endangered Languages, Multilinguality

1. Introduction

As a low-resource language, Indonesian has an increasing number of speakers and potential developments. However, research in Indonesian language often face challenges, such as difficulty in collecting standardised dataset for specific task, which is causing the issues in the reproducibility of past research. To encourage more research in Indonesian by providing publicly available language resources, we constructed a new dataset which is more representative of how Indonesian language is being used in reality.

Currently, improving the accessibility of Indonesian language resources is crucial to support various demands in Indonesia. In particular, we focus on a geoparsing task among others that deal with entities of location. The COVID-19 outbreak has drawn more attention to the dynamics between tourists and major destinations, such as Indonesia. Texts are valuable resources to analyse these dynamics as they contain information about human behaviours, experiences, and reputations of tourist spots. Such information is essential for the local government to manage and promote the country.

Considering the challenges in geoparsing, such as ambiguous entity types due to common names (e.g., whether the word 'Soetomo' refers to a road, a hospital, or other entities), we designed an annotation scheme which covers not only named expressions, but also nominal expressions. For instance, the geographic entity 'Soetomo Hospital' is sometimes referred by nominal expressions, such as 'the hospital' and 'this building'. By recognising the nominal expressions, end-user applications based on geoparsing would be more accurate in disambiguating entities mentioned.

In this work, we present an Indonesian-English comparable (having almost the same content and similar mentions of entities) travelogue dataset. We covered English articles to provide a more diverse dataset and to improve language technologies for other languages. Figure 1 shows an example of annotated texts in our dataset. Our dataset includes two main characteristics: (i) Indonesian-English comparable contents[1] and (ii) annotations of geographic expressions[2]. In particular, we annotated not only named expressions (e.g., 'Daya Station' and 'Tana Toraja'), but also nominal expressions (e.g., 'bus' and 'route'). This point distinguishes our dataset from typical datasets of named entities.

In the following, we would further elaborate on the construction of our dataset and the subsequent evaluations. We conducted experiments on our dataset to clarify the performance level of current entity analysis systems. More specifically, we trained classifiers on our dataset to recognise both named entities and nominal expressions. The results showed a promising future use of our dataset as we obtained F1-score above 0.9 for both languages. Other potential utilisations of our dataset are for comparison analysis and transfer learning, where we attempt to leverage a model trained on high-resource languages to handle low-resource languages.

We will release our annotated dataset and experimental codes at `https://github.com/naist-nlp/mtd-gem`.

[1]Our dataset is not a parallel corpus because some phrases and sentences cannot be aligned between Indonesian and English travelogues.

[2]This is the first step towards geoparsing where we covered the recognition of geographic entity mentions.

LREC-COLING 2024, pages 3759–3771

INDONESIAN

TRANS_NOM LINE_NAME FAC_NAME LOC_NAME

Kamu bisa naik bus dari Jalan Perintis Kemerdekaan atau Terminal Daya, Makassar.

TRANS_NOM LOC_NAME LOC_NAME

Beberapa operator bus yang melayani rute Makassar - Tana Toraja adalah:

ENGLISH

TRANS_NOM FAC_NAME FAC_NAME

You can catch a bus from either the Jl. Perintis Kemerdekaan Station or the Daya Station.

TRANS_NOM LINE_NOM LOC_NAME LOC_NAME

Some bus companies that serve the route from Makassar to Tana Toraja are:

Figure 1: Example of Annotated Sentences in Our Dataset

2. Related Work

Our work is motivated by the fact that Indonesian is still considered as low-resource and that we perceive a substantial utilisation of the dataset constructed from personal travel documentation. Besides having mentions of named entities and nominal expressions, travelogue articles also include information, such as the sequence of visits to places and the author's impressions. The sequence of visits could be used to determine the trajectory of travel and factuality analysis (whether a location is indeed being visited or only mentioned), which would be useful for fellow travellers in trip planning. Then, the author's impressions could be used for semantic analysis, which would be helpful in providing feedback to the government or relevant organisations for event management and site maintenance.

2.1. Low-Resource Languages

Low-resource languages (LRLs) were defined as languages spoken in the world with less linguistic resources for language technologies (Cieri et al., 2016). In the research done by Joshi et al. (2020), the distribution of language resources was further divided into six clusters. Indonesian language was put under the category of languages which were lacking in terms of labelled data collection but having a growing presence in the digital world. This corresponded with the increasing effort to develop numerous datasets and language models (Wilie et al., 2020; Ariesandy et al., 2020; Winata et al., 2023).

Moreover, with the rise of awareness to preserve language diversity, more researchers were studying the challenges faced by LRLs and their feasible solutions. As stated by Doğruöz and Sitaram (2022), LRLs suffered a consequence of compromising between the accuracy of the system and the representativeness of the dataset. Besides, Magueresse et al. (2020) also discussed that it was necessary to collect new and diverse datasets as a way to resolve the problems faced by LRLs.

Drawing from the current position of Indonesian language as an LRL with a lot of potential and is on the move, we would contribute by constructing a new language resource built from less formal texts to improve the representativeness of its actual daily use. As seen in past surveys in 2019[3] and 2022[4], we were yet to see such dataset for the task of geoparsing (see Appendix A). In the recent collaborative initiative to collect and unify existing resources for Indonesian languages called NusaCrowd (Cahyawijaya et al., 2023), we were also yet to see a dataset focusing on geographic entity (see Appendix B).Thus, our dataset would certainly add knowledge into Indonesian language learning and assist in the improvement of related technologies.

2.2. Challenges in Geoparsing

In the research by Gritta et al. (2020), geoparsing consists of two main tasks: toponym extraction (geotagging) and toponym resolution (geocoding). Geotagging is similar to the task of named entity recognition (NER), but it is more focused on reference (mention) of location (toponym) in the text. Geocoding is regarded as entity linking where we aim to disambiguate location mentions in the text using available databases.

[3] https://github.com/irfnrdh/Awesome-Indonesia-NLP

[4] https://github.com/gentaiscool/indonesian-nlp

Several challenges exist in the task of geoparsing, such as metonymy resolution (Gritta et al., 2018b) and location inference based on the surrounding context (Farzana and Hecking, 2023). Metonymy occurs when a toponym word is used to substitute for something else. For example, in the sentence 'Japan wins the 2023 World Baseball Classic', the word 'Japan' refers to the Japanese baseball team instead of the country location. Other than that, sometimes a location is not explicitly mentioned in the text. Hence, to figure out the exact location being referred to, we need to infer from the surrounding context.

Geoparsing task might be challenging if we were to solely rely on the conventional NER system. As such, we considered adding information of nominal expressions in the text, so that there would be more contextual information for the model to learn. With this in mind, we designed an annotation scheme which encompassed nominal expressions of location mentions categorised into four entity types. This categorisation would allow the model to better distinguish the types of entities being referred to. Henceforth, we expect our dataset to improve the performance of existing system for geoparsing.

3. Dataset Construction

The process of dataset construction generally followed the guidelines provided by Higashiyama et al. (2023), with some modifications for the scope of our current research. More specifically, we only used annotation labels that specifically refer to the four entity types defined.

3.1. Data Acquisition

In the beginning, we surveyed several possible sources for data collection. We determined that travelogue would fit our requirements and purposes because in personal journals, writers tend to use a more casual writing style like how they speak in daily life. Besides, travelogue would definitely contain location mentions and their nominal expressions as they were being described for the reader.

We discovered that most Indonesian blog writers preferred to have their own website rather than posting in community forums. Coupled with the issue of usage rights and recent pandemic that significantly reduced the number of travels, we only managed to obtain express consent from one author. The author wrote in two languages, namely Indonesian[5] and English[6], albeit not at the same time and not encompassing the exact same content.

Initially, we obtained 65 relevant travel blog entries in Indonesian, and then we obtained 57 articles with similar contents at a brief glance in English. As we read the articles in more detail, we only included articles with a similar structure (almost the same content, but different paragraph sequence). This was done to ensure that both pair of Indonesian-English articles were mentioning the same entities and having almost the same number of mentions and article lengths. In the end, we selected 44 articles in each language, thus making a total of 88 articles in Indonesian and English.

[5] https://nonanomad.com/

[6] https://www.littlenomadid.com/

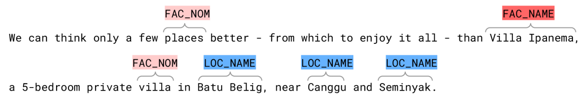

Figure 2: Sample of Annotation in English

3.2. Annotation of Named Entity

The annotation process began by manually annotating the named entities found in the text using BRAT rapid annotation tool (Stenetorp et al., 2012) from scratch. We considered using automatic annotation for named entity candidates. However, our preliminary experiment showed that the results did not meet our expectation.

In this step, we employed four named entity categories as follows:

- `LOC_NAME` for naturally existing locations, e.g., country, mountain, lake, etc.
- `FAC_NAME` for man-made structures or area, e.g., park, building, station, etc.
- `TRANS_NAME` for transportation modes or vehicles, e.g., bus, train, ship, etc.
- `LINE_NAME` for roads or waterways, e.g., street, river, route, etc.

An example of the annotation in English is shown in Figure 2. In the text, 'Villa Ipanema' is a facility because it is built by human, whereas 'Canggu', and 'Seminyak' are locations because both are the names of beach resort areas in Bali.

We were aware of ambiguities due to common names shared between entity types. In this case, we tried to determine the most probable entity type based on the surrounding context. For instance, looking back at Figure 2, 'Batu Belig' may refer to the area or the road in Bali. Since the named entities following 'Batu Belig' are clearly locations, the writer is more likely to talk about 'Batu Belig' as the area (location). Next, when we checked the address of the villa, it was not located in Batu Belig road. Thus, we confirmed that in this case, 'Batu Belig' is being referred as an area (location). Although we provided the tag `OTHER` in the case that the type of entity was really difficult to determine, we generally did not use this tag as much.

		F1	Ann1	Ann2	Both
id	Named	0.839	328	309	294
	Nominal	0.757	225	191	165
	All	0.792	553	500	459
en	Named	0.828	268	256	224
	Nominal	0.719	187	195	127
	All	0.766	455	451	351

Table 1: Inter-Annotator Agreement

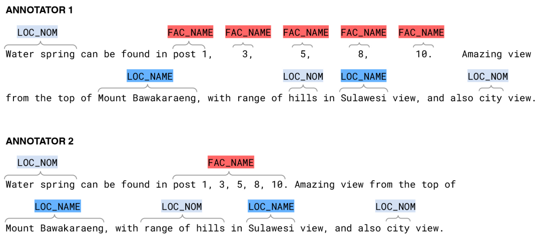

Figure 3: Sample of Span

3.3. Annotation of Nominal Expression

The next stage was annotating the nominal expressions associated with each category of the named entities. Some examples of nominal expressions are the words 'country', 'house', 'river', and such nouns. Following are the tags used: `LOC_NOM`, `FAC_NOM`, `TRANS_NOM`, and `LINE_NOM`.

At this stage, we particularly observed that `TRANS_NOM` and `LINE_NOM` had a tendency to not be associated with any named entities within the same document. We conjectured that it might be because there were many alternatives for transportation modes and routes to reach the same location, thus travellers could easily determine whichever they preferred as they took the trip.

4. Evaluation

We evaluated the sufficiency of our dataset using common methods: the inter-annotator agreement, the statistics of our dataset, and the experiments using publicly available tools. We also provide a list of known geoparsing datasets to demonstrate the contribution of our dataset (see Appendix C).

4.1. Inter-annotator Agreement

For each language covered, we involved two independent annotators with at least one native speaker. We measured the agreement scores (F1 score) for five articles selected for each language based on exact match of both the labels and the text spans. The scores are as shown in Table 1 for Indonesian and English blog entries (breakdown by each label is provided in Appendix D). In this table, we also provide the number of annotations by each annotator (Ann1 and Ann2) and the number of exact match of annotations by both annotators (Both).

The overall agreement score was higher for Indonesian articles (0.792) than that for English articles (0.766), but both scores were not that far apart. The agreement scores for named entities were higher than that for nominal expressions. Note that the selected articles happened to not have `TRANS_NAME`, hence the overall F1 scores were calculated based on macro average.

Nominal expressions were harder to recognise, and some of them were ambiguous (e.g., place, area) which made it more difficult to assign the appropriate labels. Besides, we found that the annotators marked different spans for the same nominal expressions. Since the scores were calculated based on exact match, differing spans were considered as a disagreement. An example is shown in Figure 3. We could see that both annotators recognised the nominal expression 'hills', but one annotator marked the whole span of 'range of hills'.

Number of	Total	Ave.	Total	Ave.
Sentences	1,391	31	1,914	43
Words	47,415	1,077	47,902	1,088
Named	3,937	89	2,756	62
Nominal	2,062	46	2,243	50
Named (U)	1,156	26	1,053	23
Nominal (U)	430	9	760	17

Table 2: Dataset Statistics for id (left) and en (right)

4.2. Coverage of Dataset

Another dataset based on travelogue was released formerly by Ouchi et al. (2023). We would present the statistics of our dataset in similar manner in Table 2 for both Indonesian (id) and English (en).

Both the Indonesian and the English articles had in total around 1,000 mentions of unique named entities (Named (U)) for domestic and international travel trips. Apparently, the English articles had more variety of unique nominal expressions (Nominal (U)). This might explain why English had a lower agreement score: because it was more difficult to recognise the nominal expressions.

In comparison with existing geoparsing datasets (Appendix C), there was only one dataset in Indonesian language. Moreover, most datasets have the size below 10,000 mentions, except for one dataset that we referred to. Among all these datasets, there was also only one that used travelogue as the data source. Based on this, we could see that our dataset, with a total of approximately 11,000 mentions, is of sufficient size.

4.3. Experiments

The aim of the experiments is to clarify the performance level of current entity analysis systems. We trained classifiers to recognise named entities and

		Precision	Recall	F1
	Named	0.881	0.841	0.853
id	Nominal	0.910	0.914	0.912
	Overall	0.923	0.938	0.931
	Named	0.877	0.859	0.866
en	Nominal	0.902	0.910	0.906
	Overall	0.922	0.922	0.922

Table 3: Experiment Results (Macro Ave.)

nominal expressions on our dataset using spaCy[7]. For each language, we split 44 articles into the train, validation, and test sets in the ratio of 8:1:1, giving 35, 4, and 5 articles respectively. Although the validation and test sets only contained small numbers of articles, there were about 500-600 mentions for each language. We considered that this was quite a reasonable size to evaluate the classifiers under the low-resource setting. The training was done using spaCy NER with corresponding transformers for Indonesian[8] and English[9]. The results of the experiments are shown in Table 3.

For both languages, the scores for nominal expressions were higher than that for named entities. This corresponded to the fact that there were more kinds of named entities than nominal expressions (see Table 2), hence it was easier to recognise nominal expressions. Some errors that we discovered happened when the entities were expressed in different ways. For example, the entity 'Heijo Palace' was sometimes written as 'Heijo-kyo'. Our classifier was able to recognise 'Heijo Palace' as one entity mention but separated 'Heijo' and 'kyo' as two entities. A possible reason for this is because dash (-), especially in Indonesian, is often used as a connector between two different locations (e.g., rute Makassar-Tana Toraja in Figure 1). From these results, we perceived the importance of further experiments with our dataset as well as our classifiers.

Our classifiers managed to achieve overall F1-score of 0.931 for Indonesian and 0.922 for English. However, we were aware of a possible bias in the results due to the limitation of our data source. Thus, we tried our classifiers on texts from different authors with different writing styles and covering entities which were not present in our dataset. We observed that the results corresponded to the reported scores, i.e., majority of the spans and tags were correctly identified with a few misses (especially in cases such as the use of dash or entities with longer names). This indicated that we could use this new dataset for further improvements and evaluations of currently existing models.

[7]https://spacy.io/

[8]https://huggingface.co/indolem/indobert-base-uncased

[9]https://huggingface.co/roberta-base

Simple comparisons of available NER model (spaCy en_core_web_sm) and our classifier in English are presented in the Appendices. The four kinds of text we sampled are: (i) travelogue from the same author (Appendix E), (ii) travelogue from a different author (Appendix F), (iii) Wikipedia (Appendix G), and (iv) news article (Appendix H). For spaCy, the labels related to geographic entities are:

- `FAC`: Buildings, airports, highways, bridges, etc.
- `GPE`: Countries, cities, states.
- `LOC`: Non-GPE locations, mountain ranges, bodies of water.

From the comparisons that we have done, our classifier performed well even with a rather small training data. Furthermore, among all the examples, we tried to use texts with entities that were not covered in the travelogue. The results showed that our classifier still managed to accurately recognise these references. Therefore, this proved the potential use of our dataset for futher experiments and expansion.

5. Conclusions

In this work, we have constructed an Indonesian-English dataset from travelogue articles with a new annotation scheme that included named entities and their nominal expressions. This dataset covers the first part of geoparsing: geotagging. The experiments conducted showed that classifiers trained on our dataset were able to achieve over 0.9 F-score for both Indonesian and English. This confirmed that our dataset would be useful in improving current geoparsing systems for low-resource language. As the next step towards geoparsing, we will continue to extend the coverage of our dataset for geocoding. We will release our annotated dataset to enable other researchers to conduct reproducible experiments and develop more sophisticated geoparsing systems.

Limitations

Currently, our dataset is limited because we only managed to acquire one bilingual travelogue written by one author. As a result, our findings might be biased towards the author's writing style. In the future work, we plan to increase the diversity in our dataset by adding more articles by different authors. Further analysis could be done by evaluating the model's performance with existing NER dataset. We will also extend the coverage of our dataset by including coreference resolution and entity linking, as well as other types of information, such as expressions of human behaviours and experiences.

Acknowledgments

We would like to thank the anonymous reviewers and meta reviewers for their constructive comments. This study was supported by JSPS KAKENHI Grant Number JP22H03648.

Bibliographical References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017a. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Ika Alfina, Septiviana Savitri, and Mohamad Ivan Fanany. 2017b. Modified dbpedia entities expansion for tagging automatically ner dataset. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 216–221.

Asrul Sani Ariesandy, Mukhlis Amien, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Synthetic source language augmentation for colloquial neural machine translation.

Valentina Kania Prameswara Artari, Rahmad Mahendra, Meganingrum Arista Jiwanggi, Adityo Anggraito, and Indra Budi. 2021. A multi-pass sieve coreference resolution for Indonesian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 79–85, Held Online. INCOMA Ltd.

Jessica Naraiswari Arwidarasti, Ika Alfina, and Adila Alfa Krisnadhi. 2019. Converting an indonesian constituency treebank to the penn treebank format. In *2019 International Conference on Asian Language Processing (IALP)*, pages 331–336.

Annisa Nurul Azhar, Masayu Leylia Khodra, and Arie Pratama Sutiono. 2019. Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting. In *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 35–40.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Agung Dewandaru. 2020. Event geoparsing indonesian news dataset.

Agung Dewandaru, Dwi Widyantoro, and Saiful Akbar. 2020. Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in indonesian news domain. *International Journal of Geo-Information*, 9:712.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

Sheikh Mastura Farzana and Tobias Hecking. 2023. Geoparsing at web-scale - challenges and opportunities. In *Proceedings of the First Workshop on Geographic Information Extraction from Texts (GeoExT 2023) co-located with The 45th European Conference on Information Retrieval (ECIR 2023)*.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2018a. Which Melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. 2020. A pragmatic guide to geoparsing evaluation. In *Language Resources and Evaluation*, volume 54, pages 683–712.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2018b. What's missing in geographical parsing? In *Language Resources and Evaluation*, volume 52, pages 603–623.

Yohanes Gultom and Wahyu Catur Wibowo. 2017. Automatic open domain information extraction from indonesian text. In *2017 International Workshop on Big Data and Information Security (IWBIS)*, pages 23–30.

Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. 2023. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation.

Devin Hoesen and Ayu Purwarianti. 2018. Investigating bi-lstm and crf with pos tag embedding for indonesian named entity tagger. In *2018 International Conference on Asian Language Processing (IALP)*. IEEE.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Arfinda Ilmania, Abdurrahman, Samuel Cahyawijaya, and Ayu Purwarianti. 2018. Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis. In *2018 International Conference on Asian Language Processing (IALP)*, pages 62–67.

Rini Jannati, Rahmad Mahendra, Cakra Wishnu Wardhana, and Mirna Adriani. 2018. Stance classification towards political figures on blog writing. In *2018 International Conference on Asian Language Processing (IALP)*, pages 96–101.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1287–1296, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Liputan6: A large-scale Indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608, Suzhou, China. Association for Computational Linguistics.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. XPersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and Mirna Adriani. 2018. Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task. In *Proceedings of the 9th Global Wordnet Conference*, pages 245–250, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Miftahul Mahfuzh, Sidik Soleman, and Ayu Purwarianti. 2019. Improving joint layer rnn based keyphrase extraction by using syntactical features. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE.

Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. 2017. Geographical entity annotated corpus of japanese microblogs. *Journal of Information Processing*, 25:121–130.

David Moeljadi, Aditya Kurniawan, and Debaditya Goswami. 2019. Building cendana: a treebank for informal indonesian. *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, pages 156–164.

Sri Mulyana Muhammad Fachri. 2014. Pengenalan entitas bernama pada teks bahasa indonesia menggunakan hidden markov model.

Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. 2023. Arukikata travelogue dataset.

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.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5.

Paul Rayson, Alex Reinhold, James Butler, Chris Donaldson, Ian Gregory, and Joanna Taylor. 2017. A deeply annotated testbed for geographical text analysis: The corpus of lake district writing. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities '17, page 9–15, New York, NY, USA. Association for Computing Machinery.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95.

Ken Nabila Setya and Rahmad Mahendra. 2023. Semi-supervised textual entailment on indonesian wikipedia data. In *Computational Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, March 18–24, 2018, Revised Selected Papers, Part I*, page 416–427, Berlin, Heidelberg. Springer-Verlag.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M. MacEachren, and Scott Pezanowski. 2018. Geocorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1):1–29.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 907–916, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Haryo Akbarianto Wibowo, Tatag Aziz Prawiro, Muhammad Ihsan, Alham Fikri Aji, Radityo Eko Prasojo, Rahmad Mahendra, and Suci Fitriany. 2020. Semi-supervised low-resource style transfer of indonesian informal to formal language with iterative forward-translation.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Andika William and Yunita Sari. 2020. Click-id: A novel dataset for indonesian clickbait headlines. *Data in Brief*, 32:106231.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Appendices

A. Summary of Surveys in 2019 and 2022

Tasks	Publications
Morphology Analysis	(Pimentel et al., 2021)
Part-of-Speech Tagging	(Hoesen and Purwarianti, 2018)
Named Entity Recognition	(Hoesen and Purwarianti, 2018)
Word Sense Disambiguation	(Mahendra et al., 2018)
Constituency Parsing	(Arwidarasti et al., 2019; Moeljadi et al., 2019)
Dependency Parsing	(Zeman et al., 2018)
Coreference Resolution	(Artari et al., 2021)
Chatbot	(Lin et al., 2021)
Question Answering	(Clark et al., 2020)
Summarization	(Koto et al., 2020; Kurniawan and Louvan, 2018)
Keyphrase Extraction	(Mahfuzh et al., 2019)
Natural Language Inference	(Setya and Mahendra, 2023; Mahendra et al., 2021)
Sentiment Analysis	(Purwarianti and Crisdayanti, 2019; Azhar et al., 2019; Ilmania et al., 2018)
Emotion Classification	(Saputri et al., 2018)
Stance Detection	(Jannati et al., 2018)
Hate Speech Detection	(Ibrohim and Budi, 2019, 2018; Alfina et al., 2017a)
Clickbait Detection	(William and Sari, 2020)
Style Transfer	(Wibowo et al., 2020)

Table 4: Research and Resources in Indonesian Language

B. NER Datasets in NusaCrowd (as of 2023)

Dataset Name	Year	Size	Domain	Publications
IndQNER	2022	3,118 sentences	religion	-
IndoNLU NERGrit	2020	2,090 sentences	general	(Wilie et al., 2020)
NERGrit	2020	17,437 sentences	general	-
NERP (IndoNLU Split)	2018	8,400 sentences	news	(Hoesen and Purwarianti, 2018)
NER UI (IndoLEM split)	2017	2,125 sentences	general	(Gultom and Wibowo, 2017)
Singgalang	2017	48,957 sentences	wiki	(Alfina et al., 2017b)
WikiAnn (multilingual)	2017	254,240 mentions	wiki	(Pan et al., 2017)
NER UGM (IndoLEM split)	2014	2,343 sentences	news	(Muhammad Fachri, 2014)

Table 5: NER Datasets in NusaCrowd

C. Geoparsing Datasets

Dataset Name	Year	Language	Size	Domain	Publications
ATD-MCL	2023	ja	**12K**	**Travelogue**	(Ouchi et al., 2023)
Event Geoparsing	2020	**id**	1.1K	News	(Dewandaru, 2020)
GeoWebNews	2020	en	2.4K	News	(Gritta et al., 2020)
SemEval-2019 T12	2019	en	8.4K	Science	(Weissenbacher et al., 2019)
GeoCorpora	2018	en	3.1K	Microblog	(Wallgrün et al., 2018)
TR-News	2018	en	1.3K	News	(Kamalloo and Rafiei, 2018)
GeoVirus	2018	en	2.2K	News	(Gritta et al., 2018a)
CLDW	2017	en	3.7K	Historical	(Rayson et al., 2017)
LRE Corpus	2017	ja	1.0K	Microblog	(Matsuda et al., 2017)

Table 6: Details of Geoparsing Datasets

D. Breakdown of Inter-Annotator Agreeement Scores by Label

Label	Indonesian				English			
	F1	Ann1	Ann2	Both	F1	Ann1	Ann2	Both
LOC_NAME	0.949	207	215	203	0.864	160	174	149
FAC_NAME	0.817	106	85	82	0.788	98	73	68
TRANS_NAME	-	-	-	-	-	-	-	-
LINE_NAME	0.750	15	9	9	0.833	10	9	7
LOC_NOM	0.844	88	85	79	0.551	72	82	48
FAC_NOM	0.767	86	74	62	0.633	81	76	49
TRANS_NOM	0.805	25	13	13	0.900	13	12	12
LINE_NOM	0.613	26	19	11	0.792	21	25	18

Table 7: Inter-Annotator Agreement by Label

E. Comparison on Travelogue Article from the Same Author

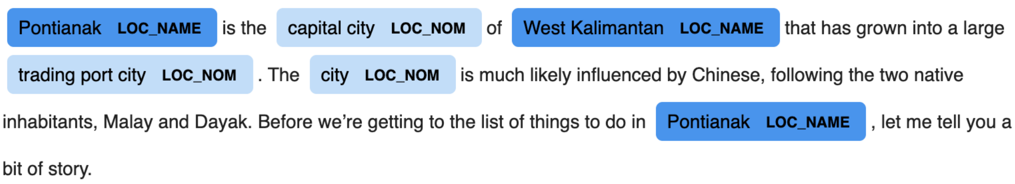

Figure 4: Our Classifier on Travelogue Article from the Same Author

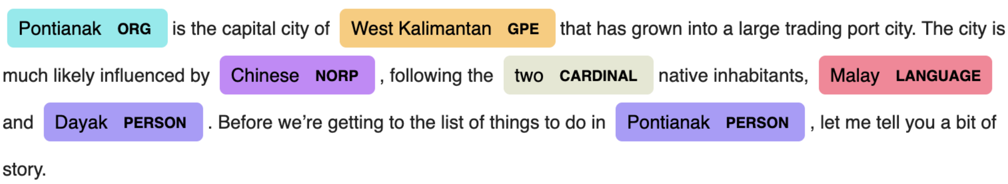

Figure 5: SpaCy Classifier on Travelogue Article from the Same Author

F. Comparison on Travelogue Article from a Different Author

Fortresses FAC_NOM and defensive walls FAC_NOM pepper the island LOC_NOM , as do churches FAC_NOM and traditional villages FAC_NOM . Being an island LOC_NOM there are plenty of coves LOC_NOM and little beaches LOC_NOM to discover. The diving is superb here, with underwater caves LOC_NOM , and plenty of wrecks FAC_NOM to discover. Go souvenir hunting around Valletta LOC_NAME , hitch a ride with a donkey on Gozo LOC_NAME , eat delicious sea-food in a small fishing village LOC_NOM , or admire the sunset from one of the many cliffs LOC_NOM .

Figure 6: Our Classifier on Travelogue Article from a Different Author

Fortresses and defensive walls pepper the island, as do churches and traditional villages. Being an island there are plenty of coves and little beaches to discover. The diving is superb here, with underwater caves, and plenty of wrecks to discover. Go souvenir hunting around Valletta GPE , hitch a ride with a donkey on Gozo GPE , eat delicious sea-food in a small fishing village, or admire the sunset from one of the many cliffs.

Figure 7: SpaCy Classifier on Travelogue Article from a Different Author

G. Comparison on Wikipedia Article

Nara LOC_NAME was the capital LOC_NOM of Japan during the Nara period from 710 to 794 as the seat of the Emperor before the capital was moved to Kyoto LOC_NAME . Nara LOC_NAME is home to eight temples FAC_NOM , shrines FAC_NOM , and ruins FAC_NOM , specifically Tōdai-ji FAC_NAME , Saidai-ji FAC_NAME , Kōfuku-ji FAC_NAME , Kasuga Shrine FAC_NAME , Gangō-ji FAC_NAME , Yakushi-ji FAC_NAME , Tōshōdai-ji FAC_NAME , and the Heijō Palace FAC_NAME , together with Kasugayama Primeval Forest LOC_NAME , collectively form the Historic Monuments of FAC_NOM Ancient Nara, a UNESCO World Heritage Site LOC_NOM .

Figure 8: Our Classifier on Wikipedia Article

Nara GPE was the capital of Japan GPE during the Nara GPE period from 710 CARDINAL to 794 CARDINAL as the seat of the Emperor before the capital was moved to Kyoto GPE . Nara GPE is home to eight temples, shrines, and ruins, specifically Tōdai GPE -ji, Saidai ORG -ji, Kōfuku NORP -ji, Kasuga Shrine PERSON , Gangō PRODUCT -ji, Yakushi-ji PERSON , Tōshōdai ORG -ji, and the Heijō Palace ORG , together with Kasugayama Primeval Forest ORG , collectively form the Historic Monuments of Ancient Nara ORG , a UNESCO World Heritage Site ORG .

Figure 9: SpaCy Classifier on Wikipedia Article

H. Comparison on News Article

Making time for winter wellness can help you weather Alaska's LOC_NAME cold. And at Salted Roots FAC_NAME in Seward LOC_NAME , set on an inlet LOC_NOM along the Kenai Peninsula LOC_NAME , guests stay in cozy A-frame cabins FAC_NOM surrounded by a spruce forest LOC_NOM for two-night winter wellness packages that include private yoga lessons and massage as well as plenty of sauna time. A newly renovated sister property FAC_NOM , Rustic Roots FAC_NAME , with rustic seaside cabins, is opening next door in January 2024.

Figure 10: Our Classifier on News Article

Making time for winter DATE wellness can help you weather Alaska GPE 's cold. And at Salted Roots ORG in Seward GPE , set on an inlet along the Kenai Peninsula LOC , guests stay in cozy A-frame cabins surrounded by a spruce forest for two-night TIME winter wellness packages that include private yoga lessons and massage as well as plenty of sauna time. A newly renovated sister property, Rustic Roots ORG , with rustic seaside cabins, is opening next door in January 2024 DATE .

Figure 11: SpaCy Classifier on News Article