LREC-COLING 2024

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

Dataset

CONTENT

1. Introduction

Low-Resource Languages (LRLs)

- Languages spoken in the world with **less linguistic resources** for technologies.
  (Cieri et al. 2016)
- Indonesian language is **lacking** in labelled data collection, but having a **growing** presence in the digital world (cluster 3).
  (Joshi et al. 2020)
- There is an increasing effort to develop Indonesian **datasets** and **language models**.
  (Willie et al. 2020, Ariesandy et al. 2020, Aji et al. 2022, Winata et al. 2023)

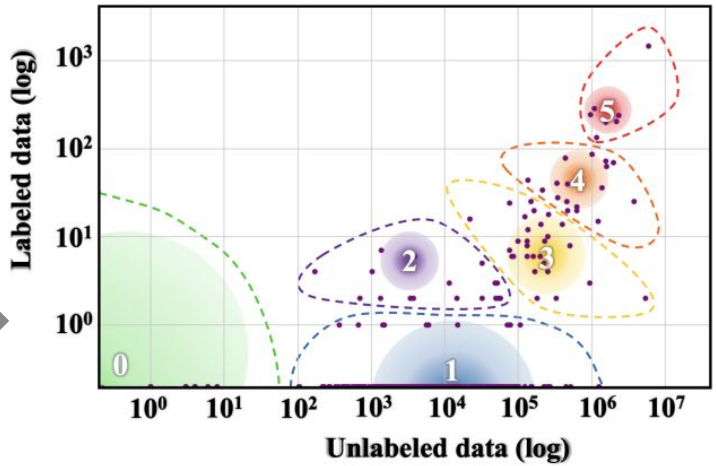

Fig. 1 - Language Resource Distribution
(Joshi et al. 2020)

Challenges Faced by LRLs

Accuracy

- ➔ Large size
- ➔ Well-documented
- ➔ Formal sources

Representativeness

(Dogruoz and Sitaram. 2022)

- New
- Diverse

(Magueresse et al. 2020)

Importance of Geographic Data in Indonesia

- With the **rise in travel and tourism** in Indonesia, more information from both locals and common tourists would be needed.
- Valuable information are often contained in texts written by **common people**.
- Collection of geographic data through the use of sensors and its reports in the form of statistics by certain organizations are often **limited**.

Image taken from: https://etimg.etb2bimg.com/photo/100758092.cms

Processing Geographic Data in Texts

Geoparsing consists of two main tasks (Gritta et al. 2020) :

1. Toponym Extraction (**Geotagging**)
   Similar to **named entity recognition (NER)**, but more focused on reference (mention) of location (toponym) in the text.
2. Toponym Resolution (**Geocoding**)
   Regarded as **entity linking (EL)**, i.e., disambiguate location mentions in the text using available knowledge bases.

Processing Geographic Data in Texts

Sample Text

```
Last week, we went to visit Heijo Palace.
We really enjoyed our time there, it was
such a big area to explore. Before that,
we also bought some drinks at Family Mart
near Yamato-Saidaiji station. We thought of
visiting Nara Park too, but it was too late.
We decided to visit the park the day after.
```

Processing Geographic Data in Texts

Sample Text

Last week, we went to visit Heijo Palace. We really enjoyed our time there, it was such a big area to explore. Before that, we also bought some drinks at Family Mart near Yamato-Saidaiji station. We thought of visiting Nara Park too, but it was too late. We decided to visit the park the day after.

Tasks:

1. Named Entity Recognition

Processing Geographic Data in Texts

Sample Text

Last week, we went to visit Heijo Palace. We really enjoyed our time there, it was such a big area to explore. Before that, we also bought some drinks at Family Mart near Yamato-Saidaiji station. We thought of visiting Nara Park too, but it was too late. We decided to visit the park the day after.

Tasks:

1. Named Entity Recognition
2. Nominal Expressions Recognition

Processing Geographic Data in Texts

Sample Text

Last week, we went to visit Heijo Palace. We really enjoyed our time there, it was such a big area to explore. Before that, we also bought some drinks at Family Mart near Yamato-Saidaiji station. We thought of visiting Nara Park too, but it was too late. We decided to visit the park the day after.

Tasks:

1. Named Entity Recognition
2. Coreference Resolution (including Nominal Expressions Recognition)

Processing Geographic Data in Texts

Sample Text

Last week, we went to visit Heijo Palace. We really enjoyed our time there, it was such a big area to explore. Before that, we also bought some drinks at Family Mart near Yamato-Saidaiji station. We thought of visiting Nara Park too, but it was too late. We decided to visit the park the day after.

Tasks:

1. Named Entity Recognition
2. Coreference Resolution
3. Entity Disambiguation/Linking

Fig. 2 - Sample Capture of Google Maps

Image taken from: https://www.google.com/maps

Challenges in Geoparsing

- **Metonymy Resolution** (Gritta et al. 2018)
  A toponym word is used to substitute for something else
  For example, “Japan wins the 2023 World Baseball Classic”
- **Location Inference** based on the surrounding context (Farzana and Hecking. 2023)
  A location is being described instead of explicitly mentioned in the text
  For example, “Famous park in Nara prefecture with many deers” [Nara Park]

Contextual Information

State of Indonesian Language Resources

Name	Year	Domain
IndQNER	2022	religion
IndoNLU NERGrit	2020	general
NERGrit	2020	general
NERP (IndoNLU Split)	2018	news
NER UI (IndoLEM split)	2017	general
Singgalang	2017	wiki
WikiAnn (multilingual)	2017	wiki
NER UGM (IndoLEM split)	2014	news

Table 1: NER Datasets in NusaCrowd
(Cahyawijaya et al. 2023)

Dataset Name	Year	Language	Size	Domain
ATD-MCL	2023	ja	12K	Travelogue
Event Geoparsing	2020	id	1.1K	News
GeoWebNews	2020	en	2.4K	News
SemEval-2019 T12	2019	en	8.4K	Science
GeoCorpora	2018	en	3.1K	Microblog
TR-News	2018	en	1.3K	News
GeoVirus	2018	en	2.2K	News
CLDW	2017	en	3.7K	Historical
LRE Corpus	2017	ja	1.0K	Microblog

Table 2: Geoparsing Datasets (Size in Mentions)

Focus on Geographic

Indonesian Travelogue

New

Diverse

Constructing Indonesian-English Travelogue Dataset

Focus on Geographic

Contextual Information

2. Dataset Construction

Scope of the Dataset

Referring to the recent Japanese travelogue dataset (Higashiyama et al. 2023, Ouchi et al. 2023)

1	2	3	4
Data Acquisition	Mention Annotation	Coreference Annotation	Entity Linking

Note:

Mention → both named entities and **nominal expressions**

Scope of the Dataset

Our Scope

1	2	3	4
Data Acquisition	Mention Annotation	Coreference Annotation	Entity Linking

Note:

Mention → both named entities and **nominal expressions**

Acquiring the Data

1

Data Acquisition

2

Mention Annotation

→ Obtained **88 articles** (44 ID + 44 EN). (see detailed statistics here)

→ Only articles with very **similar** content are included.

How similar are the ID and EN articles?

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:

Fig. 3 - Sample of Annotated Sentences in Indonesian and English

Annotation Process

1

Data Acquisition

2

Mention Annotation

→ Using brat rapid annotation tool with some tags as defined in JP dataset.

How the Annotation was Conducted

Tags included in the annotation:

1. **LOC_NAME** : naturally existing locations, e.g., country, mountain, lake
2. **FAC_NAME** : man-made structures or area, e.g., park, building, station
3. **TRANS_NAME** : transportation modes or vehicles, e.g., bus, train, ship
4. **LINE_NAME** : roads and waterways, e.g., street, river, route
5. And another four for the nominal expression of each of the above.
   LOC_NOM, **FAC_NOM**, **TRANS_NOM**, **LINE_NOM**

How the Annotation was Conducted

Overview:

1. Based on the guideline, did **trial** annotation for 15 articles.
2. **Revised** the guideline based on what was discovered in the trial round.
3. Continued the annotation and asked for other **independent** annotators.
4. Calculated **inter-annotator agreement** and observe the results.
5. Conducted some **experiments** to evaluate the dataset.

3. Dataset Evaluation

Inter-Annotator Agreement (Exact Match)

		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 3: Inter-Annotator Agreement (Exact Match)

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 4: Inter-Annotator Agreement by Labels (Exact Match)

Different Spans in Annotation

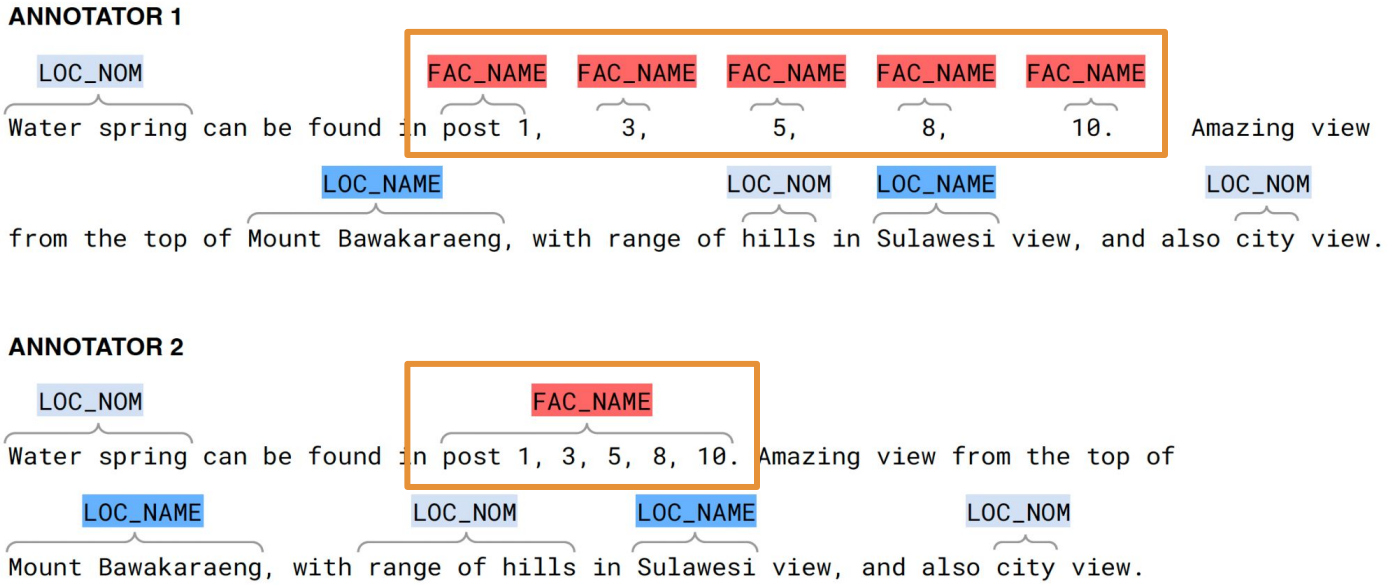

Fig. 4 - Sample of Annotated Sentences with Different Spans in English

Different Spans in Annotation

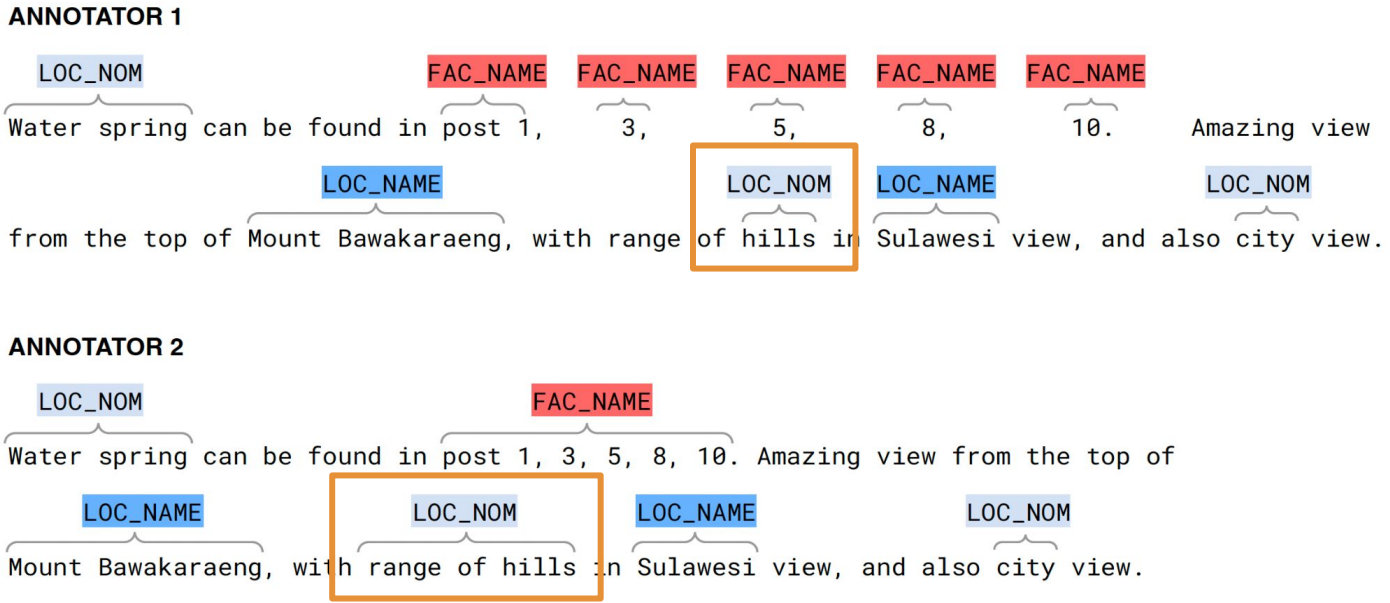

Fig. 4 - Sample of Annotated Sentences with Different Spans in English

Dataset Statistics

Number of	Indonesian		English	
	Total	Average	Total	Average
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 (Unique)	1,156	26	1,053	23
Nominal (Unique)	430	9	760	17

Table 5: Dataset Statistics for Indonesian and English

~11K mentions

Dataset Name	Year	Language	Size	Domain
ATD-MCL	2023	ja	12K	Travelogue
Event Geoparsing	2020	id	1.1K	News
GeoWebNews	2020	en	2.4K	News
SemEval-2019 T12	2019	en	8.4K	Science
GeoCorpora	2018	en	3.1K	Microblog
TR-News	2018	en	1.3K	News
GeoVirus	2018	en	2.2K	News
CLDW	2017	en	3.7K	Historical
LRE Corpus	2017	ja	1.0K	Microblog

Table 2: Geoparsing Datasets (Size in Mentions)

Experiment and Results

- Trained **classifiers** to recognise named entities and nominal expressions.
- Used **spaCy** NER with **transformers** for Indonesian and English.
- For each language, split 44 articles into the train, validation, and test set in the **ratio of 8:1:1**, giving 35, 4, and 5 articles.

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

Table 6: Results of Experiments (Macro Ave.)

3. Dataset Evaluation

Comparison with SpaCy

- Simple comparison using **different texts**:
  1. Travelogue by the same author
  2. Travelogue by a different author
  3. Wikipedia article
  4. News article
- Labels related to geographic entities in SpaCy:
  1. **FAC** : Buildings, airports, highways, bridges, etc.
  2. **GPE** : Countries, cities, states.
  3. **LOC** : Non-GPE locations, mountain ranges, bodies of water.

Comparison with SpaCy on Unseen Sample

Pontianak LOC_NAME is the capital city LOC_NOM of West Kalimantan LOC_NAME that has grown into a large trading port city LOC_NOM . The city LOC_NOM is much likely influenced by Chinese, following the two native inhabitants, Malay and Dayak. Before we're getting to the list of things to do in Pontianak LOC_NAME , let me tell you a bit of story.

no miss

Fig. 5 - Our Classifier on Travelogue Article by the Same Author

~~Pontianak ORG~~ is the capital city of West Kalimantan GPE that has grown into a large trading port city. The city is much likely influenced by Chinese NORP , following the two CARDINAL native inhabitants, Malay LANGUAGE and Dayak PERSON . Before we're getting to the list of things to do in ~~Pontianak PERSON~~ , let me tell you a bit of story.

2 misses

Fig. 6 - SpaCy Classifier on Travelogue Article by the Same Author

Comparison with SpaCy on Another 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 .

1 miss

Fig. 7 - Our Classifier on Travelogue Article by 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.

no miss

Fig. 8 - SpaCy Classifier on Travelogue Article by a Different Author

Comparison with SpaCy on Wikipedia Article

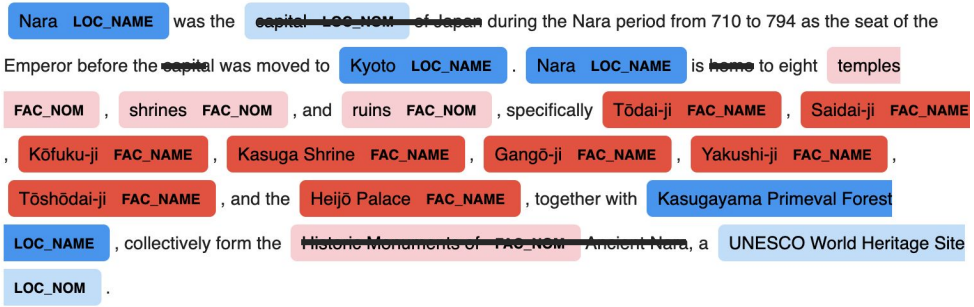

Fig. 9 - Our Classifier on Wikipedia Article

4 misses

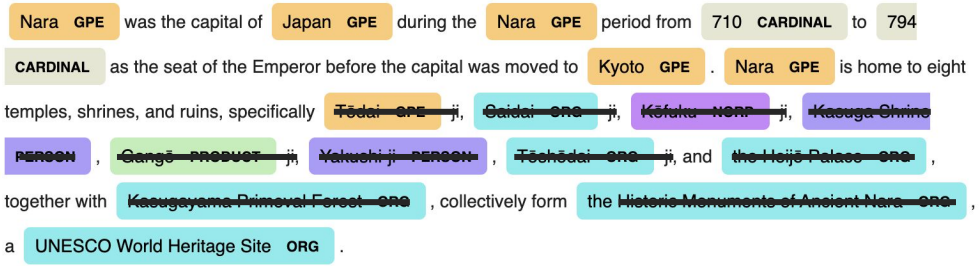

Fig. 10 - SpaCy Classifier on Wikipedia Article

10 misses

Comparison with SpaCy 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.

1 miss

Fig. 11 - 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 .

2 misses

Fig. 12 - Our Classifier on News Article

Further Analysis

- Types of **mistakes** occurring in our classifier:
  1. Not recognising new mentions
  2. Not recognising existing mentions
  3. Misclassification of mentions
  4. Incomplete span of mentions
- Further investigation is necessary to figure out the **characteristics** and **tendencies** of our classifier, i.e., whether the problem mainly lies in the dataset or the classifier itself.

4. Conclusions

SUMMARY

- We have constructed an **Indonesian-English travelogue dataset** to with a new annotation scheme that includes nominal expressions.
- We have conducted **experiments** and observe the **potential use** of our dataset for research.
- Further **expansion of scope** is necessary.

5. Future Works

5. Future Works

Expansion of the Dataset

Within the same scope of use:

- Add more travelogue articles from diverse authors (writing styles).
- Cover more languages → foreign or local languages in Indonesia.
- Conduct more experiments, e.g., comparing with NER datasets, etc.
- Improve the guideline or the tools to handle multi-spans of mention.

Beyond the current scope of use:

- Include coreference resolution and entity linking (geocoding).
- Extend the use of the dataset, e.g., including human behaviours, etc.

Dataset

End of Presentation

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.

Thank you for your attention.