Malaysian English News Decoded: A Linguistic Resource for Named Entity and Relation Extraction

Mohan Raj Chanthran[1], **Lay-Ki Soon**[1*], **Huey Fang Ong**[1], **and Bhawani Selvaretnam**[2]
[1]School of Information Technology, Monash University, Malaysia, [2]Valiantlytix Sdn Bhd
[1]{mohan.chanthran, soon.layki, ong.hueyfang}@monash.edu,
[2]bhawani@valiantlytix.com

Abstract

Standard English and Malaysian English exhibit notable differences, posing challenges for natural language processing (NLP) tasks on Malaysian English. An experiment using state-of-the-art Named Entity Recognition (NER) solutions in Malaysian English news articles highlights that they cannot handle morphosyntactic variations in Malaysian English. Unfortunately, most of the existing datasets are mainly based on Standard English, which is not sufficient to enhance NLP tasks in Malaysian English. To the best of our knowledge, there is no annotated dataset that can be used to improve the model. To address this issue, we have constructed a Malaysian English News (MEN) dataset, which contains 200 news articles that are manually annotated with entities and relations. We then fine-tuned the spaCy NER tool and validated that having a dataset tailor-made for Malaysian English could significantly improve the performance of NER in Malaysian English. This paper presents our efforts to acquire data, the annotation methodology, and a detailed analysis of the annotated dataset. To ensure the quality of the annotation, we have measured the Inter-Annotator Agreement (IAA), and any disagreements were resolved by a subject matter expert through adjudication. After a rigorous quality check, we have developed a dataset with 6,061 entities and 3,268 relation instances. Finally, we discuss spaCy fine-tuning setup and analysis of NER performance. This unique dataset will contribute significantly to the advancement of NLP research in Malaysian English, allowing researchers to accelerate their progress, particularly in NER and relation extraction.

Keywords: Annotated Dataset, Malaysian English, Named Entity Recognition, Relation Extraction, Low-Resource Language

1. Introduction

1.1. Overview

Relation Extraction (RE) is a natural language processing (NLP) task that involves identifying relations between a pair of entities mentioned in a text. This task requires identifying the entities and predicting their relations based on the context of the sentence or document. Many previous studies on RE are based on supervised learning (Swampillai and Stevenson, 2011; Chan and Roth, 2011; Sahu et al., 2019; Wang et al., 2020). This means that the performance of the RE models depends very much on the quality of the annotated dataset used for the training.

Malaysian English (ME) is a variant of English that has evolved from Standard English, incorporating local words and grammatical structures commonly used by Malaysians (Ismail et al., 2007). It is widely used in daily communications, both informal and formal, such as news reports (Ismail et al., 2007). Malaysian English is considered a creole language, where it has the influence of Malay, Chinese, and Tamil together with Standard English. ME includes morphosyntactic and semantical adaptations (Imm, 2014). Listed below are some examples of morphosyntactic adaptations, the translation and meaning of these words can be referred to in the Appendix A:

- **Loan Words:** Words adopted from Malay and Chinese. Example: *nasi lemak, amah* and *ang pow*. (Tan, 2009) has also given some examples like: *kenduri, imam, ustaz, bumiputera, orang asli, Datuk, Makcik, Dewan Negara, Menteri Besar* and *Yang di-Pertua*.

- **Compound Blends:** Words combined from two different words. Example: *tidak apa attitude* and *Chinese sinseh*. Some examples from (Tan, 2009) are: *pondok school, Orang Asli Affairs, Tabung Haji Board* and *kampung house*.

- **Derived Words:** A morphosyntactic adaptions to make new words. Example: *datukship (adding suffix -ship), Johorean (adding suffix -ean)* and *non-halal (adding prefix non)*.

Although widely used in Malaysia, ME has not received much attention in the field of natural language processing (NLP) and is considered a low-resource language. Our goal is to help develop resources and facilitate future research in this area by introducing the Malaysian English News (MEN) Dataset.

*Corresponding Author.

1.2. Motivation

As discussed in Section 1.1, the differences between Standard English and Malaysian English make the data collection process challenging. Existing datasets, which are primarily based on Standard English (American or British English), are not appropriate for this study.

An effective RE model relies on the outcome of the Named Entity Recognition (NER) model. As noted in Section 1.1, entities in Malaysian English news articles exhibit morphosyntactic variations, which requires the expansion of existing NER solutions for accurate entity extraction. To evaluate the precision of existing NER solutions for Malaysian English news articles, we carried out an experiment using several selected tools, including spaCy (Montani et al., 2022), Flair (Akbik et al., 2019), Stanza (Qi et al., 2020), and Malaya (Husein, 2018). To ensure a fair comparison, these NER models are selected because they meet two criteria: i) It was trained using the OntoNotes 5.0 dataset, which is a widely used and comprehensive dataset for NER tasks. ii) It achieved the highest F1-Score among all available NER models provided by the tool we used.

We have conducted an experiment with 30 Malaysian English sentences to predict entities. The 30 sentences were manually annotated and compared with the entities predicted by the NER models. The chosen NER models achieved a micro F1-Score of less than **0.6** and the best performing NER tool is spaCy with F1-Score of 0.58. In Appendix 5 we can see the performance of other NER tools such as Flair, Stanza, and Malaya that achieved an F1-Score of 0.55, 0.43, and 0.55, respectively. In particular, the four NER models exhibited low effectiveness in predicting entities from the *LOC*, *GPE*, and *EVENT* categories. The result of the experiment has been shared in the Appendix B. This observation strongly suggests that existing models cannot accurately predict entities in Malaysian English.

In addition to the experimental results, our survey of existing datasets revealed that no Malaysian English dataset has ever been developed. Most existing RE datasets like DocRED (Yao et al., 2019), ACE-2005 (Walker, 2005), TACRED (Zhang et al., 2017), FewRel (Han et al., 2018), and CodRED (Yao et al., 2021) are based on Standard English. Considering these two motivations, we have created our annotated dataset explicitly tailored to the Malaysian English context. An annotated dataset in Malaysian English is crucial to handle the semantic and morphosyntactic adaptations present in Malaysian English news articles.

1.3. Contribution

The main contribution of this work is a Malaysian English News (**MEN**) dataset with annotated entities and relations. 200 Malaysian English News articles have been manually annotated by four well-trained human annotators. In total, we collected 6,061 annotated entities and 3,268 relation instances from the annotations. Based on observation, around 60% of the entities mentioned by *PERSON, ORGANIZATION, ROLE, TITLE* and *FACILITY* are very localized in Malaysian contexts and share the adaptation of Bahasa Malaysia. To our knowledge, this dataset is one of its kind, focusing specifically on Malaysian English news articles. This dataset can also be used for other NLP tasks like Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), Coreference Resolution and Semantic Role Labelling. Using the dataset, we have also fine-tuned the spaCy NER model. The importance of fine-tuning spaCy model is to find if there is any improvement in entity extraction from Malaysian English context. We evaluate the performance and how the fine-tuned NER model overcomes the gaps. The dataset together with annotation guideline has been published in: `https://github.com/mohanraj-nlp/MEN-Dataset`

The paper is structured as follows. Section 2 provides an overview of existing entity recognition and relation extraction datasets that are relevant to our work. Section 3 describes the news articles that we collected for human annotation. Section 4 discusses the fine-tuning of the spaCy model with the MEN-Dataset and evaluates the performance of the NER model. Finally, Section 5 concludes our work presented in this paper and points to potential future work.

2. Related Work

2.1. Existing Low Resource Language NER Dataset

Most of the prominent high-quality NER datasets like The Message Understanding Conference 6 (MUC-6) (Grishman and Sundheim, 1996), CoNLL-03 (Tjong Kim Sang and De Meulder, 2003) and OntoNotes 5.0 (Hovy et al., 2006) are mainly focused on Standard English. It is important to understand the annotation methodology of low resource language dataset in order to improve our annotation work. In conjunction with that, we have studied on several low resource NER dataset.

Wojood is a nested entity dataset developed for the Arabic language (Jarrar et al., 2022). The entity labels to annotate Wojood are adapted based on dataset OntoNotes 5.0. To ensure the quality of annotation, the annotator has proposed to calculate the Inter-Annotator Agreement (IAA) using

F1-Score. The higher the F1-Score, the higher the agreement between the annotators. Wojood is reported to have achieved an outstanding micro F1 score of **0.976**. (Buaphet et al., 2022) published a large-scale Nested Named Entity Recognition (NER) dataset for one of the Asian low-resource languages, Thai. The dataset covers 10 coarse grained entity labels like PERSON, ORGANIZATION, NUMBER, WORK_OF_ART, NORP, MISCELLANEOUS, LOCATION, DATE, FACILITY and EVENT. Another highlight of the dataset is that it includes nested entities with a maximum depth of 8 layers. The approach of labeling nested entities has been applied when annotating MEN-Dataset.

The datasets discussed above are representative of monolithic languages as they do not exhibit significant influences from other languages. However, in our current context, we are dealing with creole languages, where there is a notable influence of other languages with Standard English. MasakhaNER is one of the large-scale creole languages NER dataset that is built on 10 underpresented African languages (Adelani et al., 2021). Development of MasakhaNER has help in the development and evaluation of NER models for the 10 languages. The dataset has annotated with four entity types PERSON, ORGANIZATION, LOCATION and DATE. The annotation methodology discussed in the paper has helped us build on our annotation guideline, with some changes based on our objective. Before the development of MasakhaNER, another creole language-based dataset called NaijaNER (Oyewusi et al., 2021). NaijaNER is developed based on 5 Nigerian Languages (Nigerian English, Nigerian Pidgin English, Igbo, Yoruba and Hausa). NaijaNER has adapted the entity labels from OntoNotes 5.0 and consists of 18 entity labels.

Our initial study on low resource and creole language NER datasets has given us the understanding on the data collection process, annotation methodology, and gaps that have been solved by these datasets.

2.2. Existing Datasets for Relation Extraction

There are numerous Relation Extraction datasets available, where ACE-2005 stands out as one of the widely-used benchmark dataset. ACE-2005 (Walker, 2005) provides English, Arabic, and Chinese annotations and 18 relation labels. The high quality and extensive annotation make ACE-2005 popular among researchers. Apart from relation, ACE-2005 also includes annotated entities and events. It may provide little detail for annotating relations spanning multiple sentences or document-level relations.

DocRED (Yao et al., 2019) is a popular dataset for inter-sentential relation extraction models. DocRED captures relations between entities over several sentences in a document, unlike other relation extraction datasets. In document-level relation extraction research, this dataset helps researchers capture relations across phrases.

Thorough studies of ACE-2005 and DocRED provide us with a solid understanding of not only the basic approach used for annotating named entities and relations, but also relevant relation labels that can be incorporated into our dataset.

3. Malaysian English News (MEN) Dataset with Annotated Entities and Relation

3.1. MEN-Dataset Acquisition

A total of 14,320 news articles were scrapped from prominent Malaysian English news portals, including New Straits Times (NST) [1], Malay Mail (MM)[2] and Bernama English [3]. This 14,320 news articles are compilled as MEN-Corpus.

	MM	NST	Bernama	Total
Nation	1,938	3,185	2,890	8,013
Business	90	2,218	1,757	4,065
Sports	61	1,474	707	2,242
Total	2,089	6,877	5,354	14,320

Table 1: Number of news articles collected across news portals and categories.

Table 1 shows the statistics of MEN-Corpus, where the articles are scrapped from news categories like Nation, Business and Sports. We have selected 200 news articles from MEN-Corpus to develop the MEN-Dataset. One of the reasons for selecting only 200 articles from the 14,320 scraped news articles for annotation is to ensure that there is a representative sample of articles. We carefully curated this smaller subset of articles to ensure that the annotated dataset accurately reflects the broader distribution of linguistic patterns, entities, and relations present in the entire corpus. By opting for a smaller annotation set, we aimed to maximize the efficiency of the annotation process while still obtaining a dataset that offers valuable insight into entity and relation extraction tasks within the context of Malaysian English news articles. Additionally, working with a smaller set allowed us to perform thorough quality checks, ensuring the production of a high-quality annotated dataset despite resource limitations. The dataset consists of 120

[1] https://www.nst.com.my/
[2] https://www.malaymail.com/
[3] https://www.bernama.com/en/

news from category Nation, 60 news articles from Business, and 20 from Sports. Diverse categories in the corpus enable the construction of more robust NER and RE models.

3.2. Annotation Process

3.2.1. Annotation Setup

Annotators who are proficient in both Malaysian English and Bahasa Malaysia are required for the annotation task. The annotators were selected after they underwent training and assessments. Four annotators who performed well in the assessments were selected. The annotators are divided into two groups and each group is assigned to annotate 100 Malaysian English news articles within the 8 weeks milestone.

Figure 1 shows the annotation process we have followed for each milestone, where the first four focus on entity annotation, and the last four on relation annotation. To maintain consistency and ensure quality, we conducted an analysis of Inter-Annotator Agreement (IAA) at each milestone. Any disagreements that arise were resolved by an adjudicator. This approach aims to prevent annotation errors and produce high-quality datasets.

The annotation guideline [4] defines entity labels and gives examples of entity mentions. For relations, we provide the definition of relation labels and possible entity labels that can be associated with the relations. The guideline was prepared in an iterative manner, where it was first produced before the annotation started, then incrementally refined based on the annotation progress, with concise instructions.

3.2.2. Entity Annotation

The entity labels used for annotation are derived from OntoNotes 5.0 (Hovy et al., 2006). OntoNotes 5.0 has 18 entity labels, which include 11 named entities and 7 value entities. As value entities do not require morphological or syntactic adaptation, we only included named entities in our annotation rules. The entity labels adapted and used for entity annotation are PERSON, LOCATION, ORGANIZATION, NORP, WORK_OF_ART, LAW, LANGUAGE, FACILITY, and PRODUCT. To support Malaysian English, we also included two additional entities:

- *TITLE*: *TITLE* refers to academic, religious and Malay royalty titles. This is added to capture the morphological adaptation, specifically for *PERSON* entities that appear with their title, as part of their name in Malaysian English news. Example of entity mentions are *Datuk, Datin* and *Tan Sri*.
- *ROLE*: *ROLE* refers to the position that a *PERSON* is holding. The *ROLE* also enables us to capture the morphological adaptation. It usually comes together with the name of the *PERSON*. Example of entity mentions are *Mentri Besar, co-founder* and *police chief ACP*.

In Appendix C we have shared the list entity labels and descriptions. Figure 2 displays news articles annotated with *ORGANIZATION* and *LOCATION*. For nested entities, the annotators were instructed to follow a hierarchical approach, clearly marking the parent and child entities. The scenario of a nested entity will only happen for the entity mentioned *PERSON*. Consider this sentence as an example: *... his decision to support Prime Minister Datuk Seri Anwar Ibrahim was ultimately done for the sake of his constituents in suburban Selangor...* The annotator will need to annotate below entities as *PERSON*:

1. *Prime Minister Datuk Seri Anwar Ibrahim*
2. *Datuk Seri Anwar Ibrahim*
3. *Anwar Ibrahim*

Once the annotators completed the entity annotation, we calculated the IAA and will later be adjudicated if there are any disagreements.

3.2.3. Relation Annotation

The annotation relation labels are adapted from DocRED (Yao et al., 2019) and ACE-2005 (Walker, 2005), with some relation labels related to DATE or TIME omitted. To facilitate the annotation, we provide the annotators with a list of possible Argument Type. Argument Type aids annotators with possible entity types for each relation labels.

In Appendix D we have shared the list relation labels, descriptions, and the corresponding argument types. In Fig. 2, a relation named *manufacturer - DocRED* is used to link *Honda Malaysia* and *City Hatchback*. Referring to the annotation guideline for the relation, we can verify that the Argument Type for *manufacturer - DocRED* is *ORGANIZATION* and *PRODUCT*.

3.2.4. Inter-Annotator Agreement

In order to ensure quality annotations, the Inter-Annotator Agreement (IAA), a crucial metric for evaluating annotation made by human annotators, was used. The agreement was calculated by comparing the number of identical labels assigned by two or more annotators working on the same task (Artstein, 2017). We calculate the IAA separately for entity and relation annotations, as it will provide a comprehensive understanding of the accuracy and reliability of the annotations. Cohen's Kappa

[4] Annotation Guideline

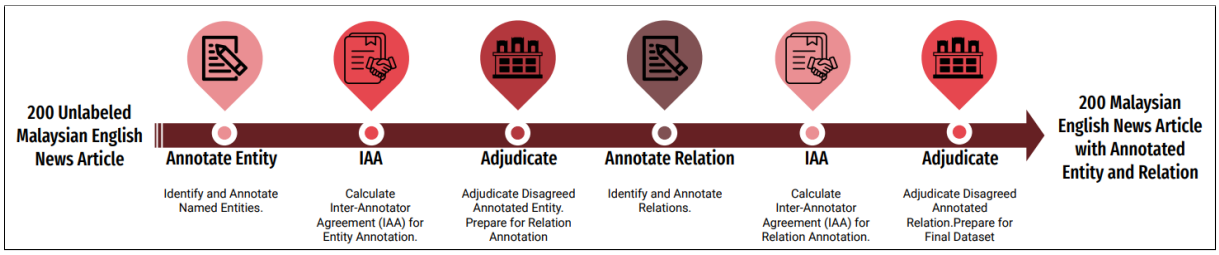

Figure 1: Phases in the annotation process to annotate news article for each milestone. This phase has helped ensure accuracy and consistency.

Figure 2: Snippet of news article that has been annotated with entities and relations. The annotated entities are underlined, and their corresponding entity labels are included below the line. Additionally, we linked the relations between two annotated entities. Each relation has a suffix of the dataset name, which indicates the dataset from which the relation label has been adapted.

was not used because it is suitable for tasks involving negative cases (Hripcsak and Rothschild, 2005; Campillos-Llanos et al., 2021). In our case, there are no negative cases in entity annotations.

In each group, an annotator is designated as the Gold Standard, and their annotations are compared to those of the other annotator in the same group to calculate the F1-Score. A higher F1-Score indicates a higher level of agreement between the two annotators in that group (Hripcsak and Rothschild, 2005).

IAA Analysis for Entity Annotation Two aspects were considered when calculating the F1-Score between the two annotators in each group: (i) the exact span of the entity mention and (ii) the entity label assigned to the entity mention. For the entity annotations, we achieved macro F1-Score of 0.818. Several datasets, such as (Jarrar et al., 2022; Brandsen et al., 2020; Jiang et al., 2022), evaluated IAA using F1-Score. (Jarrar et al., 2022) which adapted the entity labels from OntoNotes 5.0 and achieved an outstanding micro F1-Score of **0.976**. Our IAA is considered lower compared to (Jarrar et al., 2022). To assure dataset quality, we adjudicated disagreed entity annotations. A subject matter expect has been appointed as adjudicator, to evaluate and make final decision on the disagreed annotations.

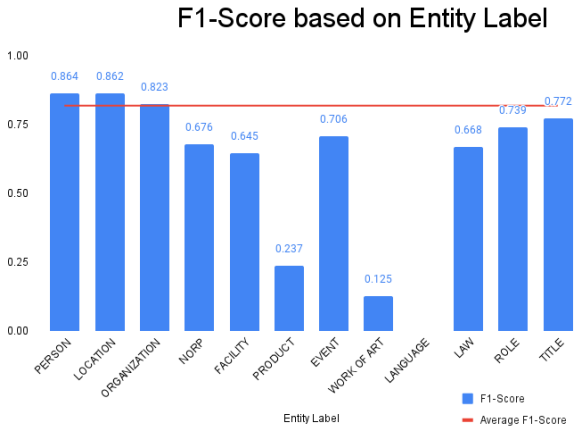

Figure 3: F1-Score calculated to measure the agreement based on entity labels.

Here are some observation of F1-Score from Figure 3:

1. For entity label *LANGUAGE*, the F1-Score shows NA. This happens because there is no entity mention that belongs to the annotated entity label. More details on the total number of entities mentioned can be found in Figure 4.

2. *PERSON* achieved the highest F1 score, which is 0.864, and the second highest is *LOCATION* with F1-Score of 0.862.

3. The entity label *WORK_OF_ART* achieved the lowest F1-Score, which is 0.125.

If we compare the F1-Score calculated across the entity labels (Fig. 3) and the total number of annotated entity mentions (Fig. 4), we can observe that entity labels with F1-Score below the average F1-Score have a total number of entity mentions less than 500 each. With limited examples to refer

to, annotators may have varying interpretations or perspectives on how to label entities correctly.

For instance, one of the entities mentioned in the entity label *WORK_OF_ART*, *Puteri Gunung Ledang* is a Fictional Character and according to the annotation guideline it should have been annotated as *PERSON*. However, the annotators annotated it as *WORK_OF_ART* because they mistaken it for the name of a film. Another example, *Most Established State in Healthcare Travel* is actually a recognition, therefore the mention is not suitable to be annotated as *WORK_OF_ART*. This issue can be resolved during adjudication. Based on the analysis, we found that the annotators had no difficulty handling annotation for an entity such as *PERSON*, *LOCATION* and *ORGANIZATION*. This is because the entity mentions that belong to these classes are clear and easily identifiable.

IAA Analysis for Relation Annotation For the IAA of Relation Annotation, we defined a criterion where both annotators have to agree on an exact match of relation instances. The macro F1-Score for overall relation annotation is **0.51**, which indicates a moderate level of agreement between the annotators in the relation annotation task.

No	Relation Label Adapted	IAA (F1-Score)
1	DocRED	0.567
2	ACE-2005	0.312
3	Overall	0.512

Table 2: F1-Score calculated to measure the agreement based on Relation Labels.

In Table 2, we examined the IAA with relation labels adapted from DocRED and ACE-2005 respectively. In particular, the IAA for ACE-2005 is lower compared to DocRED, despite ACE-2005 having fewer relation labels (17) than DocRED (84). In this research, we have also introduced an additional relation label, ***Others***, which can be used when no appropriate relation label is available for a given instance. During our post-annotation feedback session, it became apparent that the annotators encountered difficulties when annotating with relation labels from ACE-2005, in contrast to their experience with DocRED. ACE-2005 has limited relation labels available for specific scenarios and argument types. This narrower scope can make it more challenging for annotators to annotate appropriate relation labels. However, for each disagreement over the annotation of the relation, the adjudicator will make the final decision.

3.3. Dataset Statistics

In this section, we describe the dataset composition, including the distribution of annotated entities based on the labels and relations of selected entities. We also discuss the entity pairs and relations occurring together, providing valuable insights into the interconnectedness and contextual dependencies among entities and relations in MEN articles. Listed below are some statistics on the annotated MEN-Dataset:

1. Total Entities of 6,061 and 2,874 unique entities. The entities are annotated based on 12 entity labels.

2. Total Relation Instance of 4,095. We have 2,237 and 1,031 relation instances based on labels adapted from DocRED and ACE-2005 respectively. Around 827 relation instances have annotated with label Others.

Statistics for Entity Annotation Figure 4 shows the total and unique entity mentions that have been annotated. Most of the entities mentions annotated in the news articles are *PERSON, LOCATION* and *ORGANIZATION*. *WORK_OF_ART* is considered the entity label with the least mentions, and there are no entity mentions belonging to *LANGUAGE*.

From the 6,061 entity mentions that have been annotated, 67% of them are in Nation category, 22% in Business and 11% in Sports. Based on our further analysis, about 60% of the entities mentions from *PERSON, ORGANIZATION, ROLE, TITLE* and *FACILITY* are based on the context of Malaysian English and share the adaptation from the Malay language. Around 42% of the 1646 entity mentions for *PERSON* is associated together with an entity mention from *ROLE* or *TITLE*. We would like to highlight some examples of annotated entities that exhibit morphosyntactic adaptation (we will provide more samples in the Appendix E):

1. PERSON, TITLE:

   (a) Tan Sri Dr Noor Hisham Abdullah. “Tan Sri” is a loanword, it is a common honorific title given for PERSON.

   (b) Raja Permaisuri Agong: “Raja Permaisuri Agong” is a loanword, it is a Royal TITLE given for people from royal families.

2. ORGANIZATION:

   (a) Bank Negara Malaysia. Bank Negara Malaysia is considered a compound blend as it combines words from both English and Bahasa Malaysia. “Bank”, representing the institution’s. “Negara” is a Malay word meaning “nation”, “Malaysia” indicates geographical location.

3. NORP:

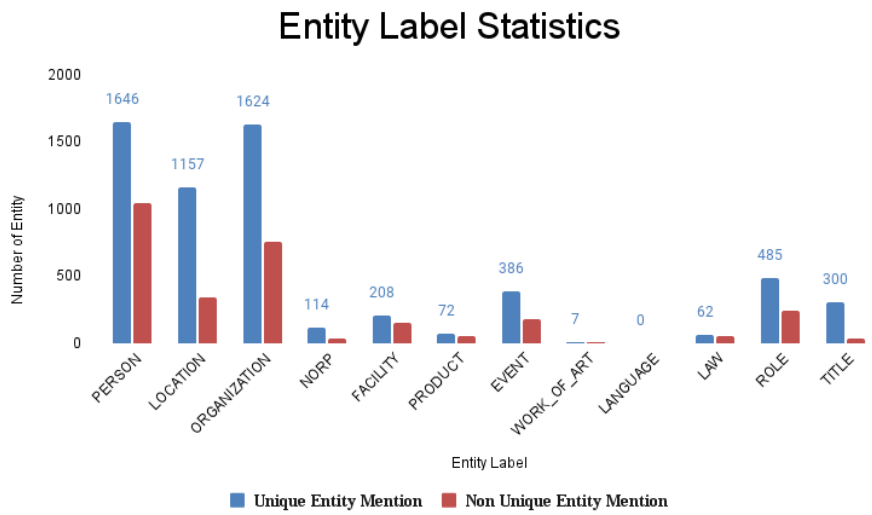

Figure 4: Total and Unique Entity Mention Annotated based on the Labels.

(a) Sarawakians. Sarawakians is a derived word that indicates the people of the state of Sarawak.

Statistics for Relation Annotation Referring to the dataset statistics listed in the start of Section 3.3, 826 instances were labeled as *Others*. Out of the 84 relation labels adapted from DocRED, only 50 unique relation labels were annotated by annotators. On the contrary, for ACE-2005, 16 out of the total of 17 relation labels were employed. Based on the distribution of annotated relation instances, 69% of the instances are based on the relation labels adapted from DocRED. In Appendix F, we have presented the distribution of relation labels in the MEN-Dataset. Since we are developing Document-Level Relation Extraction, we have designed the guidelines to ensure that the dataset covers relations from both inter-sentential and intra-sentential contexts. Out of the 3,268 relation instances, 54% are for intra-sentential relations, while the remaining 46% are for inter-sentential relations.

4. spaCy

4.1. Background

We investigate the contribution of MEN-Dataset in NER using spaCy (Montani et al., 2022). spaCy is a free and open source NLP library that supports task like Tokenization, NER, and POS Tagging. spaCy has 84 trained pipelines in several languages, including English. A pipeline is a sequence of processing steps applied to text in spaCy. Figure 5 presents the pipeline steps in spaCy. For English, spaCy has 4 trained pipeline

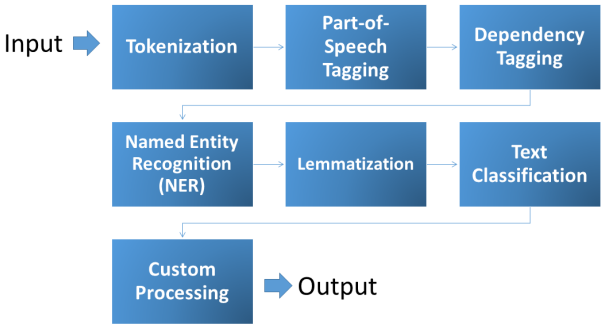

Figure 5: Processing steps in spaCy pipeline, it starts by giving the input. The outcome will include linguistic annotations after all the processing steps.

named “en_core_web_sm”, “en_core_web_md”, “en_core_web_lg”, “en_core_web_tf”.

1. en_core_web_sm: This is a small-sized model that is designed for more memory efficient. The small model does not have any word vector. For NER, the pipeline has been trained with OntoNotes 5.0 (Hovy et al., 2006) and achieved F1-Score of 0.845 [5].

2. en_core_web_md: This is a medium-sized model. The medium model has a reduced word vector table with 514k keys, 20k unique vectors. Similar to en_core_web_sm, the NER component is trained with OntoNotes 5.0 (Hovy et al., 2006) and achieved F1-Score of 0.852 [6].

[5] Model Metadata for en_core_web_sm
[6] Model Metadata for en_core_web_md

Entity Label	Total Annotated Entity in MEN-Dataset	Total Annotated Entity in Validation Set	spacy-blank	en_core_ web_sm	en_core_ web_sm-FT	en_core_ web_lg	en_core_ web_lg-FT	en_core_ web_trf	en_core_ web_trf-FT
PERSON	1646	108	0.982	0.39	0.852	0.555	0.806	0.811	0.923
LOCATION	1157	150	0.942	0.04	0.897	0.041	0.906	0.043	0.916
ORGANIZATION	1624	262	0.956	0.474	0.884	0.53	0.881	0.764	0.874
EVENT	386	30	0.892	0.15	0.771	0.205	0.779	0.25	0.842
PRODUCT	72	6	0.769	0	0.667	0	0.363	0	0.727
FACILITY	208	27	0.416	0	0.4	0	0.272	0	0.421
ROLE	485	35	0.968	0	0.361	0	0.75	0	0.925
NORP	114	5	1	0.625	0.833	0.416	0.8	0.588	1
TITLE	300	4	0.8	0	0	0	0	0	0.4
LAW	62	5	0.2	0.4	0.8	0.5	0.7	0.2	0.8
LANGUAGE	0	0	0	0	0	0	0	0	0
WORK_OF_ART	7	2	0.5	0	0.5	0	1	0	1
Total Entities	6061	634							
Overall Micro F1-Score			0.94	0.211	0.832	0.25	0.844	0.331	0.883

Table 3: Fine-Tuned model performance (based on F1-Score) on spaCy Model, calculated based on validation set for each entity labels.

3. en_core_web_lg: This is a medium-sized model. The medium model has word vector with 514k keys, 514k unique vectors. For NER, the model has been reported to achieve F1-Score of 0.854 [7].

4. en_core_web_tf: This model is based on the RoBERTa architecture. For NER, the model has been reported to achieve F1-Score of 0.90 [8].

4.2. Fine-Tuning spaCy NER Model

spaCy allows us to fine-tune the model using custom dataset. As described in Section 4.1, there are different components in the pipeline. spaCy supports training the pipeline by disabling or enabling specific components, depending on the use case. The MEN-Dataset is split into training (75%), test (10%) and validation (15%) total entities of 5065, 453 and 618 respectively. The splitted dataset is used to fine-tune spaCy NER model. In this experiment, we produce the following spaCy models:

1. spacy-blank: We loaded and trained a blank spaCy model, which has not been trained with any dataset before.

2. en_core_web_sm, en_core_web_lg, en_core_web_trf: These models have already been pre-trained with English datasets like OntoNotes 5.0 (Hovy et al., 2006).

3. en_core_web_sm-FT, en_core_web_lg-FT, en_core_web_trf-FT: The pre-trained model will be fine-tuned with our annotated MEN-Dataset.

Fine-tuning different spaCy models will allow us to identify the best performing spaCy NER model. To fine-tune, we started by converting our annotated dataset into BIO tagging scheme. Then we generated a Configuration File [9], which could be used to fine-tune NER model. Since the spaCy model does not support training for nested entities, we have adopted alternative strategies to address this limitation. For each nested entity, we will train the model separately with the article. Different nested entities within an article may exhibit unique characteristics in terms of context, structure, and relationships. Training the model separately allows each entity to be treated independently, capturing entity-specific patterns and nuances. As the objective of the research is to enrich the NER model specifically for Malaysian English, we limit the evaluation model to entity-specific evaluation. This means that we consider if different nested entities are correctly predicted in any part of the articles independently.

At the end of the training, spaCy will provide us with the best trained model (model-best) based on the use case. In our work, since we fine-tuned spaCy for NER, the best model will be selected based on the highest F1-Score achieved in the validation set.

4.3. Results and Analysis

Table 3 shows the result of evaluation on the fine-tuned spaCy NER model. To evaluate the performance of NER model, we used our validation set on en_core_web_sm, en_core_web_lg, and en_core_web_trf and consider the results as the baseline. The results were then compared with spacy-blank, en_core_web_sm-FT, en_core_web_lg-FT, and en_core_web_trf-FT. Table 3 has further details on the performance of the models by different entities labels.

[7] Model Metadata for en_core_web_lg
[8] Model Metadata for en_core_web_trf
[9] Generate Configuration File

Based on the result, we can observe that spacy-blank has achieved the highest F1-Score (0.94), while en_core_web_sm-FT has achieved the lowest F1-Score (0.832). On average, all the fine-tuned spaCy models have shown a significant improvement of +230%, which outperforms the pre-trained language model. If we compare with the fine-tuned pre-trained model, the average F1-Score of the spaCy model (F1-Score is 0.88) is higher than the further pre-trained model (F1-Score is 0.81). These are the findings learned from the experimental results:

1. Across all the fine-tuned spaCy models, there is an significant improvement on F1-Score. On average, the fine-tuned model (when compared with en_core_web_sm and en_core_web_sm-FT) has achieved F1-Score improvement of +401%.

2. All the three base spaCy models (en_core_web_sm, en_core_web_lg, en_core_web_trf) have obtained F1-Score of 0 for entity label PRODUCT, FACILITY, ROLE, TITLE, LANGUAGE, and WORK_OF_ART. It is important to note that the NER component in the spaCy model is trained on the OntoNotes 5.0 (Hovy et al., 2006) dataset, and in the MEN-Dataset we have followed 10 of 12 entity labels in OntoNotes. As such, apart from the entities labeled ROLE and TITLE, the base model was expected to predict entities for the remaining entity labels. However, it failed to do so in our experiment.

3. Referring back to the introduction of Malaysian English (in Section 1), entity label that exhibits morphosyntactic adaptions is TITLE, WORK_OF_ART, ORGANIZATION, LOCATION, and FACILITY. For these entity labels, we can see a significant improvement achieved by the fine-tuned NER models.

As the conclusion, our custom spaCy NER model called spacy-blank has achieved highest F1-Score compared to other fine-tuned spaCy models. Our fine-tuning approach has improved the performance of the spaCy NER model for Malaysian English. In Appendix G, we have presented samples of entities predicted by the spaCy NER model before and after fine-tuning with MEN-Dataset. This improvement is credited to the development of the MEN-Dataset.

5. Conclusion

This paper presents our endeavor to construct **MEN-Dataset**, a comprehensive Malaysian English news articles, annotated with entities and relations. Our thorough background studies highlight two limitations of existing state-of-the-art (SOTA) solutions related to Malaysian English. First of all, our literature review shows that there is no quality dataset on Malaysian English available. Second, the SOTA Named Entity Recognition tools are not able to fully capture the entities in Malaysian English news articles. To resolve the first limitation, we developed a human-annotated **MEN-Dataset**. MEN-Dataset contains 6,061 entities and 3,268 relation instances. Meanwhile, for the second limitation, we fine-tuned the performance of an existing state-of-the-art NER tool, namely spaCy and achieved significant improvement in NER. Fine-tuned on the MEN-Dataset, our **spacy-blank** achieves the highest F1-Score of 0.94. On average, all fine-tuned spaCy models achieve an improvement of +230%. For future work, we will be expanding the dataset using Human-In-The-Loop Annotation while ensuring that the quality of the dataset is preserved.

6. Ethical Consideration

This paper presents a new dataset, constructed following ethical research practice, as discussed below.

1. **Intellectual Property and Privacy**. The dataset is constructed using selected news articles published by a few news portals in Malaysia, namely New Straits Times (NST) [10], Malay Mail [11] and Bernama English [12]. Our institution's Human Research Ethics Committee was consulted prior to the start of the project. It was concluded that the news media do not require any ethics approval, as they were written and published for public consumption. In fact, analysis of news articles is commonly done by commentators. All entity mentions are based on the context of news articles.

2. **Compensation**. The annotators were compensated for the time they spent training, discussions, and the number of news articles they annotated.

7. Acknowledgements

Part of this project was funded by the Malaysian Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/ICT02/MUSM/02/2. We would like to express our appreciation for their support.

8. Bibliographical References

[10] https://www.nst.com.my/
[11] https://www.malaymail.com/
[12] https://www.bernama.com/en/

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

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.

Weerayut Buaphet, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Attapol Rutherford, and Sarana Nutanong. 2022. Thai nested named entity recognition corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1473–1486, Dublin, Ireland. Association for Computational Linguistics.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Med Inform Decis Mak*, 21(1):69.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, page 57–60, USA. Association for Computational Linguistics.

George Hripcsak and Adam Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12:296–8.

Zolkepli Husein. 2018. Malaya. https://github.com/huseinzol05/malaya.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

T.S. Imm. 2014. Exploring the malaysian english newspaper corpus for lexicographic evidence. 32:167–185.

Noriah Ismail, Normah Ismail, and Kamalanathan Ramakrishnan. 2007. Malaysian english versus standard english: Which is favored?

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, Henning Peters, Paul O'Leary McCann, jim geovedi, Jim O'Regan, Maxim Samsonov, Daniël de Kok, György Orosz, Marcus Blättermann, Madeesh Kannan, Duygu Altinok, Raphael Mitsch, Søren Lind Kristiansen, Edward, Lj Miranda, Peter Baumgartner, Raphaël Bournhonesque, Richard Hudson, Explosion Bot, Roman, Leander Fiedler, Ryn Daniels, kadarakos, Wannaphong Phatthiyaphaibun, and Schero1994. 2022. explosion/spaCy: v3.7.1: Bug fix for 'spacy.cli' module loading.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. 2021. Naijaner : Comprehensive named entity recognition for 5 nigerian languages.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316, Florence, Italy. Association for Computational Linguistics.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023. Finer: Financial named entity recognition dataset and weak-supervision model.

Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 25–32, Hissar, Bulgaria. Association for Computational Linguistics.

Siew Tan. 2009. Lexical borrowing in malaysian english: Influences of malay. *Lexis*, 3.

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.

Christopher Walker. 2005. *Multilingual Training Corpus LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium*.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721, Online. Association for Computational Linguistics.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4452–4472, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A. Words with morphosyntactic adaptations

Word	Meaning
kenduri	refers to feast.
iman	a PERSON who leads prayers in the mosque.
ustaz	a religious male teacher.
bumiputera	a Malaysian of indigenous Malay origin.
orang ali	a collective term for the indigenous peoples of Malaysia.
Datuk	a title denoting membership of a high order of chivalry, given to Males.
Makcik	used to call Aunty respectively.
Dewan Negara	is the upper house of the Parliament of Malaysia.
Mentari Besar	used to refer to important ministers from political party.
Yang Di Pertua	is a title for the head of state in certain Malay-speaking countries.
tiada apa attitude	refers to does not care attitude.
Chinese sinseh	a honorific term that used to refer "person born before another".
pondok school	organised religious schools for traditional Islam people.
Tabung Haji Board	the Malaysian hajj pilgrims fund board.
kampung house	a typical Malay house in villages.
Datukship	a title denoting membership of a high order of chivalry, given to Males.
Johorean	A native or inhabitant of Johor in Malaysia.
non halal	foods that not to be eaten by those observing Islamic teachings.
ang pows	a gift of money packed into a red packet.
bomohs	a Malay shaman and traditional medicine practitioner.
cheongsams	a dress worn by Chinese peoples/

Table 4: Refers to the meaning of words, examples with morphosyntactic adaptation.

B. Result of experiment conducted to evaluate performance of NER tools in Malaysian English News

API	Total Entity	Malaya		Flair		Stanza		spaCy	
Model		ontonotes-xlnet		ner-ontonotes-large		en-ontonotes-18classes		en_core_web_trf	
Evaluation		Correctly Classified	F1-Score	Correctly Classified	F1-Score	Correctly Classified	F1-Score	Correctly Classified	F1-Score
Entity Type									
PER	28	18	0.64	14	0.5	24	0.86	20	0.71
LOC	38	10	0.26	5	0.13	0	0	15	0.39
GPE	19	15	0.78	19	1	13	0.68	16	0.84
EVENT	6	4	0.75	4	0.67	1	0.17	1	0.17
ORG	21	14	0.67	18	0.86	13	0.62	14	0.67
FAC	10	6	0.60	5	0.5	3	0.3	6	0.60
WORK_OF_ART	1	0	0	0	0	0	0	1	1
NORP	7	5	0.71	7	1	2	0.29	3	0.43
Micro F1-Score	130	72	0.55	72	0.55	56	0.43	76	0.58

Table 5: Result of experimentation conducted with 30 sentences from Malaysian English News Article using NER Tool

*Note: To ensure a better readability of the Table, it has been moved to the Appendix.

C. List of Named Entity Labels

No	Entity Label	Description
1	PERSON	The Entity PERSON includes Name of Person in the text. This entity type has been adapted from OntoNotes 5.0.
2	LOCATION	LOCATION is any place that can be occupied by or has been occupied by someone in this EARTH and outside of EARTH. Entity mention that could be labelled as GPE has been labelled as LOCATION.
3	ORGANIZATION	ORGANIZATION is group of people with specific purpose.
4	NORP	NORP is the abbrevation for the term Nationality, Religious or Political group.
5	FACILITY	FACILITY refers to man-made structures.
6	PRODUCT	PRODUCT refers to an object, or a service that is made available for consumer use as of the consumer demand.
7	EVENT	An EVENT is a reference to an organized or unorganized incident.
8	WORK OF ART	WORK OF ART refers to ART entities that has been made by a PERSON or ORGANIZATION.
9	LAW	LAW are rules that has been made by an authority and that must be obeyed.
10	LANGUAGE	LANGUAGE refers to any named language.
11	ROLE	ROLE is used to define the position or function of the PERSON in an ORGANIZATION.
12	TITLE	TITLE is used to define the honorific title of the PERSON.

Table 6: List of 12 entity labels and description.

D. List of Relation Labels

No	Relation Label	Dataset Adapted	Entity Type One	Entity Type Two	Description
1	head of government	DocRED	PER	ORG,LOC	head of the executive power of this town, city, municipality, state, country, or other governmental body
2	country	DocRED	PER,ORG	LOC	sovereign state of this item (not to be used for human beings)
3	place of birth	DocRED	PER	LOC	most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
4	place of death	DocRED	PER	LOC	most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character
5	father	DocRED	PER	PER	"male parent of the subject."
6	mother	DocRED	PER	PER	"female parent of the subject."
7	spouse	DocRED	PER	PER	"the subject has the object as their spouse (husband, wife, partner, etc.)."
8	country of citizenship	DocRED	LOC	PER	the object is a country that recognizes the subject as its citizen
9	continent	DocRED	LOC	LOC	continent of which the subject is a part
10	head of state	DocRED	PER	LOC	official with the highest formal authority in a country/state
11	capital	DocRED	LOC	LOC	seat of government of a country, province, state or other type of administrative territorial entity
12	official language	DocRED	LOC,ORG	PER	language designated as official by this item
13	position held	DocRED	PER	ROLE	subject currently or formerly holds the object position or public office
14	child	DocRED	PER	PER	subject has object as child. Do not use for stepchildren
15	author	DocRED	PER	WORK_OF_ART	main creator(s) of a written work
16	director	DocRED	PER	WORK_OF_ART	director(s) of film, TV-series, stageplay, video game or similar
17	screenwriter	DocRED	PER	WORK_OF_ART	person(s) who wrote the script for subject item
18	educated at	DocRED	PER	ORG	educational institution attended by subject
19	composer	DocRED	PER	WORK_OF_ART	"person(s) who wrote the music"
20	occupation	DocRED	PER	ROLE	"occupation of a person"
21	founded by	DocRED	PER	ORG	founder or co-founder of this organization, religion or place
22	league	DocRED	ORG	EVENT	league in which team or player plays or has played in

23	place of burial	DocRED	PER	LOC	location of grave, resting place, place of ash-scattering, etc. (e.g., town/city or cemetery) for a person or animal. There may be several places: e.g., re-burials, parts of body buried separately.
24	publisher	DocRED	PER	WORK_OF_ART	organization or person responsible for publishing books, periodicals, printed music, podcasts, games or software
25	maintained by	DocRED	PER,ORG	FAC,ORG	person or organization in charge of keeping the subject (for instance an infrastructure) in functioning order
26	owned by	DocRED	PER	ORG, FAC, PRODUCT	owner of the subject
27	operator	DocRED	PER	PRODUCT,FAC	person, profession, or organization that operates the equipment, facility, or service
28	named after	DocRED	PER	FAC,ORG,EVENT	"entity or event that inspired the subject's name, or namesake (in at least one language)."
29	cast member	DocRED	PER	WORK_OF_ART	"actor in the subject production"
30	producer	DocRED	PER	WORK_OF_ART	person(s) who produced the film, musical work, theatrical production, etc. (for film, this does not include executive producers, associate producers, etc.)
31	award received	DocRED	PER, ORG, WORK_OF_ART, TITLE	WORK_OF_ART, TITLE	award or recognition received by a person, organization or creative work
32	chief executive officer	DocRED	PER	ORG	highest-ranking corporate officer appointed as the CEO within an organization
33	creator	DocRED	PER	WORK_OF_ART, PRODUCT	maker of this creative work or other object (where no more specific property exists)
34	ethnic group	DocRED	PER	ORG	subject's ethnicity (consensus is that a VERY high standard of proof is needed for this field to be used. In general this means 1) the subject claims it themselves, or 2) it is widely agreed on by scholars, or 3) is fictional and portrayed as such)
35	performer	DocRED	PER	WORK_OF_ART	actor, musician, band or other performer associated with this role or musical work
36	manufacturer	DocRED	ORG	PRODUCT	manufacturer or producer of this product
37	developer	DocRED	ORG,PER	PRODUCT,FAC	organization or person that developed the item
38	legislative body	DocRED	ORG	ORG	legislative body governing this entity; political institution with elected representatives, such as a parliament/legislature or council

39	executive body	DocRED	ORG	ORG	branch of government for the daily administration of the territorial entity
40	record label	DocRED	ORG	WORK_OF_ART	brand and trademark associated with the marketing of subject music recordings and music videos
41	production company	DocRED	ORG	WORK_OF_ART	company that produced this film, audio or performing arts work
42	location	DocRED	PER,FAC,ORG	LOC	location of the object, structure or event.
43	place of publication	DocRED	WORK_OF_ART	LOC	geographical place of publication of the edition (use 1st edition when referring to works)
44	part of	DocRED	PER	ORG,EVENT	"object of which the subject is a part (if this subject is already part of object A which is a part of object B, then please only make the subject part of object A)."
45	military rank	DocRED	PER	ROLE	"military rank achieved by a person (should usually have a ""start time"" qualifier), or military rank associated with a position"
46	member of	DocRED	PER	ORG	organization, club or musical group to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a political position, such as a member of parliament.
47	chairperson	DocRED	PER	ORG	presiding member of an organization, group or body
48	country of origin	DocRED	LOC	WORK_OF_ART, PRODUCT	country of origin of this item (creative work, food, phrase, product, etc.)
49	diplomatic relation	DocRED	ORG	ORG	diplomatic relations of the country
50	residence	DocRED	PER	FAC,LOC	the place where the person is or has been, resident
51	organizer	DocRED	PER,ORG	EVENT	person or institution organizing an event
52	characters	DocRED	PER	WORK_OF_ART	characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
53	lyrics by	DocRED	PER	WORK_OF_ART	author of song lyrics
54	participant	DocRED	PER,ORG	EVENT,ORG	"person, group of people or organization (object) that actively takes/took part in an event or process (subject)."
55	given name	DocRED	PER	PER	first name or another given name of this person; values used with the property should not link disambiguations nor family names
56	location of formation	DocRED	ORG	LOC	location where a group or organization was formed
57	parent organization	DocRED	ORG	ORG	parent organization of an organization.
58	significant event	DocRED	PER,ORG	EVENT	significant or notable events associated with the subject
59	authority	DocRED	PER	ORG	entity having executive power on given entity
60	sponsor	DocRED	PER,ORG	PER,EVENT	organization or individual that sponsors this item

61	applies to jurisdiction	DocRED	LAW	LOC	the item (institution, law, public office, public register...) or statement belongs to or has power over or applies to the value (a territorial jurisdiction: a country, state, municipality, ...)
62	director / manager	DocRED	PER	ORG	person who manages any kind of group
63	product or material produced	DocRED	PER	WORK_OF_ART	material or product produced by a government agency, business, industry, facility, or process
64	student of	DocRED	PER	PER	person who has taught this person
65	territory claimed by	DocRED	ORG	LOC	administrative divisions that claim control of a given area
66	winner	DocRED	PER,ORG	EVENT	"winner of a competition or similar event, not to be used for awards or for wars or battles"
67	replaced by	DocRED	PER	PER	"other person or item which continues the item by replacing it in its role."
68	capital of	DocRED	LOC	LOC	country, state, department, canton or other administrative division of which the municipality is the governmental seat
69	languages spoken, written or signed	DocRED	PER	LANGUAGE	language(s) that a person or a people speaks, writes or signs, including the native language(s)
70	present in work	DocRED	PER	WORK_OF_ART	this (fictional or fictionalized) entity or person appears in that work as part of the narration
71	country for sport	DocRED	PER,ORG	LOC	country a person or a team represents when playing a sport
72	represented by	DocRED	PER	ORG	person or agency that represents or manages the subject
73	investor	DocRED	PER,ORG	ORG	individual or organization which invests money in the item for the purpose of obtaining financial return on their investment
74	intended public	DocRED	PER,ORG	PRODUCT,EVENT	this work, product, object or event is intended for, or has been designed to that person or group of people, animals, plants, etc
75	partnership with	DocRED	ORG	ORG	partnership (commercial or/and non-commercial) between this organization and another organization or institution
76	statistical leader	DocRED	ORG,PER	EVENT	leader of a sports tournament in one of statistical qualities (points, assists, rebounds etc.).
77	board member	DocRED	PER	ORG	member(s) of the board for the organization
78	sibling	DocRED	PER	PER	"the subject and the object have at least one common parent (brother, sister, etc. including half-siblings)"
79	stepparent	DocRED	PER	PER	subject has the object as their stepparent
80	candidacy in election	DocRED	PER,ORG	EVENT	election where the subject is a candidate
81	coach of sports team	DocRED	PER	ORG	sports club or team for which this person is or was on-field manager or coach

82	subsidiary	DocRED	ORG	ORG	subsidiary of a company or organization; generally a fully owned separate corporation.
83	religion	DocRED	PER	ORG	religion of a person, organization or religious building, or associated with this subject
84	Physical.Located	ACE-2005	PER	FAC, LOC	Located captures the physical location of an entity.
85	Physical.Near	ACE-2005	PER, FAC, LOC	FAC, LOC	Indicates that an entity is explicitly near another entity.
86	Part-Whole.Geo	ACE-2005	FAC, LOC	FAC, LOC	Captures the location of FAC, LOC, or GPE in or at or as a part of another FAC, LOC or GPE.
87	Part-Whole.Subsidary	ACE-2005	ORG	ORG, LOC	Captures the ownership, administrative, and other hierarchical relationships between organizations and between organizations and GPEs.
88	Per-Social.Business	ACE-2005	PER	PER	Captures the connection between two entities in any professional relationships.
89	Per-Social.Family	ACE-2005	PER	PER	Captures the connection between one entity and another entity in family relations.
90	Per-Social.Lasting	ACE-2005	PER	PER	Captures the relations that invlovle personal contact (Where one entity has spent time with another entity, like classmate, neighbor), or indication that the relationships exists outside of a particular cited interaction.
91	Org-Aff.Employment	ACE-2005	PER	ORG,LOC	Captures relationship between Person and their employers.
92	Org-Aff.Ownership	ACE-2005	PER	ORG	Captures relationship between a Person and an Organization owned by that PERSON
93	Org-Aff.Founder	ACE-2005	PER,ORG	ORG,LOC	Captures relation between an entity and an organization that has been founded by the entity
94	Org-Aff.Student-Alum	ACE-2005	PER	ORG-Educational ONLY	Captures relation between Person and an educational institution.
95	Org-Aff.Sports-Affiliation	ACE-2005	PER	ORG	Captures relation between Player, Coach, Manager with their affiliated Sport ORG
96	Org-Aff.Shareholder	ACE-2005	PER, ORG, GPE	ORG, GPE	Captures the relation between an agent and an Organization
97	Org-Aff.Membership	ACE-2005	PER, ORG, GPE	ORG	Membership captures relation between an entity and organization which the entity is a member of

98	Agent-Artifact.UOIM	ACE-2005	PER, ORG, GPE	FAC	When an entity own an artifact, uses an artifact or caused an artifact to come into being.
99	Gen-Aff.CRRE	ACE-2005	PER	ORG, LOC	"When there is a relation between PER and LOC in which they have citizenship. Or when there is a relation between PER and LOC they live. Or when when there is a relation between PER and religious ORG or PER. Or when there is a relation between PER and LOC or PER entity that indicates their ethnicity"
100	Gen-Aff.Loc-Origin	ACE-2005	ORG	LOC	Captures the relation between an organization and the LOC where it is located.
101	Others		ANY ENTITY	ANY ENTITY	Can be used for any entity pair that does not have a suitable Relations Listed

Table 7: List of 101 relation labels and description.

E. More Samples from MEN-Dataset

In this appendix, we would like to provide few more samples of entity mention that exhibits morphosyntactic adaptation from MEN-Dataset.

1. PERSON, TITLE:
   (a) “Datuk Seri Najib Razak”. “Datuk Seri” is a loanword, it is a common honorific title given to "PERSON".
   (b) “Datin Norziati Othman”. “Datin” is a loanword, it is a common honorific title given to the wife of a man conferred with “Datuk”.
   (c) “Datuk Seri Tengku Zafrul Abdul Aziz”: “Tengku” is a loanword, it is a royal TITLE given to people from royal families. It is typically used to address or refer to members of royal families or individuals with noble lineage.
2. ORGANIZATION:
   (a) “Jabatan Imigresen Malaysia”, “Universiti Malaya”, “Menara Tabung Haji”, “Pakatan Harapan”, “Harimau Malaya”, “Bursa Malaysia”, “Parti Sarawak Bersatu”. These are some entities mentioned, which are considered direct borrowing and are frequently mentioned in Bahasa Malaysia and commonly found in Malaysian English articles.
   (b) “Public Mutual Berhad”, “Digital Nasional Berhad”, “Malaysian Takaful Association”. These are some entities that show compound blend, where part of the words are in Bahasa Malaysian and some are in English.
3. NORP:
   (a) "Bumiputera”. "Bumiputera” refers to the "sons of the soil" and includes ethnic groups recognized as indigenous or native to Malaysia.
   (b) "Malaysians”. "Malaysian" refers to Malaysian nationals or individuals who are citizens of Malaysia.
   (c) "Orang Asli”. "Orang Asli" refers to indigenous peoples of Peninsular Malaysia.
4. EVENT:
   (a) "Hari Keselamatan Internet 2022 – Edisi Malaysia”, "Santai Cycle Persatuan Keluarga Polis", "Aspirasi Keluarga Malaysia", "The Asset Triple A Islamic Finance Awards". These are some entities mentioned, which are considered direct borrowing and are frequently mentioned in Bahasa Malaysia and commonly found in Malaysian English articles.
5. EVENT:
   (a) "Hari Keselamatan Internet 2022 – Edisi Malaysia”, "Santai Cycle Persatuan Keluarga Polis", "Aspirasi Keluarga Malaysia", "The Asset Triple A Islamic Finance Awards". These are some entities mentioned, which are considered direct borrowing and are frequently mentioned in Bahasa Malaysia and commonly found in Malaysian English articles.
6. FACILITY:
   (a) “Sekolah Kebangsaan (SK) Parit Penghulu relief centre”, “Sri Subramaniar Temple Gunung Cheroh”, “Port Klang station”, “Sultan Ibrahim Stadium”, "Kampung Jirat bridge". These are some entities that show compound blend, where part of the words are in Bahasa Malaysian and some are in English.
   (b) "Jalan Sultan Iskandar”, "Sekolah Kebangsaan Taman Sri Muda 2", "Wisma OCM", "Dewan Serbaguna Benut". These are some entities mentioned, which are considered direct borrowing and are frequently mentioned in Bahasa Malaysia and commonly found in Malaysian English articles.

F. Statistics of Relation Labels in MEN-Dataset

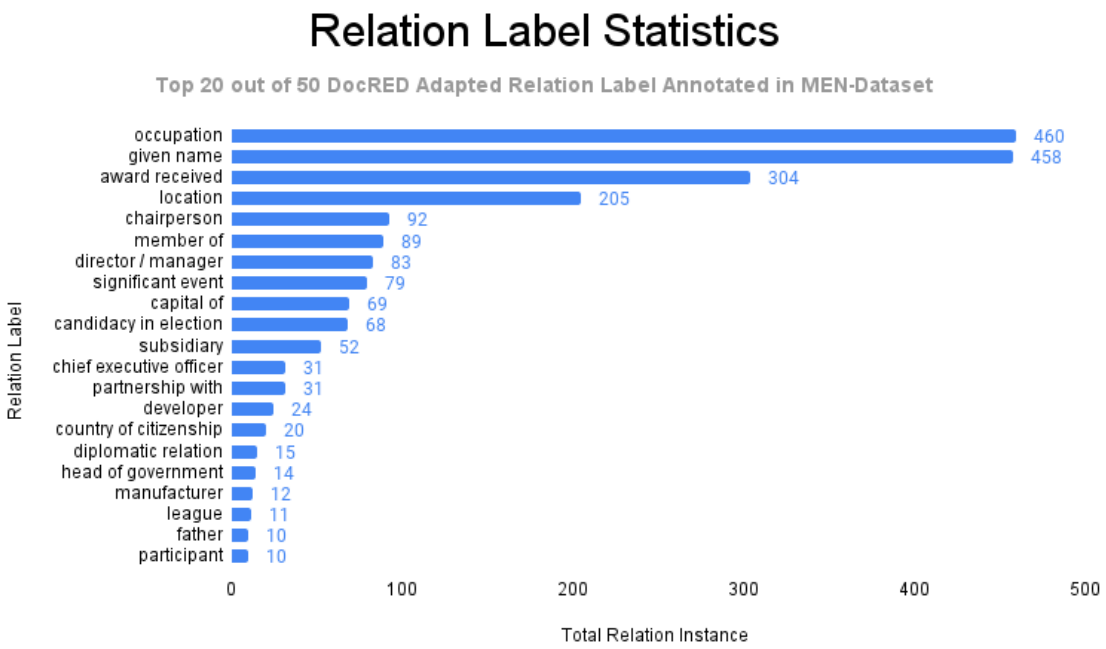

Figure 6: Top 20 Relation Labels adapted from DocRED Dataset. Out of 84 Relation Labels, only 50 Relation Labels are annotated

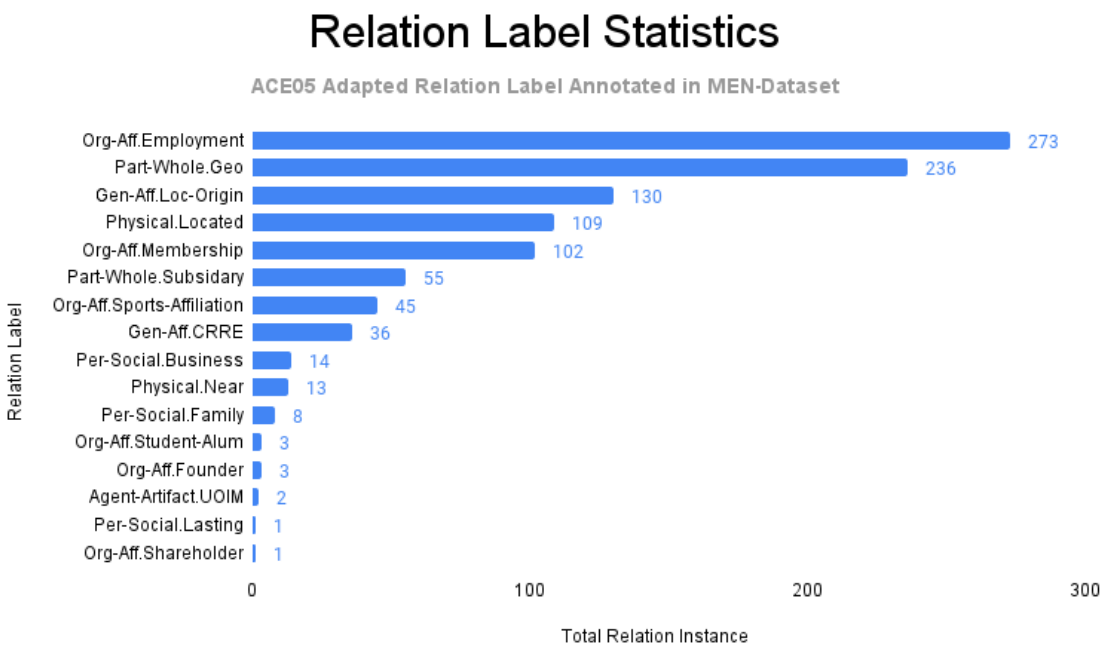

Figure 7: Relation Labels adapted from ACE-2005 Dataset

G. Sentence Samples from MEN-Dataset

Figure 8: Outcome of spacy-trf and spacy-blank when validated with news article from MEN-Dataset (Sample 1)

Figure 9: Outcome of spacy-trf and spacy-blank when validated with news article from MEN-Dataset (Sample 2)