How well ChatGPT understand Malaysian English? An Evaluation on Named Entity Recognition 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
{mohan.chanthran, soon.layki, ong.hueyfang}@monash.edu
[2]Valiantlytix
bhawani@valiantlytix.com

1 Result and Analysis

In this section, we present the outcome of the experiment that we conducted. In Section 1.1, we discuss how ChatGPT performs NER and RE on MEN-Dataset, together with the observed limitations.

1.1 *How well did ChatGPT perform in extracting entities from Malaysian English? Does it perform better?*

Figure 1 shows the experiment results using different prompt settings. Some observation made from Figure 1 are:

1. ChatGPT achieved highest F1-Score with prompt 3 Shot+Guideline+Explanation. From the overall experiment, the average F1-Score recorded was 0.488, and the highest F1-Score was 0.497. The result shows that providing a few shot samples with explanation and annotation guidelines enabled ChatGPT to do NER by complying with the instructions. Providing three-shot samples with annotation guidelines was sufficient for ChatGPT to understand the task and annotate.

2. The impact of the guidelines is significant in improving the performance of ChatGPT. Each non-consistent prompt technique with guidelines improved the performance of ChatGPT in comparison to outcome without guidelines.

3. Self-consistent technique is not effective in ensuring quality output by ChatGPT. If we compare the experiment results with and without self-consistent approach for zero-shot, the F1-Score with the self-consistent approach is lower. This shows that integrating the Self-Consistent technique with few shot learning approaches did not yield substantial improvements in all cases. However, this technique helps to ensure the consistency of the outcome.

4. Although we made multiple prompting strategies, the overall F1-score did not improve significantly. The overall difference of F1-Score recorded is 0.488 ± 0.01.

During the annotation of the MEN-Dataset, we calculated the Inter-Annotator Agreement (IAA) using the F1-Score and achieved a score of 0.81. Meanwhile, the highest F1-Score achieved by ChatGPT from this experiment was 0.497. This shows that there are still some limitations that can be observed from ChatGPT.

1.2 *What are the limitations of ChatGPT in extracting entities? Were there specific types of entity labels that ChatGPT consistently struggled to extract or misidentify?*

In Table 3, we can see the F1-Score from the perspective of entity label level. This helps us to understand more about how ChatGPT extracts the entities. We manually checked the outcome from ChatGPT to understand its limitation in extracting entities. The following findings were observed from the outcomes generated by self-consistent prompting:

1. Entity labels like PERSON, LOCATION, and ORGANIZATION have more than 1000 entity mentions annotated in MEN-Dataset. While the remaining entity labels have a total entity mention of less than 300.

2. The entity label PERSON has an average F1-Score of 0.507. Our analysis noticed that most errors happened due to Loan Words and Compound Blend found in Malaysian English news articles. Here are some examples:

   (a) Tan Sri Dr Noor Hisham Abdullah. "Tan Sri" is a loanword, a common honorific

*Corresponding Author.

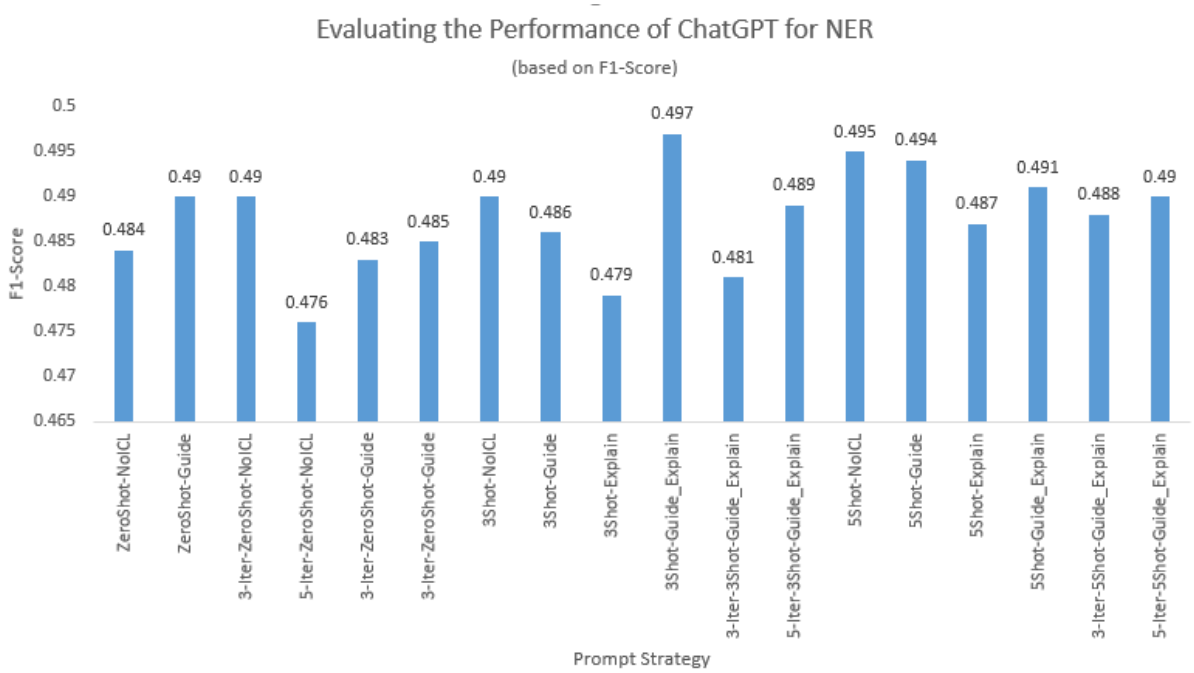

Figure 1: F1-Scores based on entities extracted by ChatGPT for Malaysian English news articles.

title for PERSON. It is often used to mention important personals. It is often used together with the name of PERSON.

(b) Datuk Seri Haji Amirudin bin Shari. "Datuk Seri" is a loanword, a common honorific title for PERSON.

Apart from the errors due to Loan Words and Compound Blend, ChatGPT did not extract any co-referring entities. For example, *Tan Sri Dr Noor Hisham Abdullah* is also used as *Noor Hisham Abdullah* in a similar article, but ChatGPT did not extract it.

3. For ORGANIZATION, we noticed the importance of providing annotation guidelines. Several entity mentions from ORGANIZATION were not extracted before including the guideline in the prompts. Examples of entity mention are: *Session Court, Public Mutual Funds, Parliment*. Furthermore, ChatGPT did not extract any abbreviations of entity mentions from entity label ORGANIZATION. Some examples:

   (a) *ATM*: The full form of ATM is "Angkatan Tentera Malaysia".

   (b) *Armada*: The full form of Armada is "Angkatan Bersatu Anak Muda".

   (c) *PN*: The full form of PN is "Perikatan Nasional".

   Similar issues observed for PERSON, where the co-reference of entity mentions was not extracted.

4. For NORP, we noticed most of the errors were due to *Derived Words*. For instance, *Sarawakians*, and *Indonesian*. The guideline included some examples for NORP, covering some frequently mentioned NORP, such as *Bumiputera, Non-Bumiputera* and *Malaysians*. The given examples were extracted correctly by ChatGPT. Apart from that, entity mentions with Loan Words like *1998 Reformasi movement* were not identified by ChatGPT correctly.

5. Most of the entities mentioned from FACILITY that were not extracted by ChatGPT are with characteristics Compound Blend. The entities mentions from FACILITY have both English and Bahasa Malaysia, such as *CIMB Bank Jalan Sagunting, Dataran Rakyat* and *Aulong Sports Arena*. In addition, ChatGPT misidentified some entity labels. For instance, the entity mentioned that was supposed to be predicted as FACILITY was mistaken as LOCATION, and vice versa. Some other examples:

   (a) *Kuala Lumpur International Airport*

should be labeled as FACILITY instead of LOCATION.

(b) *Jalan Langgak Golf* should be labeled as LOCATION instead of FACILITY.

(c) *Sibujaya public library* should be labeled as FACILITY instead of LOCATION.

6. Most of the entity mentions in WORK_OF_ART are based on local creative works, consisting of Compound Blend. Some examples are *Aku Mau Skola* and *Puteri Gunung Ledang*.

7. TITLE always appears together with the name of PERSON. In MEN-Dataset, the TITLE is annotated separately. The TITLE can be honorific or academic title. The honorific title consists of Loan Words like *Datuk, Datuk Seri, Datin, Tan Sri* and more.

In conclusion, ChatGPT did not work well in extracting entity mentions with Loan Words, Compound Blend, and Derived Words. Apart from that, ChatGPT did not extract any co-reference entity mentions. Furthermore, any abbreviations of entity mentions were also not extracted by ChatGPT.

1.3 *How accurate was ChatGPT in extracting relations between entities, and were there any notable errors or challenges?*

The MEN-Dataset was annotated based on the relation labels adapted from DocRED and ACE05. There is also a special relation label named NO_RELATION, which is annotated when no suitable relation labels exist for a particular entity pair. Due to the different characteristics of relation labels, we experimented with relation labels adapted from DocRED and ACE05 separately. We used prompt settings similar to the previous experiment.

Figure 2 shows the F1-Scores calculated based on the relations classified by ChatGPT for every entity pair. The average F1-Score for relation adapted from DocRED and ACE05 are 0.64 and 0.35 respectively. Some findings based on the results presented in Figure 2 are:

1. **In-Context Learning improved the performance of ChatGPT in identifying the relations**. In both zero-shot and few-shot scenarios, the performance of ChatGPT has improved when providing both guidelines and explanations.

2. **Explanations made limited impact.** Including explanations and a few shot samples does not improve this task's performance. This approach has somehow improved the performance of ChatGPT in extracting entities.

3. **5 Shot Learning slightly improved the performance of ChatGPT, compared to 3 Shot Learning of various prompting techniques.**

4. **Complexity of relation labels.** When comparing the performance of ChatGPT across the two datasets, it is evident that the DocRED dataset produces a higher F1-Score than the ACE dataset. This can be seen across all evaluated prompting techniques.

One interesting observation is that in MEN-Dataset, 20% of the relation triplets were labeled with NO_RELATION. However, ChatGPT labeled as high as 80% of the relation triplets as NO_RELATION. While no morphosyntactical adaptation is involved when predicting the relation, understanding the context of the news article will impact the performance of ChatGPT in predicting the relations. In conclusion, we have seen the gap of ChatGPT on RE task for Malaysian English news article. To better understand the gap between Malaysian English and the Standard English, another question that may arise is *How good is ChatGPT in NER and RE on Standard English?*

1.4 *How good is ChatGPT in predicting entities and relations from Standard English articles?*

In this experiment, we chose 195 articles with annotated entities and relations from DocRED. To ensure a valid comparison, we highlight some differences between MEN-Dataset and DocRED as follows:

1. In MEN-Dataset, we have 11 entity labels, while in the DocRED dataset, there are six entity labels. The overlapping entity labels are PERSON, ORGANIZATION, and LOCATION.

2. In MEN-Dataset, we have a total of 101 relations labels. There are 84 relation labels adapted from DocRED and 17 from ACE-05. Meanwhile, DocRED has 96 relation labels.

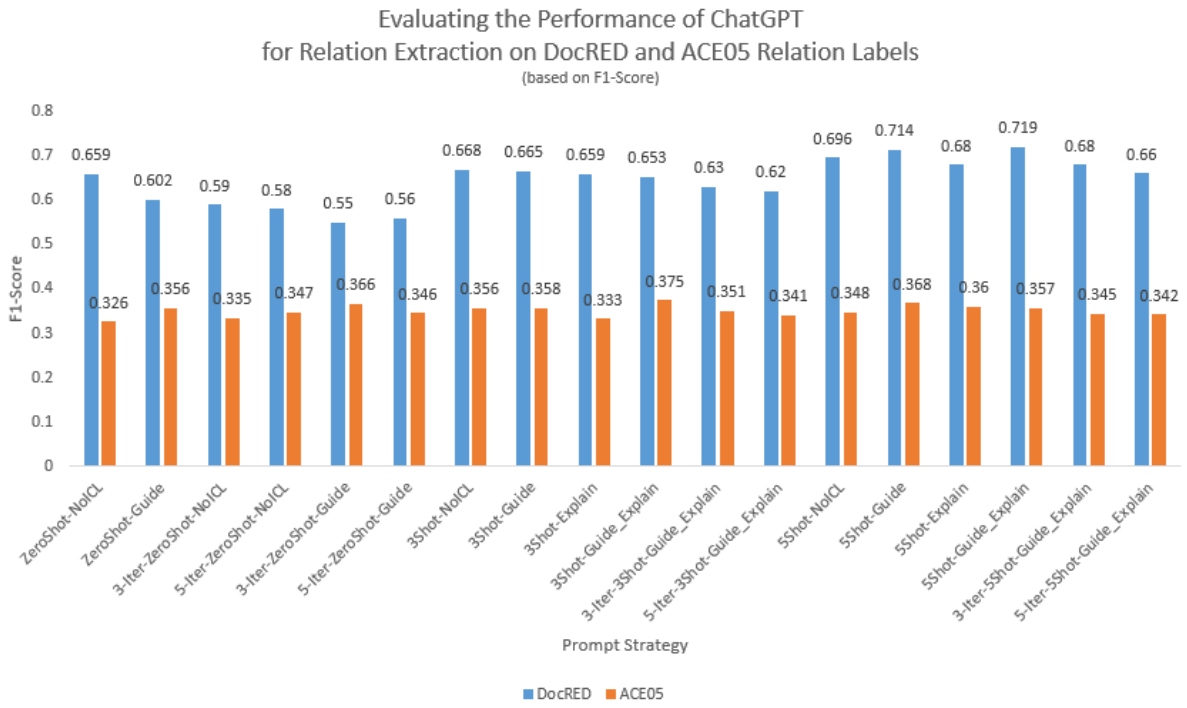

Figure 2: Performance of ChatGPT in classifying relations based on relation labels adapted from DocRED and ACE05

3. MEN-Dataset was developed from news articles while DocRED was developed using Wikipedia documents.

4. MEN-Dataset consists of news articles with a minimum of four and a maximum of 40 sentences, while the DocRED dataset has a minimum of 2 to a maximum of 20 sentences. The length of the article in DocRED is shorter than MEN-Dataset.

5. Most importantly, MEN-Dataset is based on Malaysian English, and DocRED is based on Standard English.

Both datasets feature document-based annotations and encompass both inter- and intra-sentential relations. As there are some differences between the two datasets, we made some modifications in the experiments:

1. For entity extraction, we compare the performance of ChatGPT based on entity label PERSON, ORGANIZATION, and LOCATION only.

2. For relation extraction, we compare the performance of ChatGPT based on overlapping 84 relations between MEN-Dataset and DocRED.

3. In the previous section, we evaluated the performance of ChatGPT based on 18 different prompt settings (refer to Appendix A). However, for the DocRED dataset, where the annotation guidelines for entity annotation and explanations for few-shot learning are not available, we specifically applied the following prompting techniques: ZeroShot-NoICL, 3-Iter-ZeroShot-NoICL, 5-Iter-ZeroShot-NoICL, 3Shot-NoICL, and 5Shot-NoICL (refer to Appendix A).

Prompt Name	**F1-Score (NER)**		**F1-Score (Relation Extraction)**	
	MEN-Dataset	DocRED	MEN-Dataset	DocRED
ZeroShot-NoICL	0.57	0.65	0.659	0.76
3-Iter-ZeroShot-NoICL	0.567	0.725	0.59	0.654
5-Iter-ZeroShot-NoICL	0.558	0.733	0.58	0.64
3Shot-NoICL	0.57	0.615	0.668	0.663
5Shot-NoICL	0.568	0.738	0.696	0.665

Table 1: Comparing the performance of ChatGPT between MEN-Dataset (Malaysian English) and DocRED (Standard English)

Table 1 presents the F1-Scores obtained for this experiment. It is noticeable that the performance of ChatGPT for NER varies significantly between the MEN-Dataset and DocRED datasets. For every prompt setting, the F1-Score for NER in DocRED (Standard English) is higher than MEN-Dataset

(Malaysian English). This language-specific performance could be due to the morphosyntactic adaptation that has been discussed and detailed in Section 1.2. Meanwhile, the performance of ChatGPT for Relation Extraction does not provide any significant difference between the two datasets. This could be due to the dataset's characteristics, where both were developed for inter- and intra-sentential relations. This result could also be due to morphosyntactic adaptation that can be seen in MEN-Dataset entities only, which does not impact Relation Extraction.

References

A Different Prompting Techniques

Prompt Name	Prompt Technique	ICL	Description
ZeroShot-NoICL	Zero Shot	None	Only news articles will be given to ChatGPT. Based on the existing knowledge, ChatGPT will need to extract entities and relation.
ZeroShot-Guide	Zero Shot	Guideline	Only annotation guideline will be provided to ChatGPT. ChatGPT will need to extract entities and relation based on guideline.
3-Iter-ZeroShot-NoICL	Self Consistent Zero Shot (3 Iteration)	None	Only provide news articles to ChatGPT. No additional context will be given. Based on the existing knowledge, ChatGPT will need to extract entities and relation.
5-Iter-ZeroShot-NoICL	Self Consistent Zero Shot (5 Iteration)	None	No additional context will be given to ChatGPT. The entity or relation that is consistently extract from similar news article will selected as final output.
3-Iter-ZeroShot-Guide	Self Consistent Zero Shot (3 Iteration)	Guideline	Annotation guideline will be given to ChatGPT. The entity or relation that is consistently extract from similar news article will selected as final output.
5-Iter-ZeroShot-Guide	Self Consistent Zero Shot (5 Iteration)	Guideline	Annotation guideline will be given to ChatGPT. The entity or relation that is consistently extract from similar news article will selected as final output.
3Shot-NoICL	3 - Shot Learning	None	Three news articles with entities and relation extracted will given as context to ChatGPT. ChatGPT will need to extract entities and relation based existing knowledge and provided sample news articles.
3Shot-Guide	3 - Shot Learning	Guideline	Together with three news articles, ChatGPT will be provided with annotation guideline. ChatGPT will need to extract entities and relation based existing knowledge and provided sample news articles.

3Shot-Explain	3 - Shot Learning	Explaination	Each instance of an entity and relation will be accompanied by an explanation for its extraction. ChatGPT's task will involve extracting entities and relations using the existing knowledge and information provided in the sample news articles.
3Shot-Guide_Explain	3 - Shot Learning	Guideline+Explanation	Each instance of an entity and relation will be accompanied by an explanation for its extraction. Additionally, the annotation guideline will also be give to ChatGPT. ChatGPT's task will involve extracting entities and relations using the existing knowledge and information provided in the sample news articles.
3-Iter-3Shot-Guide_Explain	Self Consistent Sampling (3 Iteration) + 3 - Shot Learning	Guideline+Explanation	Each instance of an entity and relation will be accompanied by an explanation for its extraction. Additionally, the annotation guideline will also be give to ChatGPT. ChatGPT's task will involve extracting entities and relations using the existing knowledge and information provided in the sample news articles. The entity or relation that is consistently extract from similar news article will selected as final output.
5-Iter-3Shot-Guide_Explain	Self Consistent Sampling (5 Iteration) + 3 - Shot Learning	Guideline+Explanation	Each instance of an entity and relation will be accompanied by an explanation for its extraction. Additionally, the annotation guideline will also be give to ChatGPT. ChatGPT's task will involve extracting entities and relations using the existing knowledge and information provided in the sample news articles. The entity or relation that is consistently extract from similar news article will selected as final output.
5Shot-NoICL	5 - Shot Learning	None	The explaination is similar to 3 - Shot Learning.
5Shot-Guide	5 - Shot Learning	Guideline	The explaination is similar to 3 - Shot Learning.
5Shot-Explain	5 - Shot Learning	Explaination	The explaination is similar to 3 - Shot Learning.
5Shot-Guide_Explain	5 - Shot Learning	Guideline+Explanation	The explaination is similar to 3 - Shot Learning.

3-Iter-5Shot-Guide_Explain	Self Consistent Sampling (3 Iteration) + 5 - Shot Learning	Guideline+Explanation	The explaination is similar to 3 - Shot Learning.
5-Iter-5Shot-Guide_Explain	Self Consistent Sampling (5 Iteration) + 5 - Shot Learning	Guideline+Explanation	The explaination is similar to 3 - Shot Learning.

Table 2: Different prompting techniques used to evaluate ChatGPT capabilities for NER and Relation Extraction

B Evaluating ChatGPT NER Capability with MEN-Dataset (From Perspective of Entity Label)

No	Prompt Name	PERSON (Total Entity: 1646)	LOCATION (Total Entity: 1157)	ORGANIZATION (Total Entity: 1624)	NORP (Total Entity: 114)	FACILITY (Total Entity: 208)	PRODUCT (Total Entity: 72)	EVENT (Total Entity: 386)	WORK_OF_ART (Total Entity: 7)	LANGUAGE (Total Entity: 0)	LAW (Total Entity: 62)	ROLE (Total Entity: 485)	TITLE (Total Entity: 300)
1	ZeroShot-NoICL	0.51	0.625	0.614	0.23	0.18	0.149	0.388	0	0	0.383	0.245	0
2	ZeroShot-Guide	0.503	0.632	0.615	0.265	0.22	0.139	0.399	0	0	0.464	0.266	0
3	3-Iter-ZeroShot-NoICL	0.5	0.621	0.616	0.25	0.19	0.123	0.412	0	0	0.392	0.346	0.041
4	5-Iter-ZeroShot-NoICL	0.497	0.61	0.603	0.182	0.175	0.116	0.366	0	0	0.391	0.301	0.021
5	3-Iter-ZeroShot-Guide	0.495	0.6	0.618	0.187	0.23	0.102	0.36	0	0	0.433	0.335	0.035
6	5-Iter-ZeroShot-Guide	0.51	0.617	0.618	0.29	0.21	0.138	0.356	0	0	0.364	0.176	0.032
7	3Shot-NoICL	0.51	0.615	0.615	0.172	0.23	0.115	0.364	0.054	0	0.463	0.321	0.04
8	3Shot-Guide	0.512	0.625	0.615	0.166	0.18	0.127	0.36	0	0	0.392	0.193	0.027
9	3Shot-Explain	0.511	0.62	0.603	0.193	0.211	0.129	0.325	0.031	0	0.475	0.31	0.051
10	3Shot-Guide_Explain	0.505	0.623	0.617	0.256	0.245	0.133	0.399	0	0	0.391	0.386	0.04
11	3-Iter-3Shot-Guide_Explain	0.509	0.606	0.598	0.227	0.165	0.117	0.362	0	0	0.409	0.307	0.032
12	5-Iter-3Shot-Guide_Explain	0.503	0.606	0.607	0.225	0.205	0.176	0.391	0	0	0.499	0.321	0.027
13	5Shot-NoICL	0.511	0.622	0.607	0.215	0.18	0.165	0.423	0	0	0.53	0.298	0.036
14	5Shot-Guide	0.508	0.614	0.618	0.195	0.216	0.13	0.406	0	0	0.531	0.378	0.036
15	5Shot-Explain	0.507	0.611	0.591	0.215	0.235	0.134	0.418	0	0	0.385	0.372	0.041
16	5Shot-Guide_Explain	0.51	0.623	0.609	0.201	0.263	0.136	0.381	0	0	0.374	0.305	0.066
17	3-Iter-5Shot-Guide_Explain	0.512	0.617	0.612	0.236	0.225	0.151	0.398	0	0	0.341	0.266	0.059
18	5-Iter-5Shot-Guide_Explain	0.511	0.607	0.609	0.221	0.247	0.09	0.366	0	0	0.474	0.36	0.038
	Average F1-Score	0.507	0.616	0.61	0.218	0.212	0.132	0.382	0.005	0	0.427	0.305	0.035

Table 3: The F1-Score from the perspective of entity label.