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

Abstract

Recently, ChatGPT has attracted a lot of interest from both researchers and the general public. While the performance of ChatGPT in named entity recognition and relation extraction from Standard English texts is satisfactory, it remains to be seen if it can perform similarly for Malaysian English. Malaysian English is unique as it exhibits morphosyntactic and semantical adaptation from local contexts. In this study, we assess ChatGPT's capability in extracting entities and relations from the Malaysian English News (MEN) dataset. We propose a three-step methodology referred to as ***educate-predict-evaluate***. The performance of ChatGPT is assessed using F1-Score across 18 unique prompt settings, which were carefully engineered for a comprehensive review. From our evaluation, we found that ChatGPT does not perform well in extracting entities from Malaysian English news articles, with the highest F1-Score of 0.497. Further analysis shows that the morphosyntactic adaptation in Malaysian English caused the limitation. However, interestingly, this morphosyntactic adaptation does not impact the performance of ChatGPT for relation extraction.

1 Introduction

With the recent emergence of Large Language Models (LLM), we have observed a paradigm shift in natural language processing (NLP). These LLM include PaLM (Chowdhery et al., 2022), ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023), and Llama 2 (Touvron et al., 2023). ChatGPT, inarguably the most popular LLM currently, is developed by OpenAI and has demonstrated remarkable ability in language understanding and generating coherent responses. The use of ChatGPT has been observed in various NLP tasks, including Sentiment Analysis (Wang et al., 2023; Belal et al., 2023), Topic Classification (Reiss, 2023; Gilardi et al., 2023), and Information Extraction (Wei et al., 2023; Li et al., 2023; Hu et al., 2023). There have been several research works conducted to evaluate the capabilities of ChatGPT for NER and RE (Li et al., 2023; Han et al., 2023; Chan et al., 2023). While most of the evaluation outcomes focused on Standard English, it raises a question: *Is ChatGPT capable of extracting entities and relations from Malaysian English News*?

[*]Corresponding Author.

Originating from Standard English, Malaysian English (ME) has evolved into a unique form of English incorporating local words from languages like Bahasa Malaysia, Chinese and Tamil (Ismail et al., 2007). Malaysian English exhibits usage of Loan Words, Compound Blends and Derived Words (Imm, 2014). Some example sentences with the usage of Loan Words, Compound Blends and Derived Words are provided, such as:

1. "... billion of jobs in the next five to seven years, as well as Bukit Bintang City Centre with RM600 million jobs awarded so far". From this sentence, Bukit Bintang City Centre is a compound blend where "Bukit Bintang" refers to the name of LOCATION in Bahasa Malaysia, and this entity refers to a shopping mall (LOCATION).
2. "... economy to provide higher-paying jobs in cutting-edge technology for Selangorians, he said". From this sentence, "Selangorians" is a derived word that indicates the people from the state of Selangor.
3. "KUALA LUMPUR: Prime Minister Datuk Seri Anwar Ibrahim today urged ... business tycoon Tan Sri Syed Mokhtar Albukhary ...". From this sentence, "Datuk Seri" and "Tan Sri" is a loanword, it is a common honorific title given for PERSON.

The existence of loan words, compound blends, and derived words in the usage of entity mentions has motivated us to assess the performance of ChatGPT in Malaysian English, specifically for Named Entity Recognition (NER) and Relation Extraction (RE).

Prompting techniques like Zero Shot, Few Shot, and Chain of Thought (CoT) have been proven to improve the performance of ChatGPT in various NLP tasks (Brown et al., 2020; Han et al., 2023; Chan et al., 2023; Wei et al., 2022). In-context learning helps ChatGPT to understand more about the task in hand and define the scope on the task to be completed. It has been proven effective for domain-specific tasks, such as legal reasoning (Kang et al., 2023). Keeping these in mind, we propose a novel three-step method to extract the entities and relations from Malaysian English news articles, called "*educate-predict-validate*". Section 3 discusses these three steps in details.

ChatGPT's ability to extract entities and relations is measured based on its agreement with human-annotated labels using the F1-Score. Our evaluation aims to establish a benchmark for ChatGPT's performance in Malaysian English texts. The code for this experiment is available at Github[1] for reproducibility. The contributions of this research can be summarised as follows:

1. *In-context learning for better ChatGPT performance*. A novel approach to identify and extract entities and relations from any document or text by providing sufficient contexts to ChatGPT.

2. *Comprehensive assessment of ChatGPT performance on Malaysian English News Articles*. A total of 18 different prompt settings have been carefully engineered to evaluate ChatGPT's capability in NER and RE. The output produced by ChatGPT is compared against human-annotations.

In short, the analyses reported in this paper answer these questions: a) *How well does ChatGPT perform in extracting entities from Malaysian English?*; b) *Are there specific types of entity labels that ChatGPT consistently struggle to extract or misidentified?*; c) *How accurate is ChatGPT in extracting relations between entities?*; d) *How good is ChatGPT in predicting entities and relation from Standard English?*.

Section 2 presents the evaluation done on ChatGPT for Standard English. Section 3 discusses our proposed "*educate-predict-validate*" methodology. Section 4 describes our experimental setup. Section 5 presents our experiment results and findings, including an analysis of the challenges and limitations encountered by ChatGPT when handling Malaysian English news articles. Finally in Section 6 we have concluded our work and our future work.

2 Related Work

2.1 LLM for Information Extraction

To understand the capabilities of LLM on entity and relation extraction, we have gone through some recent research on LLM for Information Extraction (IE). (Wei et al., 2023) has proposed ChatIE, a zero-shot information extraction framework using ChatGPT. The information extraction task will be conducted into two stages and it will be based on question-answering approach. In the first stage, a sentence will be passed to ChatGPT followed by a question asking whether the sentence contains any entities, relations, or event types from a predefined list. The question prompt will include the list of entity, relation, or event types. In the second stage, the prompt will be modified depending on the specific task. For NER, the entity type extracted from first stage will be given to ChatGPT to extract all entity mentions. Meanwhile, for RE, both entity type and relation type will be given to ChatGPT to identify entity mentions that match with the entity type and relation. ChatIE improves performance by an average of 18.98% compared to ChatGPT without ChatIE. However it is noticeable that the F1-score varies depend on the dataset that has been tested upon.

(Li et al., 2023) assesses the ability of ChatGPT in 7 Fine Grained IE tasks like Entity Typing, NER, Relation Classification, and RE. The prompt is formulated by considering two distinct configurations: Standard-IE settings and OpenIE settings. Compared to the baseline and SOTA models, ChatGPT's performance is less competent. For NER tasks, ChatGPT performance is lower in OntoNotes (with 18 Labels) compared to ConLL (4 Labels). For relation classification and RE, ChatGPT performance is lower in TACRED (42 Labels) compared to Se-

[1] https://github.com/mohanraj-nlp/ChatGPT-Malaysian-English

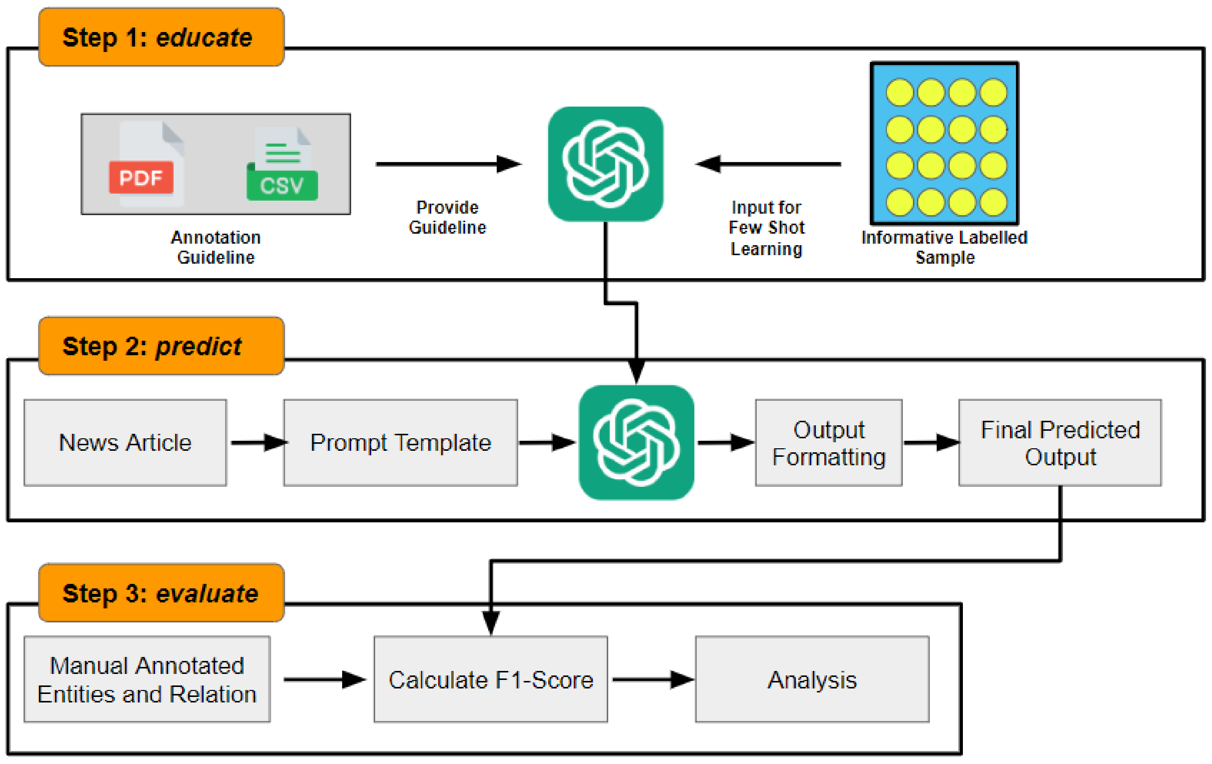

Figure 1: Detailed steps in the proposed *educate-predict-evaluate* methodology

mEval2010 (10 Labels).

(Han et al., 2023) conducted an extensive evaluation to examine the performance of ChatGPT in IE. A total of 14 subtasks related to IE were tested using 17 distinct datasets. The experimental conditions employed in this study encompass three prompt settings: zero shot prompt, few shot prompt, and few shot with CoT prompts. The experiments conducted evaluated several subtasks that are relevant to our research, including NER-Flat, NER-Nested, Relation Triplet (RE-Triplet), and Relation Classification (RE-RC). The experimental results showed that ChatGPT exhibited superior performance in the NER-Flat task as compared to the NER-Nested task. The F1-Score for RC-RE reached its lowest value at 19.47 when evaluated on the TACRED dataset under zero shot conditions. In the case of RE-Triplet, the dataset NYT-multi exhibited the lowest F1-Score, which amounted to 3.45. The experimental results also indicated that ChatGPT did poorly in relation classification for entities, with its lowest performance observed in triplet extraction.

3 *educate-predict-evaluate*

ChatGPT is one of the widely used Large Language Models. It can be easily interacted through the provided Web interface, by asking questions and make conversation with the model. Providing additional context helps ChatGPT to learn and better understand the tasks in hand. In this paper, we propose a systematic methodology called *educate-predict-evaluate*, which aims to carry out a comprehensive evaluation on ChatGPT capability in NER and RE within Malaysian English context. Figure 1 shows detailed view of proposed approach.

1. *educate*: The idea behind this is to teach ChatGPT how to extract entity and relation from Malaysian English texts. To accomplish this, we provided ChatGPT with the annotation guideline prepared while developing MEN-Dataset. This approach is also called as In-Context Learning (ICL). Appendix A shows a sample of prompt generated with annotation guideline for extracting entities. Apart from guideline, we also applied Few Shot Learning approach. In Few Shot Learning, we provided a few news articles with annotated entities and relations. In addition, we also provided some explanations that include the context, or justifications on why entities and relations are extracted from news article. These explanations were provided by the human annotators who contributed to developing and annotat-

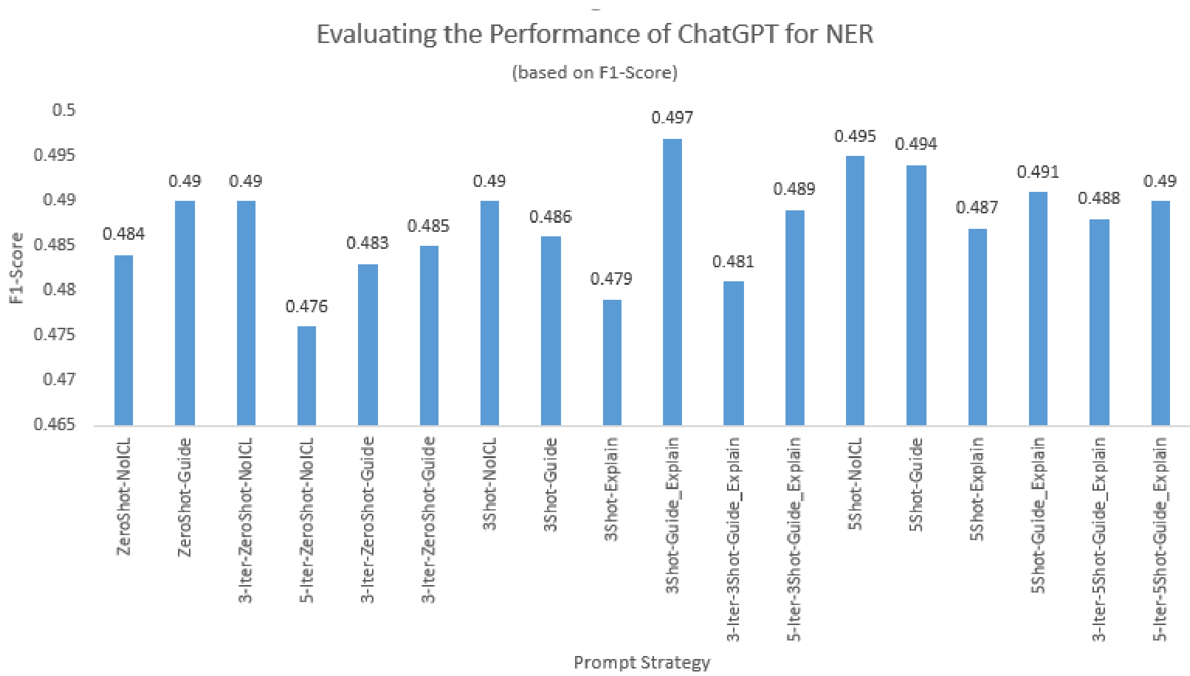

Figure 2: F1-Scores based on entities extracted by ChatGPT for Malaysian English news articles.

ing MEN-Dataset. Appendix B presents some samples of explanations given for entity extraction.

2. *predict*: We propose a Self-Consistent Few Shot Prompting Technique, together with the explanation on why each entity has been annotated by the human annotator. The explanation acts as additional context for ChatGPT to identify the entities and relations. (Wang et al., 2022) proposed the Self-Consistent prompting techniques, where the idea behind is to choose the most consistent answer as the final answer of ChatGPT. For instance, a prompt for a chosen news article will be provided to ChatGPT three times, and the entities that have been extracted more than twice will be considered as final output for the particular news article. In Table 5, we have listed all 18 different prompt settings used in this experiment. Appendix C presents the prompt used to extract entities while Appendix D presents the prompt used to identify relations from news articles.

3. *validate*: We have assessed the performance of ChatGPT on NER and RE by calculating the F1-Score with human annotation provided by the dataset.

Statistics	Frequency
Total Entities	6,061
Total Unique Entities	2,874
Total Relations	3,268
Total Relation based on DocRED Labels	2,237
Total Relation based on ACE-2005 Labels	1,031

Table 1: The statistics of total Entities and Relation annotated in MEN

4 Experiment

4.1 Dataset

We used two datasets to evaluate the performance of ChatGPT for NER and RE, which include:

1. MEN-Dataset is a Malaysian English news article dataset with annotated entities and relations. We have built the dataset with 200 news articles extracted from prominent Malaysian English news articles portals like New Straits Times (NST)[2], Malay Mail (MM)[3] and Bernama English[4]. The dataset consists of 11 entity labels, and 101 relation

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

labels. Appendix E and Appendix F contain the complete lists of entity and relation labels respectively. For entities, we have adapted the labels from dataset OntoNotes 5.0 (Hovy et al., 2006). The relation labels are adapted from ACE05 (Walker, 2005) and DocRED (Yao et al., 2019). Table 1 presents the statistics of the entities and relations annotated in the dataset.

2. DocRED: DocRED (Yao et al., 2019) is a prominent dataset designed specifically for inter-sentential relation extraction models. The dataset includes annotated entities and relations. The dataset has been chosen to facilitate a comparative analysis of ChatGPT's performance in both Malaysian English and Standard English.

While we have adapted entity labels from OntoNotes 5.0 and relation labels from ACE 05, we did not use these datasets for this evaluation. The OntoNotes 5.0 dataset is structured at the sentence level, with entity annotations specific to each individual sentence. An earlier effort showed that ChatGPT does not perform well on longer text (Han et al., 2023). To mitigate the impact of input length on ChatGPT's performance, we have opted to utilize a dataset containing longer context sequences. This decision led us to select DocRED for evaluation. It is also important to note that the MEN dataset encompasses both inter and intra-sentential relations.

4.2 Experimental Setup

The experiment was conducted in between April 2023 and August 2023. Notably, the outcome of ChatGPT exhibited variability over time (Chen et al., 2023). While OpenAI API is available, we decided to use ChatGPT[5] official website. There were several reasons for our decision, and these have been discussed in Section 8. To ensure a fair comparison, we used 195 articles for experiment. Another five articles were used for Few-Shot learning context. The In-Context Learning technique involves the integration of annotation guidelines and/or a limited set of few-shot samples as input of ChatGPT. During the process of picking few-shot samples, we implemented a filtering mechanism to identify and prioritize samples that possess the highest quantity of annotated entities or relation labels. For NER, we provided articles as input; meanwhile, for RE, we provided articles and entity pairs. For the evaluation metrics, we utilized F1-Score, and Human Validation, as mentioned in Section 5. The F1-Scores were calculated by comparing ChatGPT's predictions with human annotations in the dataset.

5 Result and Analysis

In this section, we present the outcome of the experiment that we conducted. In Section 5.1, we discuss how ChatGPT performs NER and RE on MEN-Dataset, together with the observed limitations.

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

Figure 2 shows the experiment results using different prompt settings. Some observation made from Figure 2 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 strate-

[5] https://chat.openai.com/

gies, 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.

5.2 *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 6, 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 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.

5.3 *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 3 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 3 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?*

5.4 *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.

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

4. MEN-Dataset consists of news articles with

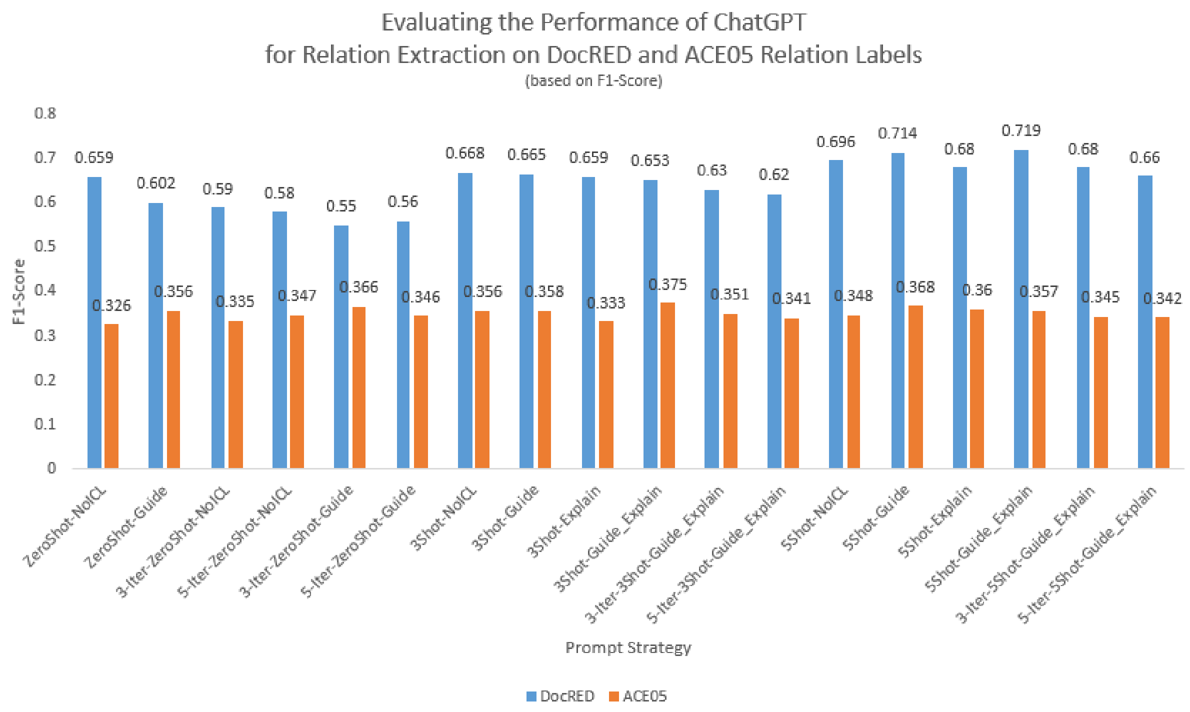

Figure 3: Performance of ChatGPT in classifying relations based on relation labels adapted from DocRED and ACE05

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 G). 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 G).

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

Table 2 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 5.2. 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.

6 Conclusion

In this paper, we comprehensively evaluated and analyzed ChatGPT's ability to extract entities and classify relations from Malaysian English news articles. Our extensive experiment was conducted with 18 different prompting approaches. The experimental results prove that morphosyntactic adaptation impacted the performance of ChatGPT in extracting entities from Malaysian English news articles. We discussed our findings from the experiments, including an analysis of the limitations of ChatGPT. ChatGPT could not achieve satisfying performance when extracting entities from Malaysian English news articles. Apart from the limitation in understanding the context of inputs, there are a few factors that influenced the performance of ChatGPT. These include the dataset's characteristics, additional contexts like guidelines and explanations, and several few-shot examples. The morphosyntactic adaptation exhibited by Malaysian English influenced the performance of ChatGPT for NER. Given the annotation of our MEN-Dataset, we could only assess the performance of ChatGPT in NER and RE. For future work, we plan to expand our evaluations by incorporating a broader range of NLP downstream tasks. Furthermore, we will extend our assessment to include other language models, such as GPT-4 (OpenAI, 2023) and Llama 2 (Touvron et al., 2023), for NER and RE tasks, specifically in the context of Malaysian English. Finally as a future work, we will also expand the coverage of our experiment with different prompting techniques to ensure our evaluation is statistically significant.

7 Ethical Consideration

In this paper, we evaluated the performance of ChatGPT in extracting entities and relations from Malaysian English news articles. The evaluation was done using news articles (from MEN-Dataset) and Wikipedia articles (from DocRED dataset). No ethics approval was required because these articles were written and published for public consumption. This decision is made after consulting our institution's Human Research Ethics Committee. Besides, ChatGPT was only used to extract information (like entities and relation) from our input and it does not require generating any responses that poses harmful or inappropriate content. As mentioned in Section 4.2, we used ChatGPT [6] official website and we sent the input one by one, without spamming the website.

8 Limitations

Here are some of the limitations in this experiment:

1. As explained in the Introduction (Section 1), various Information Extraction tasks can be done using ChatGPT. However, in this research paper, we focused only on NER and RE due to the annotation of our Malaysian English dataset. In future, we will expand our dataset to cater for other NLP tasks.

2. Secondly, we could only conduct the experiments reported in ths paper with small data size. The MEN-Dataset consists of only 200 news articles, with annotated entities and relations. The work on expanding the dataset with more annotated news articles is ongoing, and will be used for thorough experiments and analysis.

3. We used ChatGPT Web version instead of OpenAI API in the experiments, due to the following reasons:

   (a) OpenAI API does not have ability to store information about past interactions. This means, it would have been difficult to provide additional context like Annotation Guideline. However this is not the case when using ChatGPT web interface. LangChain[7] has not supported "Memory" functionality when the experiments were conducted.

   (b) Resource Constraint and Efficiency. The utilization of the OpenAI API will incur costs. Small set of data enables better and in-depth analysis ChatGPT outcome.

[6] https://chat.openai.com/
[7] https://www.langchain.com/

References

Mohammad Belal, James She, and Simon Wong. 2023. Leveraging chatgpt as text annotation tool for sentiment analysis. *ArXiv*, abs/2306.17177.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *ArXiv*, abs/2304.14827.

Lingjiao Chen, Matei Zaharia, and James Y. Zou. 2023. How is chatgpt's behavior changing over time? *ArXiv*, abs/2307.09009.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *ArXiv*, abs/2305.14450.

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*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *ArXiv*, abs/2303.16416.

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?

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? *arXiv preprint arXiv:2310.14880*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *ArXiv*, abs/2304.11633.

OpenAI. 2022. Chatgpt. `https://openai.com/chatgpt`.

OpenAI. 2023. Gpt-4. `https://arxiv.org/abs/2303.08774`.

Michael V. Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *ArXiv*, abs/2304.11085.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *ArXiv*, abs/2304.04339.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv*, abs/2302.10205.

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.

A Prompt Generated with Entity Annotation Guideline

ICL Prompt for Annotation Guideline

Template:

Act as a Data Annotator. You will be given an Annotation Guideline, understand the guideline before start annotation. Since the guideline is too long, you will be given as separate chunks.

Here is chunk number {<chunk_number>}: {<guideline_chunk>} \n If you understand, just response Yes I understand or not you may ask question and no further explanation is required.

Input:

chunk_number: 5

guideline_chunk: Md Ali]] , [Raja Permaisuri Agong [Tunku [Azizah Aminah Maimunah Iskandariah]]] , [Johor Bahru Mayor [Datuk [Mohd Noorazam Osman]]] , [ustaz [Hussain] , and [Yang di-Pertuan Agong [Al-Sultan Abdullah Ri'ayatuddin Al-Mustafa Billah Shah]] b . AVOID ROLE : i. Makcik , Pakcik , Atuk , Nenek or Anak . 6 . Names of Titles with Academic Degrees and Titles a . [Dr. [John Raj]] , and [Associate Professor [Deva]] 7 . Names and designations of God . a . [Allah] , [God] , [God Almighty] , and [Insya-Allah] there will be … 8 . Names with Family Name Affixes (like bin , binti , al- , A/L , A/P) . Example : a . [Yang Berhormat [Puan [[Kasthuriraani] a/p [Patto]]]] and [[Tun Hajah Rahah] binti [Tan Sri Haji Mohamed Noah]] Notes 1 . In some scenarios , the ROLE & TITLE will be embedded together with the entity mentioned PERSON . 2 . In some scenarios , annotators will find NESTED ENTITY . In this scenario , the annotator can annotate BOTH ENTITIES . a . Example : Tun Hajah Rahah binti Tan Sri Haji Mohamed Noah i . The annotator can annotate " Tun Hajah Rahah binti Tan Sri Haji Mohamed Noah " , " Tun Hajah Rahah " , and " Tan Sri Haji Mohamed Noah " separately . 3 . In some scenarios , the PERSON (PER) name is included as part of the LOC , ORG , and FAC names . Example : a . [Taman Tun Dr Ismail] , [Perumahan Haji Malik] . i . This type of entity should NOT be annotated as a PERSON . 4 . In some scenarios , the ROLE & TITLE will be embedded together with the entity mentioned PERSON . Example : a . [Prime Minister [Datuk Seri [Ismail Sabri Yaakob]]] i . In this scenario , the Annotator can annotate [Prime Minister Datuk Seri Ismail Sabri Yaakob] , [Datuk Seri Ismail Sabri Yaakob] , and [Ismail Sabri Yaakob] AS PERSON . ii . Also , Annotator needs to annotate : 1 . Prime Minister as ROLE 2 . Datuk Seri as TITLE iii . The annotation for the above example will look like the below :

Note: The above chunk is guideline related to entity label, PERSON

Figure 4: Prompt template used to provide entity annotation guideline as separate chunks

B Entities and Explanations

ICL Prompt for Few Shot

Template:

Now you will be provided with some sample of News Articles and the Entity Extracted from the article. This Entity are extracted based on the Annotation Guideline and Context of News Article.

Here is the news article {<news_article>}: Here is the extracted entities {<entity_extracted>}

Input:

news_article: KUALA LUMPUR, Jan 10 — Former Mara Incorporated Sdn Bhd (Mara Inc) chairman Datuk Mohammad Lan Allani was today given a discharge not amounting to an acquittal (DNAA) by the Sessions Court here on two counts of money laundering charges involving RM10 million. Judge Suzana Hussin made the decision after Malaysian Anti-Corruption Commission (MACC) deputy public prosecutor, Allan Suman Pillai, applied for Mohammad Lan's DNAA as he would be made to face the charges again at the Kota Kinabalu Sessions Court, Sabah in February or March. Lawyer Teh See Khoon, who represented the accused, appealed for bail of RM400,000 to be reduced to RM300,000 as the prosecution dropped the two charges and his client needed the money for the bail when he would be charged again in Sabah. Allan Suman did not object to the application. <truncated>

entity_extracted: {'LOCATION': ['The entity mention Sabah is labeled as LOCATION, because Sabah is a state located in Malaysia.',
'The entity mention Kota Kinabalu is labeled as LOCATION, because Kota Kinabalu is a city located in Malaysia.'],
'ORGANIZATION': ['The entity mention MACC is labeled as ORGANIZATION, because MACC is an organization, known as the Malaysian Anti-Corruption Commission.'],
'PRODUCT': ['The entity mention CIMB Bank Private Banking is labeled as PRODUCT, because CIMB Bank Private Banking is a service and product offered by CIMB Bank.'],
'LAW': ['The entity mention Section 16 (a) (A) is labeled as LAW, because Section 16 (a) (A) is a proper law act approved by the authorities.',
'The entity mention Anti-Money Laundering and Anti -Terrorism Financing Act 2001 (Act 613) is labeled as LAW, because Anti-Money Laundering and Anti -Terrorism Financing Act 2001 (Act 613) is a proper law act approved by the authorities.'],
'TITLE': ['The entity mention Datuk is labeled as TITLE, because Datuk is an honorary title used in Malaysia.',
'The entity mention Datuk is labeled as TITLE, because Datuk is an honorary title used in Malaysia.']}

Note:
1. The news_article is truncated and only some entity ention with explanation is shown in this sample.
2. The words highlighted in Green are entity mention and the words highlighted in Yellow are entity label.

Figure 5: A few examples of manually annotated entities along with explanations for why they have been annotated.

C Prompt for NER Task in ChatGPT

Example of NER Task Prompt Breakdown

Template:

Act as Data Annotator, identify and extract all the entity PERSON, LOCATION, ORGANIZATION, EVENT, WORK_OF_ART, ROLE, TITLE, NORP, FACILITY, PRODUCT, LAW, LANGUAGE from the input news article. Input: {<input>}. Provide the output as JSON.

Input (News Article):

KUALA LUMPUR: The government is bracing for a spike in Covid-19 daily cases up to 22,000 by end of March this year. According to a tweet by Health director-general Tan Sri Dr Noor Hisham Abdullah, he said this was the projection based on the current infectivity rate which continued upward to 1.20 as of yesterday. (Feb 5). 'Hence the onus is on us to comply to all the public health and social measures plus increasing vaccination and booster shot coverage to reduce the transmission and lower the Ro/Rt to less than 1,' he said today. As of yesterday, the highest Rt value was reported in Perlis at 1.49. The infectivity rate in Sabah was 1.41; Putrajaya at 1.29; Kedah at 1.22; Selangor at 1.19; Penang and Sarawak at 1.17; Johor was 1.16; Pahang at 1.14; Terengganu at 1.13; Kuala Lumpur at 1.12; and Melaka at 1.11. States with infectivity below 1.1 are Negri Sembilan and Perak at 1.1; and Kelantan at 1.08; No Rt value was assigned for Labuan with 15 cases.

Generated Output from ChatGPT


```
{'mention': 'Noor Hisham Abdullah',
  'label': 'PERSON',
  'position': {'start_offset': 177, 'end_offset': 197}},
 {'mention': 'KUALA LUMPUR',
  'label': 'LOCATION',
  'position': {'start_offset': 0, 'end_offset': 12}},
 {'mention': 'March',
  'label': 'LOCATION',
  'position': {'start_offset': 99, 'end_offset': 104}},
 {'mention': 'Feb 5',
  'label': 'LOCATION',
  'position': {'start_offset': 318, 'end_offset': 323}},
 {'mention': 'Sabah',
  'label': 'LOCATION',
  'position': {'start_offset': 638, 'end_offset': 643}},
<18 more entities>
```


Figure 6: The prompt template used to extract entities based on news article provided.

D Prompt for RE Task in ChatGPT

Example of RE Task Prompt Breakdown

Template:

You will be provided with List of Entities and News Article as Input. Only annotate the relation based on this labels: ['head of government', 'country', 'place of birth', 'place of death', 'father', 'mother', 'spouse', 'country of citizenship', 'continent', 'head of state', 'capital', 'official language', 'position held', 'child', 'author', 'director', 'screenwriter', 'educated at', 'composer', 'occupation', 'founded by', 'league', 'place of burial', 'publisher', 'maintained by', 'owned by', 'operator', 'named after', 'cast member', 'producer', 'award received', 'chief executive officer', 'creator', 'ethnic group', 'performer', 'manufacturer', 'developer', 'legislative body', 'executive body', 'record label', 'production company', 'location', 'place of publication', 'part of', 'military rank', 'member of', 'chairperson', 'country of origin', 'diplomatic relation', 'residence', 'organizer', 'characters', 'lyrics by', 'participant', 'given name', 'location of formation', 'parent organization', 'significant event', 'authority', 'sponsor', 'applies to jurisdiction', 'director / manager', 'product or material produced', 'student of', 'territory claimed by', 'winner', 'replaced by', 'capital of', 'languages spoken, written or signed', 'present in work', 'country for sport', 'represented by', 'investor', 'intended public', 'partnership with', 'statistical leader', 'board member', 'sibling', 'stepparent', 'candidacy in election', 'coach of sports team', 'subsidiary', 'religion', 'NO_RELATION'] and guideline provided earlier. Answer me in JSON format. Follow this format: { "annotations": [{ "entity_pair": { "head": "Entity", "tail": "Entity" }, "relation": "Relation Label" }, { "entity_pair": { "head": "Entity", "tail": "Entity" }, "relation": "Relation Label" }] }. Only give me the response asked, without any explaination. Here is the News Article: {<news_article>}. and the Entity Pairs: {<entity_pairs>}

Input:

news_article: Similar as in **Example of NER Task Prompt Breakdown**

entity_pairs: ["1. ('Health director-general', 'Tan Sri Dr Noor Hisham Abdullah')", "2. ('Tan Sri', 'Noor Hisham Abdullah')", "3. ('Dr', 'Noor Hisham Abdullah')"]

Generated Output from ChatGPT


```
[{'entity_pair': {'head': 'government', 'tail': 'KUALA LUMPUR'},
  'relation': 'capital'},
 {'entity_pair': {'head': 'government', 'tail': 'Health director-general'},
  'relation': 'position held'},
 {'entity_pair': {'head': 'Health director-general',
   'tail': 'Tan Sri Dr Noor Hisham Abdullah'},
  'relation': 'head of government'},
 {'entity_pair': {'head': 'Tan Sri Dr Noor Hisham Abdullah',
   'tail': 'Noor Hisham Abdullah'},
  'relation': 'spouse'},
 {'entity_pair': {'head': 'Tan Sri Dr Noor Hisham Abdullah',
   'tail': 'Dr Noor Hisham Abdullah'},
  'relation': 'child'},
 {'entity_pair': {'head': 'Dr Noor Hisham Abdullah',
   'tail': 'Noor Hisham Abdullah'},
  'relation': 'father'},
 {'entity_pair': {'head': 'Dr', 'tail': 'Noor Hisham Abdullah'},
  'relation': 'NO_RELATION'}]
```


Figure 7: The prompt template used to extract relations based on news article and entities provided.

E 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 3: Entity Labels

F 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	NO_RELATION		ANY ENTITY	ANY ENTITY	Can be used for any entity pair that does not have a suitable Relations Listed

G 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 5: Different prompting techniques used to evaluate ChatGPT capabilities for NER and Relation Extraction

H 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 6: The F1-Score from the perspective of entity label.