Z-coref: Thai Coreference and Zero Pronoun Resolution

Poomphob Suwannapichat and **Sansiri Tarnpradab** and **Santitham Prom-on**
Department of Computer Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
{poomphob.suwan, sansiri.tarn, santitham.pro}@kmutt.ac.th

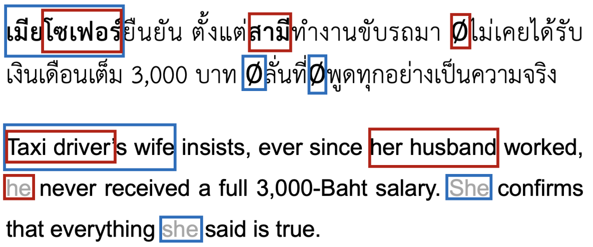
Abstract

Coreference resolution (CR) and Zero Pronoun Resolution (ZPR) are vital for extracting meaningful information from text. However, limited research and datasets pose significant challenges in Thai language. To address this, we developed an annotated joint CR and ZPR dataset. Additionally, we introduced the Z-coref model, capable of simultaneously handling CR and ZPR tasks by adjusting the span definition of a prior CR architecture to include token gaps. The proposed model trained on our dataset outperformed the state-of-the-art in resolving both coreference resolution and zero-pronoun resolution, while taking less time to train.

1 Introduction

Coreference resolution (CR) is the task of identifying and linking words or phrases referring to the same entity in a text. It is a crucial step in natural language processing (NLP) taken to determine the meaning of a text by resolving ambiguity. One of the tasks in CR is known as zero pronoun resolution (ZPR). The main goal of ZPR is to determine the reference of a missing pronoun, or so-called a zero pronoun (ZP) – a linguistic phenomenon in which a pronoun in a sentence can be omitted because its referent is clear from the context. This omission is often easily recognizable by humans but presents a challenge for machines. Zero pronoun resolution still remains a difficult task in pro-drop languages like Thai, Chinese, and Japanese.

Figure 1 illustrating a news headline written in Thai and its English translation, exemplifies the challenge of ZPs. Nouns and zero pronouns (∅) marked with blue squares refer to the wife of a taxi driver, while those in red squares refer to the taxi driver. It can be noticed that there are several occurrences of ZPs although the headline and language style are succinct. These brief sentences present a challenge for a machine to interpret.



เมียโซเฟอร์ยืนยัน ตั้งแต่**สามี**ทำงานขับรถมา Øไม่เคยได้รับเงินเดือนเต็ม 3,000 บาท Øลั่นที่Øพูดทุกอย่างเป็นความจริง

Taxi driver's wife insists, ever since her husband worked, he never received a full 3,000-Baht salary. She confirms that everything she said is true.

Figure 1: Examples of Thai news headline and the translated versions in English. ZPs are represented as '∅'. The box color scheme indicates entities with the same reference. The text in gray indicates expression that can be omitted in Thai

While there exist various baseline models and large annotated datasets for CR in English, there is a paucity of research in this area for the Thai language. Only one dataset and one baseline model by Han-coref are publicly available (Phatthiyaphaibun and Limkonchotiwat, 2023); however, neither covers the case of zero pronouns. Therefore, this study makes the following contributions: (1) We have taken the initiative to create the first dataset that combines both CR and ZPR for Thai language; (2) We introduce a novel approach, Z-coref, which is capable of handling CR in Thai while also addressing the challenges posed by ZPs, a nature of the Thai language; (3) We conducted a comparative analysis of our approach with the joint CR and ZPR model for Chinese language introduced by Chen et al. (2021). Our model not only significantly outperforms in terms of training time but also exhibits a slightly higher performance. Lastly, our source code, dataset and model are available at https://github.com/psuwannapich/z-coref.

2 Related Works

In this section, we first introduce previous works in the topic of coreference resolution, followed by zero pronoun resolution. Then, CR methods pro-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 535–542
August 11-16, 2024 ©2024 Association for Computational Linguistics

posed for Thai language along with zero pronouns will be discussed.

2.1 Coreference Resolution

A number of neural models for coreference resolution have been developed. Among them, Lee et al. (2017) is the first to introduce an end-to-end neural CR model that employs span representations. The score to consider pairs of query span q and candidate antecedent span c is denoted as $f(c, q)$, which is a combination of query mention score $f_m(q)$, candidate mention score $f_m(c)$ and joint antecedent score $f_a(q, c)$ as shown in Equation 1.

$$f(c, q) = f_m(q) + f_m(c) + f_a(q, c) \tag{1}$$

The mentions were formed using a span head layer that averages token representations of consecutive tokens. Nevertheless, given that all combinations of spans and coreferential pairs are considered, the model complexity becomes $O(n^4)$, where n is the number of tokens.

To improve the computational efficiency, Kirstain et al. (2021) performed the algorithm without using span representation (s2e-coref). The results demonstrated that the memory usage during inference time has reduced with insignificant effect on the performance.

To compute the mention score, only the representation of start token $\mathbf{m_{q_s}}$ and end token $\mathbf{m_{q_e}}$ are used, rather than all tokens in the span (Equation 2). Here $\mathbf{m_{q_s}}$ and $\mathbf{m_{q_e}}$ are the vector projections related to the mention score from the query's start token q_s and end token q_e, respectively, while $\mathbf{B}$ and $\mathbf{v}$ are parameters that the model learns during training.

$$f_m(q) = \mathbf{v}_s \cdot \mathbf{m}_{q_s} + \mathbf{v}_e \cdot \mathbf{m}_{q_e} + \mathbf{m}_{q_s} \cdot \mathbf{B}_m \cdot \mathbf{m}_{q_e} \tag{2}$$

Similarly, the antecedent score is determined using the start and end tokens of both the query span q and the candidate span c, as outlined in Equation 3. The equation includes four terms that represent the combinations of the start and end tokens from q to c. The vector $\mathbf{a}$ corresponds to the projection associated with the antecedent score for each token.

$$\begin{aligned} f_a(c, q) = {} & \mathbf{a}_{c_s} \cdot \mathbf{B}_{a_{ss}} \cdot \mathbf{a}_{q_s} + \mathbf{a}_{c_s} \cdot \mathbf{B}_{a_{se}} \cdot \mathbf{a}_{q_e} \\ & + \mathbf{a}_{c_e} \cdot \mathbf{B}_{a_{es}} \cdot \mathbf{a}_{q_s} + \mathbf{a}_{c_e} \cdot \mathbf{B}_{a_{ee}} \cdot \mathbf{a}_{q_e} \end{aligned} \tag{3}$$

Subsequently, Otmazgin et al. (2022) introduced F-coref, which exhibited enhanced performance and efficiency through the implementation of dynamic batching and knowledge distillation techniques. The transformer model for token representation calculation was modified from Longformer (Beltagy et al., 2020), which was widely used in the CR task to the more lightweight DistilRoBERTa (Sanh et al., 2019). By leveraging knowledge distillation from the LingMess model (Otmazgin et al., 2023), the size of the F-coref model was reduced without compromising its overall performance.

2.2 Zero Pronoun Resolution

In general, ZPR tasks take the location of query ZP as an input, then find any suitable antecedent for the pronoun. For instance, (Yin et al., 2018b) employed recurrent neural networks with an attention mechanism to extract the antecedent noun phrase using the input ZP query. Under the same theme, deep reinforcement learning techniques were employed for ZPR in (Yin et al., 2018a). The model's agent has actions to determine whether to consider them as coreferential based on a given pair of ZP and candidate noun phrase.

A ZPR model has also been introduced for Arabic by Aloraini and Poesio (2020), through utilization of multilingual BERT model (Devlin et al., 2018). The model also unexpectedly achieved higher performance in Chinese compared to previous state-of-the-art.

However, an iteration over all gaps between words is required to resolve all ZPs with these approaches. To address this issue, Chen et al. (2021) integrated ZPR and CR into a single task; all gaps in a document are considered as a candidate mention for CR and use an end-to-end model to resolve the coreferential.

2.3 Thai Coreference and Zero Pronouns

Currently, research in CR for Thai language is limited due to the lack of public datasets. Earlier work by Kongwan et al. (2022) used their previous dataset in Elementary Discourse Units segmentation for the task. They localized the mentions using a rule-based method on the part of speech and applied a mention-ranking model (Denis and Baldridge, 2008) to find the coreferential pair. To improve the model performance further, Han-coref (2023) used the architecture from F-coref model (2022) and added a tokenization module to handle the ambiguity of word boundary. Additionally, a coreference dataset of 1,338 documents along with an annotation guideline was created.

Resolving cross-document CR, Theptakob et al. (2023) used agglomerative clustering on pairwise entity coreference score to determine coreferences across documents. In Thai study on ZPR, Sumanakul (2022) employed a mask language model. A masked token was inserted at the ZP's location, and a pre-trained transformer model was utilized to predict the masked token. These mask predictions were considered as the coreferential answers. Additionally, they performed token classification to determine ZP types (first, second or third person). However, no research has considered ZPs together with CR in Thai yet.

3 Methodology

In this paper, we established a CR dataset that contains details of ZPs and modified the CR model's architecture to handle ZPs.

3.1 Data Annotation

We retrieve 1,338 documents from Han-coref (Phatthiyaphaibun and Limkonchotiwat, 2023) including Thai news headlines and Wikipedia. Due to the difference in scope, we need to re-annotate the dataset. We selected as annotators, Thai native speakers who were not linguists. These annotators must be fluent in Thai and have the capability to read and comprehend Thai news. The annotation guideline was written due to the ambiguity of the language to ensure the corrective of the annotators. The annotation process is divided into two steps: (1) identify mentions and (2) link the coreferential mentions. Annotators are asked to indicate mentions; words or phrases that refer to a specific person or organization. Other specific words such as items or locations are ignored, in order to maintain a manageable scope and enable non-linguistic annotators to participate more effectively.

3.2 Z-coref

Our Z-coref model employs F-coref (2022) model's architecture, a faster and smaller version of s2e (2021), incorporating knowledge distillation from LingMess (2023). The s2e model utilizes only the first and last tokens within a span, rather than all tokens in the span to create representations. Nevertheless, the s2e model lacks compatibility with ZPs because the span cannot be a gap between words without any characters. Normally, span $span(s, e)$ is a concatenation of consecutive tokens start from token s (t_s) to token e (t_e). For example, in Figure 2, $s(2, 3)$ is the span "loves dogs".

From the definition, the smallest span is a token when $e = s$. We expand this definition further by also considering the gap between two consecutive tokens $g(s-1, s)$ which is the gap between t_{s-1} and token t_s. With this modification, the span that starts from token s and ends at token $s-1$ is considered a special type of span used to represent a ZP. Therefore, both the normal span and the token's gap can be defined using the modified span definition:

$$span(s, e) = \begin{cases} [t_s; t_{i+s}; ...; t_{e-1}; t_e] & \text{if } s \leq e \\ gap(e, s) & \text{if } s - 1 = e \end{cases} \tag{4}$$

As illustrated in Figure 2, $s(5, 4) = g(4, 5)$ corresponds to the gap between "but" and "hates". Furthermore, it becomes necessary to introduce a new special token at both the document's beginning and end to effectively manage instances of ZPs occurring in those positions. This adapted definition ensures the seamless compatibility of ZPs with the concept of the s2e model. Due to non-explicit-word-boundary language, we rely on a subword tokenizer that is integrated with the base transformer model because we aim to tokenize the document into the smallest units possible, thereby preventing ZPs from being within the middle of a token.

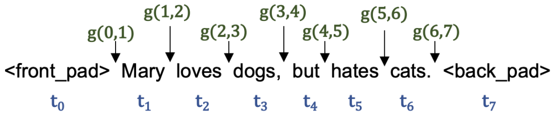


Figure 2: Tokens and gaps example. "<front_pad>" and "<back_pad>" tokens are added. Any positions between consecutive tokens are consider as gaps.

Rather than using Longformer (2020) or DistilRoBERTa (2019), we used WangchanBERTa (Lowphansirikul et al., 2021), a pre-trained transformer model on Thai corpus to extract contextual representations from tokens. The downstream pipeline is the same as F-coref (2022) model with modification for ZPs. Normally, F-coref model filters invalid spans using Equation 5

$$f_m(q) = \begin{cases} f(q_s, q_e) & \text{if } s \leq e < s + max\_length \\ -10,000 & \text{otherwise} \end{cases} \tag{5}$$

The mention score of a valid span is calculated normally using Equation 2. On the other hand, the score of an invalid span (the span that is longer than the max length or that the start token comes after the end token) is fixed to a large negative number -10,000. To add ZPs to the model, we simply changed the first condition of Equation 5 to accommodate the scenario where $s - 1 = e$, which signifies a ZP. Consequently, our Z-coref is now compatible with normal mentions and ZPs.

4 Dataset

The dataset has been annotated by eight annotators. Due to time constraints and the challenging nature of the task, each annotator was only able to annotate a subset of the dataset. However, by combining the annotations from all annotators, the entire dataset of 1,338 documents was covered with at least two annotations per document. Details of dataset are further described in Appendix A

5 Experiment and Results

This experiment aims to evaluate our proposed model against the e2e-joint-coref model developed by Chen et al. (2021) using our annotated dataset. The models were trained for 150 epochs to compare their performance and training time requirements. Both models were trained using an Nvidia GeForce RTX 3090 GPU with no other processes running concurrently during the training sessions for the fairness of time comparison. Detailed experiment setting are discussed in Appendix B

As shown in Table 1, our proposed model significantly reduces the training time compared to e2e-joint-coref. This is due to the removal of span representation in s2e model (Kirstain et al., 2021), which reduces memory usage and enables the use of larger batch sizes. Additionally, dynamic batching from F-coref (Otmazgin et al., 2022) further decreases the model training time by optimizing batch creation. These improvements allow our model, which modifies the span definition from the F-coref model, to be trained approximately 9-14 times faster than e2e-joint-coref, which uses the architecture from e2e-coref and doubles the number of tokens by considering all gaps as additional tokens. (WangchanBERTa achieving the lowest improvement at 8.8 times faster and mBERT achieving the highest at 13.8 times faster)

As illustrated in Table 2, PhayaThaiBERT encoder yields the highest F1 score for both settings.

Base encoder	e2e-joint-coref	Z-coref
WangchanBERTa	3 hr 5 min	21 min
PhayaThaiBERT	3 hr 50 min	23 min
mBERT	4 hr 35 min	20 min
XLM-RoBERTa	4 hr	21 min

Table 1: Model training time comparison.

Base encoder	e2e-joint-coref	Z-coref
WangchanBERTa	0.716	**0.744**
PhayaThaiBERT	0.730	**0.758**
mBERT	**0.702**	0.658
XLM-RoBERTa	0.677	**0.729**

Table 2: Model performance (CoNLL F1 score) comparison.

In addition, Z-coref with PhayaThaiBERT encoder exhibits superior performance compared to others. Nevertheless, when employing mBERT encoder, Z-coref is unable to surpass the performance of e2e-joint-coref. In the case of XLM-RoBERTa and WangchanBERTa, further elaboration on these results is presented in Appendix B as the performance observed in Table 2 alone may not suffice in drawing a definite conclusion.

6 Conclusion

Due to the lack of a dataset and baseline model for CR in the Thai language, as well as the nature of pro-drop languages that can cause original CR to overlook ZPs that frequently occur in informal language such as news articles, we have created the first Thai joint dataset of CR and ZPR. We also introduce Z-coref, a lightweight joint CR and ZPR model. Z-coref with PhayaThaiBERT encoder achieves higher performance than previous work from Chen et al. (2021) and significantly reduces training time.

Our method effectively resolves the majority of ZPs. However, it may face limitations when multiple ZPs occur within the same gap. For example, in the sentence "They would hit (it) (so) (it) flees", the words in parentheses can be omitted in Thai. Consequently, the gap between "hit" and "flees" contains two ZPs: the object of the first subsentence and the subject of the second subsentence. This scenario highlights a potential challenge for our approach.

References

Abdulrahman Aloraini and Massimo Poesio. 2020. Cross-lingual zero pronoun resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 90–98, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Shisong Chen, Binbin Gu, Jianfeng Qu, Zhixu Li, An Liu, Lei Zhao, and Zhigang Chen. 2021. Tackling zero pronoun resolution and non-zero coreference resolution jointly. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 518–527, Online. Association for Computational Linguistics.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Honolulu, Hawaii. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Authapon Kongwan, Farzana Kabir Ahmad, and Siti Kamaruddin. 2022. Anaphora resolution in thai edu segmentation. *Journal of Computer Science*, 18:306–315.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun and Peerat Limkonchotiwat. 2023. Han-Coref: Thai Coreference resolution by PyThaiNLP.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sumana Sumanakul. 2022. *Resolving Thai zero pronoun using masked language model*. Ph.D. thesis, Office of Academic Resources, Chulalongkorn University.

Nathanon Theptakob, Thititorn Seneewong Na Ayutthaya, Chanatip Saetia, Tawunrat Chalothorn, and Pakpoom Buabthong. 2023. A cross-document coreference resolution approach to low-resource languages. In *Knowledge Science, Engineering and Management*.

Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2018a. Deep reinforcement learning for Chinese zero pronoun resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 569–578, Melbourne, Australia. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018b. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Dataset

A.1 Format

Each document within our dataset is structured in JSON format, comprising three fields: "text", "clusters", and "clusters_strings". The "text" key contains the raw textual content, while the "clusters" key contains coreference information organized in a nested list format. Mention locations are recorded in a start-and-end character index format. In regular pronouns, the start index precedes the end index. In contrast, when dealing with ZPs, the start and end indexes are equal, representing the ZP in front of the start character. Subsequently, mentions belonging to the same coreference chain are grouped together within the same list. Lastly, a "clusters_strings" key is included for the purpose of cross-checking with the string obtained from the "clusters" key.

Figure 3 illustrates an example from the dataset. Suppose the second sentence is "She loves (him)." with the word "him" omitted.

```
{
  "text": "Mary has a friend. She loves.",
  "clusters": [
    [[0,4], [19,22]],
    [[9,17], [28,28]]
  ],
  "clusters_strings": [
    ["Mary", "She"],
    ["a friend", ""]
  ]
}
```

Figure 3: Dataset example.

A.2 Annotation agreement

To measure the agreement between annotators, the metrics both for mention detection F1 score and coreference resolution CoNLL F1 score were evaluated. Pairwise evaluation was performed between all combinations of annotators. Table 3 provides the average metrics for each annotator. These values can be utilized to indirectly evaluate the degree of agreement between a particular annotator and others. The zero-pronoun metrics for annotators 4 and 8 are lower compared to those of the other annotators. Consequently, we attempt to exclude zero pronoun annotation from these annotators.

A.3 Dataset distribution

We obtain the mentions in each type as presented in Table 4. As anticipated, the dataset contains a lot of both normal and zero mentions that refer to individuals owing to the nature of news writing, which primarily focuses on individuals and omits numerous expressions. The distribution of the number of coreference chains is shown in Figure 4. Most of the documents contain less than 5 coreference chains.

Mention Type	Mention count
PER: Noun	4477
PER: Pronoun	1386
PER: Zero	2665
ORG: Noun	1119
ORG: Pronoun	50
ORG: Zero	67
Unknown	409

Table 4: The number of mentions in each type

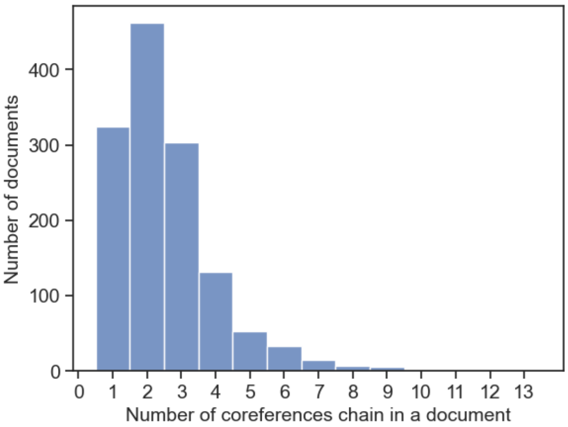


Figure 4: The number of coreference chains distribution.

B Experiment setting

To obtain a robust conclusion, random search hyperparameter tuning was conducted. Almost all the hyperparameters remained unchanged except those hyperparameters listed in Table 5. We aimed at accomplishing 40 iterations for tuning for both the models. Regrettably, e2e-joint-coref requires long training time as specified in Table 1. To ensure fairness, we aimed to allocate an equal amount of time for hyperparameter tuning for both models. As a result, we executed only 5 iterations for the e2e-coref model, which required a training time roughly equivalent to 40 iterations of the Z-coref model.

The distribution of the performance from hyperparameter tuning is visualized in histogram as shown in Figure 5. Z-coref with PhayaThaiBERT encoder exhibits superior performance compared to e2e-joint-coref. However, when employing mBERT encoders the proposed model is unable to surpass the performance of e2e-joint-coref.

Although the best performance of WangchanBERTa demonstrates that Z-coref achieves higher performance, the distribution of Z-coref still ex-

No.	Mention detection F1			Coreference Resolution F1		
	Normal	Zero	All	Normal	Zero	All
1	0.846	0.583	0.777	0.754	0.585	0.653
2	0.847	0.492	0.767	0.734	0.492	0.635
3	0.813	0.555	0.750	0.703	0.530	0.623
4	0.787	**0.250**	0.684	0.664	**0.239**	0.524
5	0.801	0.572	0.746	0.692	0.549	0.619
6	0.804	0.411	0.698	0.670	0.384	0.545
7	0.727	0.528	0.667	0.614	0.505	0.545
8	0.796	**0.305**	0.730	0.673	**0.298**	0.583
Mean	0.803	0.462	0.727	0.688	0.448	0.591

Table 3: Annotation agreement across annotator

Hyperparameter	Search space
Max length of the span	20 - 50
Proportion of unpruned spans	0.3 - 0.9
Dropout rate	0.1 - 0.6
Fully connected size	512 - 2048

Table 5: Hyperparemeters and search space.

Type	Normal	Zero	Both
Precision	0.979	0.965	0.974
Recall	0.756	0.922	0.803
F1	0.853	0.943	0.881

Table 6: Mention detection performance in each mention type

hibits high variance, and the two distributions largely overlap. This can be further analyzed by utilizing a larger sample size.

In the case of XLM-RoBERTa, only one successful experiment from e2e-joint-coref is available, as the other experiment remains diverge with zero F1 score after the training process has been completed. Although the result from XLM-RoBERTa suggests that the proposed model may outperform e2e-joint-coref, a single experiment is insufficient to draw a definitive conclusion.

C Error Analysis

After model training and hyperparameter tuning, it was observed that employing PhayaThaiBERT as an encoder resulted in the most optimal performance. We further analysis the model performance both mention detection and coreference resolution as shown in Table 6 and Table 7, respectively.

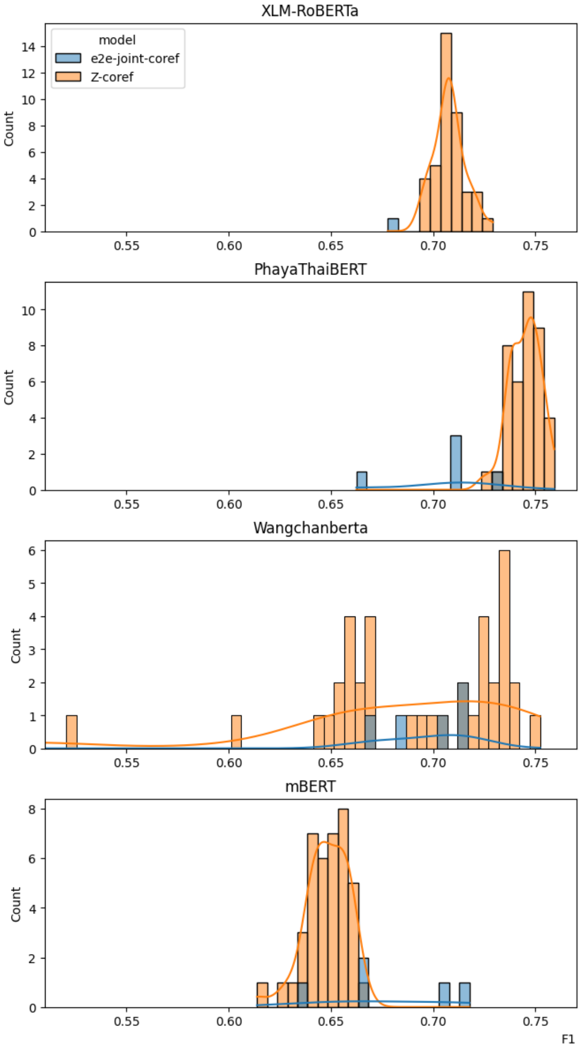


Figure 5: Model performace distribution.

Since the model's performance in detecting normal mentions is inferior to its performance in detecting zero mentions, and the recall is significantly lower than the precision, we will attempt to identify the types of normal mentions in the gold label that the model frequently fails to detect as shown in

Type	Normal	Zero	Both
Precision	0.855	0.857	0.842
Recall	0.687	0.847	0.707
F1	0.745	0.852	0.758

Table 7: Coreference resolution performance in each mention type

Mention Type	FN	TP	Recall
PER: Noun	86	371	0.810
PER: Pronoun	22	191	0.897
ORG: Noun	54	107	0.665
ORG: Pronoun	3	8	0.727

Table 8: Coreference resolution performance in each mention type

Table 8. The model exhibits higher recall in detecting pronoun mentions compared to noun mentions. This can be attributed to the greater variability observed in nouns, including names that can consist of any words. In contrast, the set of possible pronouns is limited, facilitating the model's ability to correctly identify them. Furthermore, the model demonstrates higher accuracy in detecting mentions referring to persons rather than organizations. This can be explained by the nature of the dataset, which primarily consists of news articles that predominantly focus on individuals.