A Character-Level Restoration of Sukhothai Inscriptions Using the Mask Language Model

Sujitra Tongkhum and Sukree Sinthupinyo

October 6, 2023

A Character-Level Restoration of Sukhothai Inscriptions Using The Mask Language Model

1st Sujitra Tongkhum
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
6472086421@student.chula.ac.th

2nd Sukree Sinthupinyo
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
Sukree.S@Chula.ac.th

***Abstract*—The stone inscription is one type of written literature that recorded the history story and the manifestation of cultural identity in that era through a character engraving method on the stone with sharp metal material for each character until a sentence formed. To convey the message for the readers to understand the meaning. Therefore, the completeness of that sentence is of great importance natural language processing tasks. In particular, when transcription stone inscriptions, it is found that inscriptions' parts cannot interpret. As a result of the period that elapsed, those inscriptions may have suffered deterioration from various causes, resulting in scratches over the text or faded markings, destroyed from natural disasters that making it impossible to analyze which specific characters were damaged. To address enhance the completeness of the missing sentence, this research employs a method of generating predictive models for the missing characters from the text. It utilizes the technique of incorporating a masked language model to assist in processing the experimental data, utilizing 3 types of multilingual pre-trained models as following models are used: (1) XLM-RoBERTa, (2) Bert-base-multilingual-cased, and (3) DistilBERT-base-multilingual-cased. In each training round, random characters are masked using the token "<mask>" or "[MASK]" to prompt the model to predict the missing words at the masked positions. From the experimental results, it was found that the accuracy of prediction from the three types of pre-trained models is as follows: (1) 42, (2) 53, and (3) 50 percent respectively.**

***Index Terms*—natural language processing,Bidirectional encoder representations from transformers (BERT), Transformer, mask Language model**

I. INTRODUCTION

The stone inscriptions are historical records that provide insights into the way of life and significant events of a particular period, allowing present-day society to comprehend the circumstances and culture of that era. These inscriptions are crafted by engraving letters, images, or symbols onto specific locations or objects, such as pillars, stone slabs, cave walls, and more. Their purpose is to communicate the narratives of the events that occurred during that time, as shown in Figure 1 . However, over time, these inscriptions may deteriorate due to various factors, leading to erosion and rendering it challenging to decipher the intended meaning or interpret the incomplete messages for contemporary society. Consequently, studying history through these recorded inscriptions becomes difficult.

This research study utilized a dataset from the Sukhothai period, which has been paraphrase the language used during that time into Thai, sourced from the Inscription Database of Thailand at The Princess Mahachakri Sirindhorn anthropology center. The dataset includes inscriptions from various sources Including (1) Wat Pha Mauang, (2) Nakhon Chum Inscription, (3) Pho Khun Ramkhamhaeng Inscription, (4) Wat Sri Chum Inscription, and (5) Pu Khun Chit Khun Chot Inscription. These inscriptions comprised both complete and incomplete texts by using the technique of masked language modeling (MLM), a predictive model is created to fill in the missing characters from the text and using training the model and utilizing various multilingual pre-trained models as following XLM-RoBERTa, Bert-base-multilingual-cased, and DistilBERT-base-multilingual-cased. In each training round, approximately 15 percent of the characters or text are randomly masked, replacing them with mask tokens, and the model is then tasked with predicting the missing words at the masked positions.

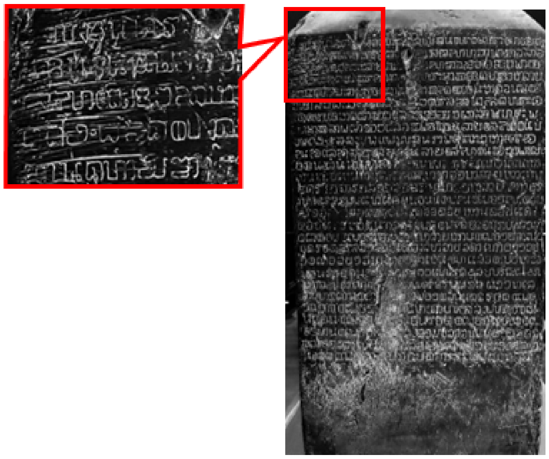


Fig. 1. The Principal Inscription of Phra Khun Ramkhamhaeng. [7]

II. RELATED WORKS

In the field of natural language processing, there have been numerous approaches proposed to address the limitations of standard language models in dealing with the challenges of data linkage and the inability to handle complex structured tasks. Examples of such tasks include question answering and fill missing letters, where contextual understanding and missing information play crucial roles. As a result, researchers have introduced a new language model known as BERT, which stands for Bidirectional Encoder Representations from Transformers [1]. The model is specifically designed for forward-looking learning in processing no label text by adjusting the learning context of the text from both the left and right sides of a sentence to enhance the fundamental aspects of fine-tuning with BERT. To address the aforementioned challenges, researchers proposed an innovative approach known as masked language modeling, inspired by a practice exercise called cloze task. In this exercise, certain words are removed from a text, and learners are required to fill in the missing words.

BERT has become a fundamental research model for developing natural language processing systems that facilitate unsupervised cross-lingual learning from large-scale data [2]. By creating a model that can learn up to 100 languages using sophisticated data handling techniques, accurate text understanding and analysis, as well as consideration of memory limitations, learning speed, and inference efficiency are achieved. This model is known as XLM-R (Cross-lingual Language Model - RoBERTa, XLM). XLM-R has demonstrated superior performance compared to the multilingual BERT (mBERT) model in cross-lingual transfer tasks and measuring the effectiveness of high-resource and low-resource languages across different levels. It also considers the impact of sample size and vocabulary size, where fixed model capacity shows that performance improves with the higher the amount of data, the better the model's performance than languages with low resource counts at one point.

These forward-looking learning models have been applied to various other research domains, such as sentiment analysis in text classification [3]. To enhance the models for sentiment classification, additional features are incorporated to capture the important contextual information, and then the Latent Dirichlet Allocation (LDA) technique is used to identify relevant topics from documents or contexts. Gibbs sampling is employed for topic modeling and random sampling. Beyond these techniques, deep learning approaches in natural language processing have also been utilized to predict the next word in a sentence [4]. For instance, LSTM (Long Short-Term Memory) and BiLSTM (Bidirectional LSTM) models have been created and evaluated using accuracy metrics. A publicly available dataset of around 6,508 articles accessible on Kaggle was used for training and testing. The LSTM and BiLSTM models were employed to predict the next word in a sentence.

III. BACKGROUND

The theories involved in this research involve a database that has been created by paraphrasing Sukhothai language into Thai. This database has been derived from the inscription database in Thailand by the Manuscriptal and Natural Language Processing (NLP) Center.

A. Inscription

Ancient archaeological evidence in the form of inscriptions can be likened to recorded narratives of human history, documented in the form of letters or images engraved on various materials such as stone tablets, wooden panels, or gold plates, among others. This is achieved through the technique of inscribing the material using sharp-edged tools known as ”stylus”. In order to communicate the events that occurred during that time, inscriptions served as important documents that reflect the historical background for the understanding of society in the present era. In Thailand, over 1200 inscriptions have been discovered and utilized for research purposes in the following.

- The Wat Pa Ma-Muang Inscription was discovered at Wat Mai (Prasat Thong), Nakhon Luang District, Phra Nakhon Si Ayutthaya Province. It was constructed during the period of the 1904 Buddhist Era. The ancient Thai script known as Sukhothai script is inscribed on this stone inscription. [5]
- The Nakhon Chum Inscription was discovered at Wat Phra Borommathat, Nakhon Chum Sub-district, Mueang District, Kamphaeng Phet Province. It was constructed during the period of the 1900 Buddhist Era. The ancient Thai script known as Sukhothai script is inscribed on this primary stone inscription, located at Wat Phra Borommathat. [6]
- The Pho Khun Ramkhamhaeng Inscription was discovered at the ancient city mound of Sukhothai, Muang Kao Sub-district, Mueang District, Sukhothai Province. It was constructed during the period of the 1835 Buddhist Era. The ancient Thai script known as Sukhothai script is inscribed on this stone pillar, which originated from the city of Sukhothai. [7]
- The Si Chum Temple Inscription was discovered within the ordination hall of Wat Si Chum, Muang Kao Sub-district, Mueang District, Sukhothai Province. It was constructed during the 19th to 20th Buddhist centuries. The ancient Thai script known as Sukhothai script is inscribed on this stone inscription. [8]
- The Phu Khun Chit Khun Chot Inscription was discovered at the right side of the base pillar, behind the main sanctuary of Wat Mahathat in Sukhothai Historical Park, Muang Kao Sub-district, Mueang District, Sukhothai Province. It was constructed during the period of the 1935 Buddhist Era. The ancient Thai script known as Sukhothai script is inscribed on this stone inscription. It is located within the Sukhothai Historical Park, in the old city of Sukhothai, Sukhothai Province. [9]

B. Natural Language Processing (NLP)

NLP is a branch of artificial intelligence and computer science that focuses on the interaction between human language and computers.it is the study of computational and mathematical models in various aspects of language, including the analysis of grammatical structures, semantics, and even speech recognition systems. NLP encompasses techniques for extracting the structure and meaning from input texts and processing the resulting output. [10]

C. Deep Bidirectional Transformers

[11] Deep Bidirectional Transformers is a deep learning model developed by Google and used for natural language processing (NLP) tasks. with the term ”Bidirectional” refers to the model’s ability to understand and process input data from both the forward and backward directions of a sentence. This means that the model can capture a better understanding of the language patterns. The Transformer is a fundamental model architecture that plays a crucial role in processing NLP data. It excels in handling sequence-to-sequence problems. BERT (Bidirectional Encoder Representations from Transformers) is one prominent example of a Deep Bidirectional Transformer model that has gained significant popularity. BERT has shown high effectiveness in tasks such as Natural Language Understanding (NLU) and Natural Language Inference (NLI) in English and other languages. Furthermore, BERT serves as a foundation for the development of various techniques and models in the present NLP landscape, such as XLM-RoBERTa and DistilBERT.

a) Model Architecture:

- Input Embedding is the initial process of the Transformer model that converts textual data into numerical vectors before entering the model, as depicted in Figure 2. This is because neural networks can effectively process numerical data, allowing the model to handle the data more easily. By transforming the input text into vector representations, the model can capture the semantic meaning and relationships between words, enabling efficient processing and analysis.

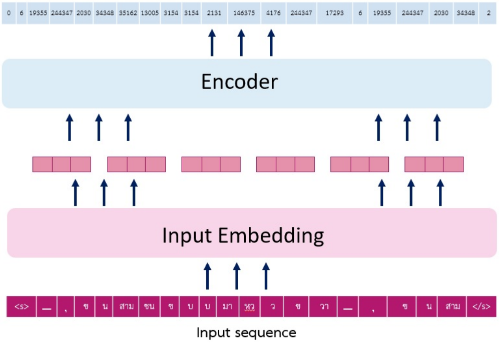


Fig. 2. The working process of the input sequence in each embedding.

- Positional Encoding is the process of adding additional information to the sequence embedding to indicate the positions of words within the input data. It helps the model understand the order or sequence of words being processed. This is achieved by utilizing a sine and cosine function, following the equation specified below:

$$PE_(pos, 2i) = \sin(\frac{pos}{10000^{\frac{2i}{dmodel}}}) \tag{1}$$

$$PE_(pos, 2i+1) = \cos(\frac{pos}{10000^{\frac{2i}{dmodel}}}) \tag{2}$$

- Attention is a function used to process query-key pairs and find the most relevant values based on their similarity. It calculates weights based on the compatibility between the query and the key and then multiplies these weights with the corresponding values to obtain the relevant output. In the context of transforming word relationships into vectors, ”Scaled Dot-Product Attention” is commonly used.In ”Scaled Dot-Product Attention,” the query, key, and value are involved. The query is multiplied with the key, and the resulting product is then summed. This sum is divided by the square root of the dimension of the key. The softmax function is applied to obtain weighted values, which represent the relevance or attention assigned to each key-value pair. This attention mechanism allows the model to focus on the most important parts of the input sequence and prioritize the relevant information for further processing.
- Multi-Head Attention enhances the capabilities of the Transformer model by using multiple sets of parallel attention computations instead of a single set. This approach enables the model to learn and understand data relationships from different perspectives and benefits from estimating linear projections of query, key, and value pairs. In Multi-Head Attention, the model considers data from various sub spaces that capture distinct aspects of the input. It simultaneously attends to the data in multiple ways, employing different learned linear projections. This allows the model to examine the input data from multiple perspectives, capturing diverse relationships and patterns.

D. Masked Language Model, MLM

The masked language model is a type of language model used in the pre-training of deep neural language models. In each training iteration, random masks are inserted into the input text. The model is then trained to understand the context and relationships within the text, allowing it to predict the original letter or vocabulary item that has been masked. This masking is represented using the tokens “<mask>” or “[MASK]” [1]. The design of the model is inspired by the concept of the Cloze task [12], which is an exercise used in language learning. In the Cloze task, certain parts of a sentence or text are removed, and learners are required to fill in the missing words or phrases. Therefore, the functioning of the model involves receiving input, randomly masking and removing words, and then converting the textual data into numerical vectors through embedding. Additional information is incorporated into the embedding of the sequence that lacks inherent order, using labeled indicators to assist the model in understanding the sequence of processed words. This process enables the model to learn the sequence of the processed words and their relationships. Next, the model performs attention by using query, key, and value in order to search and pair words together across multiple attention heads. This allows the model to predict the most relevant words by considering their contextual relationships. The technique applied to language models that incorporates this approach is called “Transformer” and the model itself is an example of a bidirectional pre-trained model. It enables forward and backward learning, allowing the model to have a comprehensive understanding of the input data.

IV. EXPERIMENT

This research utilized the method of measuring the efficiency of the model using cross-validation, which involved dividing the data into 10 equal parts. Each partition of the data, referred to as “folds” was further divided into two subsets: one for training the model and the other for testing, with a ratio of 90:10 percent, respectively dataset.Moreover, the dataset used for testing in all 10 folds contains non-overlapping data between each fold.

Evaluation criteria:

- Display the accuracy values from the data of all 10 folds.
- Perform a performance comparison with other models.

A. experimental report

- Experimental of the XLM-RoBERTa Base model.
  In Figure 3, Based on the experimental results, the XLM-RoBERTa Base model achieved the highest accuracy of 47 percent, with a character-level prediction accuracy of 43 percent. The model’s recall, indicating the likelihood of correctly predicting words, reached 47 percent. It obtained an F1-Score of 42 percent and an average accuracy of 42 percent.

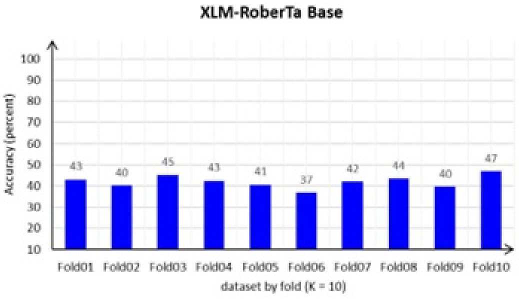


Fig. 3. The accuracy results of the XLM-RoBERTa Base model.

- Experimental of the Bert-base-multilingual-cased model.
  In Figure 4, Based on the experimental results, the Bert-base-multilingual-cased model achieved the highest accuracy of 56 percent, with a character-level prediction accuracy of 53 percent. The model’s recall, indicating the likelihood of correctly predicting words, reached 55 percent. It obtained an F1-Score of 52 percent and an average accuracy of 53 percent.
- Experimental of the DistilBERT-base-multilingual-cased model.
  In Figure 5, Based on the experimental results, the DistilBERT-base-multilingual-cased achieved the highest accuracy of 52 percent, with a character-level prediction accuracy of 55 percent. The model’s recall, indicating the likelihood of correctly predicting words, reached 52 percent. It obtained an F1-Score of 50 percent and an average accuracy of 50 percent.

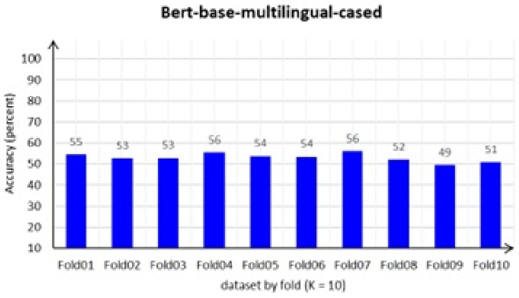


Fig. 4. The accuracy results of the Bert-base-multilingual-cased model.

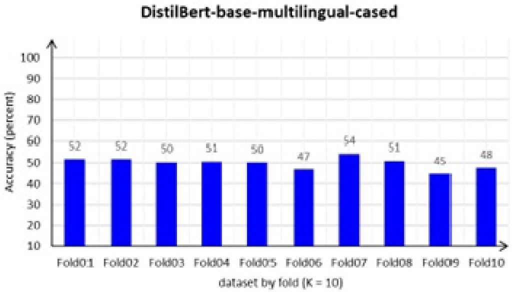


Fig. 5. The accuracy results of the DistilBert-base-multilingual-cased model.

B. The evaluation of the accuracy results for the multi-class model

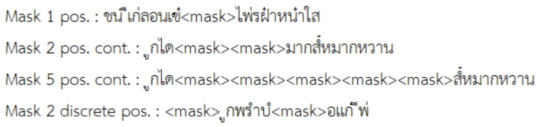
The evaluation of the multi-class model’s accuracy is divided into four methods based on the frequency of characters missing from the positions of the inscription with the highest occurrence, which are 1, 2, and 5 positions, respectively as shown in the example in Figure 6.

Methods of evaluation for the multi-class model are as follows:

- Mask 1 pos. : <mask> a character by randomly removing 1 character.
- Mask 2 pos. cont. : <mask> characters by randomly removing 2 characters, where the masked positions are consecutive.
- Mask 5 pos. cont. : <mask> characters by randomly removing 5 characters, where the masked positions are consecutive.
- Mask 2 discrete pos. : <mask> characters by randomly removing 2 characters, where the masked positions are not consecutive.

C. Perform a performance comparison with other models

From TABLE II, the results of the experiments involving multi-language pre-trained models show that all three models



Mask 1 pos. : ชนี เก่ลอนเข่<mask>ไพ่รฝ่าหน่ำใส

Mask 2 pos. cont. : ูกได<mask><mask>มากส่ำหมากหวาน

Mask 5 pos. cont. : ูกได<mask><mask><mask><mask><mask>ส่ำหมากหวาน

Mask 2 discrete pos. : <mask>ูกพรำบํ<mask>อแก ีพ่

Fig. 6. Example of masking “<mask>” token

TABLE I
THE ACCURACY RESULTS OF CHARACTER PREDICTION FOR ALL THREE MODELS.

Method	**Accuracy of models (percent)**		
	XLM-BERT	***mBERT***	***DISTILBERT***
Mask 1 pos.	47	50	51
Mask 2 pos. cont.	15	29	10
Mask 5 pos. cont.	2	6	5
Mask 2 discrete pos.	28	11	15

were trained to understand the content in the Sukhothai language. Among them, the Bert-base-multilingual-cased model achieved the highest accuracy at 53%,

V. CONCLUSION

The results of the experiments from Tables 1 and 2 in this research reveal that the performance of the models in predicting characters concealed by the <mask> token, using pre-trained models of all three types, namely (1) XLM-Roberta, (2) Bert-base-multilingual-cased, and (3) DistilBERT-base-multilingual-cased, resulted in a total of 72 classes. The accuracy values of the models, as shown in Figure 7, were 42%, 53%, and 50% respectively. Subsequently, the models that underwent learning from all three types of pre-trained models.It was found that all three models performed well in predicting masked characters using the token <mask> at a specific position (1 position randomly), with their performance closely approaching the accuracy values from the reference confusion matrix shown in Figure 8.The DistilBERT model achieved the highest prediction accuracy at 51%, followed by XLM-Roberta and Bert-base-multilingual-cased models with the accuracy of 47% and 50%, respectively.

Next, the models were put to the test to evaluate the accuracy of character prediction. From TABLE I, it was observed that all three models performed better in predicting consecutive missing characters compared to other approaches. Moreover, the prediction results demonstrated that the three models showed similar trends. For predicting two, three, and four consecutive missing characters, the accuracy values were as follows:

TABLE II
COMPARING THE PERFORMANCE OF ALL THREE MODELS.

Accuracy of models (percent)		
XLM-BERT	***mBERT***	***DISTILBERT***
42	53	50

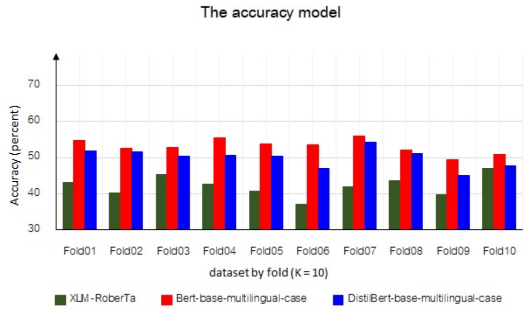


Fig. 7. The 3 model’s accuracies with confusion matrix

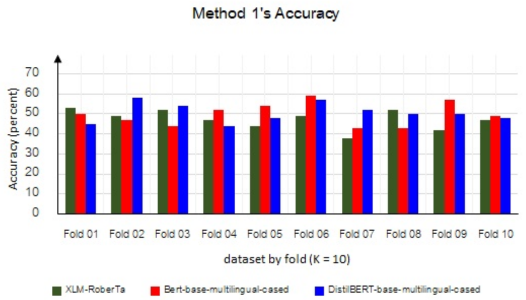


Fig. 8. The 3 model’s accuracies at a specific position

- The accuracy of the model at 2 positions of consecutive <mask>. The Bert-base-multilingual-cased model demonstrates higher proficiency in predicting masked characters compared to other models, with a margin ranging from 9% to 19%, as shown in Figure 9.

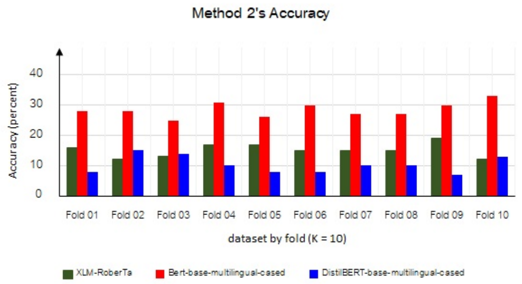


Fig. 9. The accuracy of the model at 2 positions of consecutive <mask>

- The model’s accuracy at 5 positions of consecutive <mask>.
  All three types of models exhibit the ability to predict consecutive masked characters with an accuracy ranging from 2% to 6%. This level of accuracy is considered low and insufficient to effectively predict characters , as shown in Figure 10.
- The model predicts masked characters with a gap of 2 positions, where the positions of <mask> are not consecutive.

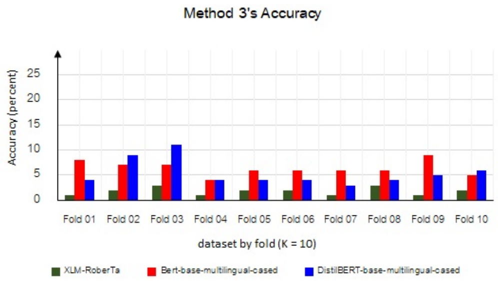


Fig. 10. The model's accuracy at 5 positions of consecutive <mask>

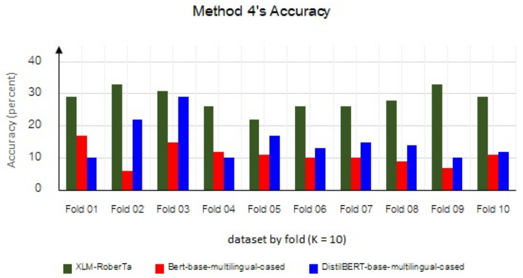


Fig. 11. The model's accuracy with a gap of 2 positions between <mask> positions should not be consecutive.

The XLM-Roberta model demonstrates better proficiency in predicting masked characters with non-consecutive positions compared to other models, as shown in Figure 11.

Acknowledgment

Sujitra Tongkhum thanks for Graduate Scholarship from the Department of Computer Engineering,Chulalongkorn University.

References

[1] Jacob Devlin, M.-W.C., Kenton Lee, Kristina Toutanova BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. 1, 1-2 DOI: https://doi.org/10.48550/arXiv.1810.04805.

[2] Conneau, A., et al., Unsupervised Cross-lingual Representation Learning at Scale. arXiv pre-print server, 2020: p. 1-3.

[3] Wu, J.-L. and W.-Y. Chung, Sentiment-based masked language modeling for improving sentence-level valence–arousal prediction. Applied Intelligence, 2022. 52(14): p. 16353-16369.

[4] Soam, M. and S. Thakur. Next Word Prediction Using Deep Learning: A Comparative Study. IEEE.

[5] Cœdès, G., wat-pa-mamuang, in WAT-PA-MAMUANG INSCRIPTIONS, V.U. Trongjai Hutangkura, Dokrak Payaksri, Editor. 1963, Trongjai Hutangkura,Vasharabhorn Ungkunshutchai,Dokrak Payaksri: THE PRINCESS MAHA CHAKRI SIRINDHORN ANTHROPOLOGY CENTRE.

[6] Y. Cœdès, G., nakhon chum, in NAKHON-CHUM INSCRIPTIONS, V.U. Trongjai Hutangkura, Dokrak Payaksri, Editor. 1983, Trongjai Hutangkura,Vasharabhorn Ungkunshutchai,Dokrak Payaksri: THE PRINCESS MAHA CHAKRI SIRINDHORN ANTHROPOLOGY CENTRE.

[7] Cœdès, G., wat-pa-mamuang, in WAT-PA-MAMUANG INSCRIPTIONS, V.U. Trongjai Hutangkura, Dokrak Payaksri, Editor. 1963: THE PRINCESS MAHA CHAKRI SIRINDHORN ANTHROPOLOGY CENTRE.

[8] Cœdès, G., wat si chum, in wat si chum inscriptions, V.U. Trongjai Hutangkura, Dokrak Payaksri, Editor. 1978, Trongjai Hutangkura,Vasharabhorn Ungkunshutchai,Dokrak Payaksri: THE PRINCESS MAHA CHAKRI SIRINDHORN ANTHROPOLOGY CENTRE.

[9] thongkhamwann, C., pu khun chit khun chot inscriptions, in pu khun chit khun chot, V.U. Trongjai Hutangkura, Dokrak Payaksri, Editor. 1957, Trongjai Hutangkura,Vasharabhorn Ungkunshutchai,Dokrak Payaksri: THE PRINCESS MAHA CHAKRI SIRINDHORN ANTHROPOLOGY CENTRE.

[10] Alpa Reshamwala, P.P., Dhirendra S Mishra, REVIEW ON NATURAL LANGUAGE PROCESSING. IRACST – Engineering Science and Technology: An International Journal (ESTIJ), 2013. 3: p. 2.

[11] Vaswani, A., et al., Attention Is All You Need. arXiv pre-print server, 2017.

[12] Taylor, W.L., "Cloze Procedure": A New Tool For Measuring Readability. Journalism quarterly, 1953. 30: p. 415-433.