Thai Sentence Completeness Classification using Fine-Tuned WangchanBERTa

Pattarapol Pornsirirung
Department of Computer and Information Science, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Khantharat Anekboon*
Department of Computer and Information Science, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Abstract:

Sentence completeness classification plays a crucial role in various natural language processing (NLP) applications, including grammar checking, text auto-completion, and language assessment. This task becomes particularly challenging in Thai due to the language's unique characteristics such as flexible word order, implicit subject omission, and the absence of explicit word boundaries. These linguistic properties make traditional rule-based and statistical approaches prone to errors when applied to Thai. To address these challenges, this research applies modern deep learning techniques, specifically leveraging pre-trained transformer models fine-tuned for Thai sentence completeness classification. This study introduces the use of WangchanBERTa, a Thai-specific adaptation of RoBERTa, pretrained entirely on Thai text. We curated a dataset of 2,000 Thai sentences, half of which were complete, and half incomplete. Each sentence was manually labeled to ensure high data quality. To enhance robustness and mitigate overfitting, we applied stratified 5-Fold Cross Validation and optimized key hyperparameters such as batch size, learning rate, and weight decay. Experimental results show that WangchanBERTa achieves an average accuracy of 99.65%, significantly outperforming mBERT, a popular multilingual baseline, which achieved only 95.82%. Notably, WangchanBERTa required just 1 hour and 15 minutes to train across all folds, compared to mBERT's 2 hours and 59 minutes. Additionally, WangchanBERTa's performance was compared with XLM-R, a state-of-the-art multilingual model, which achieved slightly higher accuracy of 99.90% but at the cost of higher computational requirements. The results emphasize the advantage of language-specific pretraining in capturing the linguistic nuances of Thai. This research highlights the importance of tailored transformer models

for low-resource languages. By demonstrating that WangchanBERTa achieves near state-ofthe-art performance with lower computational cost, this work provides a strong foundation for future Thai NLP research. Future directions include expanding the dataset, incorporating domain-specific texts, and evaluating the performance of other Thai-focused models such as PhayaThaiBERT, ultimately driving further innovation in Thai language understanding.

Keywords:

Natural Language Processing, Thai Language, BERT, Sentence Classification, Transformer Models.