NeuThai reframing: A Thai language neutral reframing dataset and model

Nanthipat Kongborrirak, Pontakorn Yuttakit
Faculty of Arts, Chulalongkorn University
{644011952, 6440156722}@student.chula.ac.th

Abstract

Sentiment transfer like reframing tasks, has been getting more attention in recent years. To contribute to this field of research, we introduce a similar task called “neutral reframing”. The goal is to eliminate negative aspects of text and generate a neutral one without changing its core meaning. To implement this task in Thai, we collected and annotated 1008 Thai negative texts to build a dataset. We then evaluate the state-of-art encoder-decoder models and a decoder-only model. Moreover, we utilize the augmentation method with our best model by translating English text to Thai. Our results indicate that the encoder-decoder models perform better than the decoder-only model overall. However, the augmentation method doesn’t improve the model’s performance.

1 Introduction

Navigating the nuances of language, particularly in the field of sentiment, presents a complex challenge. One challenging task is positive reframing, where the goal is to transform the negative text into a more positive form while retaining its original meaning. Building upon this, we introduce a task of equal significance—"neutral reframing." In this task, our objective is to neutralize negative aspects of text without changing its denotative meaning. Since this task hasn’t been explored in a Thai context before, we implement this task in Thai which lacks the abundance of data compared to English.

Due to the lack of datasets for reframing in Thai, we decided to create the dataset for this task. Comprising 1008 rows annotated using the "neutralize" strategy according to Ziems et al., 2022 (p.3684), “a strategy that involves removing or rewriting negative phrases and terms so they are more neutral (Pryzant et al., 2020)” This strategy is chosen with our belief that it best preserves the original meaning while achieving the desired reframing. For example, a negative review such as "อาหารร้านนี้รสชาติหมาไม่แดก" (This restaurant's food tastes awful) could be reframed as "อาหารร้านนี้ควรปรับปรุงรสชาติอาหาร" (This restaurant's food could use some improvement in taste).

In this work, we leverage state-of-the-art Encoder-Decoder models and compare them with the well-known Decoder model, GPT 3.5, by choosing it as a baseline model for this task. The Encoder-Decoder performed well overall and outperformed the Decoder model across all the metrics by a significant amount. However, we encounter difficulties in fully preserving meaning and eliminating negative words.

To augment our models, we use the English dataset from SALT-NLP (Ziems et al., 2022). Unexpectedly, the results of this augmentation strategy do not align with our expectations. The English data, sourced from Twitter, introduces a layer of complexity, indicating that the language used on this platform poses challenges when translated into Thai. This displays the difficulties of the augmentation process from English to Thai in the context of our reframing task. Overall, our main contributions are as follows.

1. Introduction of a Novel Task in Thai: Introduction of the challenging task of "neutral reframing" in Thai, aimed at neutralizing negative text while preserving its denotative meaning.
2. Creation of Thai Dataset: Development of a Thai dataset for neutral reframing, consisting of 1008 rows annotated using the "neutralize" strategy. This dataset addresses the lack of resources in Thai for such tasks.

3. Exploration of Augmentation Strategy: Exploration of an augmentation strategy using English data from SALT-NLP, aiming to improve the model. The unexpected outcome highlights the challenges in translating language from English to Thai, particularly in the context of reframing tasks.

2 Related Work

In natural language processing (NLP), the concept of reframing text to induce a positive perspective without changing the original content is new and a challenging task. (Ziems et al., 2022) introduced the task of positive reframing, focusing on transforming the tone of negative texts into more positive moods while retaining the original text's meaning. This task is not the same as a standard sentiment transfer task that often results in semantic changes through reframed sentiment. This research utilized the Positive Psychology Frames benchmark, which consisted of sentence pairs and structured annotations related to 7 theoretical reframing strategies.

In the study (Chakrabarty et al., 2021) focusing on the connotative meaning of words, the authors aim to reframe political arguments to be perceived as more trustworthy, while preserving their denotative meaning. They recognize that subtle changes in wording can introduce biases or manipulate emotions, and there is no parallel dataset available for training models on reframing arguments. To tackle this, the researchers introduce ENTRUST, a method that utilizes a connotation-based lexical resource to generate a dataset of arguments and their reframed counterparts. By combining controllable text generation techniques with an entailment component that ensures factual consistency, the study addresses the challenge of changing the tone of a statement without altering its factual content. Their evaluation demonstrates that ENTRUST effectively produces reframed arguments that are fluent, retain their original meaning, and are perceived as more trustworthy, outperforming strong baselines and nearing human performance in these aspects.

In our work, we focus on preserving the denotative meaning in the Thai language. Moreover, we also provide a dataset to train a supervised model in a neutral reframing task.

3 Neutral Reframing

Previous studies adopt the positive psychology frame, notably outlined by (Harris et al., 2007). Among the various strategies proposed by (Ziems et al., 2022), we've identified a challenge in retaining the original connotation for strategies other than "neutralizing." Consequently, our experiment focuses on the "neutralizing" strategy.

3.1 Neutralizing

Ziems et al., 2022 (p.3684) said “Neutralizing involves removing or rewriting negative phrases and terms so they are more neutral (Pryzant et al., 2020)”. Considering our example, “บริการที่นี่ห่วยแตกมาก” (The service here is very awful.) could be reframed to “ที่นี่ควรปรับปรุงการให้บริการ” (The service here could use some improvement.)

3.2 NeuThai dataset

Unlike existing studies that centered on English datasets, we tackle this task in the Thai context. We present our dataset “NeuThai” for the neutral reframing task. We use this dataset in both the training and evaluating process.

3.3 Annotation

We recruited annotators to reframe 1008 randomly sampled text. We have 5 annotators in total including ourselves. We asked them to reframe the text given to them by making them sound as neutral and natural as they could. Before they worked on the tasks, we provided them with some neutral reframing examples, along with annotation guidelines to help them understand the task provided in appendix B.

3.4 Augmentation

We aim not only to explore the efficacy of neutral reframing in the Thai context but also to implement augmentation experiments by translating English data from the SALT-NLP dataset (Ziems et al., 2022) to Thai. This approach anticipates potential enhancements in the model's performance.

4 Dataset

We didn’t have any parallel dataset for reframing task in Thai at first. Thus, we decided to create a dataset “NeuThai” annotated by ourselves and hired annotators. Our dataset draws from two primary sources: the Wisesight corpus

(Suriyawongkul et al., 2019) and the Thai toxic tweet dataset (Sirihattasak et al., 2019). From the Wisesight corpus, we collected 760 samples. To find texts with negative sentiments, we filtered for texts where emojis conveyed a sense of sadness. Additionally, we gathered 248 samples from the Thai toxic tweet dataset. Both datasets originated from Twitter, offering a diverse and dynamic source of textual data. In total, our dataset contains 1008 rows.

We employed an 8:1:1 ratio for data splitting, resulting in 806 samples for training, 101 for development, and 101 for testing. Our dataset is structured with two key columns: "original_text," containing negative text from the mentioned datasets, and "reframed_text," which serves as the gold standard reference for text reframed by human annotators. Finally, we removed emojis using the emoji library.

5 Model

5.1 GPT-3.5 fewshot

We accessed GPT3.5 (Brown et al., 2020) through https://chat.openai.com.We wrote the prompt “Reframe to:” with the original text before the prompt and reframed text after the prompt. We gave 5 examples of original texts and reframed text aligned with the prompt. Then, we gave a test dataset which only consisted of original text aligned with the prompt but not with reframed text. We collected the model generated texts to make the full predicted reframed text of the test dataset.

5.2 mBART

We used “facebook/mbart-large-50” (Tang et al., 2021). This model is a multilingual version of BART (Bidirectional and Auto-Regressive Transformers), which utilizes a denoising autoencoder for pre-training. mBART is pre-trained by corrupting text with an arbitrary noising function and learning to reconstruct the original text.

5.3 mBART with augmented dataset

mBART was augmented by the translated dataset from SALT-NLP (Ziems et al., 2022), filtering only labels that contain “neutralizing”. The translation process used the model “facebook/nllb-200-distilled-600M”(NLLB Team et al., 2022) to translate from English to Thai.

5.4 mT5

It’s the multilingual version of T5. mT5 is pre-trained on a large corpus of text using an objective similar to the denoising autoencoder approach, but it is more generalized. It uses a process called "text infilling," where some spans of text are replaced with a special placeholder token, and the model learns to predict the missing text. This self-supervised training task is known as "span corruption." Through this process, mT5 gains a broad understanding of language and the relationships between different parts of text. During the training process, we had a problem with mT5 tokenizer. We tried some models and found out the model that the tokenizer worked. It’s fine-tuned for summarization of mT5 “csebuetnlp /mT5_multilingual_XLSum” (Hasan et al., 2021).

6 Experiment

6.1 Experimental conditions

We chose GPT 3.5 with few-shot learning as our baseline because we want to utilize the world knowledge from pre-trained processing in this large language model. Given only 5 examples, the model will know how to generate text based on the previous examples. Instead of giving the entire train dataset, it would predict by its knowledge rather than trying to learn on dataset knowledge.

According to (Ziems et al., 2022) GPT-3 few-shot and BART both perform the best but we thought GPT may not perform well for Thai which would be found in the results section. We decided to use GPT 3.5 few-shot as our baseline and multilingual version of BART as well as another encoder-decoder model namely T5 as our methods.

6.2 Implementation

Our implementation is built upon the SALT-NLP codebase. The original code is accessible on GitHub. https://github.com/SALT-NLP /positive-frames/blob/main/run.py This code provided the foundation for utilizing neutral reframing techniques in our model. Our goal was to adapt the existing codebase to handle the nuances of the Thai language and the characteristics of our collected dataset. We used Colab Pro with the NVIDIA V100 GPU and the NVIDIA T4 to train our model. The entire training phase was completed in approximately 5 minutes.

We experimented with different learning rates, epochs, and batch sizes for both mBART and mT5 architectures.

mBART The optimal hyperparameter combination was found to be a learning rate of 1e-5, a batch size of 4, and training for 4 epochs.

mT5 The best hyperparameter for T5 involved a learning rate of 5e-5, a batch size of 4, and a training duration of 4 epochs.

Model		Automatic Evaluation						Human Evaluation		
		R-1	R-2	R-L	BLEU	∆Sentiment	Avg. Len	Meaning	Neutrality	Fluency
Decoder only	GPT-3.5 few-shot	22.47	7.80	19.54	0.3	0.41	**22.48**	3.34	3.21	2.74
Encoder-Decoder	mT5 sum	36.22	19.63	33.99	**4.2**	**0.56**	6.81	**3.40**	3.32	**4.12**
	mBART	42.37	23.10	39.56	1.9	0.46	10.60	3.35	**3.53**	4.09
	mBART augment	**42.65**	**23.45**	**40.08**	2.0	0.42	11.92	3.32	2.37	3.69
	Human	100	100	100	100	0.69	9.66	4.28	4.42	4.40

Table 1: Neutral reframing results calculated by using ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L), BLEU, and ∆Sentiment via finetuned WangchanBERTa. Meaning, Neutrality, and Fluency were manually evaluated by humans. The best performance is **bolded**.

7 Evaluation and Results

Following the previous English positive reframing work (Ziems et al., 2022), we evaluate our models for semantic similarity with the ground truth using the BLEU (Papineni et al., 2002), ROUGE (Lin, 2004). They also used TextBlob sentiment polarity ranging between -1 and 1 to find a sentiment difference. However, TextBlob doesn’t support Thai language and most Thai sentiment analysis models predict the probability of sentiment labels rather than a sentiment polarity score. Thus, we decided to normalize a predicted label with the highest probability of sentiment rank. To make sentiment ranks calculable, sentiment ranks are categorized into 1,0, and -1 for positive, neutral, and negative respectively. Sentiment rank differences are calculated by the sentiment rank of reframed text subtracted by the sentiment rank of the original text. For example, if original is negative and when reframed, its rank change will be 1 for neutral, 2 for positive, and 0 for negative. We report the average change in sentiment rank using WangchanBERTa (Lalita Lowphansirikul et al., 2021) that was fine-tuned by (Poom-sci, 2021) for a sentiment analysis task. Finally, we conducted a human evaluation to rate how well our models performed on a scale from 1 to 5. We have 3 criteria as follows, Meaning Preservation (Shang et al., 2019), Fluency of the generated text, and Neutrality which refers to how neutral the text is (Luo et al., 2019).

7.1 Automatic Evaluation

Across these metrics (Table 1) all Encoder-Decoder models outperformed Decoder only model GPT-3.5. mBART augment achieved the highest ROUGE score while mT5 sum achieved the highest BLEU and ∆Sentiment score. On the other hand, GPT-3.5 few-shot provided the highest average length of tokens but had the worst quality of reframing. We observe that both mBART augment and mBART yielded comparable results. This finding suggests that the augmentation strategy applied in the reframing task for Thai did not yield significant improvements, emphasizing the nuanced nature of language tasks in the Thai context.

7.2 Human Evaluation

mT5 yielded the highest score for meaning preserving and fluency. Contrastingly, mBART stood out with the highest score for neutrality, surpassing mBART augment. Surprisingly, mBART augment displayed a lower neutrality score than GPT-3.5 few-shot, our baseline model, showing more evidence that the augmentation strategy might not be helpful for the Thai reframing task. GPT-3.5 few-shot, despite its low fluency score, performed competitively in terms of meaning preservation and neutrality.

These human evaluation (Table 1) insights provide valuable guidance for selecting models that align with the specific requirements of the Thai reframing task, emphasizing the significance of meaning preservation and neutrality.

7.3 Error Analysis

As the score of our best model mBART could be improved further, we manually go through all the 101 texts generated by mBART to find errors. The main errors we observed are misunderstanding the meaning, contradiction in meaning, and negative words. In the case of misunderstanding the meaning, these often occurred in text that contained sarcasm or irony. For example "นี่คนนะไม่ใช่ถังขยะ ทำไมวันนี้ชอบเหวี่ยงนี่กันจัง เป็นฝ่ายถูกนะเนี่ย" was reframed to "ไม่ควรเหวี่ยงกับถังขยะ". On the other hand, the contradiction in meaning frequently occurred in the text that contained negation like "ไม่". The original meaning was changed, in the phrase like "โจ๊กไม่อร่อยแบบหมดอารมณ์กินมาก เหมือนซื้อมาทิ้ง" was reframed to "อยากกินโจ๊กไม่อร่อยจัง". Finally, negative words seldom appeared in generated texts like "อยากนอนโง่ๆอยู่บนเตียงแต่แม่งต้องตื่นมาเจอความจริงที่โคตรไม่อยากเจอ" was reframed to "อยากนอนโง่ๆอยู่บนเตียง แต่ตื่นมาเจอความจริงที่ไม่อยากเจอ".

8 Discussion

8.1 Encoder-Decoder Suitability

Our findings support that Encoder-Decoder models are better suited for the Thai reframing task compared to Decoder-only models like GPT-3.5. The reason might be the capability of Encoder-Decoder architectures to capture and represent input sequences in a meaningful way before generating the output sequence. The context-aware representation enables more coherent and contextually relevant reframed outputs, which is crucial for tasks demanding nuanced language understanding.

8.2 Limitations of Decoder-Only Models

Decoder-only models displayed limitations in generating outputs that align with the context of the reference text in the Thai reframing task. This indicates that decoding without a thorough understanding of the input sequence may lead to less relevant and contextually appropriate outputs, resulting in a worse quality of reframing.

8.3 Augmentation Strategy Challenges

Our findings indicate that the augmentation strategy, particularly in the form of translation from English to Thai, may not be as beneficial for the Thai reframing task. The challenges come from the complexity of translating English social media language, which often includes slang and diverse language variations, into Thai. The resulting translations may introduce noise or inaccuracies, leading to a negative result for the dataset augmentation.

8.4 Challenge in Reframing Thai Text

As we mentioned the main errors we found in the Error Analysis section, we analyzed and observed that the reframing task in Thai is challenging for several reasons. Firstly, in this work, our dataset was constructed from social media language, which contained a wide range of language diversity. The models may face difficulties in comprehending the languages used on social media. Finally, the size of our dataset might not be big enough for the models to learn all the negative words in Thai in order to eliminate or neutralize them.

9 Conclusion

This study focuses on the sentiment task, specifically centered on neutral reframing in the Thai language. The objective was to modify the tone of expressions from negative to neutral while retaining the core meaning. We proposed the utilization of a state-of-the-art Encoder-Decoder model, mBART and mT5, and a Decoder model, GPT 3.5, to undertake the neutral reframing tasks in the Thai language. Though the Encoder-Decoder models performed well overall, The meaning preserving and the ability to neutralize the text could use some improvement. For future directions, we first aim to enhance the Thai dataset for reframing by expanding both its size and diversity to improved model performance. Additionally, we could explore various strategies to improve our model other than augmentation, especially for the Thai reframing task.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Hesse, C. (2020). Language Models are Few-Shot Learners. *Arxiv.org*. https://arxiv.org/abs/2005.14165

Chakrabarty, T., Hidey, C., & Muresan, S. (2021, June 1). *ENTRUST: Argument Reframing with Language Models and Entailment*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.394

Harris, A. H. S., Thoresen, C. E., & Lopez, S. J. (2007). Integrating Positive Psychology Into Counseling: Why and (When Appropriate) How. *Journal of Counseling & Development*, *85*(1), 3–13. https://doi.org/10.1002/j.1556-6678.2007.tb00438.x

Hasan, T., Bhattacharjee, A., Islam, Md. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., & Shahriyar, R. (2021, August 1). *XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.413

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, & Sarana Nutanong. (2021). WangchanBERTa: Pretraining transformer-based thai language models. *CoRR, abs/2101.09635*. https://arxiv.org/abs/2101.09635

Lin, C.-Y. (2004, July 1). *ROUGE: A Package for Automatic Evaluation of Summaries*. ACLWeb; Association for Computational Linguistics. https://aclanthology.org/W04-1013

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., & Hoffman, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv:2207.04672 [Cs]*. https://arxiv.org/abs/2207.04672

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. https://doi.org/10.3115/1073083.1073135

Poom-sci. (2021). *poom-sci/WangchanBERTa-finetuned-sentiment*. https://huggingface.co/poom-sci/WangchanBERTa-finetuned-sentiment/blob/main/README.md

Sirihattasak, S., Komachi, M., Ishikawa, H., & Ishikawa, H. (2017). *Annotation and Classification of Toxicity for Thai Twitter*. www.ta-cos.org. https://www.ta-cos.org/node/25

Steven Loria. (2018). textblob documentation. *Release 0.15, 2:269*.

Suriyawongkul, A., Chuangsuwanich, E., Chormai, P., & Polpanumas, C. (2019). *PyThaiNLP/wisesight-sentiment: First release*. https://doi.org/10.5281/zenodo.3457447

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., Ai, F., & Ai, A. (2021). *Multilingual Translation from Denoising Pre-Training* (pp. 3450–3466). https://aclanthology.org/2021.findings-acl.304.pdf

Ziems, C., Li, M., Zhang, A., & Yang, D. (2022). Inducing Positive Perspectives with Text Reframing. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*(60), 3682–3700. https://aclanthology.org/2022.acl-long.257.pdf

A Example Reframes

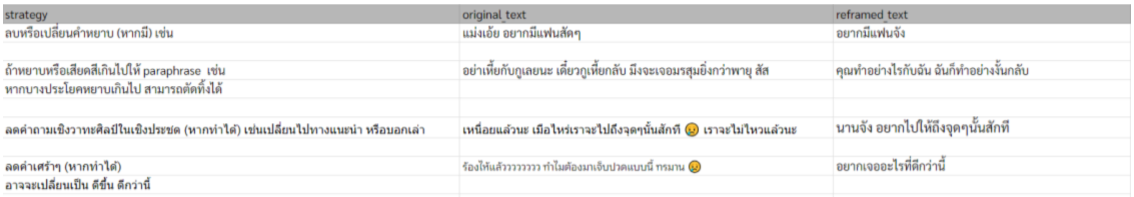
Original Text		อย่าลืมโอนเงินกันนะคะ ภายในวันที่28ค่ะไม่ตามนะไม่ให้เลื่อนด้วยนะ! นานมาแล้ว
Gold Standard		โอนเงินภายในเวลาด้วย
Generated Text	GPT-3.5 few-shot	อย่าลืมโอนเงินตามกำหนดภายในวันที่ 28 อย่างเด็ดขาดนะคะ! มานานแล้ว
	mT5	ขอให้เลื่อนไปก่อน
	mBART	อย่าลืมโอนเงินภายในวันที่ 28 นะคะ
	mBART augment	อย่าลืมโอนเงินกันนะคะ ภายในวันที่28ค่ะ

Table 2: A model comparison for reframing the same text. In this case, mBART performed better than other models.

Original Text		ต้องเริ่มอ่านหนังสืออีกแล้วหรอวะ เห้ออออออออออออออออออ
Gold Standard		ต้องเริ่มอ่านหนังสือแล้ว
Generated Text	GPT-3.5 few-shot	ความต้องการที่จะเริ่มอ่านหนังสืออีกครั้ง แม้ว่าจะเป็นเรื่องที่เหนื่อยแล้ว
	mT5	ต้องเริ่มอ่านหนังสือแล้ว
	mBART	ต้องเริ่มอ่านหนังสือจัง
	mBART augment	อยากเริ่มอ่านหนังสือจัง

Table 3: A model comparison for reframing the same text. In this case, mT5 performed better than other models and the result matched the gold standard.

B Annotation Guidelines



strategy	original_text	reframed_text
ลบหรือเปลี่ยนคำหยาบ (หากมี) เช่น	แม่งเอ้ย อยากมีแฟนสัดๆ	อยากมีแฟนจัง
ถ้าหยาบหรือเสียดสีเกินไปให้ paraphrase เช่น	อย่าเหี้ยกับกูเลยนะ เดี๋ยวกูเหี้ยกลับ มึงจะเจอมรสุมยิ่งกว่าพายุ สัส	คุณทำอย่างไรกับฉัน ฉันก็ทำอย่างงั้นกลับ
หากบางประโยคหยาบเกินไป สามารถตัดทิ้งได้		
ลดคำถามเชิงวาทะศิลป์ในเชิงประชด (หากทำได้) เช่นเปลี่ยนไปทางแนะนำ หรือบอกเล่า	เหนื่อยแล้วนะ เมื่อไหร่เราจะไปถึงจุดๆนั้นสักที 😢 เราจะไม่ไหวแล้วนะ	นานจัง อยากไปให้ถึงจุดๆนั้นสักที
ลดคำเศร้าๆ (หากทำได้)	ร้องไห้แล้ววววววว ทำไมต้องมาเจ็บปวดแบบนี้ ทรมาน 😢	อยากเจออะไรที่ดีกว่านี้
อาจจะเปลี่ยนเป็น ดีขึ้น ดีกว่านี้		

Figure 1: Annotation guidelines for neutral reframing