Transformer-based Conditional Language Models to Generate Filipino News Articles

Kenrick Lance T. Buñag
Student of BS Computer Science Program
College of Computer Studies
Angeles University Foundation
Angeles City, Philippines
kenricklance092@gmail.com

Rosanna A. Esquivel
Professor of Computer Science
BS Computer Science Program Chair
College of Computer Studies
Angeles University Foundation
Angeles City, Philippines
esquivel.rosanna@auf.edu.ph

Abstract

The study aims to develop an application that when conditioned on an image, generates a Filipino news article that is similar to a real article. It explores the capabilities of current NLP algorithms for language generation in the low-resource Filipino language. In order to carry out the study, a dataset of Filipino news articles with accompanying images will be collected. With this, the study also sets out to publish a dataset for conditional language generation in the Filipino language.

Keywords

Natural Language Processing, NLP, Natural Language Generation, Deep Learning and Machine Learning

1. Introduction

Natural language generation is a powerful AI technology that has many useful applications today. It is being used commercially with voice assistants, chatbots, automated journalism, summarizing medical records, etc. The study is investigating the capabilities of current NLP algorithms in the under researched area of text generation in the Filipino language. It specifically investigates current approaches to conditional language generation, where the computer application generates text using some context, such as with machine translation or text summarization.

There have been many previous papers in the field of NLP for the Filipino language. There has been previous work in NLP tasks for text classification (Cruz & Cheng, 2020), part-of-speech tagging (Nocon & Borra, 2016), and named-entity recognition (Cruz et al., 2018) to name a few. There has also been a few research on text generation and language modeling (Cruz et al., 2020). However, there has been no research on conditional language generation for the Filipino language. There does currently exist an unlabeled corpora for Filipino called WikiText-TL-39 and a Filipino news article dataset called NewsPH although the news articles in that do not have associated images with them. The Filipino language suffers from data scarcity which makes it difficult to develop NLP applications for it.

The research will be carried out to develop an application that will generate Filipino news articles. The research will be beneficial to future researchers in NLP as this study explores the capabilities of current state-of-the-art (SOTA) neural network architectures with a low resource language like Filipino on conditional language generation. A dataset for conditional language generation will be created for this study which can be used for further research with various other NLP tasks with the Filipino language.

1.1. Objectives

The study aims to utilize a Transformer architecture to develop an application that when given an image and category, generates a Filipino news article relating to that image and category that would be similar to a real article. A dataset consisting of Filipino news articles and associated images will be collected to be used as training data for the model. A previously pre-trained model will be fine-tuned on the collected dataset for the task of image and class conditional language generation. Two metrics will be computed to evaluate the model on these tasks. Namely, perplexity as a proxy for text generation quality and prompt ranking accuracy to evaluate prompt adherence.

2. Literature Review

The task of generating news articles is language modeling. Language models compute the probability of occurrence of words in a sequence. There have been many approaches to this. In the pre-neural network era of NLP, the most notable of which are n-gram language models (Cavnar & Trenkle, 1994). The performance of these pre-neural network methods have eventually been exceeded by algorithms that utilize word embeddings and neural networks (Mikolov et al., 2013). Algorithms based on the widely used recurrent neural networks have demonstrated great results in NLP tasks (Sundermeyer, Schlüter, & Ney, 2012). In 2017, the Transformer was proposed (Vaswani et al., 2017). Transformers now represent the SOTA in NLP tasks. GPT-2, a variant of the transformer, trains in a unidirectional language modeling setting and architectures similar to it have been shown to generate remarkably coherent text (Radford et al., 2019).

Our task which is generating news articles with the context of an image or class is conditional language modeling. Conditional language modeling is a widely explored task. Other works have demonstrated strong results in machine translation, summarization, and dialogue. However, these researches are limited to tasks where the language model is conditioned on text (Dong et al., 2019; Lample & Conneau, 2019). Pseudo-Self attention is a mechanism for conditioning a language model with any source modality even when the source modality is not text (Ziegler et al., 2019). This approach showed better results than other baseline approaches tested against it.

3. Methodology

3.1. Data Collection

Filipino news articles were web scraped from various news outlets through their websites. The news outlets in particular are: Abante, Abante TNT, Abante Tonite, Abante Tonite Archive, GMA Balitambayan, Inquirer Bandera, PhilStar Ngayon, and Pilipino Mirror. From each article in each website, the article body, article title, accompanying image, and various metadata such as date, author, category, article url, and image url were collected. After data cleaning, a total of 351,755 articles have been collected.

3.2. Data Preprocessing

As the dataset was web scraped from various websites and are not certain to be of high quality, the researchers had needed to clean the dataset. To clean the dataset, a number of filters were applied to each article.

1. Length Filter - Remove articles which have fewer than 15 words.
2. Non-latin Filter - Remove articles which consist of more than 25% non-Latin characters.
3. HTML Filter - Remove articles where the number of occurrences of HTML-related sequences of characters are more than 10% of the number of total words.

Articles where its accompanying image was not able to be read by the imaging library used, were also removed.

Data segmentation was done to divide the differently named categories of each article into 5 distinct classes: “News”, “Sports”, “Entertainment”, “Crime”, and “Other”. They are grouped according to what the researchers thought was appropriate for the category through qualitative analysis. The following shows which categories from the different news outlets were grouped to each class.

1. News - “News”, "Top News", "Local News", "balita", "talakayan", "pinoyabroad", "Bansa", "BALITA", "NEGOSYO", "PANANALAPI"
2. Sports - "Sports", "Atletiko Radar", "Sports Stories", "Sports Columnists", "Palaro"
3. Entertainment - "Entertainment", "LifeStyle", "ShowBiz", "Showbiz", "Showbiz Stories", "chikamuna", "chika"

4. Crime - "Crime", "Metro", "VisMin", "promdi", "Balitang Promdi", "Probinsiya", "METRO"
5. Other - All other categories

Special tokens were included with the inputs to the model to denote the boundaries of the different sections in the news articles such as start of the title, start of the body, and end of the article. The final text was then tokenized with a byte pair encoding (BPE) tokenizer pretrained on WikiText-TL-39 and NewsPH. The text was then truncated to a max length of 1024 tokens. The text together with the image and an integer from 0 to 4 relating to the category class is fed to the language model.

3.3. Model Architecture

As transformers now represent the SOTA in NLP tasks, a transformer based model was employed. The transformer was initially proposed as an architecture for text-to-text tasks, having an “encoder” and “decoder” section. Subsequent research saw the exploration of similar architectures to the transformer such as using only the encoder or only the decoder as the full architecture. An example of which is GPT-2. It is a decoder-only transformer that excludes the cross-attention layer of the decoder blocks. The masked self-attention layer of the decoder disables future tokens from being attended to and there is no need for the cross-attention layer as there are no encoder representations to attend to. These attributes along with various other changes to the Transformer architecture has led GPT-2 to be great at unconditional language generation tasks. Accordingly, a model architecture that is based on GPT-2 was used for this study.

Pre-training is the process of first training a model on a separate task (most often with large amounts of data) and using the learned parameters to build strong representations of language and better parameter initializations for downstream tasks. This pre-training/finetuning paradigm has been shown to be widely successful. The most popular method for pre-training for NLP tasks and the one that was used for this study is language modeling. A GPT-2 model was acquired that was trained by performing language modeling on a far larger corpus, specifically, WikiText-TL-39 and NewsPH. This model was then fine-tuned on the gathered dataset for the task of conditional language modeling.

To fine-tune the unconditional language model for conditional language modeling, PSA was employed. Our task of taking an image and category class and generating article text is of image-conditional language generation and class-conditional language generation. Pseudo-Self attention(PSA) was chosen as it demonstrated good performance for both those tasks given a pre-trained unconditional language model.

Pseudo-Self attention calls for an “encoder” that will feed contextual conditioning into the language model. The encoder for the image in this case needs to take in an image and output meaningful information about the image. A pre-trained ResNet50 was used for this. Alternatives such as the SOTA Vision Transformer were considered but a disadvantage with them was the comparatively much higher training time and parameters associated with them. Given all that, ResNet works as a good baseline CNN. Precisely, the CNN encoder is a ResNet50 without the final linear layer that goes through an additional linear layer to translate the 2048 sized vector to the specified embedding size (768 in this case) needed for use in PSA.

To inject class-conditional information, the category class indices go through an embedding layer of the specified embedding size to produce the vector representations for the categories. The image representation and the category representation is concatenated and fed to the rest of the model.

3.4. Model Evaluation

Perplexity(PPL) is used in order to tune hyperparameters and evaluate test set performance. PPL is also used as a proxy for overall text generation quality. Prompt ranking is used to measure how relevant the image and category class is to its text. We follow prompt ranking evaluation as conducted in Fan et al. (2018). To calculate prompt ranking, for each article in the test set, in the case for evaluating image prompt adherence, 9 other images are randomly sampled. The likelihood of its article text is evaluated by the model given the “true” image and the 9 other “fake” images. The rank accuracy is the percentage of which the model gives the highest likelihood to the “true” image. In the case for category class, the likelihood of its article text is evaluated by the model given the “true” class and the rest of the four classes.

4. Results and Discussion

Shown in figure 1 is a cherry-picked example of a news article that was generated by the language model. The generated article was given as context, the "News" category class and an image of a rail transit train. Sampling methods that have demonstrated to improve generation quality such as nucleus sampling, top-k sampling, repetition penalty, and temperature were used.

News

LRT-1 isinailalim sa rehabilitasyon

Isasailalim sa rehabilitasyon ang Light Rail Transit Line 1 (LRT-1) matapos magkaroon ng sunog sa isang istasyon nito na ikinasugat din ng ilang pasahero.

Ayon kay Atty. Hernando Cabrera, tagapagsalita ng LRT-1, nagkaroon ng sunog kahapon ng tanghali at idineklarang under control dakong alas-12:00 ng tanghali sa Anonas Station.

Wala namang iniulat na nasaktan o nasawi sa insidente habang inaalam pa rin kung ano ang naging sanhi ng pagsiklab ng apoy at kung magkano ang halaga ng napinsala.

Figure 1. Example of a Generated News Article

Various models were evaluated on both metrics. Pretrained and non-pretrained variants were tested. Versions that only use PSA to include either the image or the category context were also tested as well as normal GPT-2 transformers that did not incorporate PSA.

PPL is evaluated here to represent overall text generation quality. A lower PPL has the model giving a higher probability to the test set. Table 1 shows the perplexities of the tested models. The pretrained PSA Transformer with both image and category context has the lowest PPL overall from the tested models. PSA Transformers for both pretrain and non-pretrain cases outperforms their normal transformer counterparts in terms of PPL. Also not surprisingly, the non-pretrained models have higher PPLs than the pretrained models.

Table 1. Text Generation Quality as Measured by Perplexity

Model		Perplexity ↓
Pretrained		
	Transformer	10.385
	PSA Transformer w/ image	10.348
	PSA Transformer w/ category	10.347
	PSA Transformer w/ image and category	**10.343**
Non-Pretrained		
	Transformer	10.702
	PSA Transformer w/ image and category	10.642

Prompt ranking accuracy will be a measure of how relevant the image and category class is to its text. Models which do not have any way to utilize either image or category class will have a 10% rank accuracy for images and 20% rank accuracy for category class as though they were randomly picking their prompt ranking class. Table 2 shows rank accuracy of the tested models.

The pretrained PSA transformer with just image context has the highest rank accuracy for images while the non-pretrained PSA transformer with both image and category context has the highest rank accuracy for categories. It is observed that both pretrain PSA transformers with just the image context and with just the category context outperform the model with both contexts on their respective rank accuracy. This may be due to the image and category sharing the same weights in the PSA transformer with both contexts, causing them to get lower prompt accuracies compared to "specialized" models that only include one of either image or class context.

Table 2. Image and Category Adherence as Measured by Rank Accuracy

Model		Rank Acc. (Image) ↑	Rank Acc. (Category)↑
Pretrained			
	Transformer	10%	20%
	PSA Transformer w/ image	**62.71%**	20%
	PSA Transformer w/ category	10%	74.75%
	PSA Transformer w/ both image and category	58.20%	59.04%
Non-Pretrained			
	Transformer	10%	20%
	PSA Transformer w/ both image and category	57.23%	**77.41%**

4.1. Interpretation of Results

The PSA transformer models outperform the normal transformer models for perplexity. They also outperform normal transformer models when it comes to prompt ranking accuracy. With the normal transformers not utilizing the image or category class in any way.

This tells us that the Filipino news article dataset gathered for this research consists of articles wherein their images and categories are related and correlated to their text; and that modern NLP architectures are able to find these patterns. This also tells us that modern NLP architectures, in particular the Transformer and PSA are capable of working with the low-resource Filipino language in the task of conditional language generation.

4.2. Dataset

The NLP tasks explored in this research were image-conditional language generation and class-conditional language generation. The dataset gathered for this research however can be used to explore many other NLP tasks in the Filipino language and low-resource language setting.

The news articles gathered have four main components: Title, Body, Image, and Category. They can be used in many different combinations to explore different tasks. Some of these tasks are:

- Abstractive Summarization - Generation of a summary of a text from its main ideas, not by copying verbatim. Given the article body, generate a title.
- Prompt-conditional Language Generation - Generation of text from a pre-specified text prompt. Given the article title, generate a body.
- Text Classification - Assign a set of predefined categories to text. Given the article title and body, classify the category.
- Text-conditional Image Generation - Generation of an image from a text prompt. Given the article title and body, generate an image.

5. Conclusions and Future Research

We developed a Transformer-based language model that was fine-tuned on Filipino news articles and images. For both image-conditional and class-conditional language generation tasks we were able to observe improved results compared to baseline models and show that current NLP architectures are capable of working with the low-resource Filipino language for conditional language generation. Future work may employ other recent architectures that have emerged in the fields of image captioning, visual storytelling, and similar fields.

The dataset gathered for this study can be used to explore many other different NLP tasks for the Filipino language. Image-conditional language generation and text-conditional image generation in particular are two NLP tasks that this dataset has made possible that wasn't feasible to explore before with the current datasets. This research along with the dataset intends to aid future work on NLP with Filipino and other low-resource languages.

References

Cavnar, W. and Trenkle, J., N-gram-based text categorization, *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175, 1994.

Dela Cruz, B. M., Montalla, C., Manansala, A., Rodriguez, R., Octaviano. M. and Fabito, B., Named-Entity Recognition for Disaster Related Filipino News Articles, *TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South)*, pp. 1633-1636, 2018.

Cruz, J. B. C. and Cheng, C., Establishing baselines for text classification in low-resource languages, *arXiv preprint*, arXiv:2005.02068, 2020.

Cruz, J. B. C., Resabal, J. K., Lin, J., Velasco, D. J. and Cheng, C., Exploiting News Article Structure for Automatic Corpus Generation of Entailment Datasets, *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II*, pp. 86-99, 2021.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. and Hon, H., Unified language model pre-training for natural language understanding and generation, *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no.1170, pp. 13063-13075, 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J., Distributed representations of words and phrases and their compositionality, *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3111-3119, 2013.

Nocon, N. and Borra, A., SMTPOST Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging, *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters*, 2016.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., Language models are unsupervised multi task learners, *OpenAI blog*, 1.8 (2019): 9.

Sundermeyer, M., Schlüter, R. and Ney, H., From feedforward to recurrent LSTM neural networks for language modeling, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517-529, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I., Attention is all you need, *Advances in neural information processing systems*, pp. 6000-6010, 2017.

Ziegler, Z.,Melas-Kyriazi, L., Gehrmann, S. and Rush, A., Encoder-agnostic adaptation for conditional language generation, *arXiv preprint*, arXiv:1908.06938 (2019).

Biographies

Kenrick Lance Buñag is a student in the BS Computer Science Program at the College of Computer Studies - Angeles University Foundation, Philippines. His research interests include natural language processing, computer vision, and deep learning.

Rosanna Esquivel is a Professor of Computer Science and Program Chair for the Computer Science program at the College of Computer Studies - Angeles University Foundation, Philippines since 2008 up to the present where she also served as the Assistant Dean from 2014-2019. She served as the Undergraduate Thesis Coordinator at De La Salle University - Dasmariñas from 2004-2006.