Text-To-Picto Using Lexical Simplification

Notebook for the ImageCLEF Lab at CLEF 2024

Abbhinav Elliah, Ananth Narayanan P, Bhuvan S and P Mirunalini

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

Augmentative and Alternative Communication (AAC) provides a lifeline for those with language problems by employing pictograms to ensure precise message conveyance. This study fine-tunes a pre-trained translation model for text-to-picto conversion, utilizing tokenization and lexical simplification. The model aids individuals with language impairments due to genetic diseases or aphasia, showcasing its potential in simplifying complex text for effective communication. The study involves using two model: GPT-2 and Helsinki-BERT which are fine-tuned using given dataset. The Helsinki-NLP model demonstrated superior performance with a Picto-term Error Rate (PictoER) of 18.51. In contrast, the GPT-2 model had a higher PictoER of 170.81, making it prone to produce extraneous terms. These results indicate that the Helsinki-NLP model is more effective in producing accurate and contextually relevant text aligned with pictogram keywords.

Keywords

Lexical simplification, Language-specific fine-tuning, GPT-2 model, Helsinki BERT, NLP tokenizer, Keyword mapping, Picto-term Error Rate

1. Introduction

AAC provides a solution for those with language problems brought on by illnesses such as aphasia. These systems use pictograms to communicate, however, there is still a barrier in translating text or spoken language into comprehensible pictogram sequences. In the ToPicto subtask of ImageCLEF 2024 [1], the proposed system fine-tunes an existing model using the given dataset for pictogram generation via lexical simplification. This task [2] introduces two new challenges whose objective is to provide a translation in pictograms from a natural language, either from (i) text or (ii) speech understandable by the users, in this case, people with language impairments.

2. Related Works

Radford et al. in [3] offer a thorough analysis of the capabilities and performance of the GPT-2 language model in a range of natural language processing tasks. In-depth assessments of GPT-2's performance on datasets like CoQA, the CNN and Daily Mail dataset, summarization, and translation tasks are also included in this work. Pretrained models for Natural language processing (NLP) are based upon large conventional datasets, and are thus ineffective during classification, prediction tasks based on custom datasets. Neil et al. in [4] experimented Parameter-Efficient Transfer Learning for NLP in order to improve the accuracies for these tasks. This ideology is further expended to various pre-trained models such as RoBERTa by Liu et al. in [5], where a larger dataset was used along with calibrations in various hyper parameters, and Lexfit by Vulic et al. in [6] where lexical simplification is implemented.

Recent progress in natural language processing has been driven by advances in both model architecture and model pretraining. Wolf et al. in [7] introduced Transformer architectures for the same involving higher-capacity models and implementing highly-optimized tokenization library built using Rust. This is extended by the release of open-source *Transformers* library in Python. Qiang et al. in [8] proposed LSBert based on pretrained representation model Bert that is capable of using a dataset

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

abbhinav2210396@ssn.edu.in (A. Elliah); ananthnarayanan2210384@ssn.edu.in (A. N. P); bhuvan2210511@ssn.edu.in (B. S); miruna@ssn.edu.in (P. Mirunalini)

for fine-tuning during simplification and substitution of candidates via complex word identification, substitute generation, filtering and substitute ranking.

3. Approach

In this work, two distinct methods were explored for lexical simplification using pre-trained models. The first method uses GPT-2 architecture, while the second method leverages the use of Helsinki-NLP/opus-mt-ROMANCE-en model for lexical simplification.

In the initial approach, a sentence compression model is developed using a fine-tuned GPT-2 architecture with an additional linear layer for output generation. The process begins by reading source and target sentences, creating a dataset where each entry pairs a source sentence with its compressed counterpart.

A class is used to manage the data, while another class extends the pre-trained GPT2LMHeadModel[1] by adding a linear layer for mapping GPT-2's hidden states to the vocabulary size. The model is trained for 10 epochs with a batch size of 16 using the Adam optimizer. During training, sentences are tokenized with padding and the CrossEntropyLoss function is used to compute the loss between the predicted and target sequences. Padding is applied to ensure uniform sequence lengths, and the model and optimizer states are saved and reloaded for training further.

The second approach utilizes the Helsinki-NLP/opus-mt-ROMANCE-en model [2], a pre-trained translation model originally designed for Romantic languages of Europe, such as French, Spanish, or Italian. Despite its primary focus on translation tasks, the model is adapted for lexical simplification within the context of the ToPicto project. The goal is to convert complex French utterances into simplified sequences of terms linked to pictograms, thereby enhancing communication for individuals with language impairments. By fine-tuning the model on a specialized dataset, the proposed system explores its efficacy in simplifying text while preserving semantic integrity.

3.1. Data Preprocessing

The dataset for this task is provided by the ImageCLEF 2024 organizers and it is structured in JSON format, comprising training, validation, and test sets. Each entry in the dataset includes:

id: A unique identifier for each utterance.
src: The source text, an oral transcription in French.
tgt: The target sequence of simplified pictogram terms.
pictos: A list of pictogram identifiers corresponding to each term in the target sequence.

Tag	Definition	Example
id	unique identifier of each utterance	cefc-tcof-Acc_del_07-1
src	source of the utterance - text from oral transcription	tu peux pas savoir
tgt	target of the utterance - sequence of pictogram terms (tokens)	toi pouvoir savoir non
pictos	a list of pictogram identifiers linked to each pictogram terms (the size is the same as the target output).*	[6625, 35949, 16885, 5526]
		

*This information is provided for reference to give an idea of the input with the sequence of pictogram images. Each image can be obtained from the ARASAAC website as follows: https://api.arasaac.org/v1/pictograms/6625

Figure 1: Data Description

[1]GPT-2 pre-trained model documentary: https://huggingface.co/docs/transformers/en/model_doc/gpt2
[2]Helsinki-NLP repository: https://clarifai.com/helsinkinlp/translation/models/text-translation-romance-lang-english

The preprocessing involves loading the dataset, extracting the relevant fields (src and tgt), and tokenizing the text data using the Helsinki-NLP tokenizer. Tokenization converts the text into a format suitable for model processing, preserving linguistic nuances and syntax.

3.2. Proposed Model

The Helsinki-NLP/opus-mt-ROMANCE-en model is part of the Open Subtitles (opus-mt) project by the University of Helsinki. It is built on the MarianMT framework, which is a highly optimized neural machine translation (NMT) system developed by the Marian NMT group. This model is pre-trained on a vast multilingual corpus, specifically focused on Romance-languages (such as French, Spanish, Italian, Portuguese, and Romanian). The model leverages a transformer architecture, renowned for its effectiveness in handling sequence-to-sequence tasks due to its self-attention mechanisms and parallel processing capabilities.

3.2.1. Encoder-Decoder Framework

The model employs a standard transformer architecture with an encoder-decoder structure. The encoder processes the input sequence and generates contextual embeddings, which the decoder then uses to produce the output sequence.

3.2.2. Self-Attention Mechanism

Both the encoder and decoder utilize self-attention layers, allowing the model to weigh the importance of different tokens in the sequence dynamically. This mechanism helps capture long-range dependencies and contextual information. The model is fine-tuned on the ToPicto dataset, which involves adjusting its parameters to learn the mapping from complex source texts to simplified target sequences. Fine-tuning leverages the model's pre-existing linguistic knowledge, adapting it to the specific requirements of the lexical simplification task.

3.3. Methodology

The Hardware specifications of the system used for Model Training are as follows:
CPU: 12th Gen Intel(R) Core(TM) i7-12700H
GPU: NVIDIA GeForce RTX 3060

The training setup is facilitated by the Hugging Face Transformers library, which provides specialized tools and classes for sequence-to-sequence tasks. The fine-tuning process begins with loading the training datasets from JSON files, focusing on extracting the source ("src") and target ("tgt") fields. Subsequently, the source and target texts undergo tokenization via the Helsinki-NLP tokenizer to ensure they are formatted correctly for fine-tuning.

Training arguments are defined, specifying parameters such as the output directory, batch sizes (here, 4), number of epochs (here, 3), and logging frequency (here, 100) to ensure comprehensive monitoring and control of the training process, after consideration of data validity. The fine-tuning process involves optimizing the model's parameters on the training dataset, leveraging backpropagation and gradient descent to minimize the loss function and improve the model's accuracy in generating the desired sequences. During this phase, the model is iteratively trained over multiple epochs, with periodic evaluations on the validation dataset to prevent overfitting and ensure generalizability. The fine-tuned model is subsequently saved for deployment, ensuring that the trained parameters are preserved for future use. During inference, the model generates hypotheses (hyp) for the test set, which are then post-processed to ensure conformity with the expected output format and semantic coherence.

4. Results and Discussion

We have experimented with two different models as discussed above: GPT-2 and Helsinki-NLP/opus-mt-ROMANCE-en models. The following picto images are provided by ImageClef'24 organizers, and the image sequence is generated using the script file provided for the ToPicto task.

Consider for example, the source text **"ils ont un accent eux aussi euh"**.

Figure 2: Helsinki-NLP/opus-mt-ROMANCE-en model
Generated Sequence: ils avoir un dire eux

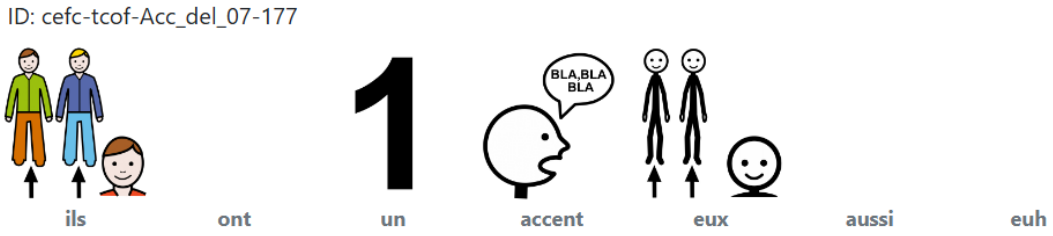

Figure 3: GPT-2 model
Generated Sequence: ils ont un accent eux aussi euh

It has been proved that Helsinki-NLP model is better in producing a more efficient and accurate output for the given source text, aligned to the meaning of the given text. Whereas, it was noted that GPT-2 model gives us comparatively less accurate result, which is due to the fact that the GPT-2 model is predominantly trained on a large number of English datasets.

The performance of the proposed architecture was evaluated using the metrics namely Picto-term Error Rate (PictoER) [9], BLEU score [10], METEOR [11]. In the evaluation of information retrieval systems based on the provided metrics, the performance of two distinct runs reveals noteworthy insights.

Table 1
Performance Comparison of Helsinki BERT and GPT-2 Models

Model	Pictoer Score	BLEU Score	METEOR Score
Helsinki BERT	18.51	68.96	83.55
GPT-2	170.81	3.93	25.57

The Helsinki BERT model demonstrates superior performance in generating French text that aligns closely with picto keywords, evidenced by its high BLEU score of 68.96, METEOR score of 83.55, and a low PictoER score of 18.51. These results indicate the model's effectiveness in producing fluent, contextually accurate text with minimal error in keyword mapping. The strong BLEU and METEOR scores highlight

the model's capability to preserve n-gram overlaps and account for synonymy, stemming, and paraphrase matching, making it highly suitable for tasks requiring precise linguistic and semantic accuracy. It's worth noting that while the Helsinki BERT model is trained on a diverse set of languages, including French, this multilingual training could contribute to a slight reduction in its BLEU score due to the broad scope of its training data.

Conversely, the GPT-2 model is significantly weaker at producing cohesive and contextually appropriate French language, as evidenced by its BLEU score of 3.93, METEOR score of 25.57, and high PictoER score of 170.81. The high PictoER score indicates considerable keyword alignment problems, while the poor BLEU and METEOR scores show the model's inability to maintain contextual relevance and fluency. The significant difference between the two shows that models must be tailored specifically to the target language and application area. In this case, the Helsinki BERT model's customized approach performs better than GPT-2's generalist capabilities. Interestingly, the majority of the English datasets used to train GPT-2 have limited its ability to perform well on French language tasks. However, the GPT-2 model did show some ability to correctly predict numbers and nouns.

5. Conclusion

In conclusion, the Helsinki-NLP model exhibits better performance in showcasing appropriate pictos for the given text over GPT-2 model. Both the model are able to predict the keywords of a given phrase with high accuracy, however the former model is able to predict the pronouns and phrase out a meaningful picto combination at a higher degree of confidence since it is pre-trained over French dataset rather than English, as in GPT-2.

6. Future Work

Subsequent research endeavors may involve the enhancement of existing models, such as further fine-tuning GPT-2, or the creation of novel models to augment their efficacy in analogous language-specific assignments. The error rates can be improved by hyperparameter tuning or by removing erroneous words, which is achieved with more data for training the model and implementing various other pre-trained models.

References

[1] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[2] C. Macaire, E. Esperança-Rodier, B. Lecouteux, D. Schwab, Overview of ImageCLEF 2024 - investigating the translation of natural language into pictograms, in experimental IR meets multilinguality, multimodality, and interaction., https://ceur-ws.org/, 2024.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[4] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: Proceedings of the 36th International

Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, 2019, pp. 2790–2799.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[6] I. Vulić, E. M. Ponti, A. Korhonen, G. Glavaš, Lexfit: Lexical fine-tuning of pretrained language models, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5269–5283.

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[8] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, X. Wu, Lsbert: A simple framework for lexical simplification, arXiv preprint arXiv:2006.14939 (2020).

[9] J. Woodard, J. Nelson, Pictoer: An information theoretic measure of speech recognition performance, in: Workshop on standardisation for speech I/O technology, Naval Air Development Center, Warminster, PA, 1982.

[10] K. Papinesi, Bleu: A method for automatic evaluation of machine translation, in: Proc. 40th Actual Meeting of the Association for Computational Linguistics (ACL), 2002, 2002, pp. 311–318.

[11] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.