A Transformer Based Approach for Text-to-Picto Generation

Notebook for the ImageCLEF Lab at CLEF 2024

Avaneesh Koushik, Jithu Morrison S, P Mirunalini and Jothir Aditya R K

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

Abstract

This study aims to develop a Text to Pictogram translation system which is used to convert a French text into its corresponding pictogram terms. The proposed system demonstrates the effectiveness of a transformer-based model in translating French text into meaningful pictogram sentences. Google-T5 is utilized and further fine-tuned on a custom dataset of French text to predict corresponding pictogram terms in French. The model underwent fine-tuning across multiple epochs to optimize performance. Additionally, the trained model was iteratively fine-tuned to enhance its translation capabilities. Metrics like PictoER score, BLEU score and METEOR score were used to assess the model's performance. The proposed model achived a PictoER score of 13.9, BLEU score of 74.3 and a METEOR score 87.0.

Keywords

Text to Pictogram, French text generation, Transformers model, Google-T5

1. Introduction

Communication is very basic for day-to-day human activities. The various modes of communication viz. speaking, listening, reading, and writing can all be hampered by a number of genetic illnesses such as Rett syndrome or by events like a stroke or auto accident, which can cause linguistic impairment. This causes a decrease in both language understanding and expression and results in a condition known as aphasia. AAC or augmentative and alternative communication in the form of pictograms can be used in certain specific situations to assist people with aphasia to communicate effectively. The goal of this task is to develop novel translation methodologies to translate text into a sequence of pictograms.

Traditionally, aiding communication for individuals with aphasia has relied heavily on manual methods and human assistance, which are both time-consuming and prone to subjective biases. The need for more efficient, scalable, and objective solutions has become increasingly evident, especially as the volume of individuals requiring such support continues to grow.

In this context, the ToPicto task of ImageCLEF 2024 represents a significant innovation in the field of assistive communication technologies. ImageCLEF is a well-known series of challenges and labs that promotes progress in multimedia information retrieval [1]. Since its start, ImageCLEF has provided a thorough evaluation framework to develop and benchmark techniques in visual information retrieval.

This research focuses on developing novel methodologies to translate text into sequences of pictograms, aiming to bridge communication gaps for individuals with aphasia. By leveraging advanced natural language processing techniques, specifically utilizing the pre-trained Google T5 model fine-tuned with the TCOF corpus of French text, this study seeks to automate the generation of pictogram sequences from French text. This approach aims to provide a means of supporting communication for

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

avaneesh2210179@ssn.edu.in (A. Koushik); jithumorrison2210564@ssn.edu.in (J. M. S); miruna@ssn.edu.in (P. Mirunalini); rkjothir2210024@ssn.edu.in (J. A. R. K)

https://jithumorrison.github.io/ (J. M. S)

those with language impairments, thereby contributing to the advancement of computational methods in assistive technologies and enhancing the quality of life for individuals with aphasia.

2. Background

Text-to-pictogram translation is a task which involves translation of natural language text into text with words for which appropriate pictograms are available. A unified approach to transfer learning in NLP tasks can be achieved by considering every text processing problem as a “text-to-text” problem, i.e. taking text as input and producing new text as output [2]. This approach was utilised for the proposed model as the dataset for this task comprises of input and output texts. Pretrained models have significantly better performance over the original T5 models [3].

A shallow linguistic analysis approach can be used to perform linguistic analysis for text to picto conversion [4]. Shallow linguistic analysis involves processing of basic linguistic units like tokenization and POS tagging, without performing deep semantic analysis. Transformers are powerful tools that helps in building more complex and effective models for sequence-based tasks. Transformer architectures have facilitated building higher-capacity models and pre-training has made it possible to effectively utilize this capacity for a wide variety of tasks [5]. The original system of text-to-picto aimed at people with an intellectual disability can be extended to various other interesting applications [6].

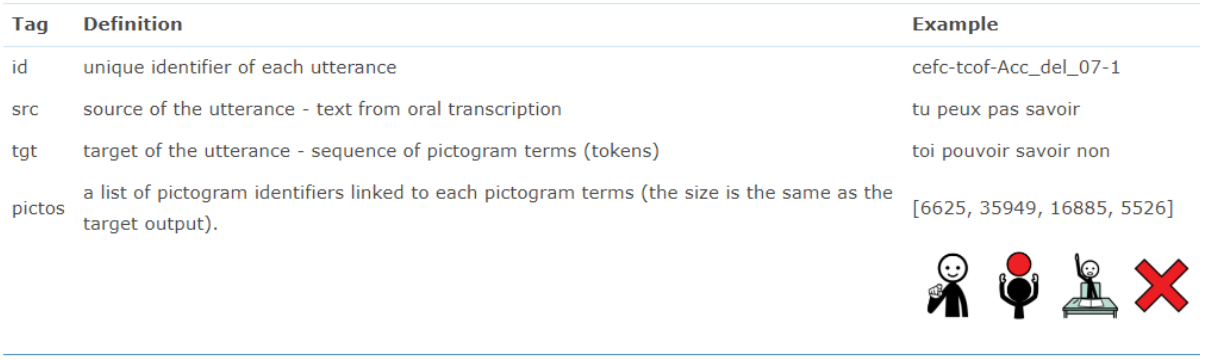
Encoder-decoder models perform better than other models on textual similarity tasks [7]. Google-T5 was found to be the better than various other approaches like LSTM and CNN for other text related tasks like hate speech detection [8] and it was found to be suitable for Question-Answer Generation [9] both of which are tasks involving pattern detection in text.

3. Approach

Google T5 or Text-to-Text Transfer Transformer is an encoder-decoder model which was pre-trained on a multi-task mixture of both unsupervised and supervised tasks. It is known to work well in tasks which require out of the box thinking. This task involves converting French text in various everyday contexts into words which are simpler and have a corresponding pictogram available. The main objective of this approach is to develop a Text-to-Picto translation system using the T5 model.

3.1. Dataset

The dataset that has been used was built from the TCOF corpus, and is stored in JSON format. TCOF contains interactions between adults, adults and children, and children themselves, covering a wide range of topics including debates, everyday situations, and medical consultations. This type of text is representative of the interactions we observe between caregivers (families, medical staff) and individuals who rely on pictograms due to language impairments [10]. Each entry in the dataset contains multiple data points, including an identifier labelled as **"id"** which is a unique identifier for the source, target pair, the source text which is an oral transcription of a sentence spoken in French labelled as **"src"**, the target sequence of simplified pictogram terms **"tgt"**, and a list that assigns a pictogram identifier to each term in the target sequence labelled as **"pictos"**.



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]

Figure 1: Dataset Description

On further analysis of the dataset, it was found that the data contained 24270 lines of French text with appropriate target text. The average size of the source lines is 54.6 words and the average size of the target text is 53.8 pictogram words.

3.2. Data Preprocessing

3.2.1. Tokenization

Tokenization is important for preparing the data for model training. Here, both the target pictogram sequence and the source French text are tokenized using the pre-trained tokenizer from the google-T5 model. The text is split into individual tokens and converted to a numerical representation. Padding and truncating limit the text sequences to have a maximum size of 256 tokens.

3.3. Model Selection

The T5 model can be adjusted for particular tasks and comes pre-trained on a large data corpus. Here, the model is adjusted to produce the simpler pictogram terms from oral transcriptions in French. The model is inherently trained in solving text-to-text tasks and hence proves to be efficient for this task. The "t5-base" variant of Google's T5 model is utilised for this task. t5-base is a snapshot of the T5 model taken after it was trained with 220 million parameters. This makes it flexible and easy to train for the given French text data.

3.3.1. Self-Attention mechanism

The self-attention mechanism enables the model to identify long-range links and dependencies in the input sequence. To be more precise, the T5 model calculates attention scores between every pair of tokens in the input sequence using self-attention layers. By assigning a different weight (attention score) to each token based on relevance to the next token, it makes every token in the sequence able to pay attention to every other token.

3.3.2. Encoder-decoder mechanism

The T5 model is based on the traditional transformer architecture comprising of an encoder-decoder structure. The input sequence is processed by the encoder, which also outputs contextual embeddings. The output sequence is subsequently generated by the decoder using these embeddings.

3.4. Methodology

The dataset is loaded from the file and the source text (src) and target sequence (tgt) are extracted for every element and stored as a list, following which it is tokenized. This is then converted to key-value pairs and fed as input to the model. TThe T5 model contains several training arguments, such as the batch size, which is set to 16 and indicates the number of training instances processed in each iteration, and the number of training epochs, which are set to 3, 5, and 6 (where the model is initially trained for 3 epochs and then further trained on the already trained model). Additionally, the save steps are set to 1,000, meaning that model checkpoints are saved every 1,000 steps to ensure training progress is recorded. The learning rate of the optimizer algorithm—such as Adam—is chosen in order to effectively update the model parameters in light of the training data. The goal of this optimization is to increase model performance by minimizing the loss function. The model is trained using the hyperparameters mentioned in Table 1

Table 1
Hyperparameters used for training the proposed model

Parameter	3 Epochs	5 Epochs	6 Epochs
model	google-t5/t5-base	google-t5/t5-base	google-t5/t5-base
max_length	256	256	256
num_train_epochs	3	5	6
per_device_train_batch_size	16	16	16
save_steps	1000	1000	1000

The fine-tuning procedure is started over the designated number of epochs by using the train method on the Trainer object to start the training process. By modifying its parameters in response to training data, the model gains the ability to translate French text to pictogram sequences efficiently.

The model, after running a few epochs is saved to the designated directory for later use after training is finished. Then, by utilizing the knowledge acquired during training, this refined model can be applied to the creation of pictogram sequences based on input French text.

The sequence of pictogram terms generated by the proposed model is converted to the corresponding pictos sequence using the resources provided by the task organisers.

For example: If the input source text is: "il y a un moment donné elle nous avait dit essayez de pas dire de mots français pendant le truc.", the model generates the following sequence of pictogram terms: "il_y_a un instant donner passé elle nous dire essayer de dire non de mot français pendant le truc". Figure 2 shows the pictos sequence corresponding to the above generated sequence of pictogram terms.

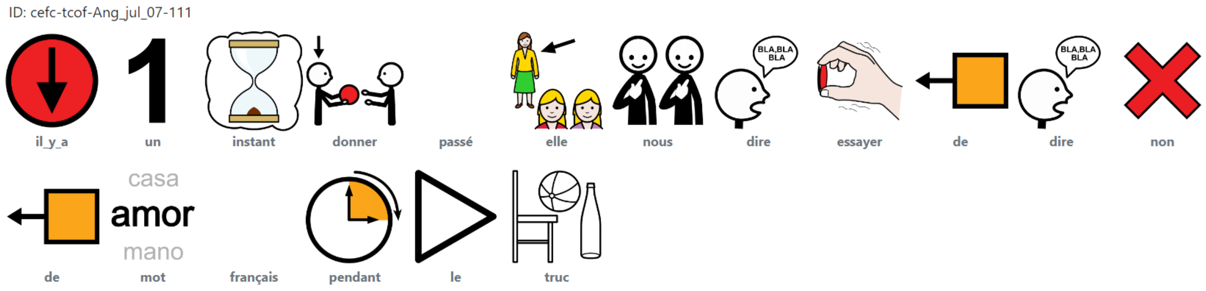


Figure 2: Generated pictos sequence

3.4.1. Resources Used

Pandas is used in the project to manipulate data, including loading data from JSON files and structuring it into dataframes. The main deep learning framework, PyTorch, makes it easier to apply and train the T5 model, which creates pictogram sequences from French text. A cloud-based Jupyter notebook environment with GPU-accelerated resources for quicker model training and inference is offered by Google Colab. Together, these tools improve productivity and efficiency by streamlining the development and experimentation process.

4. Results and Discussion

The parameters used for evaluating the model are the `Picto-term Error Rate (PictoER)` [11], BLEU [12] and METEOR [13]. The model was trained for various epochs to test the improvement in learning. These results of the proposed model for different epochs were tabulated in the following Table 2

Table 2
Performance Comparison with varying epochs

Model	PictoER	BLEU	METEOR
3 epochs (T5 model)	18.431	66.566	82.895
5 epochs(T5 model)	17.575	67.859	83.691
6 epochs (T5 model)	13.907	74.363	87.082

Comparing the outcomes from varying the epochs while training yields insightful observations. The error rate decreased from 18 to 17 when training for 5 epochs, suggesting some improvement in performance. When initially trained for 3 epochs and fine-tuned for an additional 3 epochs, the `pictoer_score` drops significantly from 17.5 to 13.9, suggesting improved generalization and performance on unseen data. This is also reflected in the BLEU score which measures the precision of n-grams and the METEOR score which focuses on word order, both of which show considerable improvement. This significant improvement underscores the effectiveness of fine-tuning in refining model parameters and enhancing its ability to capture underlying patterns in the data.

The results suggest that the model may not have been able to reach its maximum potential during the first training period. Rather, the model's representations were gradually improved through the repeated training process, which improved generalization and decreases the error rate. These findings highlight the importance of iterative training strategies and the need for careful experimentation to achieve optimal results. This shows that one may continuously enhance the model's performance and guarantee its flexibility to a variety of datasets and applications by iteratively fine-tuning it.

5. Conclusion

In conclusion, this research demonstrates the effectiveness of advanced transformer models, specifically the Google-T5, for the task of translating French text into pictogram sequences. Through iterative fine-tuning, the model consistently improved in accuracy, demonstrating its ability to handle intricate aspects of language. This was evaluated using metrics like **PictoER**, **BLEU**, and **METEOR** scores. This research emphasizes how transformer-based techniques can improve accessibility and communication for people who use augmentative and alternative forms of communication.

6. Future Scope

Researchers may enhance the model's performance and guarantee its flexibility to a variety of datasets by iteratively fine-tuning it. In order to further advance the fields of assistive technology and natural language processing, future studies could explore expanding this strategy to other languages and improving the model's adaptability to diverse linguistic contexts.

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. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.

[3] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, R. Lotufo, Ptt5: Pretraining and validating the t5 model on brazilian portuguese data, arXiv preprint arXiv:2008.09144 (2020).

[4] V. Vandeghinste, I. S. L. Sevens, F. Van Eynde, Translating text into pictographs, Natural Language Engineering 23 (2017) 217–244.

[5] 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.

[6] M. Norré, V. Vandeghinste, P. Bouillon, T. François, Extending a text-to-pictograph system to french and to arasaac, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021, pp. 1050–1059.

[7] J. Ni, G. H. Abrego, N. Constant, J. Ma, K. B. Hall, D. Cer, Y. Yang, Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, arXiv preprint arXiv:2108.08877 (2021).

[8] T. Adewumi, S. S. Sabry, N. Abid, F. Liwicki, M. Liwicki, T5 for hate speech, augmented data, and ensemble, Sci 5 (2023). URL: https://www.mdpi.com/2413-4155/5/4/37. doi:10.3390/sci5040037.

[9] S. Kumar, A. Chauhan, P. Kumar C., Learning enhancement using question-answer generation for e-book using contrastive fine-tuned t5, in: P. P. Roy, A. Agarwal, T. Li, P. Krishna Reddy, R. Uday Kiran (Eds.), Big Data Analytics, Springer Nature Switzerland, Cham, 2022, pp. 68–87.

[10] V. André, E. Canut, Mise à disposition de corpus oraux interactifs : le projet tcof (traitement de corpus oraux en français), Pratiques. Linguistique, littérature, didactique 147-148 (2010) 35–51.

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

[12] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[13] 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, Association for Computational Linguistics, 2005, pp. 65–72.