Transformer-Based Sequence-to-Sequence Modeling for English Text-to-Pictogram Translation

Notebook for the ImageCLEF Lab at CLEF 2026

Sudharshan PS

Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

Augmentative and Alternative Communication (AAC) systems support individuals with speech, language, and cognitive impairments by representing natural language through visual pictograms. This paper presents a transformer-based sequence-to-sequence approach for English Text-to-Pictogram translation using the FLAN-T5 architecture as the primary model. In addition, BART is evaluated as a secondary baseline to study differences in sequence generation performance. The goal of the proposed system is to convert natural language sentences into pictogram-oriented representations that are easier to understand in AAC environments. Both models were fine-tuned on the official English Text-to-Picto dataset using task-specific preprocessing, subword tokenization, and beam search decoding. To ensure a fair comparison, both models were trained under an identical experimental setup, including the same preprocessing pipeline, optimization strategy, hyperparameter configuration, decoding settings, and evaluation protocol. On the validation set, FLAN-T5 achieved the best overall performance with a SacreBLEU score of 51.75, a METEOR score of 71.37, and a Picto-term Error Rate (PictoER) of 0.32. The BART model also produced competitive results, obtaining a SacreBLEU score of 51.00, a METEOR score of 70.07, and a PictoER of 0.3292. The experimental results demonstrate that encoder-decoder transformer architectures are highly effective for pictogram sequence generation and AAC-oriented text translation. The comparative analysis further highlights how architectural design, tokenization strategies and decoding mechanisms collectively influence the semantic quality, structural consistency, and overall accuracy of the generated pictogram sequences.

Keywords

FLAN-T5, BART, Transformer Models, Sequence-to-Sequence (Seq2Seq), Text-to-Pictogram Translation, Augmentative and Alternative Communication (AAC), Beam Search Decoding, Subword Tokenization

1. Introduction

Augmentative and Alternative Communication (AAC) systems help individuals with speech, language, and cognitive impairments communicate using visual pictograms. Converting natural language into accurate pictogram sequences remains a difficult task because pictogram-based communication often follows simpler grammatical and semantic structures than standard written language.

The English Text-to-Pictogram task focuses on generating pictogram term sequences from natural language sentences while preserving their original meaning. Unlike traditional machine translation, the task requires semantic mapping, lexical simplification, and structural modifications to produce pictogram-oriented representations that are easier to understand in AAC environments.

In this work, a transformer-based sequence-to-sequence approach is proposed for English Text-to-Pictogram translation using FLAN-T5 as the primary model architecture. BART is also evaluated as a secondary baseline model for comparison. Both models are fine-tuned using task-specific preprocessing, subword tokenization, and beam search decoding. The proposed system aims to generate meaningful pictogram sequences that improve accessibility and communication support for AAC users.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

sudhar0705@gmail.com (S. PS)

2. Related Work

Recent progress in Text-to-Pictogram translation has been strongly influenced by advances in transformer-based sequence generation models and AAC-oriented language processing systems. Earlier pictogram generation approaches mainly relied on rule-based systems, dictionary matching, and lexical substitution methods. Although these approaches worked reasonably well for simple sentence structures, they often struggled with contextual understanding, grammatical variations, and multi-word semantic representations.

The introduction of transformer-based encoder-decoder architectures significantly improved natural language generation and translation tasks. Models such as T5 [3] and BART [5] demonstrated strong performance in sequence-to-sequence learning by capturing contextual relationships between source and target sequences through large-scale pretraining. FLAN-T5 [4] further extended the T5 framework using instruction tuning, allowing pretrained models to adapt more effectively to downstream generation tasks through task-oriented prompts.

Recent ImageCLEF ToPicto shared tasks established standardized benchmarks for generating pictogram sequences from natural language text and speech inputs for AAC applications. Previous participant systems explored multilingual transformer architectures, semantic simplification methods, encoder-decoder pipelines, and lexical normalization strategies for pictogram generation. These studies highlighted the importance of contextual understanding and structured sequence generation for improving pictogram quality and readability in AAC environments.

Subword tokenization methods also play an important role in sequence generation tasks involving pictogram terms and symbolic representations. SentencePiece tokenization [6] is widely used in T5-based architectures because of its ability to efficiently handle compound terms and uncommon word structures. In comparison, Byte-Pair Encoding (BPE) tokenization, used in models such as BART, performs well for general language generation but may introduce additional subword fragmentation for compound pictogram terms connected using underscore-based representations.

The Hugging Face Transformers framework [7] further simplified the development of transformer-based sequence-to-sequence systems by providing efficient support for encoder-decoder training, dynamic padding, beam search decoding, and large-scale pretrained language models. Motivated by the effectiveness of transformer architectures in text generation tasks, this work investigates FLAN-T5 as the primary model for English Text-to-Pictogram translation, while BART is evaluated as a secondary baseline model for comparative analysis.

3. Dataset and Task Description

3.1. Task Overview

The English Text-to-Pictogram task focuses on translating natural language English sentences into pictogram term sequences that can be used in Augmentative and Alternative Communication (AAC) systems. Unlike traditional machine translation tasks, the objective is not direct word-to-word translation. Instead, the task requires semantic mapping and structural simplification to generate pictogram-oriented representations while preserving the meaning of the original sentence. [1, 2]

3.2. Dataset Structure

The dataset for this task is provided by the ImageCLEF 2026 organizers and is structured in JSON format. The dataset consists of parallel source-target sentence pairs designed for Text-to-Pictogram generation.

Each dataset entry contains the following fields:

- **id**: A unique identifier associated with each sample. The identifiers originate from CommonVoice-based filenames and are retained for dataset consistency.

- **src**: The source English sentence provided as natural language input to the model.
- **tgt**: The target pictogram-oriented symbolic sequence corresponding to the source sentence.
- **pictos**: A list of pictogram identifiers associated with the generated target pictogram terms.

The dataset is divided into training, validation, and test splits. The training set contains 59,278 samples, while the validation and test sets contain 4,397 and 4,306 samples respectively. During preprocessing, both source and target sequences are cleaned, normalized, and tokenized before being passed to the transformer architectures. Maximum source and target sequence lengths are restricted to 128 tokens to maintain stable training and inference.

The source sentences contain natural language English text, while the target sequences are represented using textual labels linked to pictographic concepts. Multi-word semantic expressions are frequently merged into single compound representations using underscore-separated structures such as `professional_qualification` and `art_lesson`. These symbolic structures introduce additional challenges for tokenization and sequence generation models.

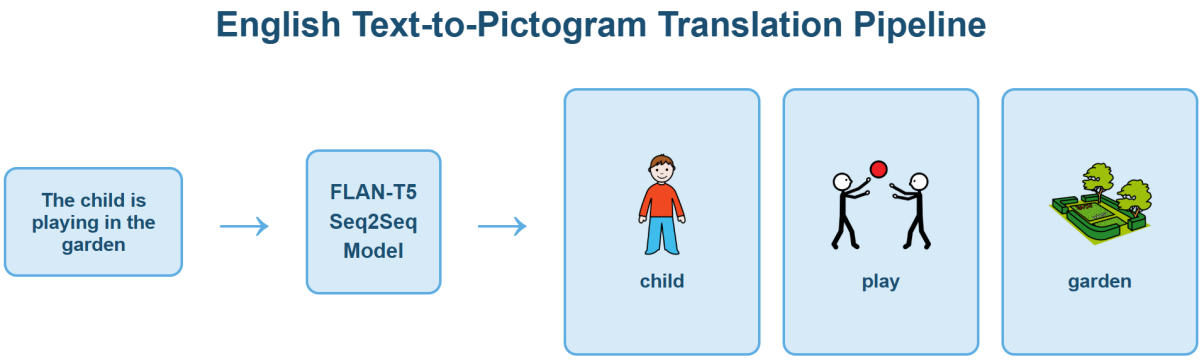

Figure 1: Example of the English Text-to-Pictogram translation

3.3. Exploratory Data Analysis

The dataset contains 59,278 training samples and consists of three primary columns: the sample identifier (`id`), source sentence (`src`), and target pictogram sequence (`tgt`). Although the sample identifiers originate from CommonVoice audio filenames, the present work focuses only on the textual source and target sequences provided in the dataset.

The dataset contains no missing values across any of the columns, indicating that all source-target pairs are complete and usable for training. Statistical analysis of sequence lengths shows that the average source sentence length is 8.03 words with a standard deviation of 2.31, while the average target pictogram sequence length is 6.40 tokens with a standard deviation of 1.43. The source sentences range from 1 to 20 words, whereas the target sequences are more constrained, varying between 3 and 8 tokens. This reduction in sequence length suggests that the pictogram representations are structurally compressed versions of the original sentences.

Vocabulary analysis further highlights the semantic compression behavior present in the dataset. The source text contains 46,156 unique tokens distributed across 475,959 total source tokens. In comparison, the target pictogram vocabulary contains only 6,721 unique terms with a total of 379,496 target tokens. This large reduction in vocabulary size indicates that multiple natural language expressions are mapped into a much smaller symbolic representation space during pictogram generation.

The analysis identified 40,906 occurrences of compound pictogram expressions and 1,987 unique compound terms in the target sequences. Examples include terms such as `professional_qualification`, and `art_lesson`. These compound representations allow multi-word semantic concepts to be compressed into single pictogram-oriented units, helping preserve contextual meaning while reducing

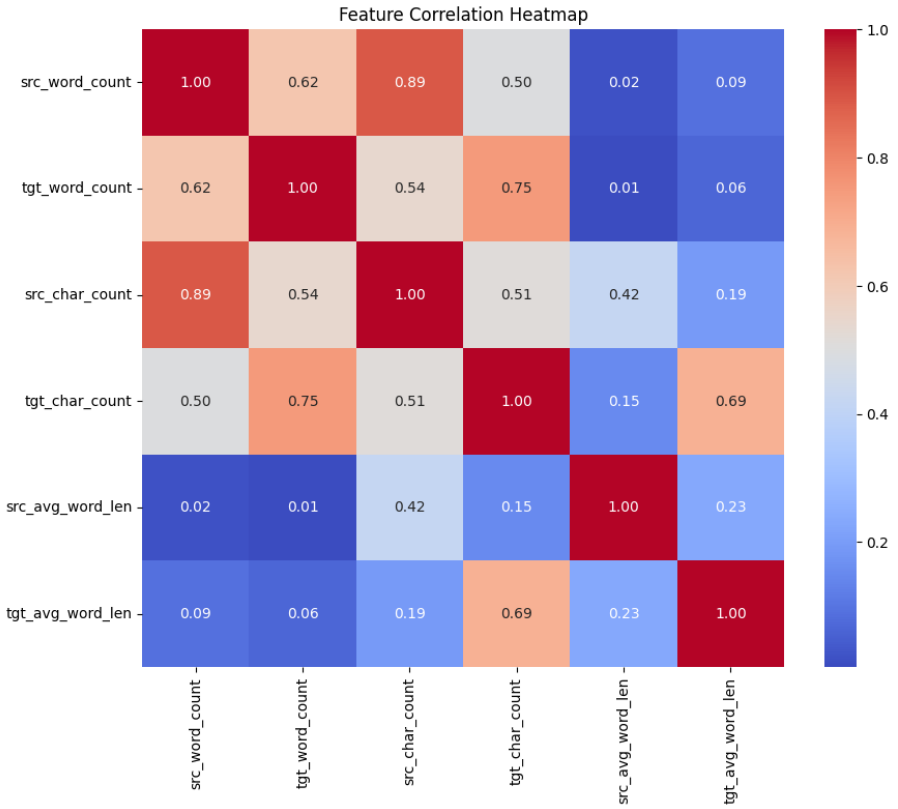

Figure 2: Correlation analysis between source and target sequence statistics. Moderate correlations between source and target lengths indicate that longer input sentences generally produce longer pictogram sequences.

sequence complexity. This characteristic also introduces additional challenges for subword tokenization and sequence generation models.

The average compression ratio between target and source sequences is approximately 0.83, indicating that the pictogram-oriented outputs are generally shorter than the corresponding natural language sentences. This observation further supports the idea that Text-to-Pictogram generation is fundamentally different from conventional machine translation tasks.

4. Proposed Methodology

The proposed framework models the English Text-to-Pictogram task as a conditional sequence generation problem using transformer-based encoder-decoder architectures. The objective is to generate semantically simplified pictogram-oriented symbolic sequences from natural language English sentences while preserving the core meaning of the source text. The two transformer architectures explored in this work are FLAN-T5-base and BART-base. Both models are trained and evaluated under a unified experimental setup to analyze the differences in architecture design and tokenizer behavior.

4.1. FLAN-T5-Based Sequence Generation

The primary model used in this work is `google/flan-t5-base`, an instruction-tuned encoder-decoder transformer architecture derived from the original T5 framework. FLAN-T5 follows a unified text-to-text learning paradigm where all tasks are formulated as sequence transformation problems.

The model contains approximately 250 million parameters and consists of stacked encoder and decoder transformer layers. The encoder processes the source English sentence and generates contextual semantic representations, while the decoder autoregressively predicts the target pictogram sequence.

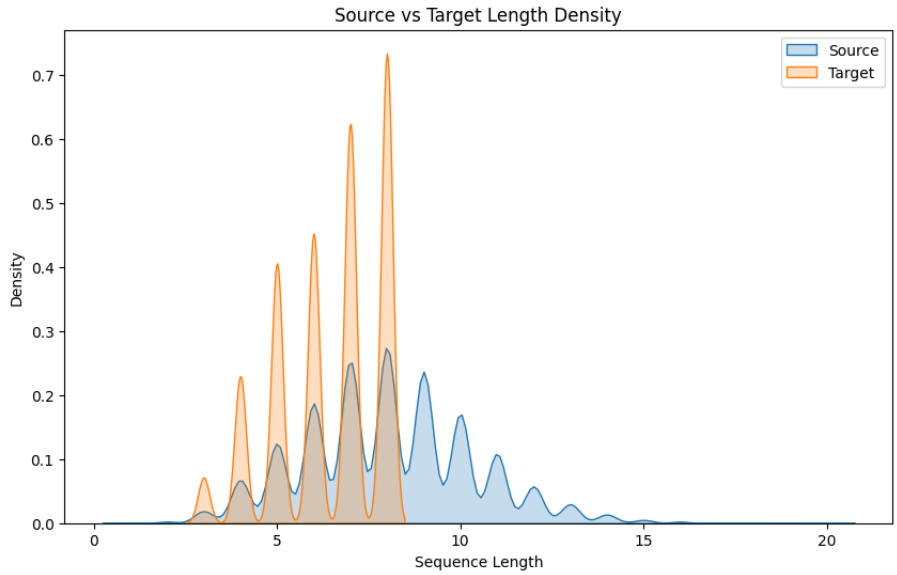

Figure 3: Distribution of source sentence lengths and target pictogram sequence lengths in the training dataset. The target sequences are generally shorter and more structurally constrained compared to the source sentences.

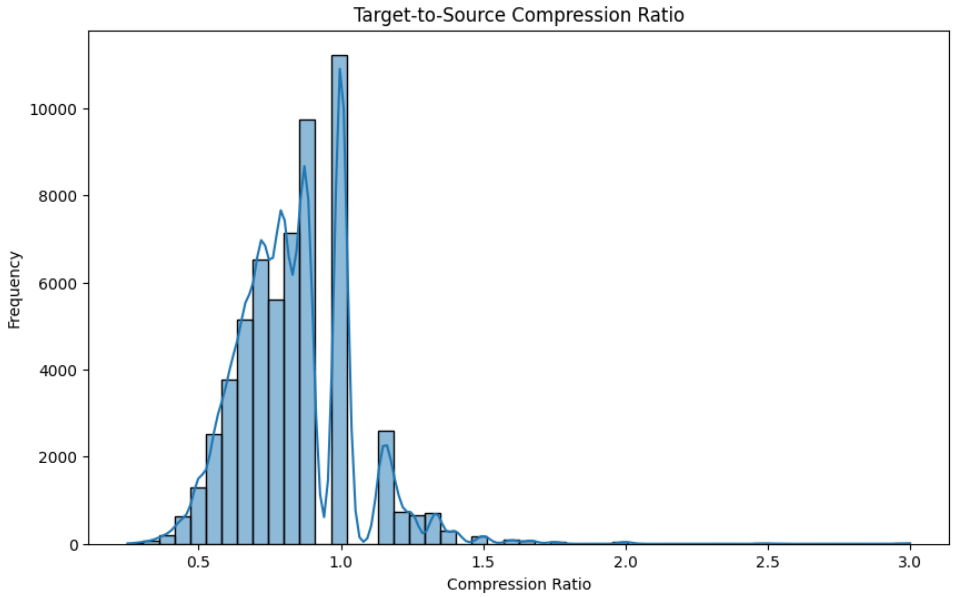

Figure 4: Distribution of target-to-source compression ratios across the training dataset. Most target pictogram sequences are shorter than their corresponding source sentences, indicating semantic compression in AAC-oriented representations.

One important characteristic of FLAN-T5 is its instruction-tuned pretraining strategy. During large-scale pretraining, the model is exposed to more than 1,800 instruction-tuning tasks. This enables stronger adaptation for structured generation problems involving semantic transformation and contextual rewriting.

To align the model with the Text-to-Pictogram objective, each source sentence is prepended with the following instruction prompt: `"translate English text to English pictogram terms:"`.

The use of explicit prompting helps the model contextualize the generation task more effectively and improves semantic alignment between source text and pictogram-oriented outputs. The architecture additionally utilizes relative positional embeddings, multi-head self-attention, feed-forward transformer blocks, and gated GELU activation functions.

Table 1
Dataset Statistics and Exploratory Analysis Summary

Metric	Value
Total Training Samples	59,278
Validation Samples	4,397
Test Samples	4,306
Dataset Columns	id, src, tgt
Missing Values	0
Average Source Sentence Length	8.03 words
Average Target Sequence Length	6.40 tokens
Maximum Source Length	20 words
Maximum Target Length	8 tokens
Minimum Source Length	1 word
Minimum Target Length	3 tokens
Source Vocabulary Size	46,156
Target Vocabulary Size	6,721
Total Source Tokens	475,959
Total Target Tokens	379,496
Average Compression Ratio	0.83
Compound Pictogram Occurrences	40,906
Unique Compound Terms	1,987
Most Frequent Source Token	the (42,675)
Most Frequent Target Token	the (46,904)

4.2. BART-Based Sequence Generation

For comparative analysis, the `facebook/bart-base` architecture is used as a secondary baseline model. BART is also an encoder-decoder transformer model but differs significantly in its pretraining methodology. Instead of instruction tuning, BART is pretrained using denoising autoencoding objectives. During pretraining, corrupted input sequences are reconstructed through objectives such as token masking, token deletion, sentence permutation, text infilling and document rotation. This training strategy enables BART to learn strong contextual sentence representations suitable for text generation tasks.

Unlike FLAN-T5, BART does not require explicit instruction prompts during inference. The source English sentence is directly passed to the encoder without additional task prefixes. The BART-base model contains approximately 139 million parameters and utilizes absolute positional embeddings, bidirectional encoder representations, autoregressive decoder generation, standard GELU activation functions.

4.3. Tokenizer Analysis

Tokenization plays an important role in the Text-to-Pictogram task because the target sequences contain large numbers of compound symbolic expressions.

FLAN-T5 uses a SentencePiece tokenizer based on a unigram language modeling strategy. SentencePiece performs subword segmentation directly on raw text and segments symbolic structures more consistently without depending entirely on whitespace boundaries. This becomes particularly useful for preserving underscore-separated compound pictogram representations during generation.

In contrast, BART uses Byte-Pair Encoding (BPE) tokenization. Although BPE performs effectively for standard natural language tasks, it occasionally fragments compound pictogram terms into multiple subword units. This increases target sequence fragmentation and can reduce semantic consistency during autoregressive decoding.

The exploratory analysis performed on the dataset identified more than 40,000 occurrences of

compound pictogram structures spanning nearly 2,000 unique symbolic terms. Efficient handling of these representations therefore becomes important for maintaining sequence-level semantic coherence.

4.4. Preprocessing Pipeline

Before training, the dataset undergoes multiple preprocessing operations to improve sequence consistency and model stability. The preprocessing pipeline involves JSON dataset loading, Source-target sequence extraction, whitespace normalization, lowercase standardization, preservation of compound symbolic structures, transformer-specific tokenization, dynamic sequence padding and tensor conversion. Both source and target sequences are restricted to a maximum sequence length of 128 tokens. Dynamic padding is applied during mini-batch construction using the Hugging Face `DataCollatorForSeq2Seq` module to improve GPU memory utilization.

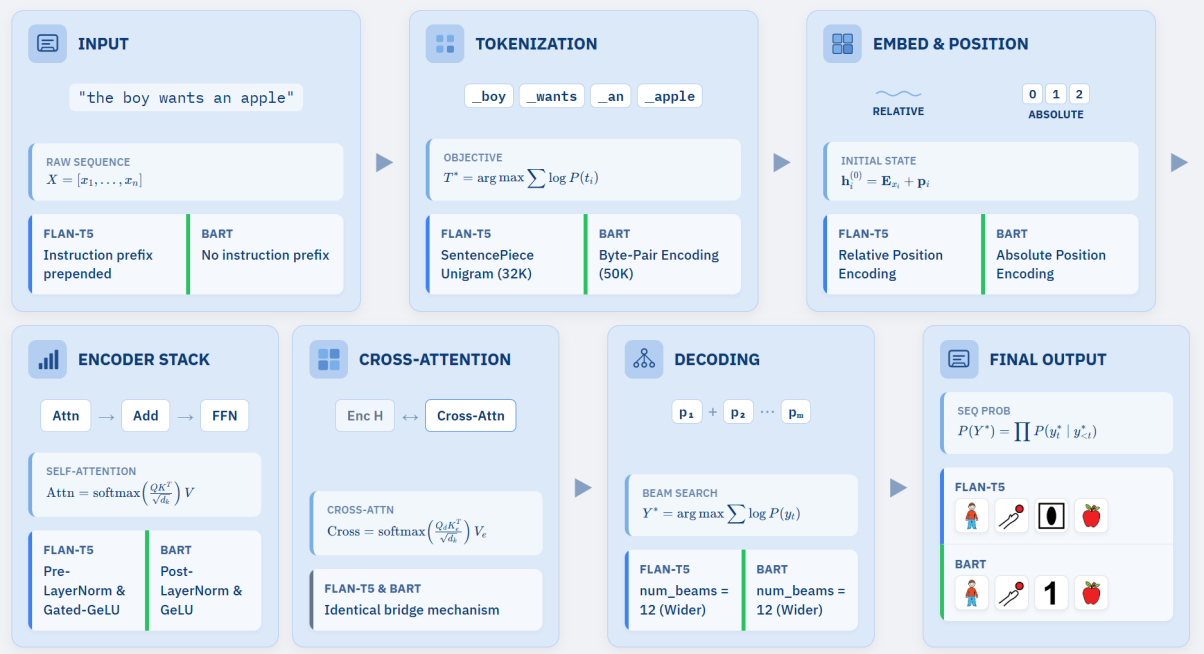

Figure 5: Comparative architecture of the FLAN-T5 and BART-based encoder-decoder transformer frameworks

4.5. Sequence Generation and Beam Search

During inference, pictogram sequences are generated autoregressively using beam search decoding. Instead of selecting only the highest-probability token at each decoding step, beam search maintains multiple candidate generation paths simultaneously and selects the most probable sequence globally.

Both FLAN-T5 and BART use a beam width of 12, allowing the decoder to explore a wider search space during generation.

4.6. Qualitative Translation Analysis

A qualitative analysis of the generated outputs shows that both transformer architectures successfully learn semantic simplification and symbolic restructuring patterns required for AAC-oriented communication. Instead of performing direct word-by-word translation, the models learn to generate compact symbolic representations that preserve the core meaning of the source sentence.

Both models correctly transform the multi-word phrase `"glass of water"` into the unified symbolic representation `"glass_of_water"`, showing that the architectures are capable of learning compound pictogram mappings from contextual sentence structures.

Table 2
Unified Comparative Analysis of FLAN-T5 and BART Architectures

Component	FLAN-T5	BART
Base Model	`google/flan-t5-base`	`facebook/bart-base`
Architecture Type	Encoder-Decoder Transformer	Encoder-Decoder Transformer
Pretraining Objective	Instruction-Tuned Text-to-Text Learning	Denoising Autoencoding
Approximate Parameters	∼250M	∼139M
Tokenizer	SentencePiece	Byte-Pair Encoding (BPE)
Tokenizer Behavior	More robust handling of compound symbolic terms	More frequent fragmentation of compound symbolic terms
Input Prompting	Explicit task prompting used	Direct source sentence input
Prompt Format	`translate English text to english pictogram terms:`	Not used
Positional Encoding	Relative Positional Embeddings	Absolute Positional Embeddings
Attention Mechanism	Multi-Head Self-Attention	Multi-Head Self-Attention
Decoder Type	Autoregressive Generation	Autoregressive Generation
Compound Symbol Handling	Strong	Moderate
Semantic Compression Capability	Strong contextual simplification	Moderate contextual simplification
Contextual Generalization	High	High
Generation Stability	More stable for long symbolic sequences	Occasionally fragmented outputs
Maximum Sequence Length	128	128
Batch Size	8	8
Learning Rate	3×10^{-4}	3×10^{-4}
Optimizer	AdamW	AdamW
Weight Decay	0.01	0.01
Beam Search Width	12	12
Framework	Hugging Face Transformers	Hugging Face Transformers
Observed Training Behavior	Faster convergence with stable decoding	Stable convergence with slight token fragmentation
Observed Output Quality	Better semantic consistency	Competitive symbolic generation
AAC-Oriented Adaptability	Strong	Moderate to Strong

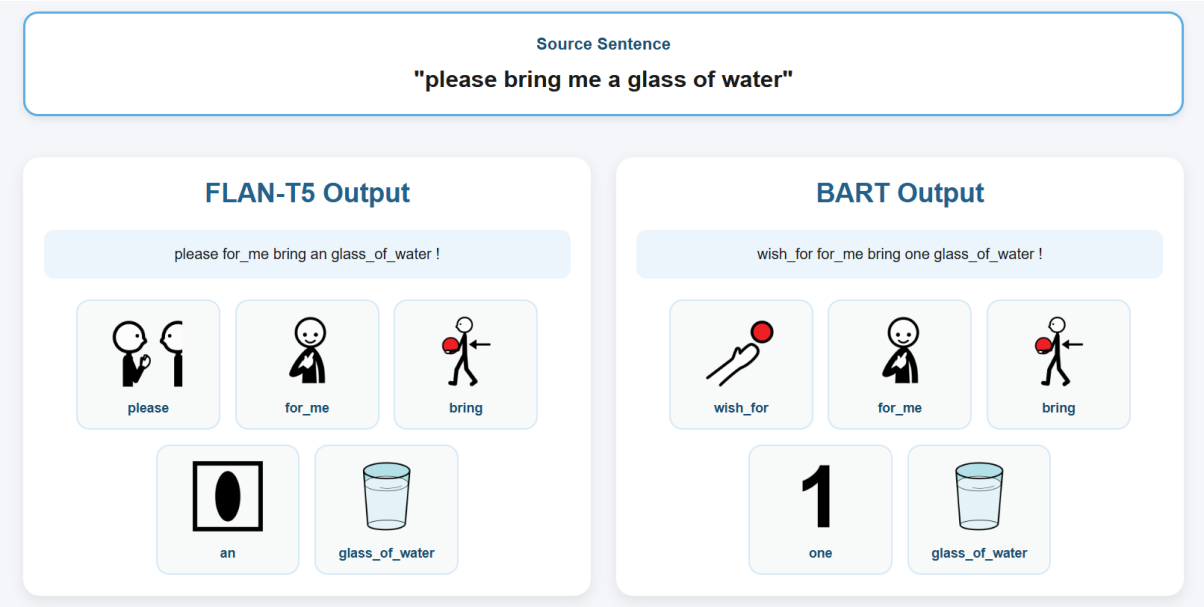

Figure 6: Qualitative comparison of pictogram-oriented symbolic sequence generation produced by the FLAN-T5 and BART transformer architectures

Both architectures reduce grammatical complexity while retaining the primary semantic concepts of the sentence. Articles and auxiliary structures are simplified, while semantically important entities and actions are preserved in the generated symbolic sequence.

These examples indicate that the transformer models learn contextual semantic mappings rather than relying solely on direct lexical substitution. The generated outputs demonstrate effective semantic compression and symbolic restructuring suitable for AAC-oriented communication environments.

5. Results and Discussion

The experimental results obtained on the validation and test datasets, summarized in Table 3, show that both transformer architectures are capable of generating semantically meaningful pictogram-oriented symbolic sequences for AAC-based communication. However, FLAN-T5 consistently achieved stronger

Table 3
Performance Comparison of FLAN-T5 and BART on the Validation and Test Sets

Model	Epochs	Validation Set			Test Set		
		SacreBLEU	METEOR	PictoER	SacreBLEU	METEOR	PictoER
FLAN-T5	5	47.44	0.6674	0.3682	50.2557	0.6947	0.3360
FLAN-T5	10	51.75	0.7137	0.3200	52.5473	0.7169	0.3148
BART	10	51.00	0.7007	0.3292	–	–	–

overall performance across both validation and official test evaluations.

Since FLAN-T5 achieved superior validation performance, only the best-performing FLAN-T5 configuration was submitted for the official hidden test evaluation. Consequently, test-set metrics for BART are not available, and the comparison between both architectures is limited to the validation set.

FLAN-T5 consistently outperformed BART under the adopted experimental configuration, demonstrating stronger semantic consistency and more coherent pictogram sequence generation. Its instruction-tuned pretraining and SentencePiece tokenization contributed to improved contextual simplification and better preservation of compound symbolic structures. Qualitative analysis further showed that FLAN-T5 generated more stable symbolic representations, whereas BART occasionally produced fragmented compound terms. Overall, the results demonstrate the effectiveness of encoder-decoder transformer architectures for AAC-oriented Text-to-Pictogram generation, with FLAN-T5 providing superior semantic preservation and symbolic sequence quality among the evaluated models.

6. Conclusion

This work presented a transformer-based approach for English Text-to-Pictogram generation for AAC-oriented communication. The study explored how sequence-to-sequence transformer architectures can convert natural language sentences into simplified pictogram-oriented symbolic representations while preserving the semantic meaning of the original text. Among the evaluated architectures, FLAN-T5 achieved the strongest overall performance on both the validation and official test evaluations. The model demonstrated better handling of compound pictogram structures, stronger contextual understanding, and more stable symbolic sequence generation during inference. Its instruction-tuned pretraining strategy and tokenizer design enabled the model to effectively adapt to semantic simplification and structured sequence generation tasks required for AAC-oriented communication systems.

The experimental and qualitative analyses further showed that the models learn contextual semantic restructuring instead of relying only on direct word-level substitution. The generated sequences were generally shorter, semantically focused, and structurally simplified, making them more suitable for pictogram-based communication environments. The study also highlighted the importance of tokenization strategies in improving symbolic sequence generation quality. Overall, the results indicate that encoder-decoder transformer architectures provide a strong foundation for Text-to-Pictogram generation and AAC-based semantic communication systems, with FLAN-T5 demonstrating the strongest performance among the evaluated models. Future work can explore larger instruction-tuned transformer models, multimodal pictogram alignment, retrieval-assisted generation, and context-aware decoding techniques to further improve semantic accuracy, symbolic consistency, and robustness across diverse AAC communication scenarios.

References

[1] M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, and D. Schwab. “Overview of the 2026 ImageCLEFtoPicto Task: Investigating the Generation of Pictogram Sequences from English Text Dataset.” In: *CLEF 2026 Working Notes*, Proceedings of the Seventeenth International Conference of the

CLEF Association (CLEF 2026), edited by E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, and J. M. Struß, Jena, Germany, September 21–24, 2026.

[2] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Ştefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W.-W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C.-N. Florea, and M. Ivanovici. "Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science (LNCS), Jena, Germany, September 21–24, 2026.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[4] H. W. Chung, L. Hou, S. Longpre, et al. "Scaling Instruction-Finetuned Language Models." *arXiv preprint arXiv:2210.11416*, 2022.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880, 2020.

[6] T. Kudo and J. Richardson. "SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018.

[7] T. Wolf, L. Debut, V. Sanh, 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*, pp. 38–45, 2020.

[8] K. Papineni, S. Roukos, T. Ward, and 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 (ACL)*, pp. 311–318, 2002.

[9] M. Post. "A Call for Clarity in Reporting BLEU Scores." In: *Proceedings of the Third Conference on Machine Translation (WMT)*, pp. 186–191, 2018.

[10] S. Banerjee and 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*, pp. 65–72, 2005.

[11] ARASAAC. "ARASAAC: Augmentative and Alternative Communication Pictograms." Available at: `https://arasaac.org`

[12] Abbhinav Elliah, Ananth Narayanan P, Bhuvan S, and P. Mirunalini. "Text-To-Picto Using Lexical Simplification." In: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024. Available at: `https://ceur-ws.org/Vol-3740/paper-146.pdf`

[13] Avaneesh Koushik, Jithu Morrison S, P. Mirunalini, and J. Aditya R K. "A Transformer Based Approach for Text-to-Picto Generation." In: *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, volume 3740 of CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024. Available at: `https://ceur-ws.org/Vol-3740/paper-154.pdf`