SSN Tokatrons at Text to Picto 2026: Sequence-Trie Constrained Decoding with Seq2Seq for English-to-Pictogram Translation*

Notebook for the ImageCLEF Lab at CLEF 2026

Varghese K James[1], Sree Krishna S[1], Sujith M[1] and Prabavathy Balasundaram[1]

[1]Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract

We present a Constrained Decoding method with a finetuned Sequence to Sequence model for the ImageCLEF 2026 ToPicto Task. Additionally, training was done with memory optimisations for a 12GB VRAM environment. We evaluate two encoder-decoder models: the standard T5-Base and the modern T5-Gemma-2-270m-270m, which adapts a decoder-only model via tied embeddings and merged attention for improved memory efficiency and robust semantic comprehension. During inference, a prefix tree (Trie) and beam search dynamically mask logits to guarantee vocabulary compliance. Our experimental results demonstrate a critical divergence in efficacy dependent on the underlying tokenizer. For T5-Base, constrained decoding improved all metrics, achieving a SacreBLEU of 46.49 and reducing the Picto-Error Rate to 36.37. Conversely, applying the identical constraint to T5-Gemma-2 triggered a severe performance collapse, dropping the SacreBLEU score to 33.76. This degradation is fundamentally attributed to likelihood misalignment and structural incompatibilities between the deterministic Trie paths and the complex pre-spacing rules of Gemma-2's SentencePiece Byte-Level BPE tokenizer. The findings indicate that while constrained decoding successfully enforces absolute dictionary compliance in standard models, integrating it with modern LLM tokenizers requires highly specialized alignment strategies to prevent generation failures.

Keywords

Augmentative and Alternative Communication, Semantic Mapping, Transformer Models, T5Gemma-2, Constrained Decoding, Natural Language Processing, Multimodal Communication

1. Introduction

Communication is a fundamental human right, but people with severe language or cognitive impairments often face significant barriers to expression. Augmentative and Alternative Communication (AAC) encompasses therapeutic approaches and tools ranging from manual signs to digital symbol boards designed to replace or supplement spoken and written language. Pictograms serve as highly effective visual communication tools, providing foundational linguistic access for individuals in early developmental stages or those with cognitive disabilities.

The motivation behind our participation in the ImageCLEF 2026 ToPicto task [1, 2] is to bridge the communication gap between AAC users and the broader society. By automating the translation of natural English text into standardized pictogram sequences, we aim to facilitate more seamless interactions. A critical challenge in this task is the rigid success parameter: the output vocabulary is strictly limited to the ARASAAC database[3]. If a model generates a perfectly valid English synonym that does not exist in the database, the translation fails entirely because the required image cannot be retrieved. To solve this, we propose fine-tuned encoder-decoder models optimized with deterministic constrained decoding at inference, ensuring the output perfectly complies with the approved dictionary without destroying valuable pretrained weights during the fine-tuning process.

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

*Code and notebooks are publicly available at https://github.com/vicfic18/text2picto.

vicfic@pm.me (V. K. James); sreekrishnassk166@gmail.com (S. K. S); Sujithmaris@gmail.com (S. M); prabavathyb@ssn.edu.in (P. Balasundaram)

0009-0002-9977-768X (V. K. James); 0009-0000-2362-165X (S. K. S); 0009-0001-0671-6195 (S. M); 0000-0003-3903-0410 (P. Balasundaram)

2. Related Works

The automation of text-to-pictogram translation has transitioned from rule-based syntactic parsing systems to sophisticated neural network approaches. Previous iterations of the ToPicto task, such as ImageCLEF 2024, focused heavily on translating French text into pictograms. In those tasks, encoder-decoder models demonstrated significant superiority over decoder-only architectures. For instance, Elliah et al. [4] from the previous year found that a fine-tuned Helsinki-NLP model achieved a Picto-term Error Rate (PictoER) of 18.51, greatly outperforming a GPT-2 model with a PictoER of 170.81. GPT-2 struggled primarily because its training dataset was predominantly English, limiting its ability to process French syntax. Anand et al. [5] implemented the Transformer model utilizing CamemBERT embeddings fused with contrastive learning, whereas Koushik et al. [6] fine-tuned Google-T5, achieving a significantly lower PictoER of 13.9.

Building upon the established efficacy of T5 architectures, recent efforts for the ImageCLEF 2025 ToPicto task have further optimized text-to-text frameworks. Hjuler and Fabre [7] fine-tuned the Google T5-large model on the new French corpus, achieving state-of-the-art test scores of 93.0 for SacreBLEU, 95.7 for METEOR, and 3.4 for PictoER in the Text-to-Picto task.

Our approach advances these methodologies by shifting focus to the English requirements of the 2026 [1] task and integrating modern, smaller and efficient encoder-decoder models with deterministic decoding constraints, ensuring the output strictly adheres to the structures required by visual AAC interfaces.

3. Dataset

The dataset provided for the 2026 ToPicto [2] task consists of 59278 natural English source sentences mapped to target sequences of semantic base-form text and their corresponding ARASAAC pictogram integer identifiers [1]. The structural foundation of this task traces its roots to resources like the TCOF corpus, which was prevalent in the French subsets of previous years [8]. Such corpora encompass a wide range of daily conversational scenarios, including debates, everyday situations, and medical consultations. This text is highly representative of real-world interactions between caregivers and individuals utilizing AAC.

Several distinct features did stand out in the dataset. The pictogram grammar was similar to a simplified English and this trend of losing detail in translation was common. An example was the word `half_body`, widely present in the training data, which acted as a stand-in for specific names. Such replacements do indicate the heavy contextual information that is required outside of the translation alone for pictogram based communications.

Src:	he	counts	among	his	students	philip gyselaer	and	cornelis van engelen
Tgt:	he	count	among	his	student	half_body	and	half_body
Src:	it's	the	heart	of	the	confederation burebasaga		
Tgt:	this be	the	heart	of	the	parliamentary_group		

Table 1
Sequence-to-sequence alignments highlighting morphological/syntactic changes (blue) and entity replacement (purple).

In addition to this, we went through the full 13733 English tagged pictograms in the ARASAAC dataset [3] and made a file with just the words separated by space. This was meant to be used as the word bank from which the Trie was to be constructed. Some pictograms had phrases as description, which was spaces which were replaced by underscores, similar to how it was done in the given dataset.

Because the ARASAAC database utilizes a highly structured and finite vocabulary, any generated token outside this established set constitutes a fatal error in the context of AAC software. To contextualize the sequence-to-sequence generation during training and inference, we uniformly prepended all

Tag	Definition	Example
id	unique identifier of each utterance	cefc-tcof-Acc_del_07-1
src	source of the utterance - text from oral transcription	you cannot know
tgt	target of the utterance - sequence of pictogram terms (tokens)	you be_able_to know not
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: The format of the dataset.

source English text with the specific prompt directive: “translate English to Picto: ”.

4. Methodology

4.1. Model Selection and Architectural Differences

We evaluated two distinct encoder-decoder architectures optimized for sequence-to-sequence tasks: the standard T5-Base (∼220M parameters) and the newer T5-Gemma-2-270m-270m (∼540M total parameters) [9]. The encoder-decoder architecture remains superior [4, 6] for translation tasks requiring simplification because the encoder processes the entire input sequence simultaneously, building a rich, bidirectional understanding of the context before the decoder begins autoregressive generation.

Unlike standard T5 models, T5-Gemma-2 is built by taking a powerful pretrained decoder-only model (Gemma 2 or 3) and adapting it into an encoder-decoder format via continued pretraining using a UL2 objective [9, 10]. To optimize efficiency and reduce memory bloat, T5-Gemma-2 utilizes tied embeddings which shares all word embeddings across the encoder and decoder and merged attention, which unifies the decoder’s self-attention and cross-attention into a single joint module. Furthermore, while standard early T5 models were heavily English-dominant, T5-Gemma-2 inherits robust inherent multilingual and multimodal capabilities from its base model, offering a richer baseline semantic comprehension. Both models were trained in bfloat16 precision to balance computational efficiency with translation performance, and input sequences were uniformly truncated to a maximum of 128 tokens.

4.2. Constrained Decoding

Standard Large Language Model tokenizers rely on subword algorithms to handle compound words and multilingual requirements, naturally drawing from their entire training vocabulary [11]. However, this task requires strict adherence to the finite ARASAAC vocabulary. Crucially, this constraint must be applied during inference, not during fine-tuning [12]. Blocking out-of-vocabulary predictions during fine-tuning distorts the probability distribution, preventing the cross-entropy loss function from accurately measuring errors and breaking the gradient descent process [13].

To enforce compliance at inference, a prefix tree (Trie) was constructed from the valid pictogram identifiers within the ARASAAC vocabulary file. During the generation loop, a custom `prefix_allowed_tokens_fn` intercepts the models’ raw logits and applies a large penalty (negative infinity) to any token that is not a valid child of the current Trie node. To combat “likelihood misalignment” where the model gets stuck because all allowed dictionary terms have a low natural probability, we utilized a 4-beam search. This allows the model to explore multiple valid paths simultaneously, eliminating out-of-vocabulary errors while preserving grammatical structure.

Algorithm 1 Retrieval of Valid Next Tokens under Vocabulary Constraints

procedure TRIE-GET-ALLOWED-TOKENS(T, G, *eos*)
  ▷ G is the sequence of already generated tokens $\langle g_1, g_2, \ldots, g_m \rangle$
  if $G.length == 0$ **then**
    return $\{k \mid k \in T.root.children\}$
  end if
  $u \leftarrow T.root$
  for $i \leftarrow 1$ **to** $G.length$ **do**
    $t \leftarrow G[i]$
    if $t \in u.children$ **then**
      $u \leftarrow u.children[t]$
      ▷ If phrase ends, allow transitioning back to start of any vocabulary phrase
    else if $u.is\_end ==$ TRUE **and** $t \in T.root.children$ **then**
      $u \leftarrow T.root.children[t]$
    else
      return $\{eos\}$
    end if
  end for
  $A \leftarrow \{k \mid k \in u.children\}$
  if $u.is\_end ==$ TRUE **then**
    $A \leftarrow A \cup \{eos\} \cup \{k \mid k \in T.root.children\}$
  end if
  if $A == \emptyset$ **then**
    return $\{eos\}$ ▷ Default fallback to prevent decoding deadlock
  end if
  return A
end procedure

4.3. Training Setup

The training was done on a consumer laptop with an Intel Core Ultra 9 275HX CPU and NVIDIA GeForce RTX 5070 Ti Mobile GPU (12 GB VRAM). The specific model sizes chosen were in light of this constraint. This implementation of Constrained Decoding was a single threaded process hence CPU bound at inference, while training was GPU accelerated.

Models were fine-tuned on 59278 training samples validated on 4397 cases. To minimize memory overhead while maintaining gradient stability, we utilized the Adafactor [14] optimizer in place of standard AdamW. Adafactor factors its momentum matrices, yielding sub-linear memory scaling that reduces the VRAM footprint significantly. Weight decay was set to 0.01.

Both models were fine-tuned over 7 complete epochs using a peak learning rate of 3×10^{-4}, regulated by a cosine scheduler following a 100-step warmup period to prevent early gradient explosion. We utilized a per-device batch size of 32 combined with three gradient accumulation steps, yielding an effective batch size of 96. Model checkpoints were evaluated at the end of each epoch, utilizing the SacreBLEU metric to determine the best model weights for final restoration.

5. Results

The evaluation was done using sacreBLEU, METEOR, and the Picto-term Error Rate (PictoER), all normalised to a scale between 0 to 100. In sacreBLEU and METEOR higher is better, and in PictoER lower is better.

Table 2
Model Evaluation Results for the validation split (n=4397 examples).

Model	Mode	SacreBLEU	METEOR	Picto-ER	Exact Match
T5-Gemma-2-270m-270m	unconstrained	45.63	64.73	38.74	**13.35**
T5-Gemma-2-270m-270m	constrained	33.76	60.92	50.63	6.39
T5-base	unconstrained	46.15	66.39	36.88	12.62
T5-base	constrained	**46.49**	**66.63**	**36.37**	12.92

6. Discussion

For the standard T5-Base model, enforcing the vocabulary constraint during decoding produced expected improvements across all translation metrics. Specifically, SacreBLEU increased by 0.34 points, METEOR increased by 0.24, and the Picto Error Rate (Picto-ER) improved by 0.51. By strictly confining the output space, the Trie successfully prevented the generation of out-of-vocabulary or hallucinated pictogram terms without disrupting the model's generation trajectory.

Conversely, enforcing the identical Trie constraint on T5-Gemma-2 drastically degraded performance. SacreBLEU dropped by 11.87 points, Picto-ER worsened by 11.89, and the Exact Match rate was more than halved (from 13.35 to 6.39). This performance collapse is fundamentally attributable to a discrepancy in tokenization strategies. The Gemma-2 architecture employs a SentencePiece-based Byte-Level BPE model characterized by unique pre-spacing rules (e.g., prepending spaces or specific subword tokens).

Standard phrase encoding prefixes such as `self.tokenizer.encode(phrase)` and `self.tokenizer.encode(" " + phrase)` align well with the standard T5 tokenizer. However, they create mismatched token paths when applied to the Gemma-2 tokenizer. Because the deterministic constraint forces the model's logits to conform to predefined Trie paths, any structural misalignment between these paths and the Gemma-2 model's natural subword predictions forces the network into selecting incorrect token sequences or prematurely generating the End-of-Sequence (EOS) token, severely penalizing overall translation quality.

7. Future Work

There is potential for an alternative implementation of Sequence-Trie based approach to constrained vocabulary translation methods such as English to Pictogram, with the caveat that it has to be tokenizer-specific. T5-Gemma-2 performs well on multilingual and reasoning benchmarks [9], but does not appear to show that in its English focused post-training scores [15]. The T5 which was mainly pretrained on English corpus performed better. A purely English focused encoder decoder model with the features of T5-Gemma-2 with a better constrained decoder algorithm can produce a viable solution to this task.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3.1 Pro in order to format and draft the research findings. After using the service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. 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. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, 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, 2026.

[2] M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, 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), Jena, Germany, 2026.

[3] S. Palao, Arasaac: Portal aragonés de la comunicación aumentativa y alternativa, https://arasaac.org, 2007. Accessed: 2026-05-26.

[4] A. Elliah, A. N. P, S. Bhuvan, P. Mirunalini, Text-to-picto using lexical simplification., in: CLEF (Working Notes), 2024, pp. 1579–1584.

[5] B. Anand, J. Themozhi, S. S. R, P. Charumathi, T. Mirnalinee, Ssn-mlrg at text to picto 2024: A bert-based approach for mapping french sentences to pictogram terms., in: CLEF (Working Notes), 2024, pp. 1474–1479.

[6] A. Koushik, J. Morrison, P. Mirunalini, J. A. R. K, A transformer based approach for text-to-picto generation., in: CLEF (Working Notes), 2024, pp. 1656–1661.

[7] M. J. Hjuler, I. Fabre, Parlez-vous picto? a transformer-based approach for text-to-picto and speech-to-picto translation in french, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[8] 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.

[9] B. Zhang, P. Suganthan, G. Liu, I. Philippov, S. Dua, B. Hora, K. Black, G. Martins, O. Sanseviero, S. Pathak, et al., T5gemma 2: Seeing, reading, and understanding longer, arXiv preprint arXiv:2512.14856 (2025).

[10] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, et al., Ul2: Unifying language learning paradigms, arXiv preprint arXiv:2205.05131 (2022).

[11] T. Kudo, 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, 2018, pp. 66–71.

[12] C. Hokamp, Q. Liu, Lexically constrained decoding for sequence generation using grid beam search, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1535–1546. URL: https://aclanthology.org/P17-1141/. doi:10.18653/v1/P17-1141.

[13] S. Geng, M. Josifoski, M. Peyrard, R. West, Grammar-constrained decoding for structured NLP tasks without finetuning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 10932–10952. URL: https://aclanthology.org/2023.emnlp-main.674/. doi:10.18653/v1/2023.emnlp-main.674.

[14] N. Shazeer, M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in: Interna-

tional conference on machine learning, PMLR, 2018, pp. 4596–4604.
[15] T. A. Chang, C. Arnett, Z. Tu, B. Bergen, When is multilinguality a curse? language modeling for 250 high-and low-resource languages, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 4074–4096.