

HULAT-UC3M at ImageCLEFtoPicto 2026: English Text-to-Pictogram Generation with T5

Notebook for the ImageCLEF Lab at CLEF 2026

Ismael Guzman-Colomina¹, Jose Luis Lopez-Cuadrado^{1,*}, Jesus Leonardo Lopez-Hernandez¹ and Israel Gonzalez-Carrasco¹

¹Computer Science and Engineering Department, Universidad Carlos III de Madrid, Av. de la Universidad 30, 28911 Leganes, Madrid, Spain

Abstract

The use of pictograms can improve linguistic comprehension and expression for individuals with complex communication needs. Within this framework, the ImageCLEFtoPicto 2026 track seeks to advance translation methods that automate the transformation of text or speech into pictogram sequences. The HULAT-UC3M approach submitted to this task addresses the generation of pictogram-term sequences from English sentences. It is based on a fine-tuned T5-base model. The submitted configuration was selected on the validation set and uses a task prefix together with greedy decoding. On the official test set, the proposed approach obtained a WER of 0.3199, a BLEU score of 51.9991, and a METEOR score of 0.7102.

Keywords

Augmentative and Alternative Communication, Text-to-pictogram generation, ImageCLEFtoPicto, T5, Pictogram sequences

1. Introduction

The ImageCLEFtoPicto 2026 task focuses on the automatic generation of sequences of pictographic tokens from English sentences [1]. This task is part of the ImageCLEF 2026 evaluation campaign [2] and provides a common setting for comparing systems that transform text into pictographic representations.

The motivation for this task comes from Augmentative and Alternative Communication (AAC), where pictograms are used to support the comprehension and expression of messages in people with communicative, linguistic or cognitive difficulties [3, 4]. In this context, the automatic generation of pictographic sequences can help reduce the effort required to adapt textual content into more accessible formats.

In contrast to conventional automatic translation, the generation of text into pictograms does not aim to produce a fluent sentence in another natural language. The expected result is a controlled sequence of pictographic concepts. This implies that the system must learn, from the training data, which parts of the input sentence are relevant for pictographic representation, which elements can be ignored, and how specific, infrequent expressions or proper names should be assigned to the available pictographic tokens. Previous work on text-to-pictogram translation and semantic resources for pictograms has addressed this problem using linguistic rules, lexical-semantic databases and links to resources such as WordNet [5, 6, 7].

This paper describes the participation of the HULAT group at Carlos III University of Madrid in the ImageCLEFtoPicto 2026 task. Our approach consisted of formulating the problem as a supervised text-to-text generation task and fine-tuning encoder-decoder models from the T5 family [8]. During the experiments, different training and decoding configurations were compared, including model variants, random seeds, learning rates, effective batch size, the use of task prefixes and generation strategies.

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

*Corresponding author.

✉ iguzman@inf.uc3m.es (I. Guzman-Colomina); jllopez@inf.uc3m.es (J. L. Lopez-Cuadrado); jelopezh@inf.uc3m.es (J. L. Lopez-Hernandez); igcarras@inf.uc3m.es (I. Gonzalez-Carrasco)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The best-submitted approach was based on a T5-Based model. The official run was selected based on the performance achieved on the validation set, taking into account both automatic metrics and the behaviour of different decoding configurations.

The rest of the paper is organised as follows. Section 2 reviews related work on AAC, pictographic resources, and neural models applied to pictograms. Section 3 describes the dataset used in the task. Section 4 presents the proposed methodology. Section 5 describes the evaluation metrics. Section 6 reports the validation results and discusses the main findings. Finally, Section 7 summarises the main conclusions.

2. Related Work

Automatic translation of text into pictograms has been studied primarily in the context of Augmentative and Alternative Communication (AAC). Initial approaches were based on linguistic rules and lexical-semantic resources. A representative example is the Text-to-Pictograph system, initially developed for Dutch, which converts text into sequences of pictograms using surface linguistic analysis and semantic mapping [5]. This line of research was subsequently extended to other languages, including English and Spanish, through interlinguistic links and semantic resources [6].

Another related line of research has focused on enriching pictographic resources with semantic information. For example, the connection between ARASAAC and WordNet allows pictograms to be associated with synsets, facilitating the disambiguation and automatic processing of pictographic concepts [7]. These resources are relevant because text-to-pictogram generation does not consist only of replacing words with symbols. In many cases, the system must select the most appropriate pictographic concept based on the context.

Recent work has also begun using neural models for AAC-related pictogram tasks. PictoBERT and PrAACT, for example, use Transformer-based models to predict pictograms on communication boards and assist users while composing messages [9, 10]. These systems differ from ImageCLEFtoPicto in that they focus on prediction within AAC interfaces rather than the complete generation of pictograms from text. Nevertheless, they demonstrate that Transformer models can be useful for modelling pictographic sequences.

Within the context of ImageCLEF, recent work has approached pictogram generation as a sequence-to-sequence task. In particular, Hjuler and Fabre explored a Transformer-based approach for Text-to-Picto and Speech-to-Picto translation in French using Google-T5 models trained on the task’s data [11]. Their work shows that T5-based models constitute a solid approach for generating sequences of pictograms from text. Our work follows this general formulation but focuses on the ImageCLEFtoPicto 2026 task with English input and compares different T5-based configurations to select the final run to submit.

3. Dataset

The data used in the ImageCLEFtoPicto 2026 task comes from the Common Voice v15 corpus [12]. Common Voice is a multilingual corpus of transcribed speech, recorded by users and designed primarily for the development of automatic speech recognition technologies. According to the official description of ImageCLEFtoPicto 2026, the texts are sourced from various public domain sources and, in the context of ToPicto, are associated with a sequence of terms linked to pictograms [1].

Although the original data source is speech-related, this work focuses only on the Text-to-Picto setting. Therefore, the model input is a textual phrase in English, and the expected output is a sequence of pictographic terms in English. Each term in the target sequence is linked to an ARASAAC pictogram identifier [13, 14]. In this way, the problem can be formulated as a supervised text-to-text generation task, where each input sentence is mapped to a sequence of pictographic terms.

The training set contains 59,278 phrase–pictogram pairs, whereas the validation set contains 4,397 examples. Both sets include the input phrase and the corresponding target sequence. The test set

Table 1

Example of an instance from the ImageCLEFtoPicto 2026 dataset.

Field	Description	Example
id	Unique identifier of the instance	common_voice_fr_32182614.mp3
src	Source sentence in English	the seat of the county is watertown
tgt	Target sequence of pictogram terms	the seat of be city
pictos	List of ARASAAC pictogram identifiers aligned with the target sequence	[8476, 36505, 7074, 36480, 2704]

contains 4,306 instances and provides only the identifier and the input phrase. For each test instance, the system must generate a hypothesis with the predicted sequence of pictographic items.

The official submission format consists of a JSON file with one prediction per test instance. Each prediction contains the phrase identifier and the system-generated hypothesis, stored in the `id` and `hyp` fields. Prior to submission, it was verified that our file contained predictions for all 4,306 test instances, with no empty hypotheses and the expected JSON structure.

4. Methodology

The ImageCLEFtoPicto 2026 task was treated as a supervised text-to-text generation problem. Each training example consists of an English sentence and its corresponding sequence of pictogram terms. The model is therefore trained to map the input sentence directly to the target pictogram sequence.

4.1. Text-to-text formulation

The task was addressed using encoder-decoder models from the T5 family [8]. This family of models allows different natural language processing tasks to be represented within a single text-to-text framework, in which both the input and output are text sequences. In our case, the input is the English phrase provided by the dataset, and the output is the expected sequence of pictographic terms.

Since T5 represents the input as a text sequence, the task instruction can be included directly before the source sentence. Two input formats were evaluated. The first used the original English sentence without modification. The second added the prefix `translate English to pictograms:` before the source text. For example, the sentence `the seat of the county is watertown` was provided either in its original form or as `translate English to pictograms: the seat of the county is watertown`. The submitted system used the prefixed input format.

4.2. Training setup

Several configurations were trained during development, mainly using T5-base. The experiments varied the random seed, learning rate, effective batch size, and the use of a task prefix. FLAN-T5-base was also used as an initial exploratory model, but the main comparison focused on T5-base configurations, since these runs were the most competitive and had comparable validation decoding results.

Table 2 summarises the main configurations considered during development. The table reports the checkpoint selected according to validation loss during training. This checkpoint was not always the final model submitted, as the final run was selected later based on validation-set generation metrics.

Training was configured for a maximum of 20 epochs. Early stopping was enabled during training, and validation loss was monitored, following the usual strategy of stopping training when validation performance no longer improves [15]. Models were evaluated and saved at the end of each epoch, and the best model according to validation loss was loaded at the end of training. The early stopping patience was set to 5, with a threshold of 10^{-4} in the training setup. Since the monitored metric was `eval_loss`, lower values were considered better.

Table 2

Main model configurations evaluated on the validation set. The last column identifies the checkpoint used in the reported validation comparison.

ID	Model configuration	Prefix	Seed	LR	Eff. batch	Evaluated checkpoint
C1	FLAN-T5-base	yes	42	10^{-4}	32	epoch 8
C2	T5-base, seed 777, effective batch 16	yes	777	10^{-4}	16	epoch 8
C3	T5-base, seed 123	yes	123	10^{-4}	32	epoch 9
C4	T5-base, seed 123, effective batch 16	yes	123	10^{-4}	16	epoch 6
C5	T5-base, no prefix	no	42	10^{-4}	32	epoch 10
C6	T5-base, seed 777	yes	777	10^{-4}	32	epoch 8
C7	T5-base, lower learning rate	yes	42	$5 \cdot 10^{-5}$	32	epoch 16
C8	T5-base, seed 123, lower learning rate	yes	123	$5 \cdot 10^{-5}$	32	epoch 15
C9	T5-base, seed 2026	yes	2026	10^{-4}	32	epoch 7

However, the final submitted run was not selected based only on validation loss. After training, the main candidate models and selected checkpoints were evaluated on the validation set using generation metrics. This distinction is important because the checkpoint with the lowest validation loss does not necessarily produce the best generated sequence, as measured by SacreBLEU, METEOR, or PictoER.

For the best submitted approach, the checkpoint with the lowest validation loss was reached at epoch 7. Nevertheless, the checkpoint at epoch 8 was selected for the official submission because it achieved the best validation performance after generation and decoding. This model corresponded to the T5-base model with seed 777, a learning rate of 10^{-4} , an effective batch size of 16, and a prefixed input.

4.3. Decoding and model selection

Once the models were trained, the main candidates were decoded on the validation set using different generation settings. Greedy decoding and beam-search variants were tested. Beam sizes of 1, 2, 4, 6, and 8 were considered, together with length penalties of 0.8, 1.0, and 1.2 for the beam-search variants. The best submitted approach used greedy decoding.

The generated sequences were then evaluated against the validation references. The approach submitted to the official test set was chosen from these validation experiments. It corresponded to the epoch-8 checkpoint of the T5-base model trained with seed 777, an effective batch size of 16, a learning rate of 10^{-4} , and prefixed input. Greedy decoding was used for the final test generation.

4.4. Implementation and reproducibility

To support reproducibility, the source code, ARASAAC English lexicon, and submitted T5 checkpoint are made available in a public repository.¹ The checkpoint is stored with Git LFS under `models/submitted_t5_base_checkpoint29640/`. The official ImageCLEFtoPicto task data are distributed by the competition organisers and are therefore not included in the repository; authorised users should place `train.json`, `valid.json`, and `test.json` in a local `data/` directory.

The English ARASAAC lexicon is used only in the post-generation output-normalisation step. It is included as `resources/arasaac_english.json` and was obtained from <https://api.arasaac.org/v1/pictograms/all/en>. This step normalises generated terms that are not present in the training/validation pictogram-term vocabulary by using ARASAAC keyword-to-pictogram links and the pictogram-term alignments observed in the training and validation data. Test references are not used.

Table 3 summarises the main implementation and reproducibility details of the submitted system.

¹Source code, reproducibility instructions, ARASAAC English lexicon, and submitted checkpoint: <https://github.com/hulat-group/clef-2026-topicto-hulat-uc3m>.

Table 3

Main reproducibility details for the submitted system.

Parameter	Value
Model	t5-base
Selected checkpoint	checkpoint29640
Training script	src/train_submitted_t5_base.py
Inference script	src/generate_official_submission.py
Seed	777
Learning rate	10^{-4}
Optimiser	AdamW
Maximum epochs	20
Effective batch size	16
Maximum source length	64 tokens
Maximum target length	32 tokens
Final decoding	Greedy decoding
Python	3.10
PyTorch	2.5.1+cu121
Transformers	5.8.0
Datasets	4.8.5
GPU	NVIDIA GeForce RTX 3060, 12 GB

5. Evaluation

The approaches were evaluated on the validation set by comparing the generated sequence of pictogram terms with the reference target sequence. For each validation instance, the model produced one hypothesis, which was compared with the corresponding `tgt` field.

The evaluation included the metrics from the ImageCLEFtoPicto task: SacreBLEU, METEOR, and Picto-term Error Rate (PictoER). SacreBLEU is a standardised implementation of BLEU and measures n-gram overlap between the generated hypothesis and the reference sequence [16, 17]. METEOR provides a complementary score based on a more flexible alignment between hypothesis and reference [18]. PictoER is derived from Word Error Rate and measures token-level errors over pictogram terms, as described in the task overview [1].

In the validation experiments, SacreBLEU and METEOR were computed directly over the generated and reference pictogram-term sequences. PictoER was approximated by mapping each generated pictogram term to the most frequent pictogram identifier observed for that term in the training and validation data, and then computing the edit distance with respect to the reference pictogram identifier sequence. This approximation was used only for internal validation, since the official evaluation was performed by the task organisers.

Higher SacreBLEU and METEOR values indicate better performance, whereas lower PictoER values indicate fewer pictogram-term errors. The final test submission followed the official format, with one generated hypothesis per test instance using the fields `id` and `hyp`.

6. Results and Discussion

6.1. Validation results

Table 4 reports the validation results for the configurations described in Table 2. For each configuration, the reported decoding strategy corresponds to the setting that obtained the highest SacreBLEU score in its validation grid. METEOR and $\text{PictoER}_{\text{approx}}$ are included as complementary measures. As explained in Section 5, $\text{PictoER}_{\text{approx}}$ was computed internally by mapping generated pictogram terms to pictogram identifiers using the available training and validation alignments.

Configuration C2 obtained the best overall result on the validation set. This configuration corresponds

Table 4

Validation results for the configurations defined in Table 2. Higher SacreBLEU and METEOR are better; lower PictoER_{approx} is better.

ID	Decoding	SacreBLEU	METEOR	PictoER _{approx}
C2	greedy	50.798	0.704	0.304
C3	greedy	50.180	0.697	0.310
C4	greedy	50.084	0.697	0.311
C5	greedy	49.909	0.698	0.310
C6	beam2	49.806	0.699	0.306
C7	greedy	49.302	0.694	0.312
C8	greedy	49.054	0.690	0.311
C9	greedy	48.775	0.692	0.313
C1	beam2	48.757	0.686	0.324

Table 5

Official test results for the submitted HULAT run in ImageCLEFtoPicto 2026.

Rank	Run ID	WER	BLEU	METEOR
2	1070	0.3199	51.9991	0.7102

to T5-base with a prefixed input, seed 777, a learning rate of 10^{-4} , and an effective batch size of 16. It achieved the highest SacreBLEU and METEOR scores and the lowest PictoER_{approx} among the evaluated systems. Although the checkpoint with the lowest validation loss in this training run was at epoch 7, the checkpoint at epoch 8 produced the best validation generation result and was therefore used for the final submission.

The comparison between C2 and C5 provides information about the use of the input prefix. C5, which was trained without the instruction `translate English to pictograms:`, obtained a lower SacreBLEU score than C2. Although the configurations also differ in seed and effective batch size, this result supports retaining the prefixed formulation in the submitted approach.

The experiments also show that reducing the learning rate did not improve performance. Configurations C7 and C8, trained with $5 \cdot 10^{-5}$, obtained lower SacreBLEU and METEOR scores than the best configuration trained with 10^{-4} . Differences among runs trained with different random seeds also indicate some variability in the fine-tuning process: C2 achieved the best result, while C9 had the lowest SacreBLEU score among the reported T5-base configurations trained with a learning rate of 10^{-4} and prefixed input.

Regarding the model family, FLAN-T5-base (C1) obtained lower validation scores than the best T5-base configurations. This result led to a focus on the subsequent development of the T5 base. Concerning decoding, greedy decoding was selected for the submitted model. Although beam-search variants were evaluated, the best validation result for C2 was obtained with greedy decoding.

6.2. Official test results

The selected approach was submitted to the official ImageCLEFtoPicto 2026 evaluation. Table 5 reports the scores provided by the organisers for the submitted HULAT run. In the official leaderboard, the approach ranked second.

The official scores are consistent with the validation-based selection procedure: the configuration chosen during development also obtained a competitive result on the unseen test set. The difference with the top-ranked system was small, with the first-ranked run obtaining a WER of 0.3148, BLEU of 52.5473, and METEOR of 0.7169. These results show that a fine-tuned T5-base model, together with validation-based selection of training and decoding settings, provides a baseline for English text-to-pictogram generation.

7. Conclusions

This paper describes HULAT-UC3M's participation in the ImageCLEFtoPicto 2026 task. The task was approached as a supervised text-to-text generation problem, using models from the T5 family to generate sequences of pictogram terms from English sentences.

Several configurations were evaluated during development, varying the model, input format, random seed, learning rate, effective batch size, and decoding strategy. The best validation result was obtained with a T5-base model using prefixed input, seed 777, a learning rate of 10^{-4} , and an effective batch size of 16. Although the lowest validation loss was achieved at epoch 7, the checkpoint from epoch 8 produced better generation results on the validation set and was used for the official submission. The final predictions were generated using greedy decoding.

The submitted run ranked second in the official ImageCLEFtoPicto 2026 evaluation, with a WER of 0.3199, a BLEU score of 51.9991, and a METEOR score of 0.7102.

Further work will examine errors in the generated sequences, particularly omissions, unnecessary terms, and named entities. Another aspect to explore is whether restricting generation to the pictogram-term vocabulary improves the final output.

8. Acknowledgments

This work is part of the grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models – HUMAN_AI), funded by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check and improve writing style. The authors also used ChatGPT (GPT-5) in order to: paraphrase and reword, improve writing style and formatting assistance.

No Generative AI tool was used to conduct the reported experiments, train the models, compute the metrics, generate images, or create the official submitted run. The experimental design, implementation decisions, model training, evaluation, and final manuscript content were reviewed and validated by the authors, who take full responsibility for the paper.

References

- [1] M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, Overview of the 2026 imageclef-to-picto task: Investigating the generation of pictogram sequences from english text dataset, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Jena, Germany, 2026.
- [2] 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, et al., 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.
- [3] D. R. Beukelman, J. C. Light, Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs, 5 ed., Paul H. Brookes Publishing, 2020.
- [4] M. Ronski, R. A. Sevcik, Augmentative communication and early intervention: Myths and realities, *Infants & Young Children* 18 (2005) 174–185.

- [5] V. Vandeghinste, L. Sevens, I. Schuurman, F. Van Eynde, Translating text into pictographs, *Natural Language Engineering* 23 (2017) 217–244.
- [6] L. Sevens, V. Vandeghinste, I. Schuurman, F. Van Eynde, Extending a dutch text-to-pictograph converter to english and spanish, in: *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 110–117.
- [7] D. Schwab, P. Trial, T. Vaschalde, L. Vial, Providing semantic knowledge to a set of pictograms for people with disabilities: A set of links between wordnet and arasaac, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 113–119.
- [8] 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.
- [9] J. A. Pereira, D. Macêdo, C. Zanchettin, A. L. I. Oliveira, R. N. Fidalgo, Pictobert: Transformers for next pictogram prediction, *Expert Systems with Applications* 202 (2022) 117231.
- [10] J. A. Pereira, C. Zanchettin, R. N. Fidalgo, Praact: Predictive augmentative and alternative communication with transformers, *Expert Systems with Applications* 240 (2024) 122417.
- [11] 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*, Madrid, Spain, 2025.
- [12] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 4218–4222.
- [13] N. Janssen, E. Navarrete, C. Perret, A. Marful, D. Paolieri, Norms for a pictographic system: The aragonese portal of augmentative/alternative communication (ARASAAC) system, *Frontiers in Psychology* 9 (2018) 2538. doi:10.3389/fpsyg.2018.02538.
- [14] Centro Aragonés de Tecnologías para la Educación, ARASAAC: Portal aragonés de la comunicación aumentativa y alternativa – public api, <https://arasaac.org/developers/api>, 2024. Accessed: 2026-05-20.
- [15] L. Prechelt, Automatic early stopping using cross validation: Quantifying the criteria, *Neural Networks* 11 (1998) 761–767. doi:10.1016/S0893-6080(98)00010-0.
- [16] 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*, 2002, pp. 311–318.
- [17] M. Post, A call for clarity in reporting bleu scores, in: *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, 2018, pp. 186–191.
- [18] 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.