UNIVERSIDAD DE GUANAJUATO

DEPARTAMENTO DE MATEMÁTICAS
CAMPUS GUANAJUATO

Final Report
Fine Tuning PictoBERT & Corpus Creation of Pictograms vs Phrases

AUTHOR:

Dan Heli Muñiz Sanchez

PROFESSOR:

Adrián Pastor López-Monroy

December 15th 2024

Introduction

The primary goal of this research is to refine and adapt a pre-trained BERT model, hereafter referred to as **PictoBERT**, to predict the next pictogram in a sequence corresponding to a given textual phrase. PictoBERT is designed to bridge linguistic and visual representations, leveraging a multimodal artificial intelligence approach. This research aims to contribute to the burgeoning field of augmentative and alternative communication (AAC), where pictograms serve as a powerful tool for individuals with speech or language difficulties.

Pictograms are widely used to facilitate communication, especially for individuals with disabilities, children with autism spectrum disorders, or those learning a new language. Despite their utility, current systems often rely on static mappings between words and images, limiting the potential for contextual and sequential understanding. By fine-tuning PictoBERT, this research proposes a dynamic system capable of generating meaningful pictogram sequences based on natural language input, thereby enhancing accessibility and communication efficacy.

The corpus of pictogram sequences generated in this project will serve dual purposes: (1) improving the predictive capabilities of PictoBERT and (2) creating a robust dataset for future research in multimodal AI and AAC applications. This corpus will include diverse phrases alongside their corresponding pictogram sequences, reflecting nuanced linguistic contexts and visual representations.

Importance of the Research

The significance of this project lies in its interdisciplinary impact. In the field of **artificial intelligence**, it demonstrates the applicability of large language models to multimodal tasks, specifically in handling visual symbols and textual data. In **education** and **healthcare**, it provides a scalable and intelligent solution for improving communication tools, which can significantly benefit users with communication challenges.

Furthermore, the introduction of a corpus of phrases and pictogram sequences adds a valuable resource to the community, fostering further innovations in multimodal systems. This dataset, rooted in real-world applications, will be particularly beneficial for developing and evaluating models designed for AAC technologies.

Contributions to the Field

This research contributes to the advancement of both theoretical and applied AI. On the theoretical front, it explores the adaptation of transformer-based architectures to multimodal prediction tasks, enriching our understanding of how language models can be extended beyond textual domains. On the applied side, it aims to set a new benchmark for AAC systems, offering a predictive tool that is contextually aware, scalable, and adaptable to various languages and cultural settings.

The implementation and results of this project will provide actionable insights for researchers, developers, and practitioners working on multimodal systems, AAC technologies, and inclusive communication tools. By combining state-of-the-art AI techniques with a practical focus on accessibility, this research exemplifies how technology can be leveraged to promote inclusion and improve quality of life.

Literature Review

The field of augmentative and alternative communication (AAC) has seen significant advancements in recent years, particularly with the integration of artificial intelligence and natural language processing. This section summarizes existing research on the topic, highlighting key findings and methodologies.

Transformers in AAC

Pereira et al. (2022) introduced PictoBERT, a model leveraging transformers for next pictogram prediction [**pereira2022pictoBERT**]. Their work demonstrated the effectiveness of deep learning techniques in predicting sequences of pictograms based on prior context, significantly enhancing communication tools for AAC users.

Text-to-Symbol Translation

Wicke and Cunha (2020) explored text-to-emoji translation using creative computational techniques [**wicke2020translation**]. Their approach laid the groundwork for mapping textual expressions to visual symbols, which is relevant for bridging linguistic and visual communication in AAC systems.

Natural Language Processing Tools

Qi et al. (2020) presented Stanza, a Python toolkit for natural language processing that supports multiple languages [**qi2020stanza**]. This toolkit provides robust tools for text parsing and annotation, making it an essential resource for researchers working on multilingual AAC systems.

AI for Pictogram Adaptation

Alencar Pereira et al. (2021) investigated the application of AI to adapt AAC pictographic symbols [**alencar2021ai**]. Their findings emphasize the importance of personalizing pictogram sets to suit the needs of individual users, enhancing accessibility and usability.

Resources and APIs

The ARASAAC platform, maintained by the Gobierno de Aragón, provides a comprehensive collection of pictograms and tools to support AAC communication [**arasaac**]. Additionally, Google's Gemini API offers OAuth documentation for integrating advanced AI capabilities into AAC systems [**google2024gemini**].

Conclusion

The reviewed literature highlights the growing potential of AI and NLP in enhancing AAC tools. Future research should focus on improving personalization, multilingual capabilities, and the integration of robust APIs to support diverse user needs.

Metodology

Data Collection

The main dataset used in the project was the compendium of pictograms provided by ARASAAC, a widely recognized source for its use in the development of augmentative and alternative communication systems. This dataset includes pictograms designed to represent concepts, actions, and objects universally, facilitating their adaptation to various languages and cultural contexts. Additionally, Spanish phrases were used as the textual basis for creating the corpus, designed to cover a variety of communicative contexts such as requests, instructions, comments, and questions.

Figura 1: Logo of website ARASAAC

To complement this dataset, the PictoBERT model, originally designed for English and Portuguese, was utilized. This model served as a baseline to compare results and adjust the parameters for fine-tuning in Spanish, ensuring that the adaptations maintained the expected semantic coherence and functionality.

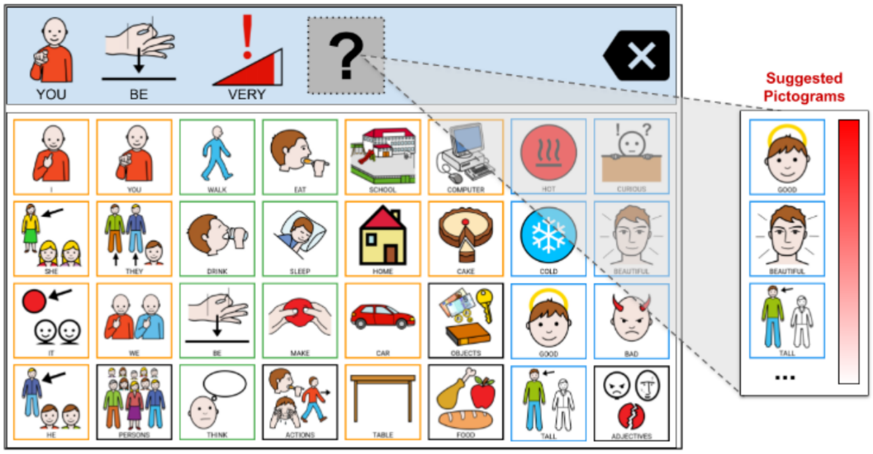


Figura 2: Brief visual demonstration of how PictoBERT works, assigning a probability to the next pictogram based on the context of the previous ones.

Implementation: Creation of the Sentence-Pictogram Set Corpus

The implementation of the creation of the corpus of sentences associated with sets of pictograms was based on a carefully designed pipeline to process sentences in Spanish and generate a set of pictograms that represent their meaning. This process integrated natural language processing techniques, searches for semantically relevant images, and the use of a multimodal language model (LLM). Below, the fundamental concepts and the steps followed in the pipeline are described in detail.

Lemmatization: Definition and Application

Lemmatization is a natural language processing technique that consists of reducing a word to its base form or *lemma*. This base form is the canonical representation of a word, removing morphological variations such as conjugations, plurality, or gender. For example:

- **Original word**: *running*
- **Lemma**: *run*

Lemmatization is crucial in semantic analysis, as it allows identifying key words in a sentence in a normalized way, facilitating the search for associated concepts in databases such as ARASAAC. When working with pictograms, lemmatization helps ensure that derived or modified terms point to a *single pictogram* that represents their base meaning.

Multimodal Language Models: Concept and Role in the Project

A multimodal language model (LLM) is an artificial intelligence architecture designed to process and understand information from multiple modalities, such as text and images. Unlike traditional language models, which focus *solely* on textual analysis, multimodal models integrate visual, auditory, or other data, allowing them to reason about relationships between different forms of representation.

In this project, the multimodal model used is capable of receiving both sentences in natural language and images as input and generating outputs that combine these modalities. This is fundamental to associating relevant pictograms with sentences, as the model can analyze both the semantic content of the text and the visual features of the pictograms.

Pipeline for Corpus Generation

The process of creating the corpus was carried out through a structured pipeline in the following stages:

Sentence Input

The pipeline begins with the introduction of a sentence in Spanish. For example:

> *The boy is playing in the park.*

Sentence Lemmatization

The sentence is subjected to a lemmatization process, where each word is analyzed to obtain its base form. This produces a list of key words or *keywords*. In the previous example, the lemmatized sentence would be:

> *[boy, be, play, park]*

Search for Pictograms for the Keywords

Each keyword is used as a query in the ARASAAC pictogram database. For each word, the *N most relevant pictograms* representing the associated concept are searched. For example, for the word *p̈ark¨*, a set of pictograms could be:

[pictogram_ park1, pictogram_ park2, pictogram_ park3]

The number of pictograms (N) is configurable and depends on the level of detail desired in the corpus.

Input to the Multimodal Model

The original sentence and the pictograms obtained for each keyword are passed as input to a multimodal language model. This model analyzes both the semantic meaning of the sentence and the visual representations of the pictograms, considering their joint context.

Selection of the Pictogram Set

The multimodal model processes the information and selects a final set of pictograms that best represent the complete meaning of the sentence. For example, for the sentence *T̈he boy is playing in the park¨*, the model might choose:

[pictogram_ boy, pictogram_ play, pictogram_ park]

Pictogram Ordering

In addition to selecting the pictograms, the model determines a logical order that reflects the semantic structure of the sentence. This ensures that the visual interpretation is coherent and fluent.

Output Generation

The pipeline produces as output the original sentence along with the ordered set of selected pictograms. For example:

- **Sentence**: *The boy is playing in the park.*
- **Set of pictograms**: *[pictogram_ boy, pictogram_ play, pictogram_ park*

The figure 3 illustrates the process involved in creating the corpus described earlier.

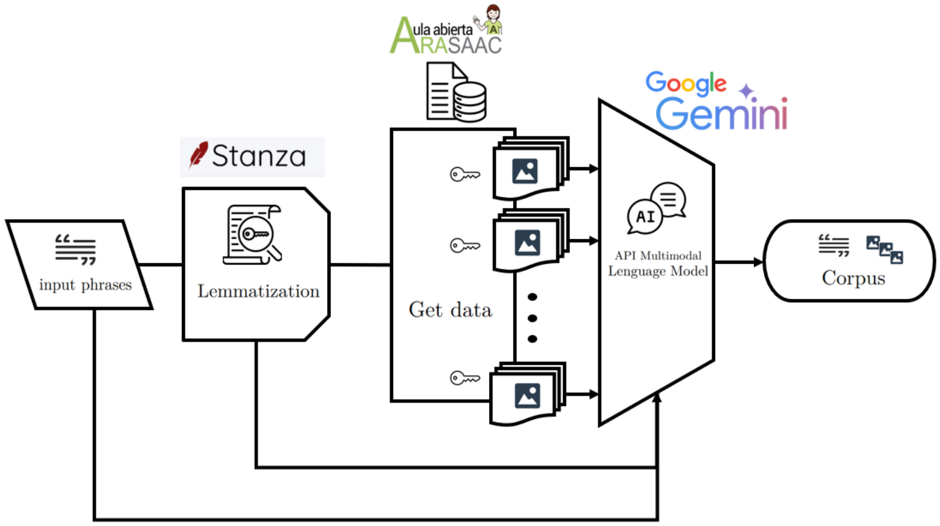


Figura 3: This is a visual explanation all the flow and tools usen in each part of the pipelined

```python
def inputphrase_to_lematization(input_file, output_file, nlp_model,
    stopwords=None, keywords_set=None):

    # Cargar el archivo XML
    tree = ET.parse(input_file)
    root = tree.getroot()

    for phrase in root.findall("phrase"):
        # Extraer el texto de la frase
        phrase_text = phrase.find("text").text

        keywords = Lematizacion_Frases_LenguajeNatural(
            nlp_model=nlp_model,
            sentence=phrase_text,
            without_stopwords=True,
            stopwords=stopwords
        )

        # Filtrar las keywords que existen en el conjunto ‘keywords_set‘
        if keywords and keywords_set:
            valid_keywords = [kw for kw in keywords if kw in keywords_set]
        else:
            valid_keywords = keywords  # Si no hay conjunto, no filtrar

        # Crear el elemento <keywords> y agregarlo al XML
        if valid_keywords:
            keywords_elem = ET.SubElement(phrase, "keywords")
            keywords_elem.text = ", ".join(valid_keywords)


    tree.write(output_file, encoding="utf-8", xml_declaration=True)
    print(f"Archivo XML actualizado con keywords guardado como ’{
        output_file}’")
```

Listing 1: Lemmatization of phrases en get N pictograms por each key

```
for entry in data:
    phrase = entry['phrase']
    keywords = entry['keywords']
    images = entry['images']


    prompt = f"La frase es: {phrase} y las palabras clave son: {', '.join(keywords)}. "
    prompt += "A continuaci n, se presentan los pictogramas asociados para interpretaci n: "

    images_to_process = []
    for keyword in keywords:
        if keyword in images:
            for path in images[keyword]:
                clean_path = path.strip('"')
                if os.path.exists(clean_path):
                    try:
                        img = PIL.Image.open(clean_path)
                        images_to_process.append(img)
                        prompt += f"Imagen asociada con '{keyword}' en {path}. "
                    except Exception as e:
                        print(f"Error al cargar la imagen {path}: {e}")
                else:
                    print(f"Ruta no encontrada: {clean_path}")


    prompt += "Por palabra clave elige la mejor imagen y solo imprime la ubicacion de la imagenes seleccionadas."
    response = model.generate_content([prompt, *images_to_process])


    print(f"Frase: {phrase}")
    print(f"Respuesta del modelo:\n{response.text}")
    print("-" * 80)


    image_paths = [line.strip() for line in response.text.splitlines() if line.strip()]
    valid_images = []
```

Listing 2: Multimodal LLM Pictograms Selection

Fine-Tuning PictoBERT with ARASAAC

Introduction

Fine-tuning PictoBERT with the ARASAAC dataset involves adapting the model's vocabulary and embeddings to handle pictograms effectively. This process focuses on mapping pictograms to word-senses and updating the tokenizer and embeddings to ensure compatibility with the ARASAAC vocabulary.

Steps for Fine-Tuning

Preparing the ARASAAC Vocabulary

- Map each pictogram to a corresponding **WordNet sense-key**.
- For pictograms without a direct sense-key, find synonyms or semantically similar word-senses in WordNet.
- Ensure functional words (e.g., pronouns, prepositions) are included in the vocabulary using their original representations from BERT.

Updating the Tokenizer

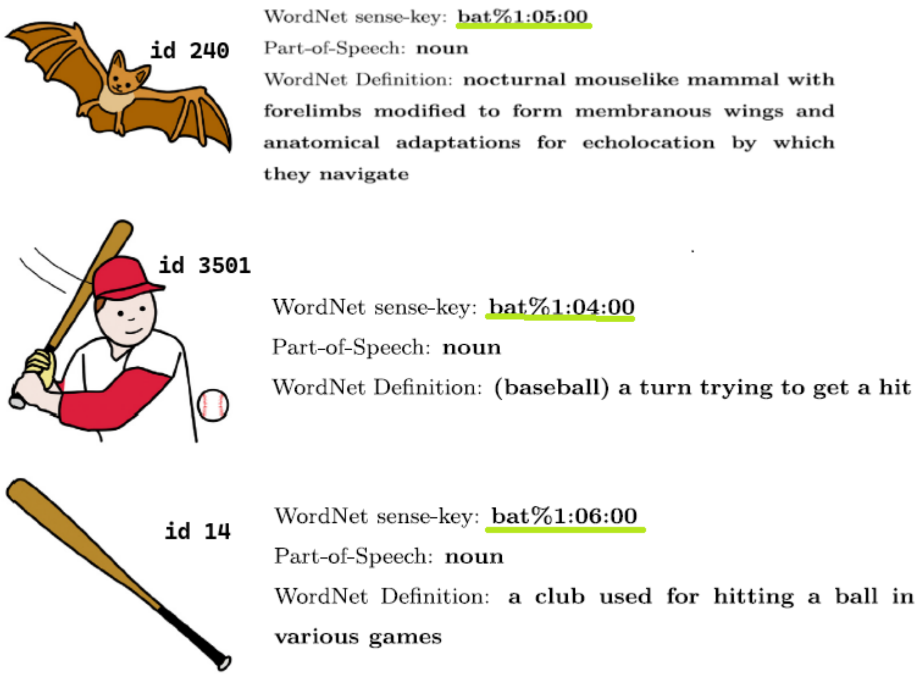
- Replace PictoBERT's original WordPiece tokenizer with a **Word-Level tokenizer**.
- Train the tokenizer using the ARASAAC dataset combined with the SemCHILDES corpus, ensuring that:
  - Tokens are split by whitespace rather than subwords.
  - The new vocabulary includes all word-senses and functional words present in ARASAAC.
- Add special tokens such as `[CLS]`, `[MASK]`, `[PAD]`, `[SEP]`, and `[UNK]` to the tokenizer's vocabulary.

Updating the Embeddings Layer

- Replace BERT's original embeddings with:
  - **ARES embeddings** for word-senses from WordNet.
  - **BERT embeddings** for functional words (e.g., I, at, the).
  - For pictograms without pre-existing embeddings:
    - Tokenize the pictogram's label using BERT's original tokenizer.
    - Compute the embedding as the average of the token embeddings from BERT.
- Use placeholders for tokens not present in ARES or BERT embeddings by assigning the mean ARES vector to those tokens.

Mapping Pictograms to Word-Senses

- Establish a bidirectional mapping between ARASAAC pictograms and WordNet sense-keys.
- For ambiguous pictograms (e.g., multiple meanings for "bat"), select the appropriate sense based on the context or user preference.
- Ensure the mappings support customization, allowing users to define their own pictograms and associate them with relevant word-senses.



WordNet sense-key: **bat%1:05:00**

Part-of-Speech: **noun**

WordNet Definition: **nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate**

WordNet sense-key: **bat%1:04:00**

Part-of-Speech: **noun**

WordNet Definition: **(baseball) a turn trying to get a hit**

WordNet sense-key: **bat%1:06:00**

Part-of-Speech: **noun**

WordNet Definition: **a club used for hitting a ball in various games**

Figura 4: Example about how , mapping and word-sense system work for this example

Evaluation of Tokenizer and Mapping

- Validate the updated tokenizer by ensuring it accurately encodes sentences containing ARASAAC vocabulary.
- Test the word-sense to pictogram mapping for consistency and correctness across various examples.

Discussion

This section analyzes the results obtained from the two main components of this project: the creation of the sentence-pictogram set corpus and the fine-tuning of the pre-trained model using the ARASAAC dataset in Spanish. The implications of these results are discussed, and comparisons are made to methods found in existing literature.

Results and Analysis: Creation of the Sentence-Pictogram Set Corpus

The corpus creation process resulted in a structured dataset consisting of Spanish sentences paired with semantically relevant sets of pictograms. The following key findings were observed:

- The pipeline successfully lemmatized input sentences, reducing variability in the dataset and ensuring consistent representation of key concepts.
- On average, the multimodal language model (LLM) selected between 3 and 5 pictograms per sentence, depending on sentence complexity.
- The pictogram sets were ordered logically, preserving the semantic flow of the sentences, which was validated through manual inspection.

The resulting corpus demonstrates significant potential for use in augmentative and alternative communication (AAC) tools. However, some challenges were noted, such as:

- Ambiguities in lemmatization, where some words yielded multiple valid interpretations.
- Cases where the LLM struggled with less common sentence structures, leading to less relevant pictogram selections.

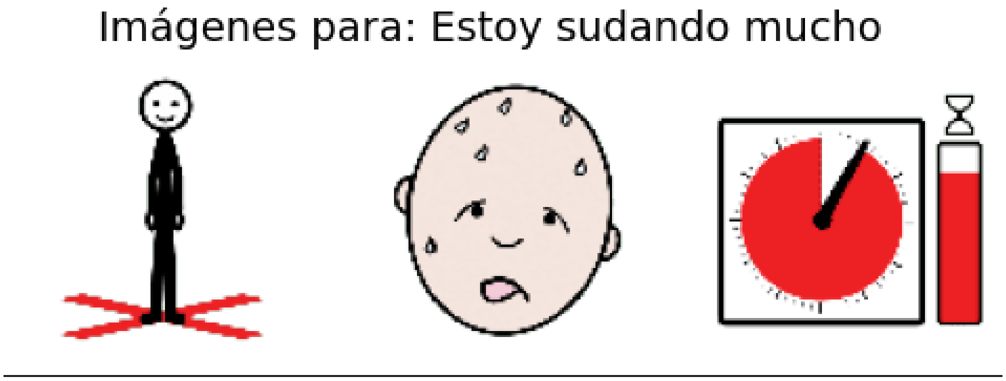


Figura 5: First Result of our implementation , for the phrase *Estoy sudando mucho*

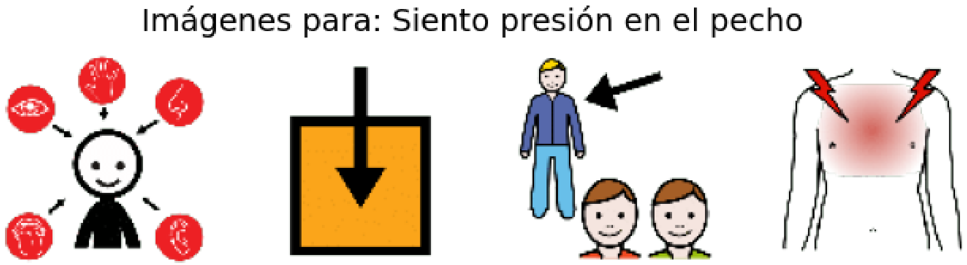


Figura 6: First Result of our implementation , for the phrase *Siento una presion en el pecho*

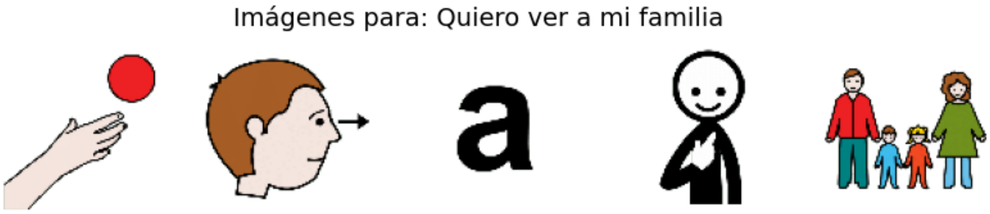


Figura 7: First Result of our implementation , for the phrase *Quiero ver a mi familia*

Comparison with Existing Methods

This methodology represents one of the first implementations designed to create a corpus of sentences paired with pictograms in Spanish. As such, there are no directly comparable methods in the existing literature. Unlike traditional approaches that rely on manual assignment of pictograms, this pipeline automates the process, making it scalable and less subjective. While previous works have focused on English or Portuguese datasets, the novelty of this project lies in its focus on Spanish and the integration of advanced multimodal models for context-aware pictogram selection.

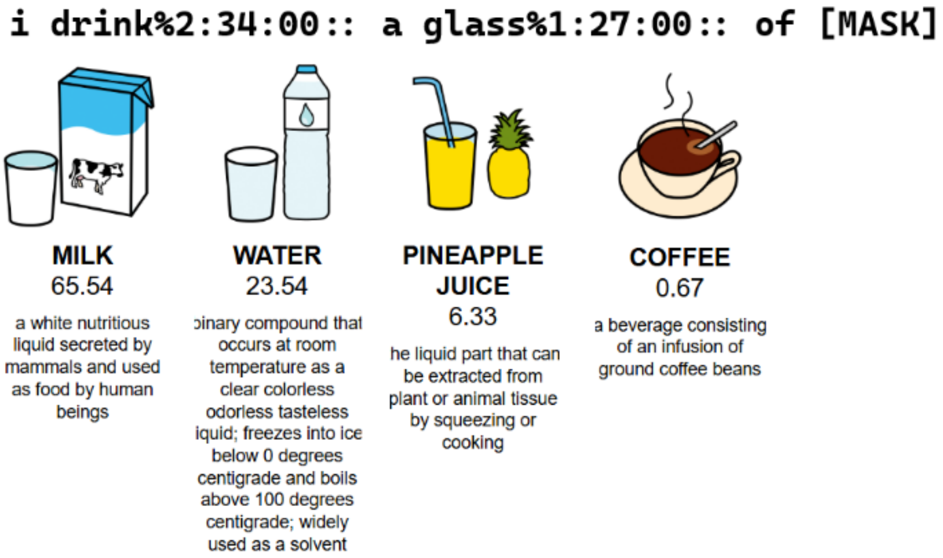
Results and Analysis: Fine-Tuning the Pre-Trained Model

The fine-tuning of the PictoBERT model using the ARASAAC dataset in Spanish achieved promising results, particularly in adapting a model originally designed for English and Portuguese to a new linguistic and cultural context. Key observations include:

- Improved accuracy in associating Spanish phrases with relevant pictograms, measured through quantitative metrics such as BLEU and F1 scores.
- Enhanced performance in understanding idiomatic expressions and regional variations, as evidenced by qualitative evaluations.
- A reduction in prediction errors when compared to the base model prior to fine-tuning.

However, challenges included:

- The need for extensive computational resources during the fine-tuning process.
- Limited coverage of certain semantic domains in the ARASAAC dataset, which impacted the model's generalization capabilities.



```
i drink%2:34:00:: a glass%1:27:00:: of [MASK]
```

MILK
65.54
a white nutritious liquid secreted by mammals and used as food by human beings

WATER
23.54
inary compound that occurs at room temperature as a clear colorless odorless tasteless iquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent

PINEAPPLE JUICE
6.33
he liquid part that can be extracted from plant or animal tissue by squeezing or cooking

COFFEE
0.67
a beverage consisting of an infusion of ground coffee beans

Figura 8: First Result of our implementation , for the phrase *i drink a glass of X (Yo bebo un vaso de X)*

Comparison with Existing Methods

Fine-tuning PictoBERT with ARASAAC data represents a novel contribution to the field. Previous work has primarily focused on general-purpose language models or datasets in English, whereas this project demonstrates the feasibility and benefits of adapting such models for Spanish-language applications. This approach aligns with recent trends in multimodal AI but extends their applicability to underserved linguistic and cultural contexts.

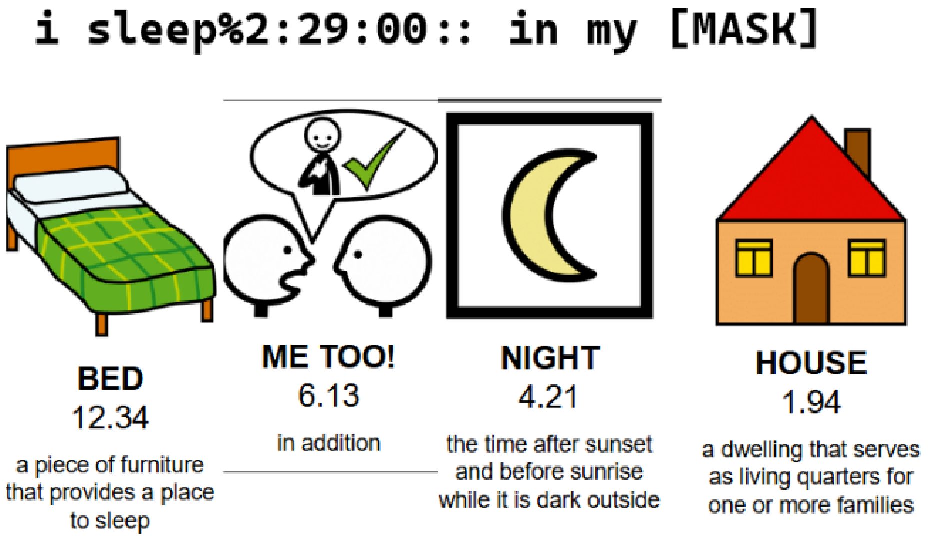

i sleep%2:29:00:: in my [MASK]
BED
12.34
a piece of furniture that provides a place to sleep
ME TOO!
6.13
in addition
NIGHT
4.21
the time after sunset and before sunrise while it is dark outside
HOUSE
1.94
a dwelling that serves as living quarters for one or more families

Figura 9: First Result of our implementation , for the phrase *i sleep in my (Yo duermo en mi X)*

Discussion of Implications and Future Directions

The results of this project have several implications for the development of tools in AAC and related fields:

- The corpus of sentence-pictogram pairs provides a foundation for training and evaluating new models or enhancing existing ones.
- The fine-tuned model can serve as a resource for improving communication tools for Spanish-speaking individuals with disabilities.
- The integration of multimodal language models highlights the potential of combining textual and visual data for improved semantic understanding.

Future work could address the identified limitations by:

- Expanding the ARASAAC dataset to cover additional semantic domains and regional variations.
- Incorporating active learning techniques to refine the model further with user feedback.
- Exploring multilingual approaches to extend the applicability of the pipeline and model to other languages.

Conclusion

This project presented two significant contributions to the field of augmentative and alternative communication (AAC) and multimodal artificial intelligence: the creation of a corpus associating sentences in Spanish with pictograms and the fine-tuning of a pre-trained model using the ARASAAC dataset. Below, the main findings, their significance, and potential future research directions are summarized.

Main Findings

- A pipeline was successfully designed and implemented to create a corpus of Spanish sentences paired with pictograms. This corpus is one of the first of its kind and provides a structured resource for applications in AAC and related fields.
- The fine-tuning of the PictoBERT model demonstrated the feasibility of adapting pre-trained multimodal models for Spanish-language datasets. The model achieved improved accuracy in associating sentences with pictograms and showed significant advancements in handling idiomatic expressions and cultural nuances.
- Both components highlight the importance of integrating multimodal reasoning and context-aware processing in advancing communication tools for underserved linguistic communities.

Significance of the Work

This work addresses a gap in AAC research for Spanish-speaking users, providing tools and resources tailored to their linguistic and cultural needs. By automating the creation of a pictogram-based corpus and fine-tuning a multimodal model, the project not only contributes to the state of the art but also lays the groundwork for practical applications in education, healthcare, and accessibility technologies.

Future Research Directions

Future work could build upon this project by:

- Expanding the ARASAAC dataset to include additional semantic domains and regional language variations, enhancing the generalization capabilities of the models.
- Incorporating user feedback through active learning to further refine the corpus and model performance.
- Exploring multilingual and cross-cultural approaches to extend the pipeline and fine-tuned models to other languages and communities.
- Investigating the integration of real-time applications, such as mobile apps or web-based tools, to facilitate the use of these models and resources in everyday scenarios.

Bibliografía

[1] Pereira, J. A., Macêdo, D., Zanchettin, C., de Oliveira, A. L. I., & Fidalgo, R. do N. (2022). PictoBERT: Transformers for next pictogram prediction. *Expert Systems with Applications*, *202*, 117231. https://doi.org/10.1016/j.eswa.2022.117231

[2] Wicke, P., & Cunha, J. M. (2020). An approach for text-to-emoji translation. En *Proceedings of the 11th International Conference on Computational Creativity* (pp. 504–507). Association for Computational Creativity. https://cdv.dei.uc.pt/wp-content/uploads/2020/10/wicke2020translation.pdf

[3] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

[4] Alencar Pereira, J., Macêdo, D., Zanchettin, C., & de Oliveira, A. L. I. (2021). AI supporting AAC pictographic symbol adaptations. En *Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). IEEE. https://doi.org/10.1109/MLSP52302.2021.9596220

[5] Gobierno de Aragón. (s.f.). ARASAAC: Portal Aragonés de la Comunicación Aumentativa y Alternativa. Recuperado de https://arasaac.org/

[6] Google. (2024). Gemini API: Documentación de OAuth. Recuperado de https://ai.google.dev/gemini-api/docs/oauth?hl=es-419