Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Towards Domain-Specific Spoken Language Understanding for a Catalan Voice-Controlled Video Game

Alex Peiró-Lilja[1,2], *Rodolfo Zevallos*[1], *Carme Armentano-Oller*[1], *Jose Giraldo*[1], *Cristina España-Bonet* [1,4], *Mireia Farrús*[2,3]

[1]LangTech Lab, Barcelona Supercomputing Center, Spain; [2]CLiC, Universitat de Barcelona, Spain; [3]UBICS, Universitat de Barcelona, Spain; [4]DFKI GmbH, Saarland Informatics Campus, Germany

{alexandre.peiro, rodolfo.zevallos, carme.armentano, jose.giraldo, cristina.espana}@bsc.es, mfarrus@ub.edu

Abstract

We design a voice-controlled video game to integrate Catalan into gaming using speech technologies developed under the Aina[1] project. The game is designed to elicit natural speech commands from players. However, a significant challenge in this endeavor is the limited availability of Catalan-language Spoken Language Understanding (SLU) datasets, especially those covering specialized linguistic domains relevant to interactive gaming environments. To address this, we implement a cascading SLU system that combines automatic speech recognition (ASR) with roBERTa-based models previously trained in Catalan. The latter was finetuned as a multi-task classifier by generating synthetic transcriptions from a small set of human-written examples. With acceptable accuracy and time inference, our goal is to evaluate its performance in-game and gather feedback from users.

Index Terms: spoken language understanding, language model, video gaming, Catalan

1. Introduction

Video games are one of the most immersive forms of entertainment, allowing players to interact with dynamic environments and characters. Significant progress has been made in using speech recognition and large language models to enable near-human interactions. However, these models are often too large, making their integration difficult in small projects that require local, low-cost implementation. The key is to map natural spoken language to specific values that the code can interpret to generate the appropriate interaction, similar to intent classification. However, the major challenge is the lack of Spoken and Natural Language Understanding (SLU-NLU) datasets for specific domains, especially in Catalan. Existing SLU-NLU Catalan datasets are created for everyday use cases: SLURP [1] –localized to Catalan– and NLUCat[2] developed under the Aina project. To address this in our designed video game, we leverage speech recognition and BERT-based models trained in Catalan as a cascading SLU system. By using a high performing ASR for Catalan, only the BERT-based model needed finetuning. Once all possible categories and values were defined, we generated the data by combining a set of human-written templates of simulated game commands. We aim to extend previous Catalan resources for video gaming [2] showing our methodology and models.

2. Video game

The main objective of the game is to place all the decorative pieces in their correct positions on the facade of a chapel before time runs out. The scenario includes natural elements such as trees, windows, a main door, gargoyles, and the scaffolding itself (Figure 1). Two workers —each with an assigned name— are placed on the scaffold, and the player, who has a full view of the scenario, must issue commands to the characters using voice instructions. The player is free to decide how to give the command to the workers, but is expected to use the environmental elements as reference points when specifying positions on the scaffolding. By defining visual reference elements, we narrow down the domain of elements and the vocabulary the player is likely to use. The idea is to allow the player to give instructions naturally—through a "walkie-talkie"—just as they would in real life, using the scene's clear visual elements as reference points (e.g., *Please, Dani, go and carry the shield that is located next to the left tree on the second level*). The speech command must include the character's name. Then one of four actions (move, pick up, carry, or place), the target object, and a location description —either one of nine objects or one of fifteen scaffold positions (a 3×5 matrix)—. Then, the command content is mapped to four categories with fixed value sets (see Table 1).

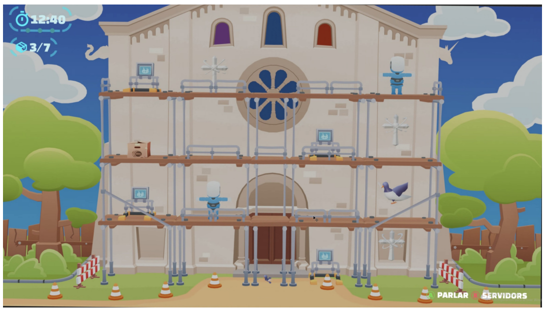

Figure 1: *Video game main scene.*

3. Two-model cascaded SLU

Choosing an end-to-end SLU approach for this problem would involve creating, in addition to texts, synthetic speech clips. Alternatively, we leverage pre-trained and robust Catalan models for a cascade system: an ASR to generate reliable transcriptions and a fine-tuned RoBERTa model as NLU. As part of the Aina project, a Catalan large version of the acoustic model Whisper [3] was finetuned with 710 hours of data for ASR task [4]. We

[1] https://projecteaina.cat/

[2] https://huggingface.co/datasets/projecte-aina/NLUCat

Table 1: *Categories and values*

NAMES	Eva, Dani, Guillem, Andreu, Raquel, Helena, NO_NAME
ACTIONS	to_place, to_move, to_grab, NO_ACTION
OBJECTS	pigeon, moon, cross, shield, sun, salamander, moses, angel, NO_OBJECT
LOCATIONS	A1, A2, A3, A4, A5, B1, B2, B3, B4, B5, C1, C2, C3, C4, C5, NO_LOCATION

converted it to Faster-Whisper large[3] and tested previously in a noisy real-world environment. Predicted transcriptions are processed by the finetuned RoBERTa-ca[4], a continually pretrained RoBERTa-based model [5] with 95 GB of Catalan data using vocabulary adaptation from a multilingual RoBERTa. Despite its size –125M params– its performance is close to XLM RoBERTa Large –561M params– in many downstream tasks.

4. NLU finetuning with generated data

4.1. Generating samples

An internal survey was conducted among the unit's collaborators, asking them to write 30 examples each, combining the various available elements. A screenshot of the scenario was provided for guidance. 13 volunteers participated, although not all completed the 30 examples. From the results of this survey, we collected 94 different templates, which combined the four possible informative categories. Also we extracted all the possible equivalent expressions and used them as synonyms to create lexically rich synthetic sentences. We have created a first batch of 87K labeled sentences from templates. To obtain a more realistic dataset, in 19430 of the examples at least one of these information slots was missing. For each of the informative slots, several possible expressions have been used. Thus, for example, to refer to the object *salamander*, we used expressions such as the hypernym *reptile* or the words referring to similar animals *dragon* and *lizard*. The selection of one or the other equivalent expression in a sentence has been done randomly. The training dataset will expand by incorporating new examples from future human input.

4.2. RoBERTa-ca to NLU

We fine-tuned RoBERTa-ca using a multi-task learning approach. Four classification layers—one per category—were connected to the pooled output, each returning a set of logits corresponding to the number of possible values. The total loss was computed as the sum of the individual cross-entropy losses. After running multiple training experiments, we found that freezing all layers of RoBERTa-ca except the last one resulted in the highest global accuracy.

5. Demo

5.1. SLU

Our SLU, deployed on an Nvidia H100 GPU, achieved an average inference time of 259 ± 113ms for transcription and 6.5 ± 16.3ms for category prediction. Table 2 shows accuracy for each category based on 22 sentences. The model seems robust to repetitions and bad structured sentences, and it is highly accurate in detecting character name. Lower accuracies are expected to prompt players to repeat and vary intonation, as one would when a colleague does not understand and you have to insist.

Table 2: *Accuracy (%) in detecting correct values per category*

	NAME	ACTION	OBJECT	LOCATION
Accuracy	100.0	86.4	59.1	86.4

5.2. Presentation

We will show a development version of the video game. The SLU predictions after giving a speech command will be displayed on-screen while observing the character's action (e.g., moving one of the characters to the left side of the main door). The game was developed with Unity game engine and will run on an MSI Intel Core i7 12th Gen with 64GB of RAM, but can run on a lower-spec device. Interaction will take place via a Corsair HS35 headset with a microphone or a similar model. For versatility, the SLU system will be called via an endpoint, allowing us to test the game's cross-platform compatibility.

6. Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project. Also funded by FairTransNLP-Language (PID2021-124361OB-C33), MICIU/AEI/10.13039/501100011033/FEDER,UE, and 2021 SGR 00313.

7. References

[1] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7252–7262.

[2] A. Peiró-Lilja, J. Giraldo, M. Llopart-Font, C. Armentano-Oller, B. Külebi, and M. Farrús, "Multi-speaker and multi-dialectal catalan tts models for video gaming," in *Interspeech 2024*, 2024, pp. 999–1000.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[4] C. Hernández-Mena, C. Armentano, S. Solito, and B. Külebi, "3catparla: A new open-source corpus of broadcast tv in catalan for automatic speech recognition," in *IberSPEECH 2024*, 11 2024, pp. 176–180.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019, cite arxiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[3] `https://huggingface.co/projecte-aina/faster-whisper-large-v3-ca-3catparla`

[4] `https://huggingface.co/BSC-LT/RoBERTa-ca`