Alma Mater Studiorum - Università di Bologna
Scuola di Scienze

A privacy-preserving AI-based Intent Recognition engine with Probabilistic Spell-Editing for an Italian Smart Home Voice Assistant

Paola Persico

Relatore:
Chiar.mo Prof.
Danilo Montesi

Controrelatore:
Chiar.mo Prof.
Maurizio Gabbrielli

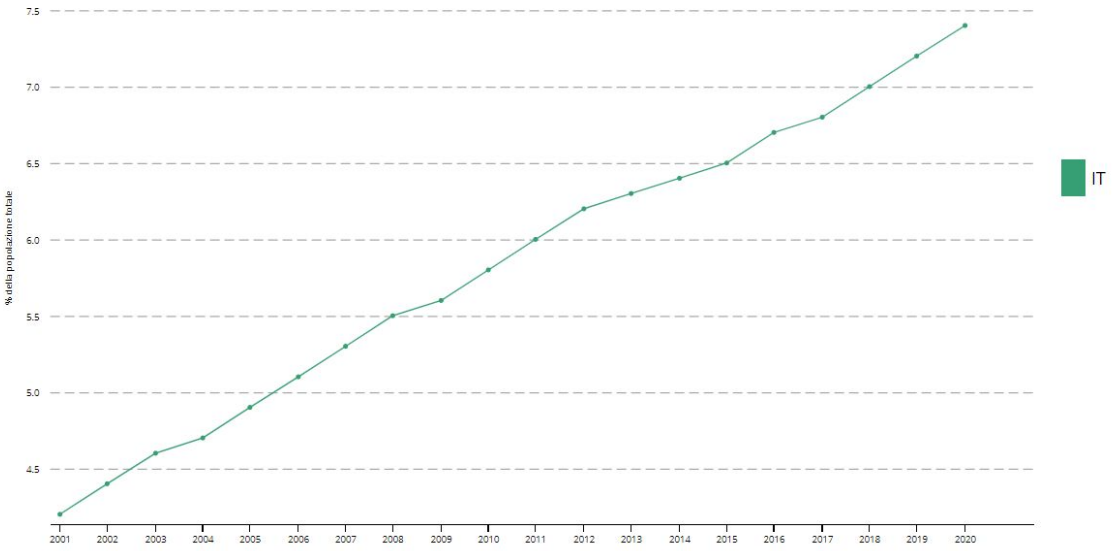
Corso di Laurea Magistrale in Informatica
A.A. 2022/2023 - Sessione II

Summary

- Motivation
- State of the Art of Voice Assistants for a Smart Home
  - Proprietary Solutions
  - Open-Source Solution: Home Assistant
- Our Proposal: Converso
  - Spelling Correction
  - Intent Recognition
- Experimental results
- Conclusions

Motivation

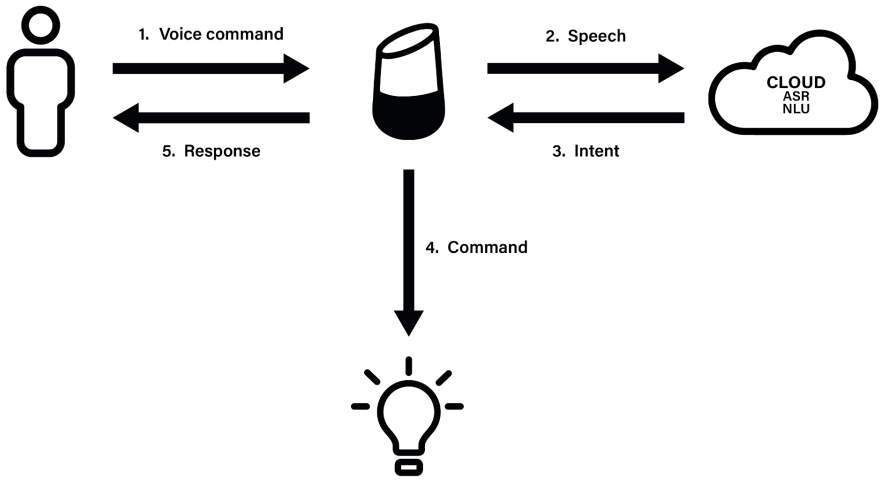
- The market of **smart appliances** is expanding
- Smart home control via **Voice Assistant** is especially useful for:
  - people who temporarily require a hands-free interaction
  - people with disabilities
  - old people



% of people aged > 80 in Italy

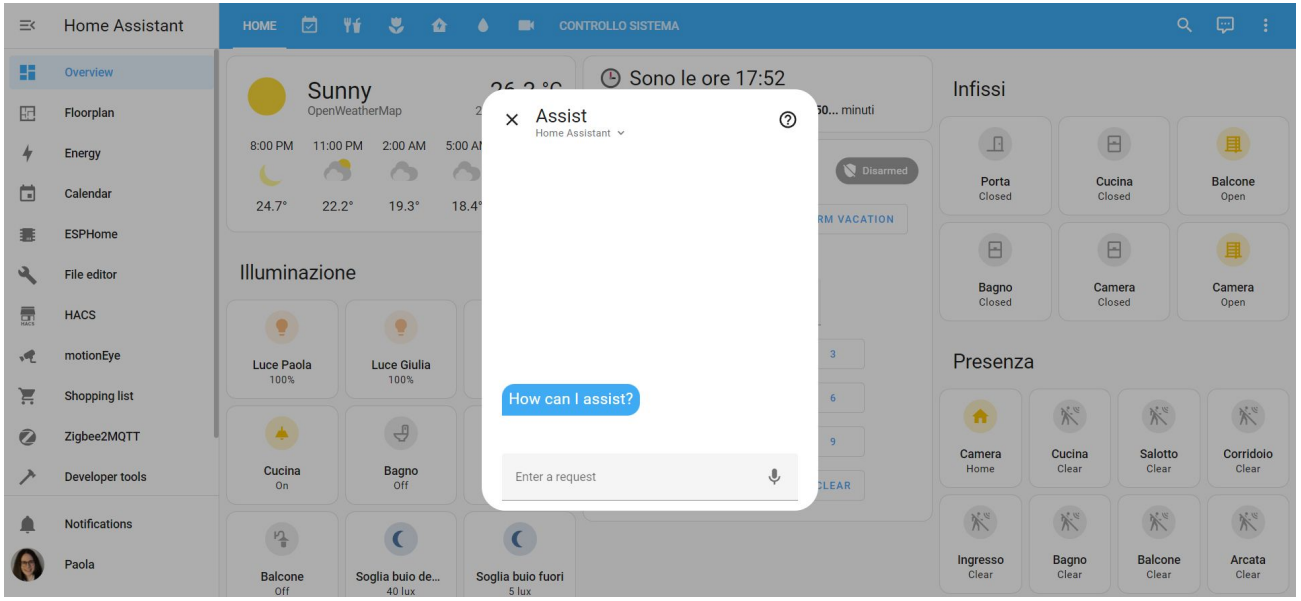
State of the Art: Proprietary Solutions

- Most popular solutions:
  - Amazon Alexa, Google Home, Apple HomeKit
- The **privacy** issue: speakers are always listening
  - audio signals are sent to the *cloud*
  - possible data leakage and abuse of recordings
  - possible accidental triggers



State of the Art: Open-source solution

- **Home Assistant**
  - *“Open source home automation that puts local control and privacy first [...] Perfect to run on a Raspberry Pi or a local server.”*
  - Abstraction: appliances are *entities* with *state* + *attributes*



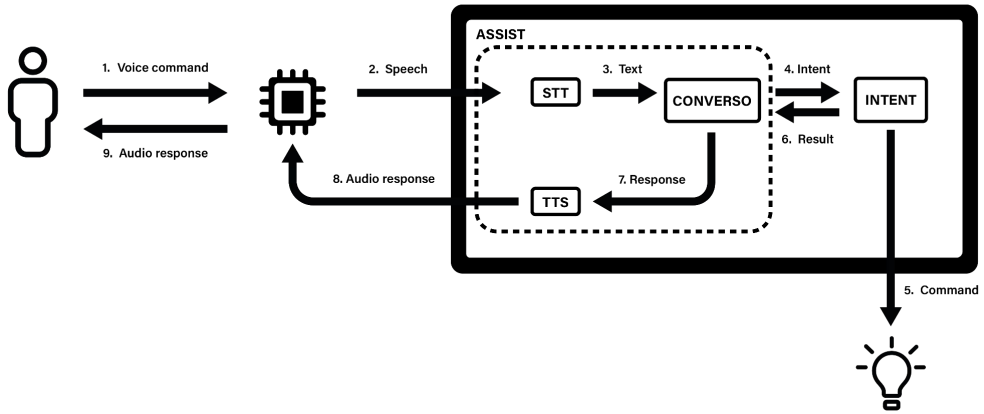
State of the Art: Open-source solution

- **Assist Pipeline**
  - STT (e.g. Whisper)
  - Intent Recognition (e.g. Hassil)
  - TTS (e.g. Piper)
- The **accuracy** issue
  - STT models struggle with the Italian language
  - Intent Recognition via Template Matching is not flexible

```
language: "it"
intents:
  HassTurnOn:
    data:
      - sentences:
          - "<turn_on> [tutte] [l(a|e)] (luc(e|i)) [(<of>|<in>)] <area>"
          - "<turn_on> [<in>] <area> [tutte] [l(a|e)] (luc(e|i))"
        slots:
          domain: light
```

Our proposal: Converso

- A new Intent Recognition engine for a Voice Assistant for the control of a smart home in the Italian language
- **Requirements**:
  - high accuracy (> 60%)
  - low delay (< 5 s)
  - privacy-preservation (no data on cloud)
    - local → low resource consumption



Converso: Spelling Correction

- **Error detection**: unlikely unigrams and bigrams
  - Domain vocabulary (dataset + custom entities)
  - Frequency hash-tables (WaCky corpus)
- **Candidates generation**: Damerau-Levenshtein edit distance
  - 3 stages: increasing edits/vocabulary
- **Candidate selection**:
  - max log likelihood
  - improvement threshold

$$P(w_{1:N}) \approx \sum_{i=1}^{N} log[(1-\lambda)\underbrace{\frac{count(w_{i-1}w_i)}{count(w_{i-1})}}_{P(w_i|w_{i-1})} + \lambda\underbrace{\frac{count(w_i)}{\sum_{j=1}^{N} w_j}}_{P(w_i)}]$$

Converso: Intent Recognition

- **Multi-Class Multi-Label classification**:
  - Intent, Slots (Domain, Device Class, State), Response
- **Synthetic dataset** generated via FB CF Grammar
  - 200+ productions, e.g.:

    Climate[NUM=sg, GEN=m, ART=il] ⟶ 'riscaldamento' | 'condizionatore' | 'termosifone'

  - 42k text commands (230 unique classes)
- **Preprocessing**
  - tokenization → stop-word removal → Word2Vec → scaling
- **Grid search** with k-fold cross-validation
  - Best model: Linear SVM

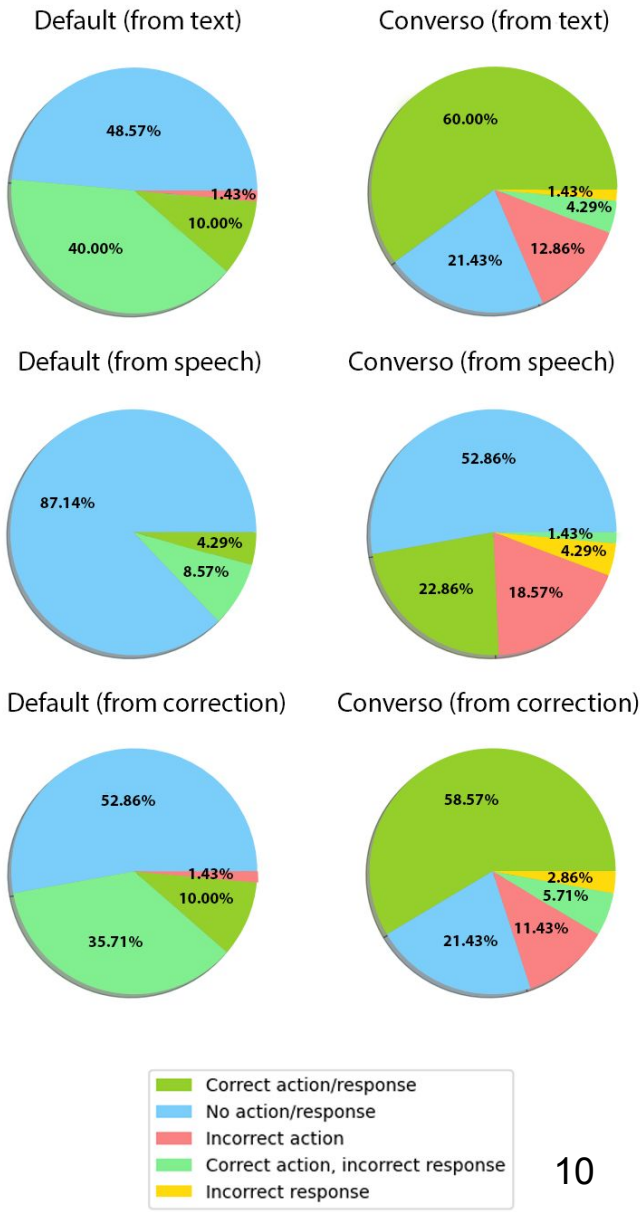
Experimental results

- **Experiment**
  - 10 participants
  - STT: Whisper base-int8
  - ESP-32 satellite
  - 360 commands
- **Results**
  - Spell-editing → lower WER

    “Pegni lucci camera dune”

  - Embeddings → higher flexibility

    “Spegni lampadina”



Conclusion

- Converso is more **privacy-preserving** than market solutions
  - no recording is sent to the Internet
- Converso is more **accurate** than open-source solutions
  - due to spell-editing and embeddings
- Future work
  - generate and gather **more data** to improve accuracy
  - reduction of **delay** of spell-editing
  - add **more intents**

Thank you for your attention