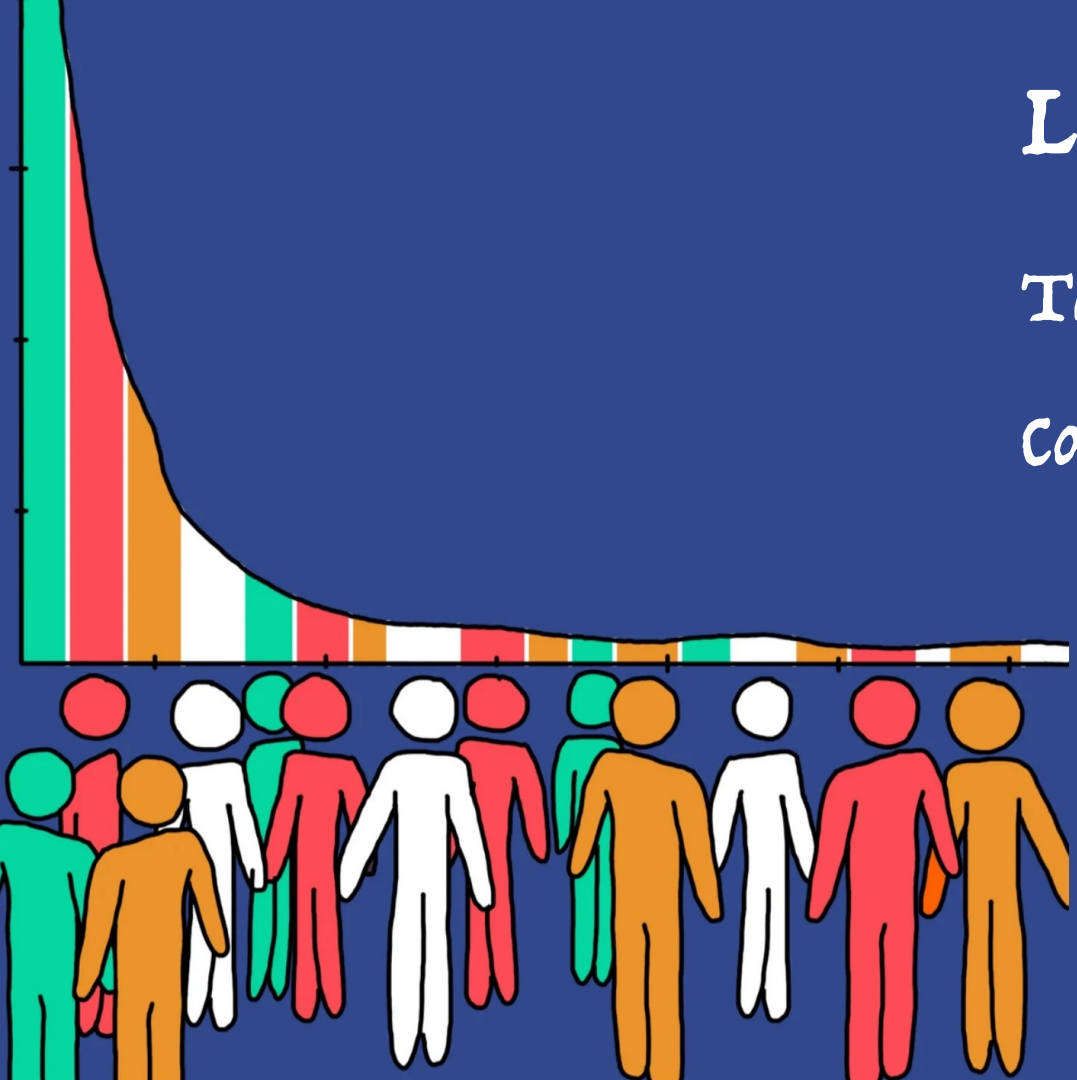
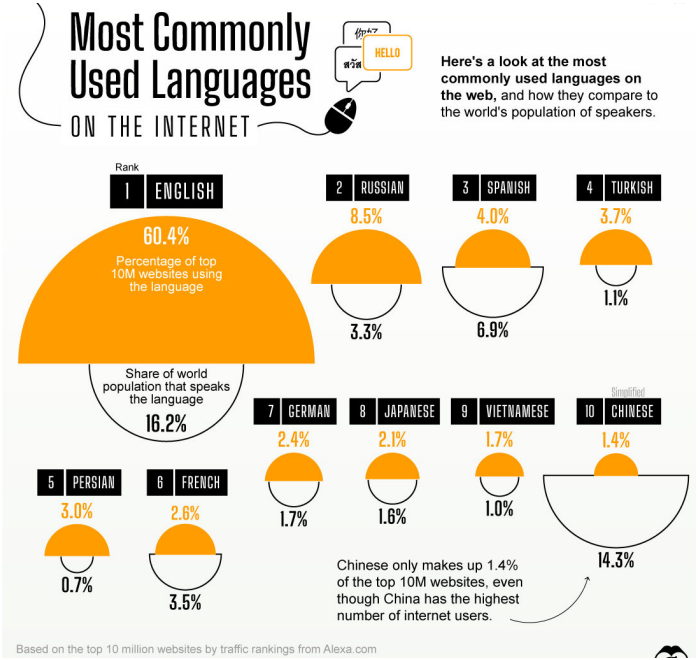


LOWRECORP:

The Low-Resource NLG Corpus Building Challenge

Khyathi Raghavi Chandu (AI2)
David Howcroft
Dimitra Gkatzia
Yi-Ling Chung
Yufang Hou
Chris Emezue
Pawan Rajpoot
Tosin Adewumi

Acknowledge the Gap



Gap in technological advancements:

- 7000 languages in the world
- 10-20 well covered in NLP

Why the gap?

- Speak local languages
- Users mix languages
- Chicken/Egg Problem:
  - Expensive to scale SOTA multilingual models

Towards Equitable Technological Landscape

Introducing LowReCorp

A challenge to collectively gather datasets and resources for LRLs

Towards Equitable Technological Landscape

Introducing LowReCorp

A challenge to collectively gather datasets and resources for LRLs

Empowering Stakeholders: Placing Language Proficiency in Their Hands

Enriching local variants by speakers and people familiar with the languages

Towards Equitable Technological Landscape

Introducing LowReCorp

A challenge to collectively gather datasets and resources for LRLs

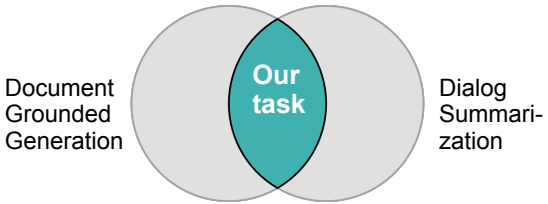
Empowering Stakeholders: Placing Language Proficiency in Their Hands

Enriching local variants by speakers and people familiar with the languages

Language as Identity and Communication

Grounded in local and indigenous sources

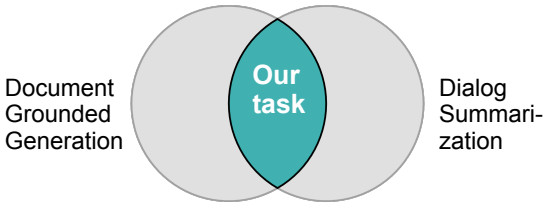
Framework Design (1/2)



Multi-purpose

- **Challenge**: Maintaining qualified annotators for LRLs
- **Ideas**: Intersect multiple tasks like document grounded response generation and dialog summarization

Framework Design (1/2)



Multi-purpose

- **Challenge**: Maintaining qualified annotators for LRLs
- **Ideas**: Intersect multiple tasks like document grounded response generation and dialog summarization

- Different text generation problems in the same dataset (multi-use)
- Same data for multiple tasks
  - opposed to multiple datasets for the same task
- Specific problems:
  - **Summarization** → **Multi-sentence generation**
    - Content relevance
    - Structural coherence
  - **Response Generation** → **Grounding**
    - Content relevance

Framework Design (2/2)

Partial Information Asymmetry and Personalization

- **Challenge**: Complete information symmetry or asymmetry is not natural in real scenarios and often, we want responses to be personalized
- **Ideas**: Design annotation tasks with partial asymmetry and controllable aspects

Framework Design (2/2)

Partial Information Asymmetry and Personalization

- **Challenge**: Complete information symmetry or asymmetry is not natural in real scenarios and often, we want responses to be personalized
- **Ideas**: Design annotation tasks with partial asymmetry and controllable aspects

- **Complete information Symmetry** → Uniformity in knowledge
  - In-domain exchange of knowledge is not information seeking
- **Complete information Asymmetry** → Teacher-student relation
  - More monologue and little conversation
- **Partial Symmetry**:
  - Participant A knows some parts of information
  - Participant B knows overlapping yet different parts of the information

Setup:

- **Set Up:**
  - Participant A (Responder): access to document/image
  - Participant B (Questioner) : access to grounded content/keywords
    - Cognitive effort & Context
  - Sub-headings & Keywords are anchor points
    - 4 sub-headings; 3-4 keywords in each

Setup:

- **Set Up:**
  - Participant A (Responder): access to document/image
  - Participant B (Questioner) : access to grounded content/keywords
    - Cognitive effort & Context
  - Sub-headings & Keywords are anchor points
    - 4 sub-headings; 3-4 keywords in each
- Context for keywords
  - May not be needed for regional topics
  - Consistency with multiple languages data

Setup:

- **Set Up:**
  - Participant A (Responder): access to document/image
  - Participant B (Questioner) : access to grounded content/keywords
    - Cognitive effort & Context
  - Sub-headings & Keywords are anchor points
    - 4 sub-headings; 3-4 keywords in each
- Context for keywords
  - May not be needed for regional topics
  - Consistency with multiple languages data
- Fine-grained control of entities discussed for evaluation

Unimodal Version

Information to the RESPONDER (document content for each subtopic)

Early Life:

Cristiano Ronaldo dos Santos Aveiro was born in Funchal, Madeira Islands to Maria Dolores dos….. Ronaldo was diagnosed with a racing heart (Tachycardia) at age 15,....

Personal Life:

Ronaldo's father, José Aveiro, died of liver disease at age 52 in September 2005. Ronaldo was 20 years old at the time… first fashion boutique under the name "CR7"

….

Information to the QUESTIONER (keywords and their brief introduction)

Early Life:

- **Funchal, Madeira Islands:** Madeira is a Portuguese archipelago that lies between 32°22.3′N 16°16.5′W …..
- **racing heart (Tachycardia):** Tachycardia is the fast beating of the heart.

……

Personal Life:

- **liver disease:** Liver disease is any disease affecting the liver.
- **fashion boutique:** A boutique is a small shopping outlet, especially one

……

Example conversation and Summary on "Early Life" sub-topic

Where were they born ?

Funchal, Madeira Islands

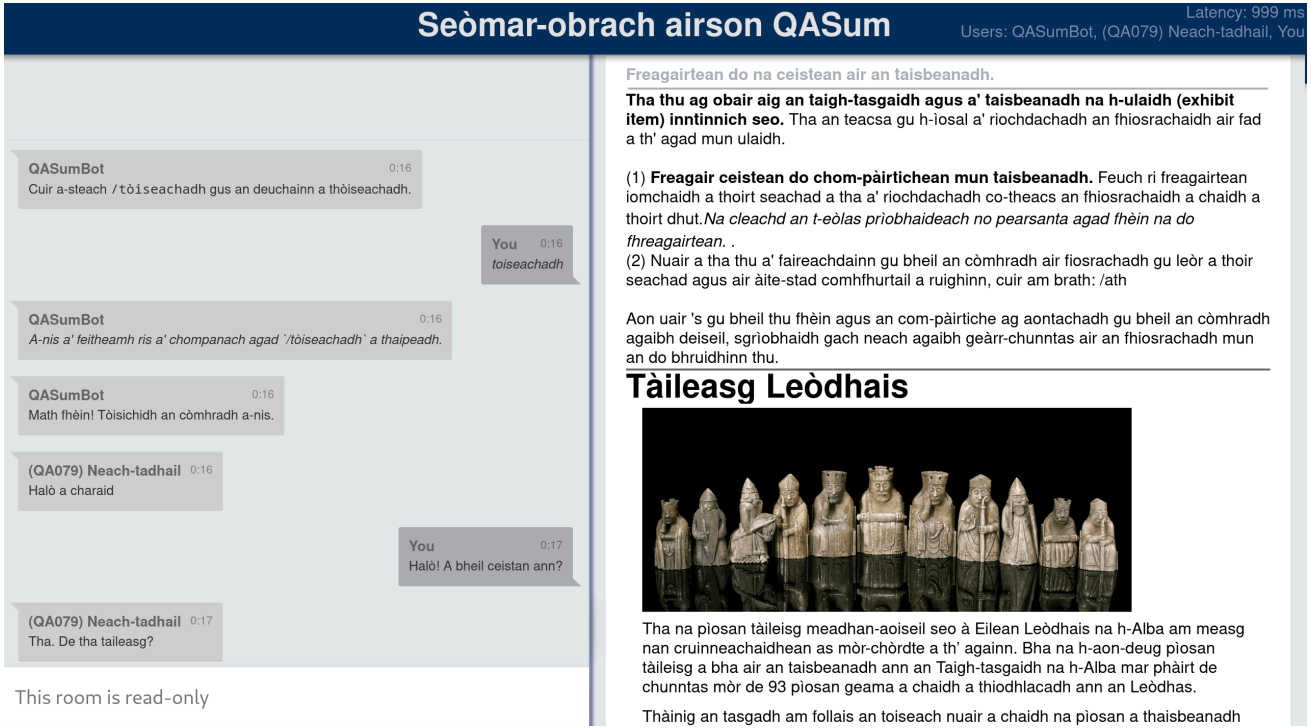
Who were their family members ?

Their mother and father were Maria Dolores dos Santos and José Dinis Aveiro. They had three siblings, one brother Hugo and two sisters, Katia and Elma.

Summary

X was born in Funchal, Madeira Islands. X's parents were Maria Dolores dos Santos and José Dinis Aveiro. They had three siblings, one brother Hugo and two sisters, Katia and Elma. X was diagnosed with a racing heart (Tachycardia) at age 15 but managed to get surgery to treat the same. ...

Multimodal Version



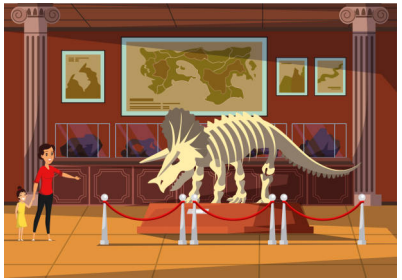
Building a dual dataset of text- and image-grounded conversations and summarisation in Gàidhlig (Scottish Gaelic)

Examples of socially beneficial use cases

Social relevance in building these tools

Example 1:

Interacting about art pieces in Museums



Example 2:

Simulate teaching languages

Implementation Strategies

Implementation Strategy	Technological Access/ Literacy	Data Volume	Quality Control

Implementation Strategies

Implementation Strategy	Technological Access/ Literacy	Data Volume	Quality Control
Online across network	High	High	Low

Implementation Strategies

Implementation Strategy	Technological Access/ Literacy	Data Volume	Quality Control
Online across network	High	High	Low
In lab or field	Moderate	Moderate	High

Implementation Strategies

Implementation Strategy	Technological Access/ Literacy	Data Volume	Quality Control
Online across network	High	High	Low
In lab or field	Moderate	Moderate	High
Offline in the lab	Can be low	Moderate	High

Submission Material ~ Assessment

Assessment:

- Submit a paper to special session
- Not a single metric-based to encourage open-ended creativity.
- Open evaluations for quality control.

Contact us at **lowrecorp@googlegroups.com**

Submission Material ~ Assessment

Assessment:

- Submit a paper to special session
- Not a single metric-based to encourage open-ended creativity.
- Open evaluations for quality control.

Metric or Corpus Property
Grounding material size, complexity, diversity
Conversation length & duration
Lexical diversity (e.g. TTR, bigram TTR)
Corpus & vocabulary size
Lexical and/or syntactic diversity (if possible)
Language typology, geography, community
Creative grounding sources or interface use

Contact us at **lowrecorp@googlegroups.com**

THANK YOU!!

Questions & feedback welcome!!