LOWRECORP: the Low-Resource NLG Corpus Building Challenge

Khyathi Raghavi Chandu[1]   **David Howcroft**[2]   **Dimitra Gkatzia**[2]
Yi-Ling Chung[3]   **Yufang Hou**[4]   **Chris Emezue**[5]   **Pawan Rajpoot**[6]   **Tosin Adewumi**[7]

[1]Allen Institute of AI, [2]Edinburgh Napier University, [3]The Alan Turing Institute,
[4]IBM Research, [5]Technical University of Munich, [6]MUST Research, [7]Luleå University of Technology
khyathic@allenai.org
lowrecorp.github.io

Abstract

Most languages in the world do not have sufficient data available to develop neural-network-based natural language generation (NLG) systems. To alleviate this resource scarcity, we propose a novel challenge for the NLG community: low-resource language corpus development (LOWRECORP). We present an innovative framework to collect a single dataset with dual tasks to maximize the efficiency of data collection efforts and respect language consultant time. Specifically, we focus on a text-chat-based interface for two generation tasks – *conversational response generation* grounded in a source document and/or image and *dialogue summarization* (from the former task). The goal of this shared task is to collectively develop grounded datasets for local and low-resourced languages. To enable data collection, we make available web-based software that can be used to collect these grounded conversations and summaries. Submissions will be assessed for the size, complexity, and diversity of the corpora to ensure quality control of the datasets as well as any enhancements to the interface or novel approaches to grounding conversations.

1 Introduction

Around the world, people speak about 7000 different languages and nearly all of these have very weak support in language technologies. While about 100 languages are included in recent large language models (e.g. Xue et al., 2021; Devlin et al., 2019), most languages do not have good resources. The situation is especially dire when we examine task-specific datasets, such as for response generation, summarisation, and other forms of natural language generation (NLG).

To address this problem, we propose a new shared task on dataset creation for NLG: LOWRECORP[1] challenge, which invites participants to collect a new dataset combining dialogue

---

[1] pronounced as <low> <re> /loʊ ri/, or <Lowry> /laʊ ri/ followed by <corp> /kɔɹp/.
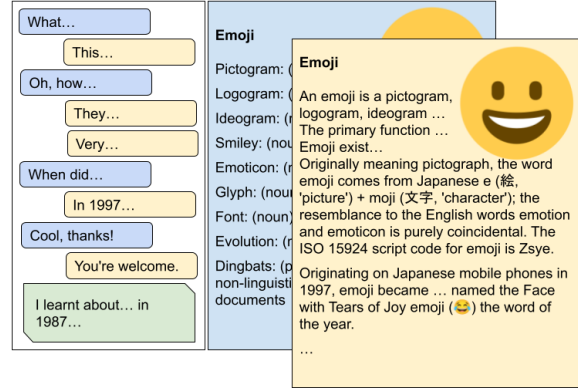

Figure 1: Schematic of the data collection interface. Participants ask questions about a topic (blue, left text bubbles) and answer them (yellow, right), each with access to the same title & image but different grounding text (keyword definitions for the QUESTIONER, a full article for the RESPONDER). After each conversation, each summarises the discussion (green, under chat).

grounded in documents and/or images and dialogue summarisation for a low-resource language (LRL) of their choice. Although conversational response generation and summarization have distinct objectives, they both aim to generate coherent output based on a given context. Drawing from this symbiosis, a new task is proposed that combines the challenges of both tasks into a single framework. The first task is **grounded response generation** and the second is **dialogue summarization**. *Grounded response generation* is the task of generating a conversational response grounded in a context such as documents, images, or other modality, to provide more contextually relevant information (Reddy et al., 2019; Kim et al., 2020; Wu et al., 2021). Similarly, *dialogue summarization* aims to generate a summary of a conversation while preserving its main ideas, and it is particularly useful in scenarios such as meeting notes and doctor-patient conversations where there may be a lot of irrelevant exchange that can obscure informative segments, making the task challenging for traditional approaches (Ghosal et al., 2022).

To facilitate this challenge, we developed a new interface for slurk (Götze et al., 2022) which allows paired participants to view different grounding materials such as text, vision, audio, etc., relating to the same topics to engage in a conversation and dialogue summarization task (Figures 1 and 2). Combining these tasks enables the efficient use of participant time and energy, which is especially important when working with LRLs which often have small communities of use, and will serve broader research efforts in linguistic analysis, language documentation, and preservation efforts (Anastasopoulos, 2019). In contrast to mitigation strategies such as data augmentation and multilingual language models (Hedderich et al., 2021; Chandu and Geramifard, 2022), our dual-task design focuses on efficiency during corpus creation. To our knowledge, this is the first work combining both grounded response generation and dialogue summarization to maximize benefits for LRLs data collection.

2 Task Description

Engaging human participants to build or annotate corpora typically takes one of two forms. In the first, bulk annotations[2] for a single task are collected via crowdsourcing, but this is challenging to replicate in LRL settings as recruiting, training, and maintaining qualified annotators is more difficult. In the second, annotation tasks are built incrementally, which requires recruiting multiple pools of participants or having the same participants return for multiple tasks – for example multiple sessions to collect annotations for retrieval, localization, and comparison (e.g. Hessel et al., 2022). In contrast, we propose a third approach that combines two tasks within the same experimental session to alleviate challenges in recruiting proficient speakers of low-resourced target languages. The two tasks are:

Task 1: Grounded Response Generation Given an image or document and the conversation history as context, respond to the previous utterance.

Task 2: (Dialogue) Summarization Given the full grounded conversation, summarize the important points with the goal to maximize reconstructing the original grounding context.

These 2 tasks of short and long form generation complement each other well, for instance, using

---

[2]Throughout, we will use 'annotations' equally to refer to annotations on or based upon existing texts as well as the creation of new texts for a corpus and other tasks performed by participants in the process of creating and annotating a corpus.

summarization to identify argument facets in a dialogue (Misra et al., 2015) or dialogue acts for summarization (Goo and Chen, 2018). Note that there can be more such pairs of complementary generation tasks.

2.1 Setup

Each instance of data collection involves a context (image/text), keywords with brief descriptions, and 2 participants. We pair participants in the roles of QUESTIONER and RESPONDER. The RESPONDER is given access to a full context, which can be an image, a document, or both. The QUESTIONER is given access to a list of keywords related to the context in order to familiarize the participants so they can ask meaningful and relevant questions. This partially serves the function of world knowledge, as though a participant knows something about the topic but not a lot. Note that the descriptions and the keywords can be in any chosen language.

Both participants are expected to familiarize themselves with their context (the full context for the RESPONDER and the keyword context for the QUESTIONER) at the start of a session. The QUESTIONER then needs to ask questions to learn more about the topic, and the RESPONDER answers those questions *based on the information provided to them in their context*. After the conversation is complete, both participants write a brief summary of the topic discussed, based on their conversation. Figure 2 depicts how our chat interface is used by the QUESTIONER and the RESPONDER to converse (left) about the context (right). An example of task setup is provided in Appendix A.

3 Implementation Strategies

The proposed dual-task design provides a framework for efficiently collecting complementary datasets. Noting that many large language models today are trained on a substantial proportion of the data found online in any given language and that these models will likely be used as components in future systems trained on the datasets our participants create, we encourage participants to focus on sources of grounding material that are less likely to be in the training data for such models already (e.g. information about museum exhibits, children's stories, etc.). We invite participants to leverage this framework to gather datasets in indigenous and local languages grounded on topics of local interest, deciding exactly *what* data to collect and *how* to
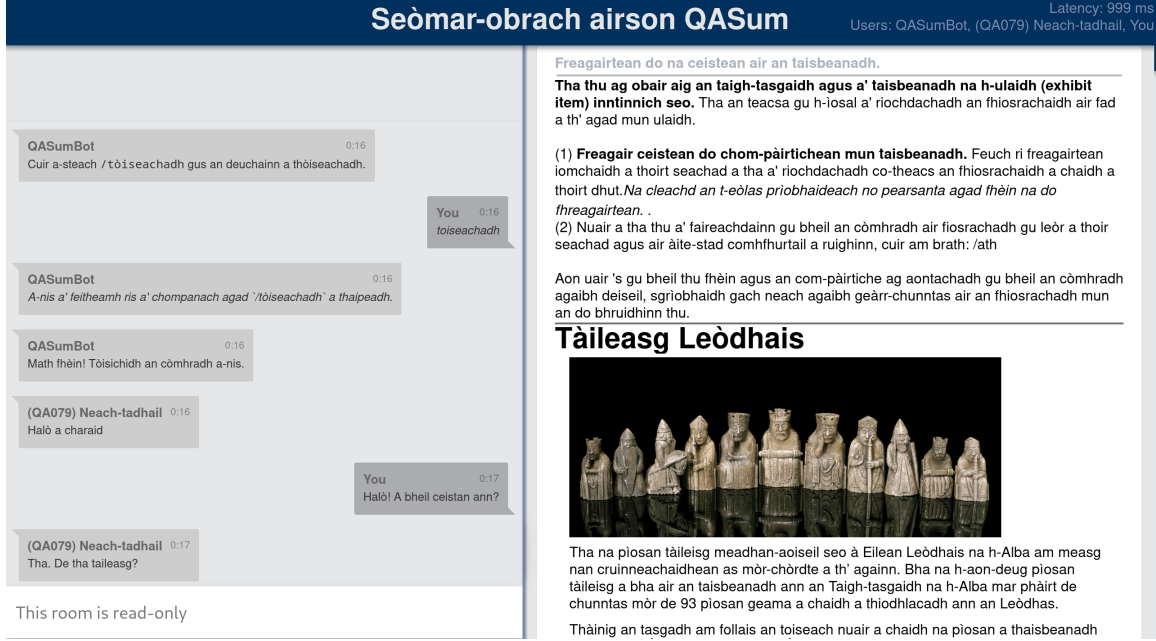
Figure 2: A screenshot of the chat interface used for data collection in Gaelic (Howcroft and Gkatzia, 2023). The chat area is on the left, and the summary of the instructions for the RESPONDER and the beginning of an entry to be discussed on the right.

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

Table 1: Tradeoffs of different implementation strategies

collect it based on the considerations of the target language and its community. We foresee at least three possible approaches to data collection, highlighted in Table 1 along with their tradeoffs.

Online across the network: In this approach, researchers use a webserver to deliver contextual materials along with an audio, video, or text chat interface. This allows researchers to potentially run multiple pairs of participants simultaneously and provides digital representations of the resulting dialogues and summaries from the start. However, this method relies on linguistic literacy to provide instructions (and potentially for data collection, if using a text-based chat interface) and on technological literacy, since participants must be familiar with using a computer, tablet, or smartphone and feel comfortable interacting with the software.

Computer-based in the lab or field: With this approach, researchers are no longer limited to participants with access to technology at home and can be on-hand to answer questions from participants about the interface or troubleshoot any problems.Researchers can use the same kind of software that they would normally deploy online, hosted on

a single laptop. In cases where the aim is to collect spoken dialogues, this also allows the researcher to control the quality of the audio equipment instead of relying on participants to use their own hardware. This method also ensures that the data is immediately available in a digital format and may help address management and/or privacy concerns since data will not need to be transmitted over the internet. This approach requires access to a lab or another controlled space and incurs travel costs either for the participants (to come to a lab of the researcher's choosing) or the researchers (to meet the participants where they are).

Offline in the field: Developing corpora for LRLs can be limited by factors such as participant literacy (Hirmer et al., 2021), lack of availability of technology, systemic societal issues (Ahia et al., 2021), and unrepresentative user bases for crowdsourcing. Therefore, it is possible to implement this dialogue+summarization task fully offline, using in-person methods. Researchers can provide contextual materials (keyword explanations and source materials) to participants on printed sheets of paper and use a microphone to record conversa-

tions and summaries. This method may be most helpful when working with participants with very low technological literacy or in communities where electricity or connectivity is limited. In order for participants to use the source materials to answer questions, however, they will need to be literate or to have materials provided to them in a visual format. This approach will also generally require transcription in addition to the kind of normalization and data cleaning required by the other approaches.

Recruitment Considerations Recruiting participants for LRLs for co-designing, corpus collection, and system evaluation can be challenging due to the small number of speakers. For example, Scottish Gaelic (*Gàidhlig*) has about $57,000$ speakers living in Scotland, [3] implying the difficulty even for a wealthy country with good internet access. In such cases, it can be helpful to collaborate with researchers and institutions that are already involved with the target community. These contacts provide access to community leaders and information exchange structures like community centers and newspapers to promote experiments to potential participants. Being able to reach audiences using spoken or signed media may be especially crucial for LRLs with lower literacy rates (Wu, 2014).

4 Submissions and Assessment

Submission Details More details about the task and the interface are available at `lowrecorp.github.io`. Researchers interested in participating in the challenge can contact us at `lowrecorp@googlegroups.com`, where participating teams can interact and receive updates and support from the organizers. Researchers will upload their data in May 2024 to allow sufficient time for reporting at INLG 2024. Each team is expected to submit a paper to a special session that discusses their chosen subject, target language, any innovations in their approach, and key corpus statistics along with a data card (Gebru et al., 2021).

Assessment A strict evaluation of the diverse submissions expected for this challenge would limit the creativity and flexibility of the teams. Hence, we intend to focus on open evaluations aimed primarily at quality control and description of the data, adapting the data-to-text corpus description guidelines of Perez-Beltrachini and Gardent (2017) to our dialogue and summarisation tasks. For exam-

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

Table 2: Parameters for quality control and evaluation

ple, participants will report the number of different grounding contexts used, the size of those grounding contexts[4], the number of conversations, and the number of summaries. Conversations should be characterized in terms of duration (time; the number of turns) and corpus statistics such as type-token ratio, vocabulary, and corpus sizes. Measures of lexical difficulty or syntactic diversity and complexity available for the language being studied should also be included. Participants should also prepare a datasheet following Gebru et al. (2018).

We will recognize and celebrate submissions based on a variety of dimensions, such as corpus size, lexical diversity, language rarity, most creative grounding source, etc. (cf. Table 2). Submissions involving creative use of local grounding sources especially in areas where technological reach is limited are particularly recognized and commended.

5 Conclusion

This challenge aims to bring together researchers interested in corpus building for LRLs to work on a shared, streamlined, and vetted protocol (tested in Scottish Gaelic) to build interesting corpora. We hope that our challenge will contribute toward recent efforts in addressing geographically diverse NLP (Fan et al., 2021; Nekoto et al., 2020; Aldabe et al., 2022) by aiding in the creation of new datasets for NLG tasks from a wider variety of languages. We aim to encourage cross-pollination of ideas and ideally set up for future generation challenges in a variety of low-resourced languages which make use of the data collected. The resulting (future) task will serve as a text-and-image-grounded complementary task to efforts like the semantic-web-grounded WebNLG challenge[5] and the syntactically-grounded Multilingual Surface Realization challenge (Mille et al., 2020).

---

[4]e.g. number of keywords, length of keyword descriptions, length of full texts for RESPONDER's, image size/complexity
[5]`https://github.com/WebNLG/2023-Challenge`

Ethical implications

While we believe that our proposed dual-task framework can maximize annotation effort, particularly for low-resource languages, data collection should be handled carefully. We highlight in this section several ethical considerations when collecting data in low-resource languages.

Bias The problem of dataset bias is, often, inevitable and can lead to false conclusions and poor generalization of learning methods trained on a given dataset, regardless of modalities (Tommasi et al., 2017). It is important that equal representation is used in the data collection, such as inclusive language, gender, race, and religion (Dhamala et al., 2021). While our approach aims at balancing data availability for low-resource languages by collecting additional data, additional countermeasures would help, for instance, a data statement outlining the data collection process and annotator demographics (Bender and Friedman, 2018). To promote application fairness, researchers are encouraged to quantify dataset bias (Adewumi et al., 2023) and measure the risks of unintended bias.

Privacy While our task is not privacy-demanding, we advocate that the resulting dataset/annotation should adhere to privacy policies such as GDPR data privacy mandates from European Union (Europe, 2019). To reduce privacy risks, several measures should be considered. For instance, when collecting conversations between questioners and answerers in this task, it is preferred that annotators address each other in a way that does not disclose their private information. Private (or personally identifiable) information, such as names and social security numbers, can expose individuals to potential harm and should not be captured in data collection unless absolutely necessary (Sokolova and Matwin, 2016). If sensitive information is collected, anonymization and/or pseudonymization techniques should be applied to protect participants (Terrovitis et al., 2012).

Responsible innovation Responsible innovation, or responsible AI, refers to careful consideration of the potential impacts and benefits of introducing a new product or service. In the context of research in low-resource languages, researchers will need to consider the impact of using online resources that might be copyrighted (e.g., digital media from museum websites). The societal impact will need to be considered such as the privacy of speakers of low-resource languages which might be compromised for instance if a dialect is only spoken by a small number of speakers.

Recruitment and Exploitation When recruiting participants, high priority should be given to first-language (L1) speakers of the languages of interest. This ensures that the data will be representative of how the language is used by its primary language community and fulfills the inclusiveness principle. Researchers should also consider whether proficiency, regardless of L1 or L2 speaker status, is adequate for inclusion in the corpus collection efforts, depending on the goals of the research. It is important that participants are not overloaded with a high volume of keywords and documents per time, as this may affect the quality of the data collected negatively in addition to being an unreasonable amount of work. Adequate compensation should be established, at a minimum adhering to industry standards or regulatory provisions but preferably aiming at providing a 'living wage'. In situations where their contribution is based on a voluntary basis, researchers will need to take extra care to ensure that participants' contributions are freely given and that their needs are respected. One option worth considering for language communities with small numbers of speakers is offering participants the option of being a named contributor to the project, to acknowledge their contribution to the preservation and technological development of their language, which they may appreciate.

Leveraging Translation Although translation is not the primary goal of the task, human translation from a high-resource language or one LRL to multiple LRLs can be used as a collection strategy. While it serves several benefits (Adewumi et al., 2022) such as reducing cost while maintaining the correctness of the task, it suffers from challenges such as entrainment (Mizukami et al., 2016; Chandu et al., 2018). Besides the difficulty of recruitment from possibly a low population, another challenge is the representation of local entities in the target languages. Some ways of solving this challenge include replacing such entities with local ones by using the knowledge of native speakers while keeping in mind that semi-automatic alterations of such technologies might include biases from the high resource languages reducing the naturalness of the data (Chandu et al., 2017).

References

Tosin Adewumi, Mofetoluwa Adeyemi, Aremu Anuoluwapo, Bukola Peters, Happy Buzaaba, Oyerinde Samuel, Amina Mardiyyah Rufai, Benjamin Ajibade, Tajudeen Gwadabe, Mory Moussou Koulibaly Traore, et al. 2022. Afriwoz: Corpus for exploiting cross-lingual transferability for generation of dialogues in low-resource, african languages. *arXiv preprint arXiv:2204.08083*.

Tosin Adewumi, Isabella Södergren, Lama Alkhaled, Sana Sabah Sabry, Foteini Liwicki, and Marcus Liwicki. 2023. Bipol: Multi-axes evaluation of bias with explainability in benchmark datasets. *arXiv preprint arXiv:2301.12139*.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Itziar Aldabe, Jane Dunne, Aritz Farwell, Owen Gallagher, Federico Gaspari, Maria Giagkou, Jan Hajic, Jens Peter Kückens, Teresa Lynn, Georg Rehm, German Rigau, Katrin Marheinecke, Stelios Piperidis, Natalia Resende, Tea Vojtěchová, and Andy Way. 2022. Overview of the ELE project. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 353–354, Ghent, Belgium. European Association for Machine Translation.

Antonios Anastasopoulos. 2019. *Computational Tools for Endangered Language Documentation*. Ph.D. thesis.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Khyathi Raghavi Chandu and Alborz Geramifard. 2022. Multilingual multimodality: A taxonomical survey of datasets, techniques, challenges and opportunities. *CoRR*, abs/2210.16960.

Khyathi Raghavi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Kumar Chinnakotla, Eric Nyberg, and Alan W. Black. 2018. Code-mixed question answering challenge: Crowd-sourcing data and techniques. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 29–38. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Sai Krishna Rallabandi, Sunayana Sitaram, and Alan W. Black. 2017. Speech synthesis for mixed-language navigation instructions. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 57–61. ISCA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *ACM FAccT 2021*.

Interactive Advertising Bureau Europe. 2019. Gdpr transparency and consent framework.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2018. Datasheets for Datasets. In *Proc. of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2022. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). *SIGIR Forum*, 55(2).

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *Proceedings of 7th IEEE Workshop on Spoken Language Technology*.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk interaction server framework: Better data for better dialog models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 558–575. Springer.

Stephanie Hirmer, Alycia Leonard, Josephine Tumwesige, and Costanza Conforti. 2021. Building Representative Corpora from Illiterate Communities: A Reviewof Challenges and Mitigation Strategies for Developing Countries. In *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2176–2189. Association for Computational Linguistics.

David M. Howcroft and Dimitra Gkatzia. 2023. Building a dual dataset of text- and image-grounded conversations and summarisation in gàidhlig (scottish gaelic). In *Proceedings of the 16th International Conference on Natural Language Generation*, Prague, Czech Republic and virtual meeting. Association for Computational Linguistics.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Simon Mille, Anya Belz, Bernd Bohnet, Thiago Castro Ferreira, Yvette Graham, and Leo Wanner. 2020. The third multilingual surface realisation shared task (SR'20): Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, pages 1–20, Barcelona, Spain (Online). Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. 2015. Using summarization to discover argument facets in online ideological dialog. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 430–440. The Association for Computational Linguistics.

Masahiro Mizukami, Koichiro Yoshino, Graham Neubig, David R. Traum, and Satoshi Nakamura. 2016. Analyzing the effect of entrainment on dialogue acts. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 310–318. The Association for Computer Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Laura Perez-Beltrachini and Claire Gardent. 2017. Analysing data-to-text generation benchmarks. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 238–242, Santiago de Compostela, Spain. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marina Sokolova and Stan Matwin. 2016. Personal privacy protection in time of big data. *Challenges in computational statistics and data mining*, pages 365–380.

Manolis Terrovitis, John Liagouris, Nikos Mamoulis, and Spiros Skiadopoulos. 2012. Privacy preservation by disassociation. *arXiv preprint arXiv:1207.0135*.

Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A Deeper Look at Dataset Bias, pages 37–55. Springer International Publishing, Cham.

Bin Wu. 2014. Embedding research in local contexts: local knowledge, stakeholders' participation and fieldwork design. *Field Methods Research Lab Blog*.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski,

Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14085–14093.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A   Example task setup

The roles of QUESTIONER and RESPONDER are designed to converse about the given context. Figure 2 is a screenshot of the developed interface in usage to collect the desired data in Gaelic language. The context here is both image and text. An example set up for text based on Wikipedia is demonstrated in Figures 3 (textual context) and 4 (conversation between the QUESTIONER and RESPONDER).

Figure 3 illustrates how information is presented to the QUESTIONER and RESPONDER in textual mode. The RESPONDER has access to the document and its sub-topics (left), while the QUESTIONER is provided with keywords and brief descriptions for each of them (right). The QUESTIONER's keywords are accompanied by the first sentence of the corresponding Wikipedia page to provide more context but does not have the details contextualized with the main topic at hand. This creates an asymmetry in the amount of information available to the two parties.

The QUESTIONER and RESPONDER discuss the document after studying the information provided to them. An example conversation is shown in Figure 4. The questioner begins the conversation by asking about the keywords starting in the first sub-topic. The order of the keywords within the sub-topic can be randomized but the order of subtopic itself cannot be changed. This is to increase diversity in the data without affecting the inherent flow of the whole topic. The role of contextualization is further enhanced here as in this example, asking for the birthplace is not natural when we look just at the place, however, in the context of the sub-topic of early life, it is possible to guess that the context is about a person and hence the questioner asks about the place of birth. The conversation includes factual and descriptive answers. At the end, the questioner summarizes the sub-topic using the information gathered from the conversation. **Data validation can be performed at any stage by an additional annotator.**

This framework offers several advantages such as producing trustworthy and grounded responses, learning surface form style differences, generating multi-sentence long-form responses, and extensibility to multilingual and cross-lingual scenarios with multilingual data. Finally, this framework also offers considerable flexibility, as it can be adapted based on available technological and linguistic resources (cf. Sec. 3). While the example we have provided here uses text as context, other approaches to grounding participant responses fit within this framework as well. For example, in addition to the keywords and source text seen by the QUESTIONER and RESPONDER, respectively, they can be presented with an image relating to the topic to make the topic more concrete and provide some shared context in addition to the individual materials they have available. Although having the advantage of gathering dual task annotations within the same session is advantageous, this framework also has some limitations that can be improved in future iterations. First, the keywords are only approximately cover the content. In the future, the plan is to explore metrics that evaluate summaries around only the keywords. Second, identifying keywords in multimodal contexts is more complex than in textual contexts.
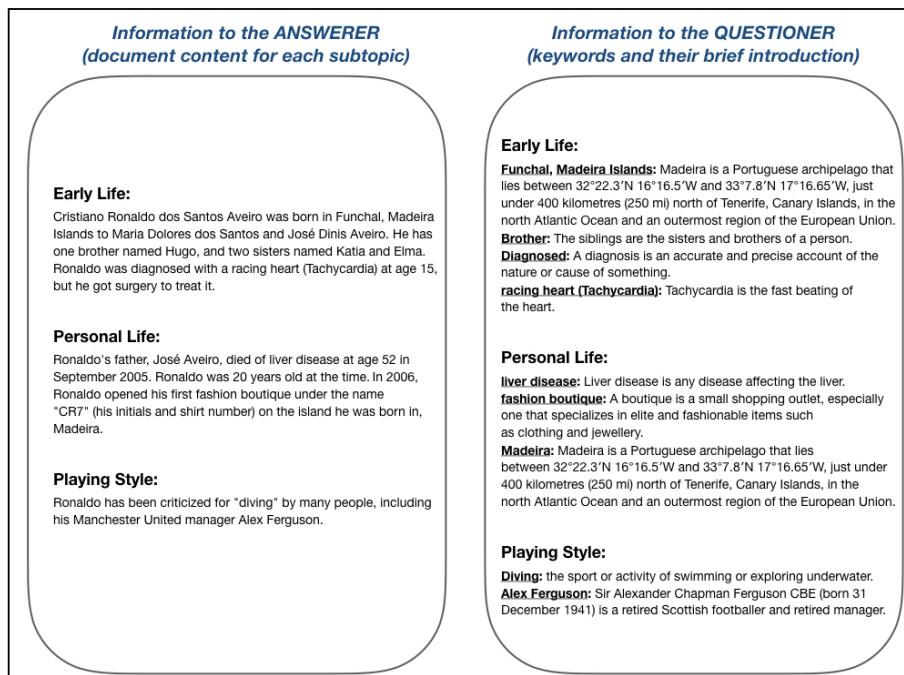
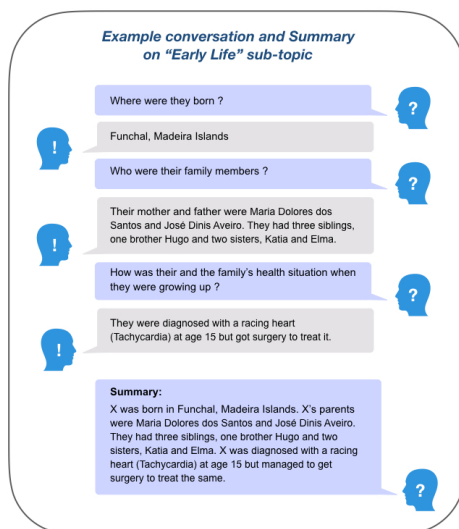Figure 3: The information provided to the Questioner and the Answerer

Figure 4: The conversation between the Questioner and the Answerer on an example sub-topic along with the summary.