UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D1.1 – UTTER Data Management Plan

Nature	DMP	**Work Package**	WP1
Due Date	31/03/2023	**Submission Date**	29/03/2023
Main authors	Barry Haddow (UEDIN)		
Co-authors			
Reviewers	Vlad Niculae, Laurent Besacier, André Martins		
Keywords	data, resources, management		
Version Control			
v1.0	**Status**	Final	2023-03-27

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

Abstract

In this deliverable we describe UTTER's plans for data management. Principally this is about management of linguistic data (corpora) produced during our research, but we also refer to the management of software and models. Through the use of public repositories for the long-term storage of data (in the CLARIN network) as well other repositories (such as those from HuggingFace) which enable easy reuse of data and models, and open-source licences, we will ensure compliance with the FAIR principles of data management.

1 Data Summary

Within UTTER, the main type of data that we will use and generate will be linguistic data, in the form of either *text* or *speech*, and normally collected into a *corpus*. This linguistic data can be collected from the web, in which case we will be mindful of copyright restrictions and fair use exceptions, or it may be recorded/written specifically for the research (for example recordings of conversations and meetings). Data may also be the output of our models, for example translations produced by a machine translation system.

The linguistic data may have some additional structure placed on it, or it may be annotated in some way. Some examples of the types of data we will use/generate are:

- Meeting recordings. We record the audio from all project meetings for use in our meeting assistance research.
- Parallel corpora. This is where we have text (or speech) in two or more languages with an alignment between mutual translations, normally at the sentence/utterance level.
- Annotated corpora. This means a collection of sentences or documents with some human annotation, for example we could have sentences annotated for emotional content.
- Summarisation data. This would be a collection of documents or speeches with accompanied summaries.

In addition to creating novel data sets, we expect to produce new versions or new combinations or selections of existing data sets in the course of our research.

In NLP research, linguistic data is used for exploration and analysis, to help in understanding the problem, for training NLP models and for evaluating those models. The data we produce will be useful for NLP researchers and practitioners (in academia and industry) for the same reasons. It is also of documentary value, making it easier for researchers to understand and reproduce our work.

Textual data used in NLP research is typically stored in (compressed) plain text files, UTF-8 encoded. Simple annotation can use tab-separated values (TSV) format, whilst richer annotation will use XML, and in both cases the annotation and metadata format is project specific. For audio data, we favour the use of uncompressed formats (eg WAV) where possible.

In the first three months of UTTER, we surveyed the existing data sets available for our use-cases (meeting assistance and customer care) in the languages of interest to the project (namely English, German, French, Portugese, Dutch and Korean). This was collected in a spreadsheet, and fulfilled Milestone 2 of the project. The spreadsheet listing the data sets will be used to guide the research efforts in the project, and will be updated as the project progresses. In the UTTER project plan, we have three "releases of UTTER data", at months 12, 24 and 36. The composition of these data releases is not specified, but they will comprise data sets that we use (or intend to use) in our research, and will be a mixture of linguistic data that we collect or annotate specifically for our research, and collections or remixes of existing corpora, that we redistribute (respecting the original licence) for the convenience and benefit of researchers building on our work.

We do not propose a "one size fits all" approach to data management. Each research data set has its own needs, and these needs may be served by more than one repository. The important repositories that we will use for using, publishing and storing data are:

- The CLARIN network (https://www.clarin.eu/). This provides long-term, publicly funded, stable storage for linguistic data, and allows for download by future researchers.
- Shared tasks organised by WMT (Kocmi et al., 2022) and IWSLT (Anastasopoulos et al., 2022). These provide an important source of parallel data, and other data related to translation tasks (for example evaluation). If we co-organise translation-related shared tasks, then we will use these organisations to distribute the data.
- The Opus repository (https://opus.nlpl.eu/). This is the largest repository for freely available parallel texts, and now features an API for easy access, as well as a set of MT models built from the Opus data.
- HuggingFace data sets (https://huggingface.co/docs/datasets/index). This is an extremely useful repository of all types of NLP data, with a convenient web interface (with searchable metadata) and an easy-to-use API. Whilst there is a risk in relying on a single company for publishing data, this repository has widespread usage in the NLP community, and publishing data here is an important way of ensuring impact.
- University datastores. The Universities involved in the project provide facilities for long-term storage of research data (e.g. University of Edinburgh's DataStore[1]). This is a location we would consider for data that needs reliable storage, but is not yet ready for public release, or perhaps cannot be released for privacy reasons.
- Github (https://github.com/). This is a code repository, providing version control, but is often used for data releases if the data set is not too large. The advantage is that it supplies all the usual facilities of a code repository, such as versioning, issue-tracking and commenting, documentation, as well allowing the code for using the data to be stored along with the data. A research paper can be accompanied by a github repository, containing all code created for the paper, data sets used, and system outputs.
- The ACL Anthology (https://aclanthology.org/). This is the publishing venue for the vast majority of NLP papers, and is fully open access, with no embargo. It provides the facility to release or link data sets to research papers.

2 FAIR Data

2.1 Making data findable, including provisions for metadata

For long-term storage and indexing of corpora, we will use the CLARIN network, which provides stable URLs for language resources. This repository will be supplemented by other mechanisms which are more widely used by NLP researchers, to increase the findability of our data. All papers and research reports published by UTTER should contain links to the data sets used and created, since this is often how researchers navigate to data sets. In addition we will make use of well-known repositories like HuggingFace datasets (for annotated NLP data) and Opus (for parallel texts), which are widely used by NLP researchers and practitioners.

[1] https://www.ed.ac.uk/information-services/research-support/research-data-service/during/data-storage

There are no universal standards for metadata in our field, but instead we will follow the standards used by the repositories. For example, resources in CLARIAH (the Dutch node of the CLARIN network) has a fact sheet for each corpus which includes research domain, types of annotations, format, languages covered, and tags. This metadata is indexed and searchable. Other CLARIN nodes have similar metadata requirements. HuggingFace datasets uses the idea of a "Dataset card" for each data set listed which has a variety of fields expressing the data's origins, formats, usage, licensing, limitations etc. These cards are searchable, and indexed using defined sets of keywords.

2.2 Making data accessible

As stated in the previous section, we prefer repositories in the CLARIN network for long-term storage with persistent identifiers, in addition to NLP repositories such as HuggingFace datasets and Opus in order to make data sets easy for researchers and practitioners to use. Both HF datasets and Opus have APIs which make it straightforward to use their corpora in further research and applications, as well as to access the metadata programmatically.

We will share all research data except in the cases where either it conflicts with a partner's commercial interests, or it contains personal identifiable information (PII). The latter would apply, for instance, to meeting recordings that we are collecting in the project for use in research. These can be released as anonymised transcripts, but the recordings and the raw transcripts will have to stay within the consortium. We will always aim to evaluate our research on public data sets. In cases where we need to evaluate on non-public data sets for specfific reasons, we should include additional evaluation on public data sets.

The metadata will be released under CC0, as per the Grant Agreement, and will include access details. Using a CLARIN node for persistent storage means that the metadata will be available as long as the data is available, and in principle this is forever. Our data sets will not generally require specific software for access, since the formats used (typically text, tsv, xml and wav) can be processed with generic tools. However the use of HuggingFace datasets and Opus allows us to provide software support within their open-source tools.

2.3 Making data interoperable

The straightforward nature of the formats used in NLP data means that interoperability at the file format level is not a problem. Essentially, NLP data sets consist of either text or speech, with possibly some annotated labels. If we ensure that the text is UTF-8 encoded, and the speech is in a standard file format, and uncompressed, then interoperability at this level will be unproblematic. If the text is preprocessed in any way (for example tokenised or split to subwords) then it becomes harder to combine with other data sets, or to use in applications which require different preprocessing. In general, we would discourage any type of preprocessing, unless it is necessary for annotation (for example, if the annotation is to be at the token level) in which case the preprocessing steps should be clearly documented.

2.4 Increase data re-use

All data releases should contain a README which will contain information about how the purpose of the data, how it was created, the format of the data, and licensing information. The README should also contain a link to the accompanying paper(s) that describe or use the data set.

For licensing, we prefer the Creative Commons CC-BY licence, which permits any form of reuse (commercial or non-commercial), as long as the user attributes the original source.

3 Other research outputs

The main research outputs (besides data) that we expect to produce in UTTER are software and models. For research papers and other outputs, we will release software and models produced within the research, in order to increase reproduciblity and impact. Our policy will be to release wherever possible (taking into account commercial interests and third-party copyright) and wherever these will be useful to other researchers.

Our software will be released through github[2], where we have created an `utter-project` organisation[3]. To increase findability we will list UTTER software releases on the project website, as well as ensuring that software is linked from the related research paper(s). All software should be released using an open-source licence in order to enable reuse. We prefer the use of a "permissive" licence such as BSD, Apache or MIT. Minimally, software should contain a README with basic installation and usage instructions, together with links to the accompanying research.

If models cannot be shared through github (because of size) they will be released through the UTTER website, as well as being shared through HuggingFace models[4] to enable easy reuse. Models releases should be accompanied with a "model card", with information about the model creation, usage and limitations.

4 Allocation of Resources

We do not envisage any extra resource usage specifically for data management. The preparation of data, software, code and models for release is part of the normal cycle of research so is included with research costs. The long-term storage of data will be provided by the CLARIN network, a public funded network providing long-term storage for linguistic data.

5 Data Security

For internal project data, we will use a secure gitlab instance hosted by the University of Amsterdam. Access security and backup for this instance is provided by the University, for the benefit of projects run by University employees. This repository can be used to host data being prepared for public release, and especially data which contains PII and therefore cannot be released until it is anonymised. For publicly available data, we use the secure, trusted repositories in the CLARIN network.

[2] https://github.com/
[3] https://github.com/utter-project
[4] https://huggingface.co/models

6 Ethics

Some of our data will include voice recordings, for instance we are recording project meetings in order to assist with research on meeting assistance. Since voice recordings are considered to be PII, and project meeting transcripts would contain PII, we need to ensure correct handling of this data to comply with GDPR, including collection of informed consent. This will be addressed in more detail in our annual ethics reviews (D9.1, D9.2 and D9.3).

7 Other Issues

Since UTTER is part-funded by the UKRI, we will adhere to the principles on data management required by that agency[5]. These are in line with the Horizon Europe policy on data management, requiring open release of research data wherever possible, and adherence to the FAIR principles.

[5] https://www.ukri.org/publications/guidance-on-best-practice-in-the-management-of-research-data/

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.10. URL https://aclanthology.org/2022.iwslt-1.10.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.1.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D1.1 UTTER Data Management Plan