Datasheet for the Auslan-Daily

Paper ID: 47

June 11, 2023

1 Datasheets for Auslan-Daily

This datasheet document of Auslan-Daily contains motivation, composition, collection process, recommended uses, and so on. The motivation behind the dataset is the lack of a large-scale dataset for Australian Sign Language (Auslan) translation. Compared to existing sign language translation datasets, Auslan-Daily has two main features: (1) the topics are diverse and signed by multiple signers, and (2) the scenes in our dataset are more complex, *e.g.*, captured in various environments, gesture interference during multi-signers' interactions and various camera positions. Hence, we hope this dataset will contribute to the development of Auslan and the advancement of sign languages worldwide in a broader context.

2 Template

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Considering different geographic regions generally have their own native sign languages, it is valuable to establish corresponding sign language translation datasets to support related communication and research. Auslan, as a sign language specific to Australia, still lacks a dedicated large-scale dataset for sign language translation. The main task of Auslan-Daily is sign language translation (SLT). Meanwhile, it can investigate other sign language-related tasks, including signer detection (SD), fingerspelling detection (FD), sign spotting (SS), isolated sign language recognition (ISLR) and sign language alignment (SLA). To the best of our knowledge, Auslan-Daily is the first publicly available large-scale Auslan translation dataset.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is collected by the UQ-CV

group from the University of Queensland.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The project is funded by Google and Australian Research Council (ARC).

Any other Comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Auslan-Daily is a multi-grained dataset for different sign language-related tasks. The are three different fine-grained visual and language pairs, including (1) video ↔ fingerspelling, (2) video ↔ gloss, and (3) video ↔ sentence.

How many instances are there in total (of each type, if appropriate)?

In Auslan-Daily, there are 2k video ↔ fingerspelling, 3k video ↔ gloss, and 25k video ↔ sentence pairs.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Auslan-Daily covers Auslan sign videos including various topics. It does not involve instances from another existing dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Auslan-Daily is a multi-grained dataset for different sign language-related tasks. The are three different fine-grained visual and language pairs, including (1) video ↔ fingerspelling, (2) video ↔ gloss, and (3) video ↔ sentence.

Is there a label or target associated with each instance? If so, please provide a description.

Yes. In Auslan-Daily, we ask experts to try to align video ↔ fingerspelling, video ↔ gloss, and video ↔ sentence pairs.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All the instances are independent of each other.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Our dataset split is determined for the Sign Language Translation task (refer to Sec. 3.3 for more details). Meanwhile, we will release the recommended data splits used in our experiments for other Auslan sign language-related tasks.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

In the labelling process, despite the annotation by experts, there may be minor errors in the frame boundaries of two consecutive signs. These errors can occur due to the heavy workload and the indistinct time boundaries of continuous sign language.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and we provide the source links.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Auslan-Daily is performed by deaf people and Australian sign language experts.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No. For the signer detection task, we just detect a signer instead of a person's identity.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals

racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

None.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

We collect online Australian Sign Language videos and corresponding subtitles, and obtain video ↔ fingerspelling, video ↔ gloss, and video ↔ sentence pairs after alignment by experts.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We invite Auslan experts to label the sign data. We design the annotation interface ourselves to facilitate the annotation of data with different granularities.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Auslan-Daily is not from a larger set.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

For alignment sign data and sign spotting/finger spelling, Auslan experts are paid at a rate of $ 75 per hour. For labelling performing signers in sign video clips, annotators are the authors.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Our project begins in May 2022 and the current dataset covers “ABC News with Auslan” through May 2023. We will continue to collect more data.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Auslan-Daily is performed by deaf people and Australian sign language experts.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data of Auslan-Daily is sourced exclusively from online. We design an annotation interface for Auslan experts and annotators to label the videos.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Yes.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes. The videos have been taken from online sources with standard licensing and no restrictions to use.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Yes, this is guaranteed by experts

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

In our research, we utilize specific data cleaning methods to filter independent subtitle data (see Sec. 3.2). Additionally, experts manually adjust or delete subtitle data during the annotation process.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, we provide the raw data on our website.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

For sign language alignment, we adopt our designed annotation interface.

Any other comments?

None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No, the dataset is newly proposed by us. The experiments shown in the paper are the only results available for the datasets.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, we provide the link to all related information on our website.

What (other) tasks could the dataset be used for?

Our dataset is versatile, as it is not exclusively used for sign language translation but also caters to an array of other sign language-related tasks.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

We employ Alphapose for human tracking, followed by manual annotation of the signers. This method facilitates the accurate and efficient collection of sign language data in multi-person scenarios.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The usage of this dataset should be limited to the scope of Auslan or sign language-related tasks.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be accessible through our website.

When will the dataset be distributed?

The dataset will be released to the public upon acceptance of this paper. We provide private links for the review process.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

We release our dataset under Creative Commons BY-NC-ND 4.0 license .

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The first author (the name will be released to the public upon acceptance of this paper).

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

E-mail addresses are at the top of the paper. (E-mail addresses will be released to the public upon acceptance of this paper).

Is there an erratum? If so, please provide a link or other access point.

Currently, no. As errors are encountered, future versions of the dataset may be released and updated on our website.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Our project begins in May 2022, and it is ongoing. We are still collecting new Auslan data to enrich Auslan-Daily.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, older versions of the benchmark will be maintained on our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, errors may be submitted to us through email.

Any other comments?

None.