| Paper | Elt | Notes |
| --- | --- | --- |
| [Aligning Subtitles in Sign Language Videos](https://arxiv.org/abs/2105.02877) | Central ideas | * Composition of the data:   + A video of continuous signing   + Subtitles corresponding to the audio content   + Ground truth that temporally aligns the sign and the translation * Method proposed:   + For embeddings: Bert for subtitles and CNN video representation learnt for sign recognition   + Transformer architecture: the two signals interact through attention layers     - * MAIN TASK: aligning a particular known subtitle within a given temporal window [in other words we know what is said in sign language but we want to find it’s correct location in the video aka not a translation task but ALIGNEMENT one]       * Contributions:         + Encoding the subtitle text as input to the alignment model significantly improves the temporal localisation quality         + The use of transformers to execute the alignment task       * Challenges: * The order of subtitles varies between spoken and sign languages * The duration of a subtitle differs considerably between signing and speech * The signing corresponds to a translation of the speech as opposed to a transcription |
| Method | * Problem formulation:   + INPUT:     - Token embeddings of the subtitle text (the elements that we want to align with frames)     - A sequence of video features extracted from a sign language video segment     - Prior estimation of the temporal boundaries (aka the assumed position of the subtitle which is aligned with the audio)       * + OUTPUT: a binary vector of the same length as the video sequence denoting the temporal location of the subtitle   SUBTITLE ALIGNER TRANSFORMER:  Encoder:  Input: text embeddings (in these we already have positional information, no additional positional encoding is used)  Architecture: stack of transformer layers, each containing a multihead attention mechanism followed by a feedforward network  Decoder:  Input: the prior and the video sequence (positional encodings are added to the inputs)  Architecture: stack of transformer layers that attend on the encoded sequence the final layer is linear layer with sigmoid activation  output : T prediction in the range[0,1] **for each frame** values that are above the threshold correspond to the predicted temporal location of the queried subtitle text   * + Pre Training strategy: pretrained on single text units using the available data in other to initialize the weights   + Text Features: Encoded using Bert pretrained on a text corpus for language modeling task then linear projection to match the input dimension of the transformer   + Video Features: extracted from the I3D sign classification model (publicly available) the input is a sequence of length T   + Prior position encoding: subtitle position estimate encoded as binary vector of length T. The prior and the video are fused using concatenation after being linear projected into the same dimension   + THE LOSS: cross entropy between the pred and ground truth     - * DTW dynamic time wrapping:The main issue of this method can be overlap since we don’t take the entirety of the video into account. |
| Other notes | * encoder and decoder architecture: 2 identical Transformer layers with 2 heads and size dmodel = 512 * The prior is the audio aligned text shifted by 3.2 seconds * To avoid overfitting for 50% of the samples we shuffle the words of delete up to two words |
| [Automatic dense annotation of large-vocabulary sign language videos](https://arxiv.org/abs/2208.02802) | Central idea | * The idea is constructing DENSE annotations instead of SPARSE ones. The bigger goal is to improve the annotations of sign language datasets in order to advance research in this area * Dataset : BOBSL BSL (same structure as the one above) * Contributions:   + Improve annotations using synonyms and subtitle-signing alignment   + Pseudo labeling using sign recognition model as a way of sign spotting |
| Method | * Definition: DENSIFICATION |
| [BBC-Oxford British Sign Language Dataset](https://www.robots.ox.ac.uk/~vgg/data/bobsl/)  BASELINE TO REPRODUCE | Model | Dataset:  The majority of episodes are either 30 minutes or 60 minutes in duration  Signers’ identities do no overlap in the train/val/test split  SUBTITLE ALIGNER TRANSFORMER: good news it’s the model from the first paper  Implementation:  The encoder takes as input BERT token embeddings of the text query we wish to align.  The decoder takes as input a sequence of video features from a continuous sign language video segment extracted from an I3D model trained with SIGN-TRAIN-M,D on sign classification and +2.7s S\_audio  First pretrain the SAT : on word-level boundaries from SIGN-TRAIN\_MD where we predict a 1-second interval centered at the automatic mouthing or dictionary sign instance annotation in a randomly chosen 20- second search window around the annotation No prior here  Second: add random shift in order to finetune the sentence alignment plus same data augmentation to avoid overfitting.  Third: finetune the model on sentence-level boundaries from SENT-TRAIN\_H |
| Other notes | * Results differ from those reported in “the first paper”, as we use sentences rather than subtitle texts. |
| [Code](https://github.com/hannahbull/subtitle_align) | The code is very readable |