Topic K: Aligning Text to Sign Language Video

Imane Elbacha Rajae Sebai

MVA 22-23

09/01/2023

Content

1. Introduction
   - Problematic and motivation
   - Training procedure
2. Project milestones
   - Baseline reproduction
   - Context improvement
3. Conclusion

Introduction

Problematic and motivation

The goal of this paper is to perform subtitle alignment for sign language videos (BOBSL dataset) in order to provide a more suitable dataset for future research in this field

Introduction

Model

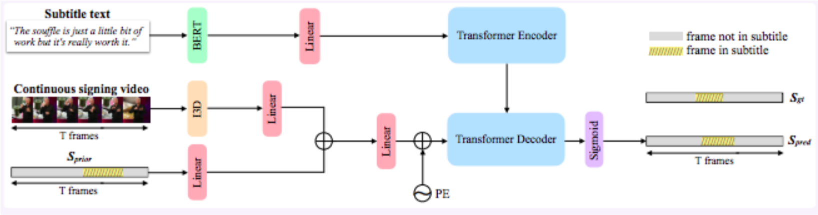


Figure – SAT model overview (Bull et al.)

Introduction

Training procedure

The model constitutes 3 main training steps

1. **Word level pretraining :** Having access to sparse sign annotations from mouthings and dictionary exemplars, the model is trained to spot the single sign occurrence on the entire BOBSL dataset.
2. **Training on audio-aligned subtitles :** the ground-truth is considered to be the coarsely aligned (shifted by +2.7s) audio-subtitles.
3. **Finetune using manually aligned subtitles :** training the SAT on the manually-aligned BOBSL subset.

Introduction

Training procedure

The model constitutes 3 main training steps

1. **Word level pretraining :** Having access to sparse sign annotations from mouthings and dictionary exemplars, the model is trained to spot the single sign occurrence on the entire BOBSL dataset.
2. **Training on audio-aligned subtitles :** the ground-truth is considered to be the coarsely aligned (shifted by +2.7s) audio-subtitles.
3. **Finetune using manually aligned subtitles :** training the SAT on the manually-aligned BOBSL subset.

Introduction

Training procedure

The model constitutes 3 main training steps

1. **Word level pretraining :** Having access to sparse sign annotations from mouthings and dictionary exemplars, the model is trained to spot the single sign occurrence on the entire BOBSL dataset.
2. **Training on audio-aligned subtitles :** the ground-truth is considered to be the coarsely aligned (shifted by +2.7s) audio-subtitles.
3. **Finetune using manually aligned subtitles :** training the SAT on the manually-aligned BOBSL subset.

Introduction

Plan

The structure of the project is :

1. Bibliography analysis
2. Baseline reproduction
3. Context improvement

Introduction

Plan

The structure of the project is :

1. Bibliography analysis
2. Baseline reproduction
3. Context improvement

Introduction

Plan

The structure of the project is :

1. Bibliography analysis
2. Baseline reproduction
3. Context improvement

Project milestones

Baseline reproduction

Method	#train	#val	#test
SAT (Albanie et al.)	1658	32	250
Ours	100	5	50

Table – BOBSL dataset split for pretraining steps (using word spottings and coarsely aligned subtitles).

Method	#train	#val	#test
SAT (Albanie et al.)	16	4	35
Ours	16	4	35

Table – BOBSL dataset split for the fine-tuning step (using manually aligned subtitles).

Project milestones

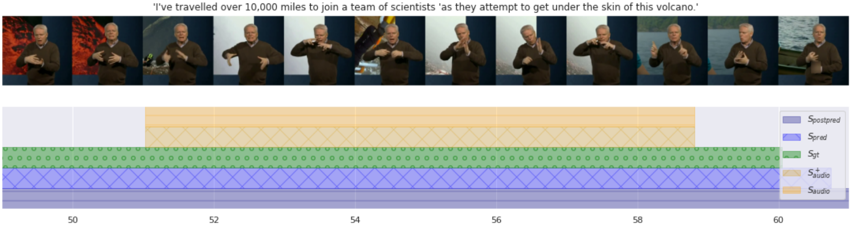
Baseline reproduction

Method	frame-acc	F1@.10	F1@.25	F1@.50
SAT (Bull et al.)	55.62	70.95	61.55	41.46
SAT (Albanie et al.)	70.37	73.33	66.32	53.18
Ours	69.16	72.11	64.54	50,79

Table – Performance results on the manually aligned BOBSL test set.

Project milestones

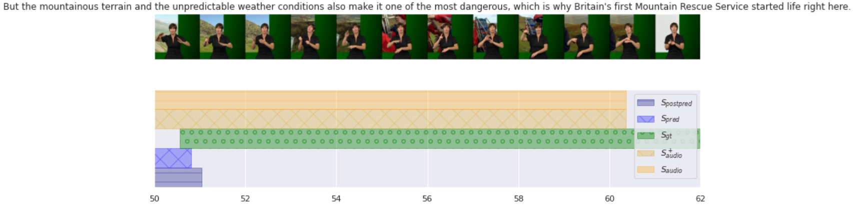
Qualitative results



'I've travelled over 10,000 miles to join a team of scientists 'as they attempt to get under the skin of this volcano.'

Project milestones

Qualitative results



But the mountainous terrain and the unpredictable weather conditions also make it one of the most dangerous, which is why Britain's first Mountain Rescue Service started life right here.

Project milestones

Context improvement strategy

One of the limitations of the training procedure presented by the paper is the loss of context elements in the training process. As an improvement strategy we suggest :

Training the model to achieve the task of aligning 3 subtitles at a time

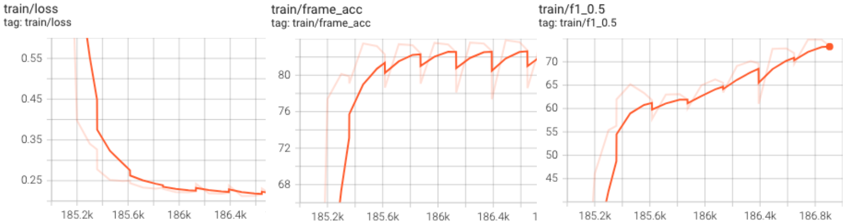
Project milestones

Context improvement strategy

- Adapt the original code to this task : adjusting the Dataloader
- Executing the training procedure
- Comparing the performance scores

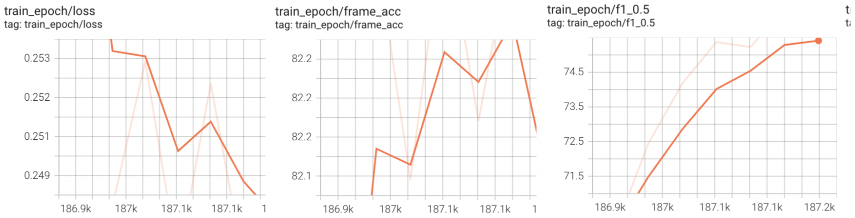
Project milestones

Train coarse subtitles : logs



Project milestones

Finetune model : logs



Conclusion

Difficulties

- Adapting to the computational cost of training on a large dataset (Trained for over 20 hours, all steps included, on a 1 x NVIDIA V100 +200$)
- Manipulating virtual machines
- Conciliate the results on the baselines of two different papers

Conclusion

Improvements for the final report

- Improving the finetuning for the context task
- Tests to examine overfitting
- Learning the BSL would have helped !