Topic K: Presentation

Aligning Text to Sign Language Video

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni and Andrew Zisserman

Sébastien Meyer

January 16, 2023

Overview

1. **The subtitle alignment task**

2. **Available packages**

3. **My modifications**

1. The subtitle alignment task

- **Problem statement: Aligning subtitles to sign language videos**
  - **Providing similar tools to the Deaf community** (automatic subtitling, etc.) than for written languages
  - **Increasing** the size of **available datasets for machine translation** (and relative tasks)

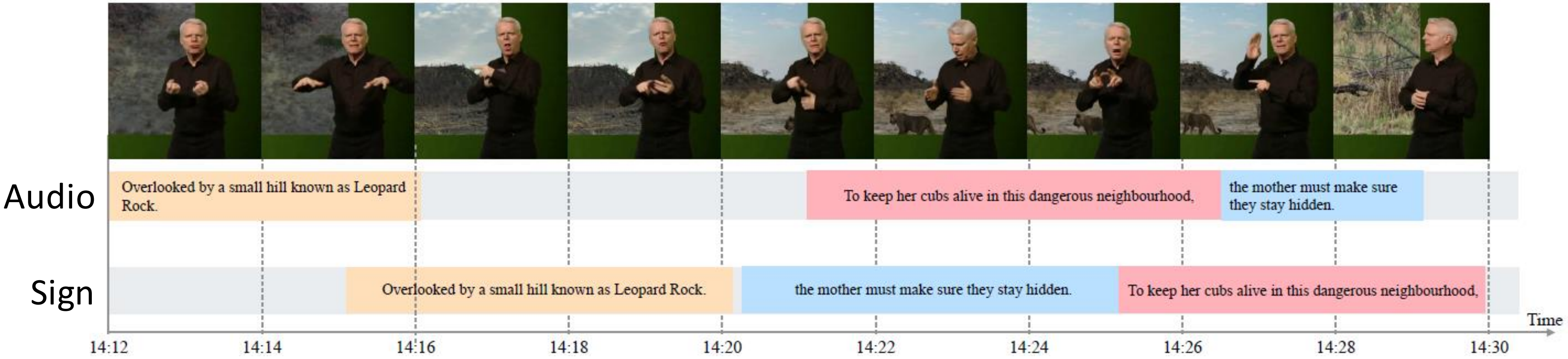

Audio
Overlooked by a small hill known as Leopard Rock.
To keep her cubs alive in this dangerous neighbourhood,
the mother must make sure they stay hidden.
Sign
Overlooked by a small hill known as Leopard Rock.
the mother must make sure they stay hidden.
To keep her cubs alive in this dangerous neighbourhood,
Time
14:12
14:14
14:16
14:18
14:20
14:22
14:24
14:26
14:28
14:30

Fig. 1: Subtitle alignment task. (from [3])

- **Sign languages have their own speed and grammar so alignment can't be infered from audio**
  - The **ordering** of subtitles might be different (see blue/red examples)
  - The **duration** of subtitles is different
  - The signing corresponds to a **translation** of the subtitles (not transcription)

2. Available packages

- **[1] BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues (2021)**
    - Trains a model to (1) find probable signing windows and (2) use mouthing cues to sharpen predictions
    - Publicly releases a dictionary of **word annotations called M** in the following

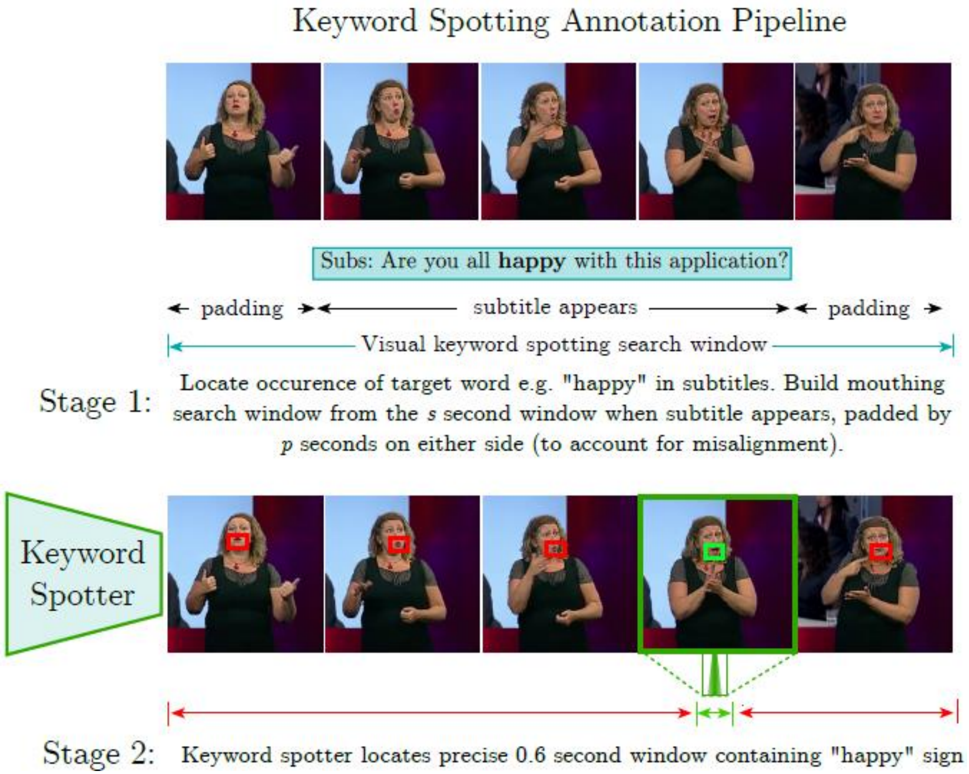

Keyword Spotting Annotation Pipeline
Subs: Are you all **happy** with this application?
padding
subtitle appears
padding
Visual keyword spotting search window
Stage 1:
Locate occurence of target word e.g. "happy" in subtitles. Build mouthing search window from the s second window when subtitle appears, padded by p seconds on either side (to account for misalignment).
Keyword Spotter
Stage 2:
Keyword spotter locates precise 0.6 second window containing "happy" sign

Fig. 2: Mouthing cues model. (from [1])

- **[2] Watch, read and lookup: learning to spot signs from multiple supervisors (2020)**
    - Uses (1) sparsely annotated sequences, (2) subtitle text and (3) available sign language dictionary
    - Publicly releases a dictionary of **word annotations called D** in the following

2. Available packages

- **[3] Aligning Subtitles in Sign Language Videos (2021)**
  - *Subtitle Aligner Transformer:* **predicting a range of frames where signing subtitles appear**
  - **Pretraining on sign spotting** task with words from dictionaries (M, D, …) and 1-second durations

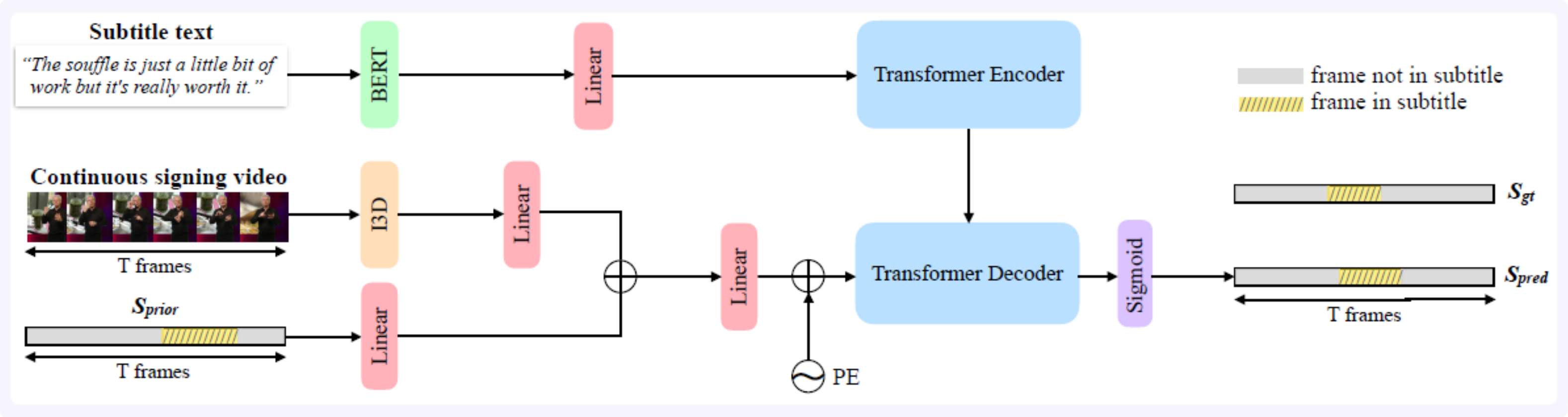
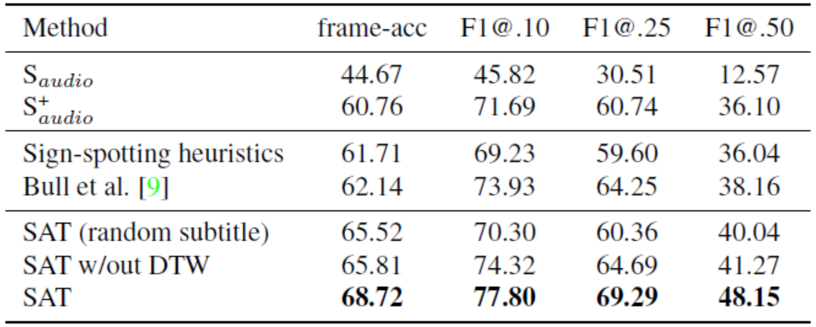


Figure 2: **SAT model overview:** We input to our model (i) token embeddings of the subtitle text we wish to align, (ii) a sequence of video features extracted from a continuous sign language video segment and (iii) the shifted temporal boundaries of the audio-aligned subtitle, S_{prior}. Using these inputs, the model outputs a vector of values between 0 and 1 of length T. Its first and last values above a threshold τ delimit the predicted temporal boundaries for the query subtitle. The location of the subtitle with respect to the window is represented in dashed yellow.



Method	frame-acc	F1@.10	F1@.25	F1@.50
S_{audio}	44.67	45.82	30.51	12.57
S^{+}_{audio}	60.76	71.69	60.74	36.10
Sign-spotting heuristics	61.71	69.23	59.60	36.04
Bull et al. [9]	62.14	73.93	64.25	38.16
SAT (random subtitle)	65.52	70.30	60.36	40.04
SAT w/out DTW	65.81	74.32	64.69	41.27
SAT	**68.72**	**77.80**	**69.29**	**48.15**

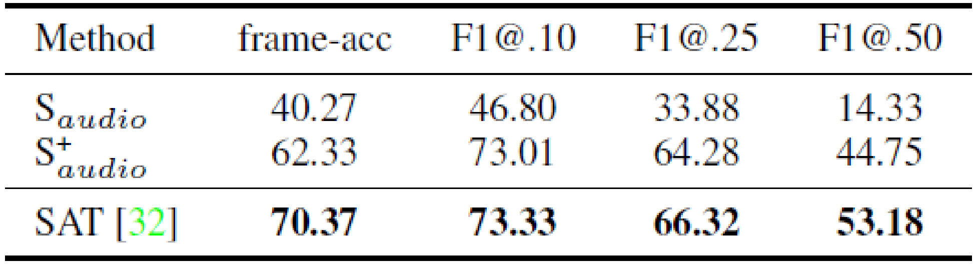
Fig. 3: *SAT* architecture and results. (from [3])

- **[4] Automatic dense annotation of large-vocabulary sign language videos (2022)**
  - Improves word annotations **M -> M*** (new *Transpotter* architecture) and **D -> D*** (using *SAT*)
  - Creates new word annotations **P** (sign classification), **E** (re-spot existing signs) and **N** (score maps)

3. My modifications

- **Forked repo**
  - See code: https://github.com/sebastienmeyer2/subtitle_align

- **Minor fixes and test**
  - Enable training on **Windows**, **run test.sh** with available data and checkpoint
  - **Print sanity checks** such as width of ground truth and prior subtitles



Method	frame-acc	F1@.10	F1@.25	F1@.50
S_{audio}	40.27	46.80	33.88	14.33
S^{+}_{audio}	62.33	73.01	64.28	44.75
SAT [32]	**70.37**	**73.33**	**66.32**	**53.18**
GitHub	**70.89**	**74.08**	**66.78**	**53.22**

Fig. 4: *SAT* results when trained on BOBSL (**M***, **D***) and from available GitHub package. (from [5])

- **Adding anchors / spottings priors to *SAT* architecture**
  - **Pretraining on 60 videos** (10 val videos) **for 40 epochs**: no use of anchors / spottings prior
  - **Training** is done **on 60 videos** (10 val videos) **for 40 epochs**
  - **Finetuning** is done on manually aligned train/val splits **for 50 epochs**
  - **Testing** is done on manually aligned test split – **with Dynamic Time Warping** post-processing

3. My modifications

Adding anchors prior

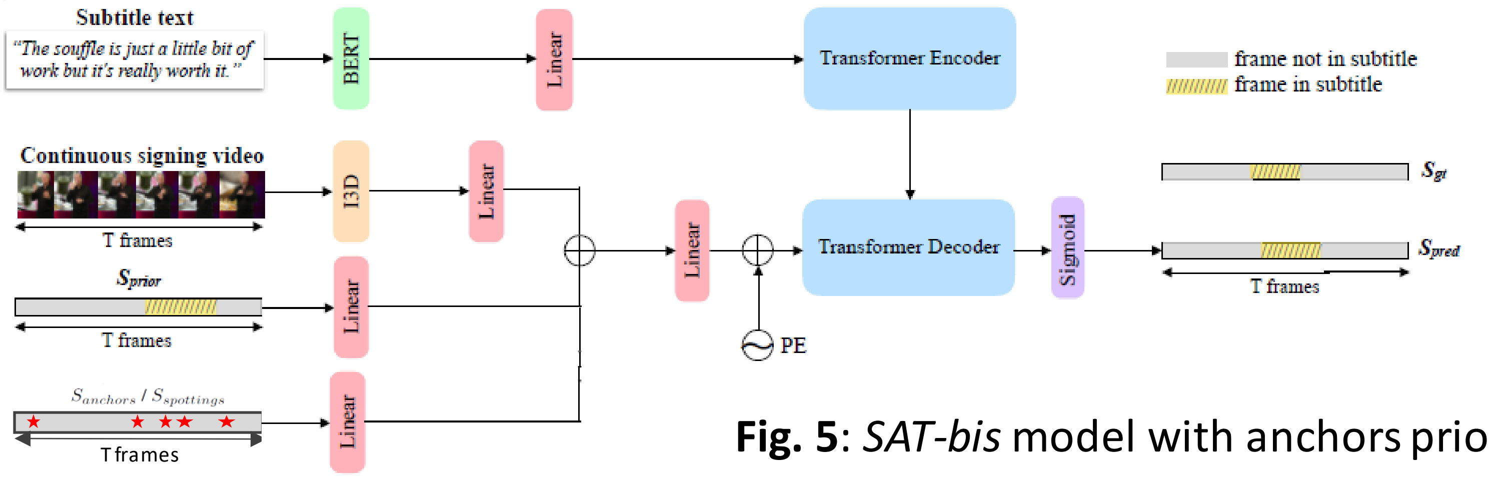


Fig. 5: *SAT-bis* model with anchors prior.

- **Adding corresponding probabilities from word annotations as anchors prior**
  - Add the **probability** value **if** the **word belongs to the subtitle** text
  - **Very sparse**: 0.66% vs. 21.14% (ground-truth) of non-zero frames
  - Results are **not better than the base model**, however they are very similar

Model	frame-acc	F1@.10	F1@.25	F1@.50
Baseline	69.02	72.23	63.83	50.27
Anchors prior	69.46	70.51	62.36	49.58

3. My modifications

Adding spottings prior

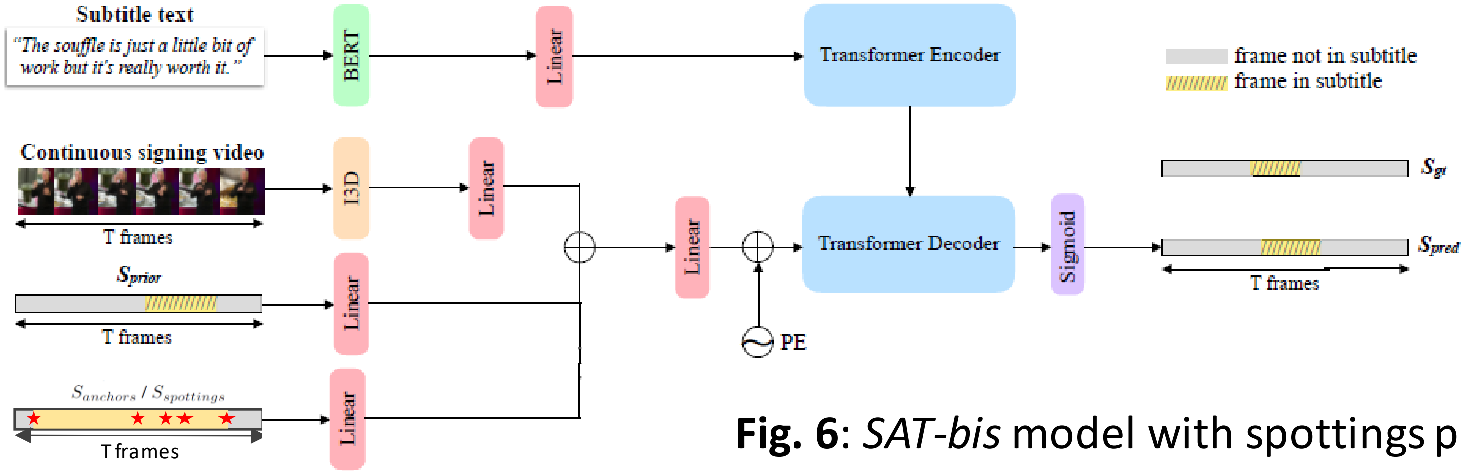


Fig. 6: *SAT-bis* model with spottings prior.

- **Filling range span of spottings to create a spottings prior**
  - **Range** of spottings for which words belong to the subtitle text are **set to ones**
  - Has **large width**: 33.46 vs. 26.43 (ground-truth) vs. 24.11 (audio) number of frames
  - Results are **worse than the base model** and worse than when using anchors prior

Model	frame-acc	F1@.10	F1@.25	F1@.50
Baseline	69.02	72.23	63.83	50.27
Spottings prior	68.82	68.99	60.61	47.35

3. My modifications

Adding adjusted spottings prior

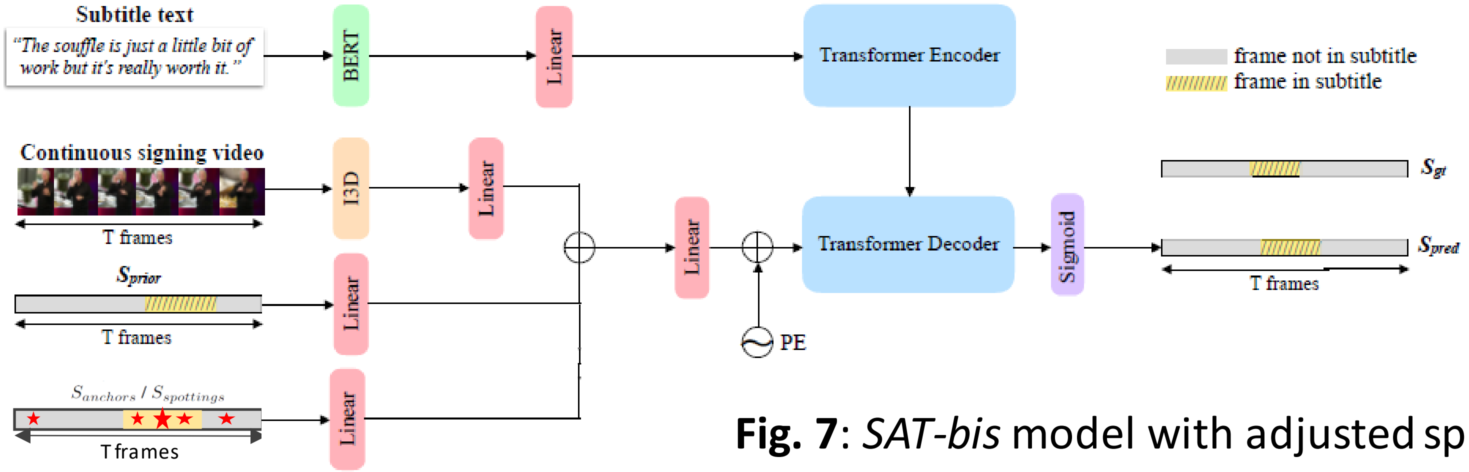


Fig. 7: *SAT-bis* model with adjusted spottings prior.

- **Since the spottings prior is too wide, adjust its size to the one of audio prior**
  - Select the **median** (robust to outliers) of spottings, and create a **vector of ones of audio prior's width**
  - However, it is **often empty**: 54.39% of spottings priors are empty vectors
  - The results are a bit better than without adjustment, however it is **still worse than the base model**

Model	frame-acc	F1@.10	F1@.25	F1@.50
Baseline	69.02	72.23	63.83	50.27
Adj. spot. prior	69.18	69.29	61.26	48.51

3. My modifications

Adding adjusted spottings prior w. PEN

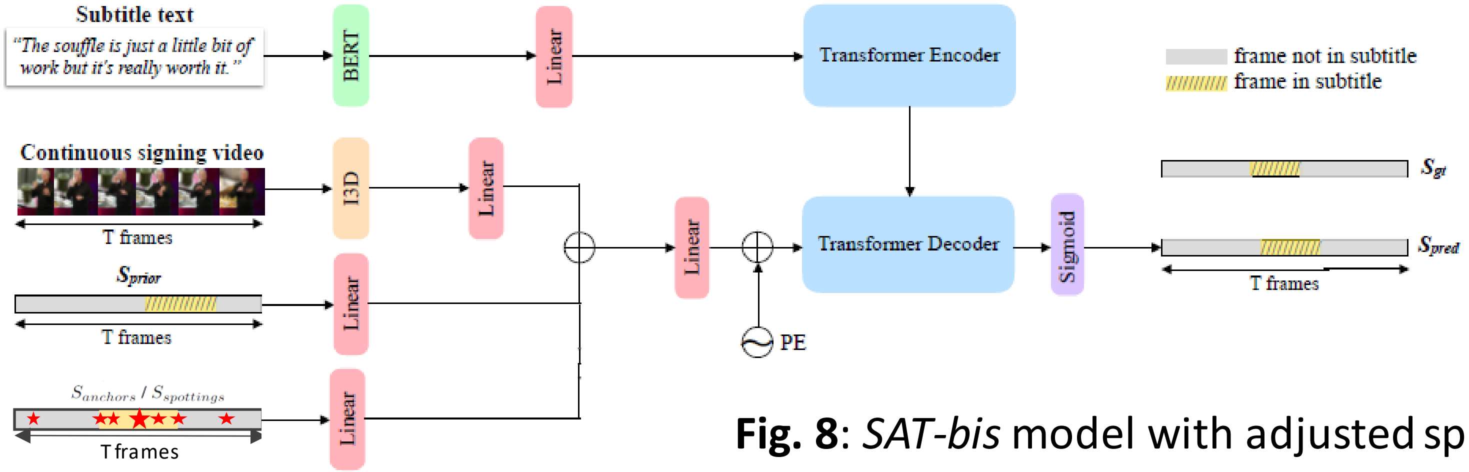


Fig. 8: *SAT-bis* model with adjusted spottings prior (+ PEN).

- **In order to get more non-empty spottings priors, use word annotations P, E and N**
  - **Less often empty**: 23.13% of spottings priors are empty vectors
  - With all these improvements, we get a better frame accuracy and **similar F1-scores to the base model**

Model	frame-acc	F1@.10	F1@.25	F1@.50
Baseline	69.02	72.23	63.83	50.27
Adj. spot. (PEN)	69.03	70.60	62.57	49.51

Concluding remarks (1/2)

Results and limitations

- **We summarize our results in the following table, last row combines all our ideas**

Model	frame-acc	F1@.10	F1@.25	F1@.50
Baseline	69.02	**72.23**	**63.83**	**50.27**
Anchors prior	**69.46**	70.51	62.36	49.58
Spottings prior	68.82	68.99	60.61	47.35
Adj. spot. prior	69.18	69.29	61.26	48.51
Baseline (PEN)	69.59	73.47	65.74	51.55
Anchors (PEN)	**71.25**	**74.09**	**66.46**	**52.72**
Adj. spot. (PEN)	70.43	73.48	65.46	52.05

- **Increasing the number of spottings will improve the results** (1 in 5 spottings priors are empty vectors)
- *Limitation*: we do not use these priors during pretraining, so it **requires to train a new reprojection layer**

Concluding remarks (2/2)

Some sample subtitles

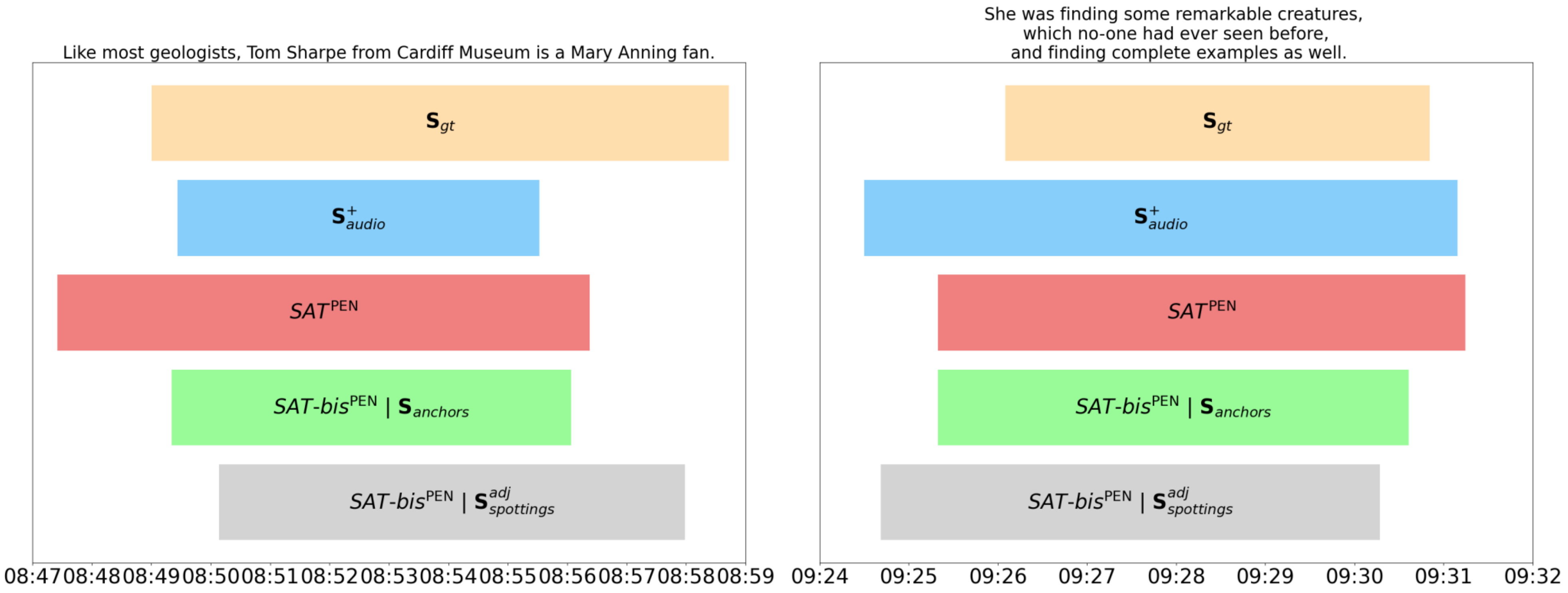


Fig. 9: two examples of sign language alignment during testing for comparing the base *SAT* model with our *SAT-bis* model, either using anchors prior or spottings prior.

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox and Andrew Zisserman. *BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues.* (Available at: https://www.robots.ox.ac.uk/~vgg/research/bsl1k/)

[2] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. (Available at: https://www.robots.ox.ac.uk/~vgg/research/bsldict/)

[3] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni and Andrew Zisserman. Aligning Subtitles in Sign Language Videos. (Available at: https://www.robots.ox.ac.uk/~vgg/research/bslalign/)

[4] Liliane Momeni, Hannah Bull, K. R. Prajwal, Samuel Albanie, Gül Varol and Andrew Zisserman. Automatic dense annotation of large-vocabulary sign language videos. (Available at: https://www.robots.ox.ac.uk/~vgg/research/bsldensify/)

[5] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland and Andrew Zisserman. BBC-Oxford British Sign Language Dataset. (Available at: https://www.robots.ox.ac.uk/~vgg/data/bobsl/)