Supplementary Material for A Neural Multi-sequence Alignment TeCHnique (NeuMATCH)

Pelin Dogan[1,4] Boyang Li[2] Leonid Sigal[3] Markus Gross[1,4]
[1]ETH Zürich [2]Liulishuo AI Lab [3]University of British Columbia [4]Disney Research
{pelin.dogan, grossm}@inf.ethz.ch, albert.li@liulishuo.com, lsigal@cs.ubc.ca

In this supplementary material, we first give details on the segmentation of videos into clips. Next, we show more alignment results computed by our approach on the datasets HM-1, HM-2, and YMS that require one-to-many matching and contain clips that do not match any sentences (i.e., *null* clips). For illustration purposes, each figure below represents only a small portion (6-12 consecutive clips) of the entire aligned sequence. Each frame represents a video clip. The aligned sentences are shown with wide brackets below or above the clips.

1. Implementation Details

As discussed in Sec. 3.2 in the main paper, we customize the action inventory using knowledge of the dataset. For one-to-one matching with *null* video clips, we use the actions Pop Clip and Match. For one-to-many matching with *null* video clips, we use Pop Clip, Pop Sentence, and Match-Retain Sentence. For all the experiments, action decoding is done greedily.

For the joint pre-training, we use 500 dimensions for the LSTM sentence encoder and 300 for the joint embeddings. The dimensions of the word and image embedding are 300 and 4096, respectively, while the margin in the ranking objective function is $\alpha = 0.05$. L_2 regularization is used to prevent over-fitting. The batch size is set to 32 and the number of contrastive samples is 31 for every positive pair. The model is trained with the Adam optimizer using a learning rate of 10^{-4} and gradient clipping of 2.0. Early stopping on the validation set is used to avoid over-fitting.

The alignment network uses 300 dimensions for the video and text stacks, 20 dimensions for the matched stack and 8 for the history stack. Optionally, we feed two additional variables into the fully connected layer: the numbers of elements left in the video and text stacks to improve the performance on very long sequences in the YMS dataset. The alignment network is first trained with the encoding networks fixed with a learning rate of 0.001. After that, the entire model is trained end-to-end with a learning rate of 10^{-5}. For HM-0, HM-1, and HM-2, we use the original data split of LSMDC. For YMS, we use a 80/10/10 split for training, validation and test sets.

Details of Video Segmentation The video segmentation can be achieved using any shot boundary detection algorithm. In this work, we segment the input videos into video clips by a Python/OpenCV-based scene detection program[1] that uses threshold/content on a given video. For the parameters, we choose the *content-aware* detection method with the *threshold* of 20 and *minimum length* of 5 frames. Having a low threshold and minimum length usually results in over-segmentation. However, NeuMATCH can handle this resulting over-segmentation with the ability of one-to-many matching.

2. Alignment Results

2.1. Successful results for Hollywood Movies 1 (HM-1)

The video sequences in HM-1 contain clips from other movies that are inserted into the original sequence, as explained in the main paper.

[1]https://github.com/Breakthrough/PySceneDetect

Figure 1: From the movie *Jack and Jill* in dataset HM-1. The fifth frame is from the movie *This is 40*, which is successfully assigned as *null*. Note the last two frames have very similar content (two women in dresses) to the sentence *"With a fuzzy shawl and cap, and a ruffled skirt."*, but our algorithm was able to identify them correctly.

Figure 2: From the movie *Blind Dating* in dataset HM-1. The third frame is from the movie *Inside Man*, and the fifth frame is from the movie *Jack and Jill*, which are correctly assigned to *null*.

Figure 3: From the movie *Juno* in dataset HM-1. The one-to-many assignment for the last three clips is correctly identified even when there is a significant perspective and content change through the clips.

2.2. Successful results for Hollywood Movies 2 (HM-2)

Each video sequence in HM-2 consists of consecutive clips from a single movie, where some sentences were discarded in order to create *null* clips. It still requires one-to-many matching of the sentences and the assignment of *null* clips.

Figure 4: From the movie *Harry Potter and the Prisoner of Azkaban* in dataset HM-2

Figure 5: From the movie *Bad Santa* in dataset HM-2

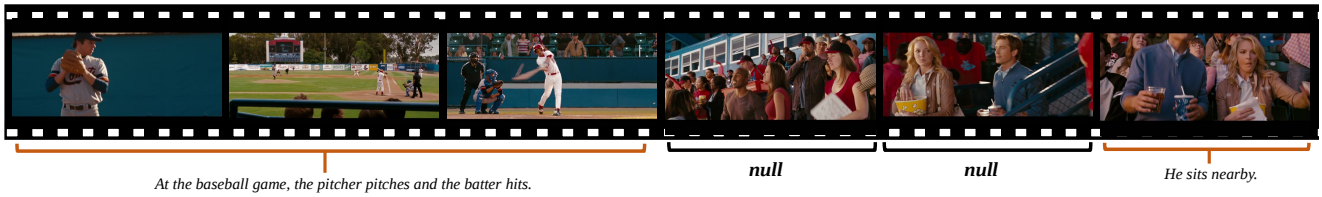

Figure 6: From the movie *The Ugly Truth* in dataset HM-2. The third clip contains a vodka bottle, which is mentioned in first sentence. The fourth and the fifth clips are very similar. However, the algorithm finds the correct alignment.

Figure 7: From the movie *Super 8* in dataset HM-2. The boy and the bicycle are visible in both the second and the third clips, but the headstones only appear in the third clip. The algorithm makes the correct decision.

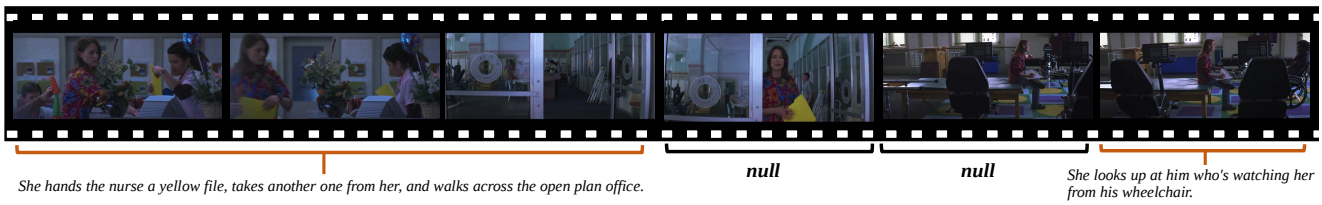

Figure 8: From the movie *Unbreakable* in dataset HM-2. The wheelchair is only visible in the last clip and the algorithm successfully picks that up.

2.3. Successful results for YouTube Movie Summaries (YMS)

In the YMS dataset, the sentences are longer than HM-1 and HM-2, and they tend to describe multiple events. We asked the annotators to break them down into small units, which allows them to precisely align the text with the video sequence. These sequences tend to be much more complex than HM-1 and HM-2.

2.4. Failure Cases

We present two failure cases below. The ground truth is shown with green brackets and NeuMATCH's predictions are with orange brackets.

Figure 9: From the movie *Doctor Strange* in dataset YMS. The original video is available at https://www.youtube.com/watch?v=fZeW-KUXHKY

Figure 10: From the movie *It (1990)* in dataset YMS. The original video is available at https://www.youtube.com/watch?v=c-sIoODkpuU

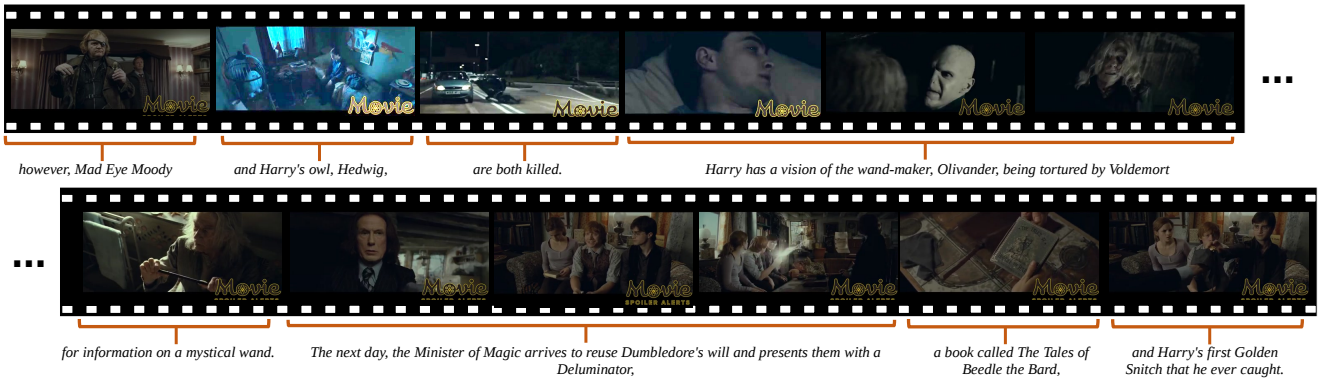

Figure 11: From the movie *Harry Potter and the Deathly Hallows* in dataset YMS. The original video is available at https://www.youtube.com/watch?v=nfuRErj9TkY

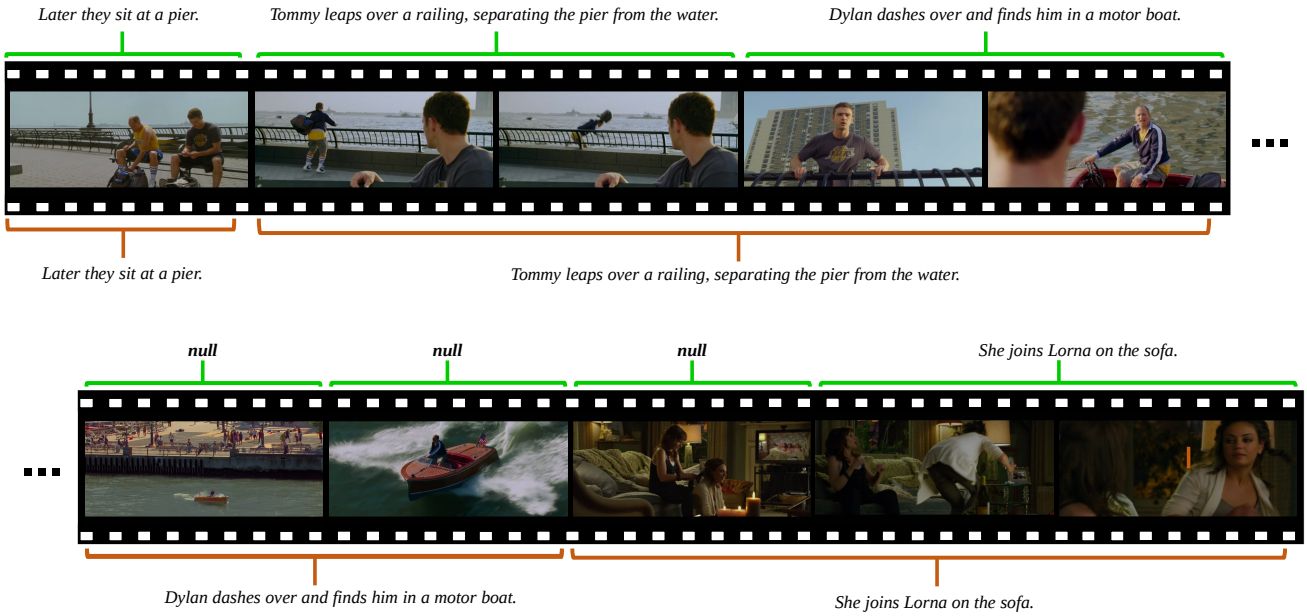

Figure 12: From the movie *Friends with Benefits* in dataset HM-2. The assignments with green marks represent the ground truth while orange marks represent our result. The first failure is that the second sentence is matched with two more clips, but the additional clips also contain the "railing" and the "water", which may have confused the algorithm. Similarly, the boat appears in the sixth and seventh clips, which may have caused the wrong alignment with the third sentence.

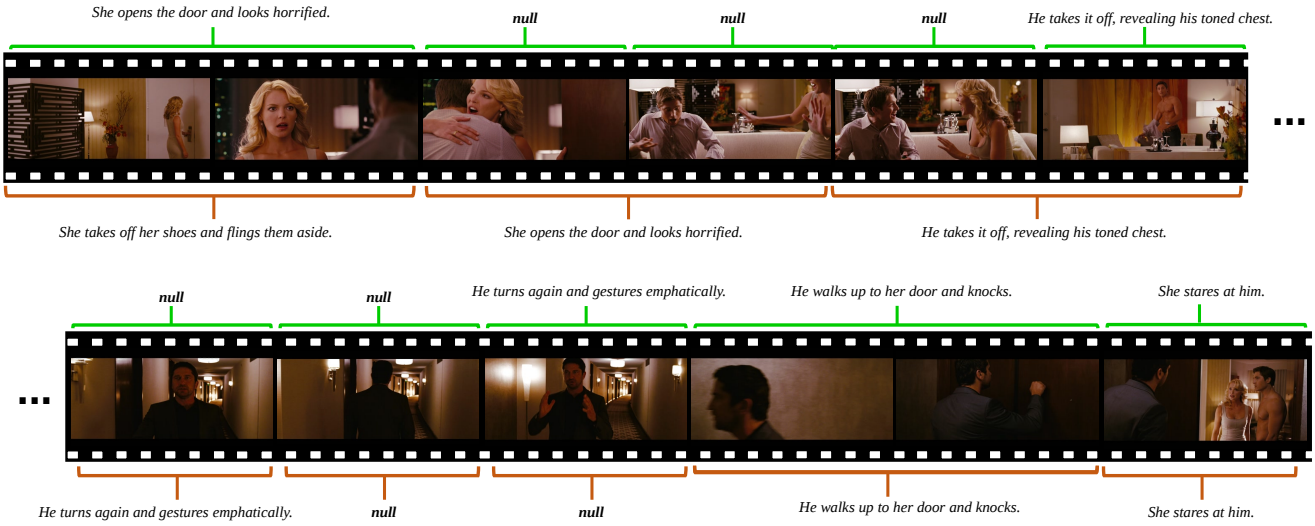

Figure 13: From the movie *The Ugly Truth* in dataset HM-2. The assignments with green marks represent the ground truth while orange marks represent our result.