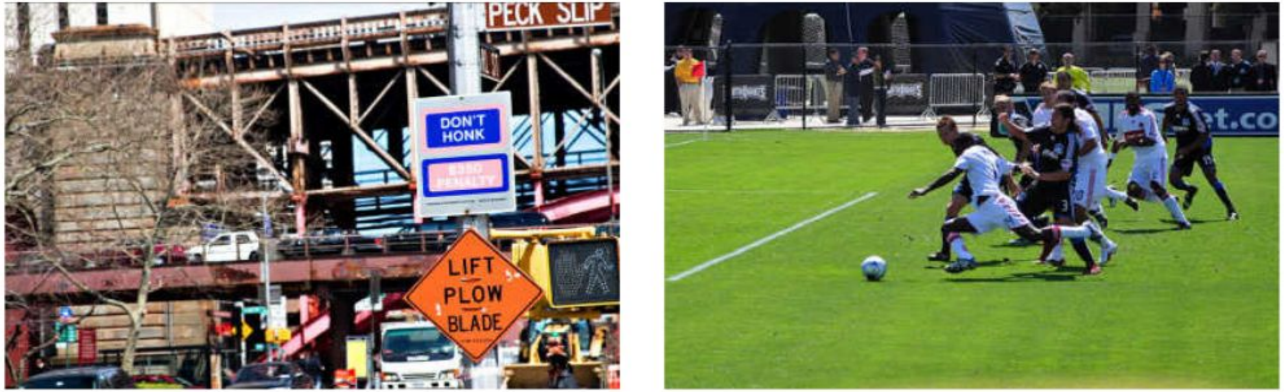
# WAT 2019 Multi-Modal Shared Task Overview

Shantipriya Parida
Idiap Research Institute, Switzerland

Ondřej Bojar
Charles University, Czech Republic

# Overview

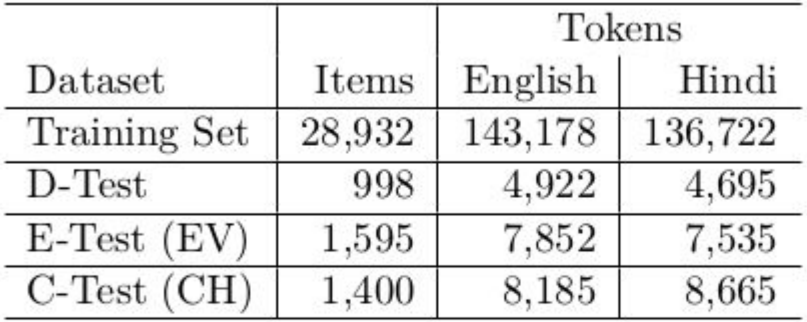
- English→Hindi multimodal translation task is based on the first English-Hindi multi-modal corpus (Hindi Visual Genome, HVG in short).
- Multi-modal task is introduced first time in WAT 2019.
- Four teams participated with twenty submissions.



Street sign advising of penalty. The penalty box is white lined.

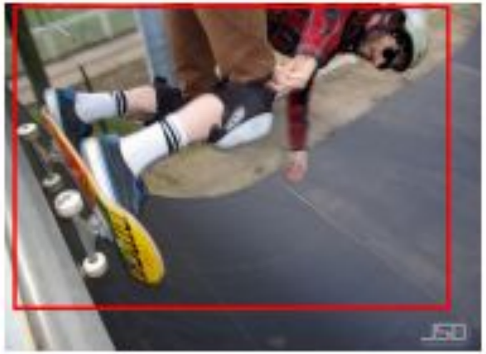
An illustration of two meanings of the word “penalty” exemplified with two images (Hindi Visual Genome)

# Dataset



| Dataset | Items | Tokens |  |
|---|---|---|---|
|  |  | English | Hindi |
| Training Set | 28,932 | 143,178 | 136,722 |
| D-Test | 998 | 4,922 | 4,695 |
| E-Test (EV) | 1,595 | 7,852 | 7,535 |
| C-Test (CH) | 1,400 | 8,185 | 8,665 |

Data for the English→Hindi multi-modal translation task. One item consists of source English sentence, target Hindi sentence, and a rectangular region within an image. The total number of English and Hindi tokens in the dataset also listed. The abbreviations EV and CH are used in the official task names in WAT scoring tables.



Source Text : Man stand of skateboard
Reference : आदमी स्केटबोर्ड पर खड़ा है

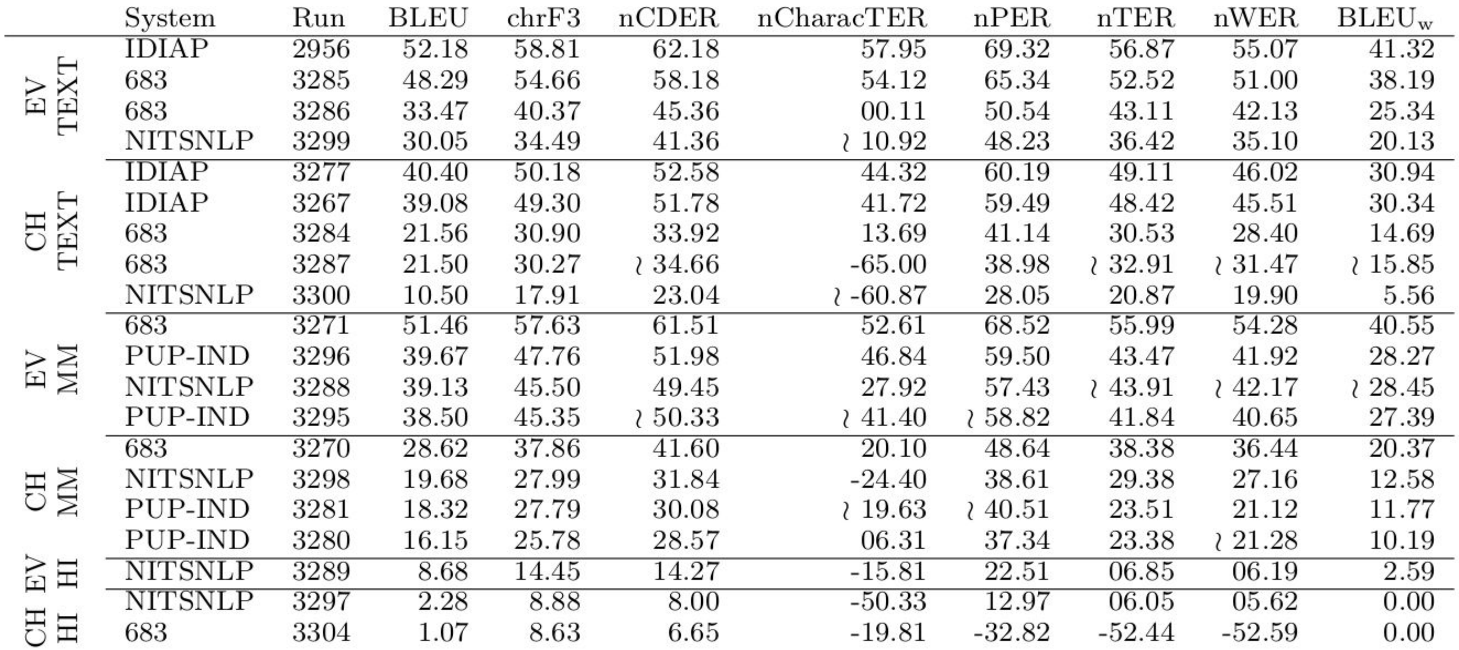
Illustration of an item

# Tracks

- **Text-Only Translation (labeled “TEXT” in WAT official tables) :** The task is to translate short English captions (text) into Hindi. No visual information can be used.  ( need to be specified other resources if used in the corresponding system description paper).
- **Hindi Captioning (labeled “HI”):** The task is generate captions in Hindi for the given rectangular region in an input image.
- **Multi-Modal Translation: (labeled “MM”):** Given an image, a rectangular region in it and an English caption for the rectangular region, the task is to translate the English text into Hindi. Both textual and visual information can be used.

# Results (Automatic Evaluation)

Several automatic metrics used for automatic evaluation

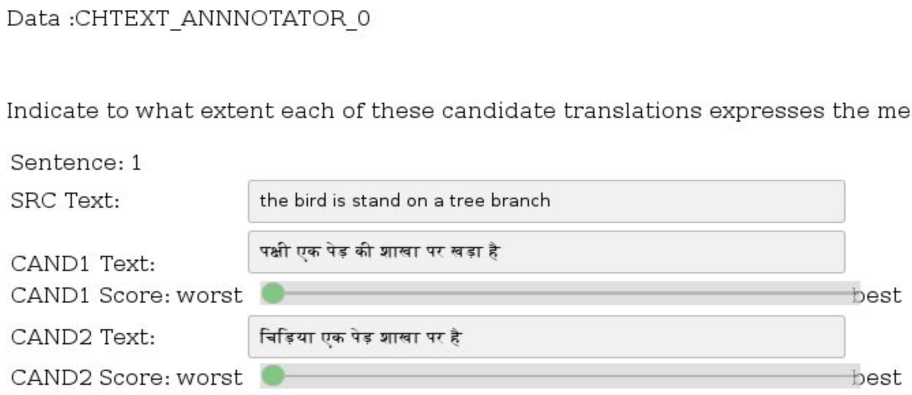


| | System | Run | BLEU | chrF3 | nCDER | nCharacTER | nPER | nTER | nWER | $BLEU_w$ |
|---|---|---|---|---|---|---|---|---|---|---|
| EV TEXT | IDIAP | 2956 | 52.18 | 58.81 | 62.18 | 57.95 | 69.32 | 56.87 | 55.07 | 41.32 |
| | 683 | 3285 | 48.29 | 54.66 | 58.18 | 54.12 | 65.34 | 52.52 | 51.00 | 38.19 |
| | 683 | 3286 | 33.47 | 40.37 | 45.36 | 00.11 | 50.54 | 43.11 | 42.13 | 25.34 |
| | NITSNLP | 3299 | 30.05 | 34.49 | 41.36 | ≀ 10.92 | 48.23 | 36.42 | 35.10 | 20.13 |
| CH TEXT | IDIAP | 3277 | 40.40 | 50.18 | 52.58 | 44.32 | 60.19 | 49.11 | 46.02 | 30.94 |
| | IDIAP | 3267 | 39.08 | 49.30 | 51.78 | 41.72 | 59.49 | 48.42 | 45.51 | 30.34 |
| | 683 | 3284 | 21.56 | 30.90 | 33.92 | 13.69 | 41.14 | 30.53 | 28.40 | 14.69 |
| | 683 | 3287 | 21.50 | 30.27 | ≀ 34.66 | -65.00 | 38.98 | ≀ 32.91 | ≀ 31.47 | ≀ 15.85 |
| | NITSNLP | 3300 | 10.50 | 17.91 | 23.04 | ≀ -60.87 | 28.05 | 20.87 | 19.90 | 5.56 |
| EV MM | 683 | 3271 | 51.46 | 57.63 | 61.51 | 52.61 | 68.52 | 55.99 | 54.28 | 40.55 |
| | PUP-IND | 3296 | 39.67 | 47.76 | 51.98 | 46.84 | 59.50 | 43.47 | 41.92 | 28.27 |
| | NITSNLP | 3288 | 39.13 | 45.50 | 49.45 | 27.92 | 57.43 | ≀ 43.91 | ≀ 42.17 | ≀ 28.45 |
| | PUP-IND | 3295 | 38.50 | 45.35 | ≀ 50.33 | ≀ 41.40 | ≀ 58.82 | 41.84 | 40.65 | 27.39 |
| CH MM | 683 | 3270 | 28.62 | 37.86 | 41.60 | 20.10 | 48.64 | 38.38 | 36.44 | 20.37 |
| | NITSNLP | 3298 | 19.68 | 27.99 | 31.84 | -24.40 | 38.61 | 29.38 | 27.16 | 12.58 |
| | PUP-IND | 3281 | 18.32 | 27.79 | 30.08 | ≀ 19.63 | ≀ 40.51 | 23.51 | 21.12 | 11.77 |
| | PUP-IND | 3280 | 16.15 | 25.78 | 28.57 | 06.31 | 37.34 | 23.38 | ≀ 21.28 | 10.19 |
| EV HI | NITSNLP | 3289 | 8.68 | 14.45 | 14.27 | -15.81 | 22.51 | 06.85 | 06.19 | 2.59 |
| CH HI | NITSNLP | 3297 | 2.28 | 8.88 | 8.00 | -50.33 | 12.97 | 06.05 | 05.62 | 0.00 |
| | 683 | 3304 | 1.07 | 8.63 | 6.65 | -19.81 | -32.82 | -52.44 | -52.59 | 0.00 |

Multi-Modal Task automatic evaluation results. For each test set (EV and CH) and each track (TEXT, MM and HI), we sort the entries by our BLEU scores. The symbol "≀" in subsequent columns indicates fields where the other metric ranks candidates in a different order. BLEU w denotes the WAT official BLEU scores.

# Results (Manual Evaluation)

- Manual Evaluation follow Direct Assessment (DA) technique by asking annotator to assign 0-100 for each candidate.
- Collected DA scores averaged for each system and track (denoted “Ave”).
- Standardized per annotator and then averaged (denoted “Ave Z”).
    - Scores are scaled, so average score of each annotator is 0 and standard deviation is 1.



Data :CHTEXT_ANNNOTATOR_0

Indicate to what extent each of these candidate translations expresses the meaning of the English source text (independently of the other candidate).

Sentence: 1

SRC Text: the bird is stand on a tree branch

CAND1 Text: पक्षी एक पेड़ की शाखा पर खड़ा है

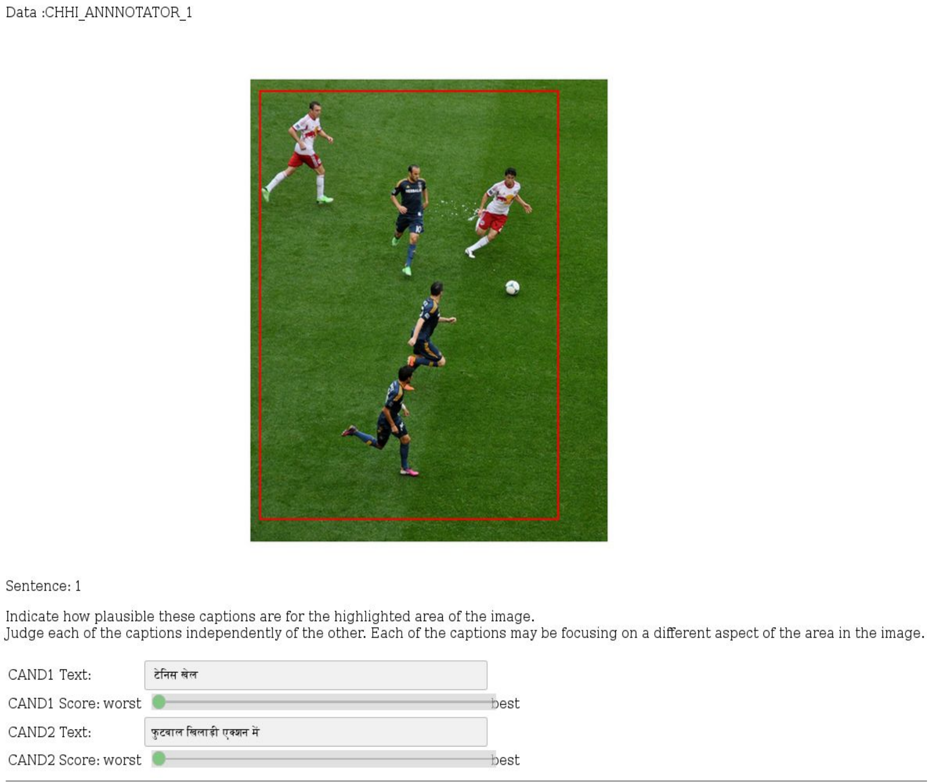
CAND1 Score: worst ● best

CAND2 Text: चिड़िया एक पेड़ शाखा पर है

CAND2 Score: worst ● best

Manual evaluation of text-only translation.

# Results (Manual Evaluation)



Data :CHHI_ANNNOTATOR_1

Sentence: 1

Indicate how plausible these captions are for the highlighted area of the image.
Judge each of the captions independently of the other. Each of the captions may be focusing on a different aspect of the area in the image.

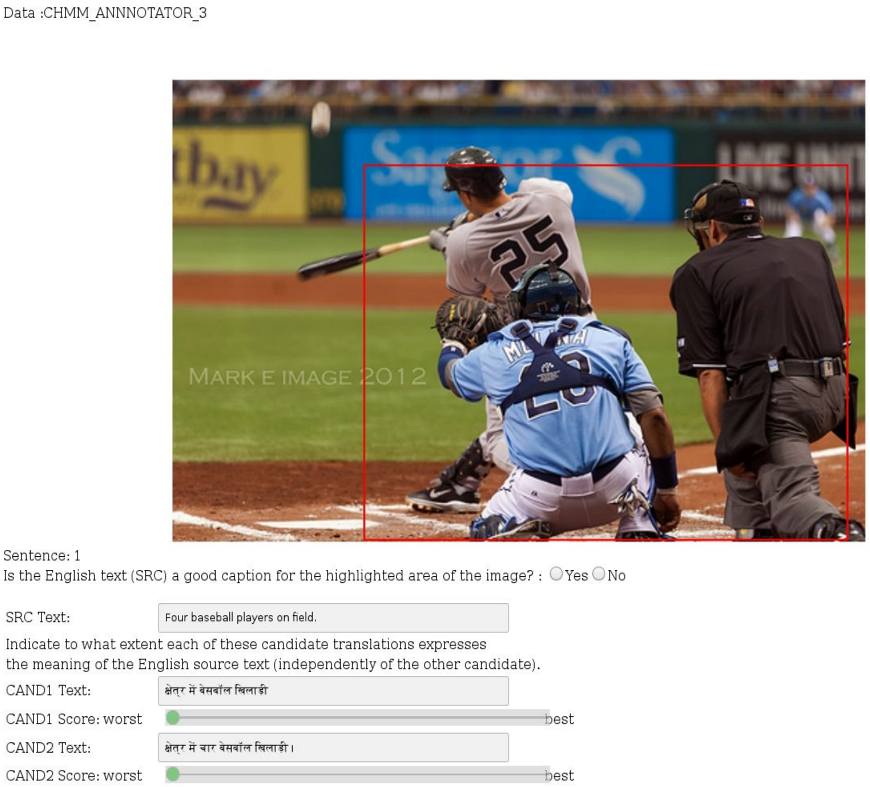
CAND1 Text: टेनिस खेल

CAND1 Score: worst ● best

CAND2 Text: फुटबाल खिलाड़ी एक्शन में

CAND2 Score: worst ● best

Manual evaluation of Hindi captioning.



Data :CHMM_ANNNOTATOR_3

Sentence: 1

Is the English text (SRC) a good caption for the highlighted area of the image? : ○ Yes ○ No

SRC Text: Four baseball players on field.

Indicate to what extent each of these candidate translations expresses
the meaning of the English source text (independently of the other candidate).

CAND1 Text: क्षेत्र में बेसबॉल खिलाड़ी

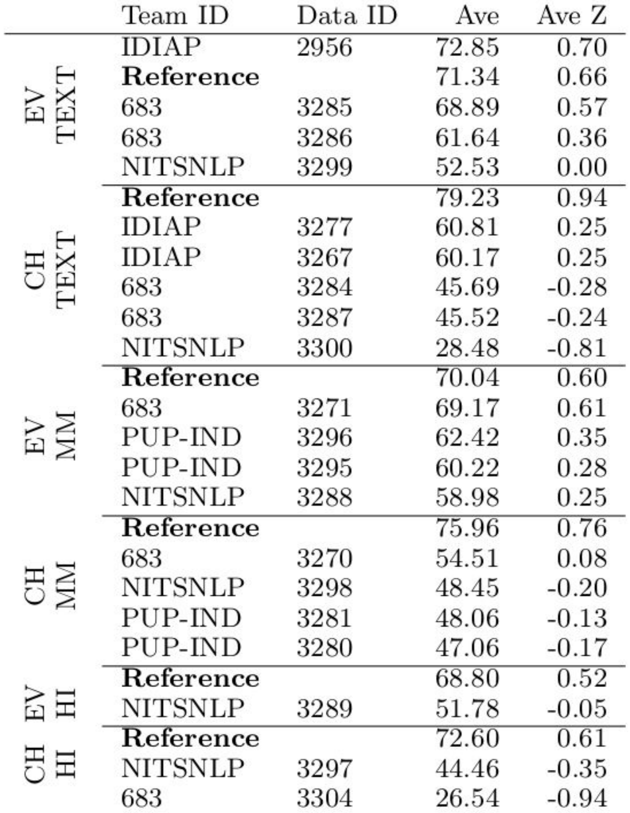
CAND1 Score: worst ● best

CAND2 Text: क्षेत्र में चार बेसबॉल खिलाड़ी।

CAND2 Score: worst ● best

Manual evaluation of multi-modal translation.

# Results (Manual Evaluation)

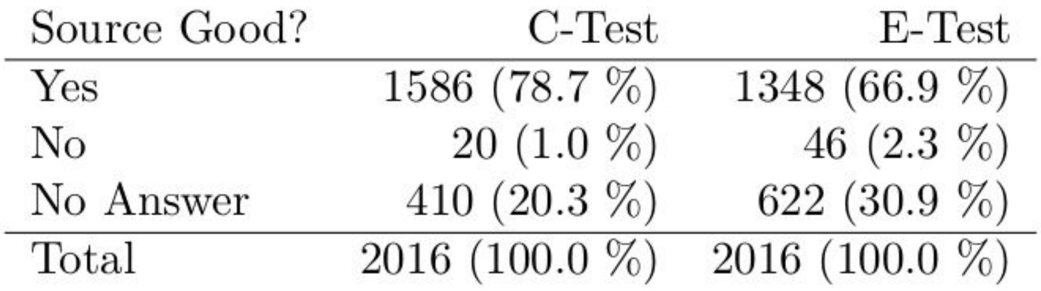


| | Team ID | Data ID | Ave | Ave Z |
|---|---|---|---|---|
| EV TEXT | IDIAP | 2956 | 72.85 | 0.70 |
| | **Reference** | | 71.34 | 0.66 |
| | 683 | 3285 | 68.89 | 0.57 |
| | 683 | 3286 | 61.64 | 0.36 |
| | NITSNLP | 3299 | 52.53 | 0.00 |
| CH TEXT | **Reference** | | 79.23 | 0.94 |
| | IDIAP | 3277 | 60.81 | 0.25 |
| | IDIAP | 3267 | 60.17 | 0.25 |
| | 683 | 3284 | 45.69 | -0.28 |
| | 683 | 3287 | 45.52 | -0.24 |
| | NITSNLP | 3300 | 28.48 | -0.81 |
| EV MM | **Reference** | | 70.04 | 0.60 |
| | 683 | 3271 | 69.17 | 0.61 |
| | PUP-IND | 3296 | 62.42 | 0.35 |
| | PUP-IND | 3295 | 60.22 | 0.28 |
| | NITSNLP | 3288 | 58.98 | 0.25 |
| CH MM | **Reference** | | 75.96 | 0.76 |
| | 683 | 3270 | 54.51 | 0.08 |
| | NITSNLP | 3298 | 48.45 | -0.20 |
| | PUP-IND | 3281 | 48.06 | -0.13 |
| | PUP-IND | 3280 | 47.06 | -0.17 |
| EV HI | **Reference** | | 68.80 | 0.52 |
| | NITSNLP | 3289 | 51.78 | -0.05 |
| CH HI | **Reference** | | 72.60 | 0.61 |
| | NITSNLP | 3297 | 44.46 | -0.35 |
| | 683 | 3304 | 26.54 | -0.94 |

Manual evaluation result for WAT Multi-Modal Tasks.

# HVG Validation

- One of the participant team spotted few error in the HVG dataset.
- We made use of the manual annotations to validate English sources in HVG.



| Source Good? | C-Test | E-Test |
|---|---|---|
| Yes | 1586 (78.7 %) | 1348 (66.9 %) |
| No | 20 (1.0 %) | 46 (2.3 %) |
| No Answer | 410 (20.3 %) | 622 (30.9 %) |
| Total | 2016 (100.0 %) | 2016 (100.0 %) |

Appropriateness of source English captions in the 4032 assessments collected for the multi-modal track.

# Discussion

- The automatic evaluation score for the “Hindi caption” is very very low as compared to other sub-tasks (“text-only” and “multi-modal” translations).
    - While the automatic scores are comparable across tasks, the Hindi-only captioning (“HI”) must be considered separately.
    - Without a source sentence, both humans and machines are very likely to come up with highly varying textual captions.
- BLEU scores by WAT main organizers and us differ a lot.
    - The reason is probably different tokenization rules.
    - The message to take is that **no scores are comparable**, unless calculated by the exact same implementation of a metric on the exact same set of sentences.
- A text-only submission (IDIAP) outperformed multi-model submissions.
    - As of now, more text data are more important than having access to the image.

# Conclusions

- Multi-modal task attracted four teams across the three tracks.
- Automatic and manual evaluation are generally in line.
    - (With a small exception for multi-modal track on E-Test.)
- Text-only system with larger data outperformed multi-modal systems.
    - ...and it also seems to have outperformed the reference translation.
- Captioning cannot be evaluated with a single reference caption.

Plans for Future:

- Revise HVG sources to remove the various errors we spotted.
- Create a new challenge test set where the image would be indeed *required* for the disambiguation.
- Add a Question Answering (QA) setup.