TREVCID 2020:
Video to Text Description

Asad Anwar Butt
NIST; Johns Hopkins University

George Awad
NIST; Georgetown University

Yvette Graham
Dublin City University

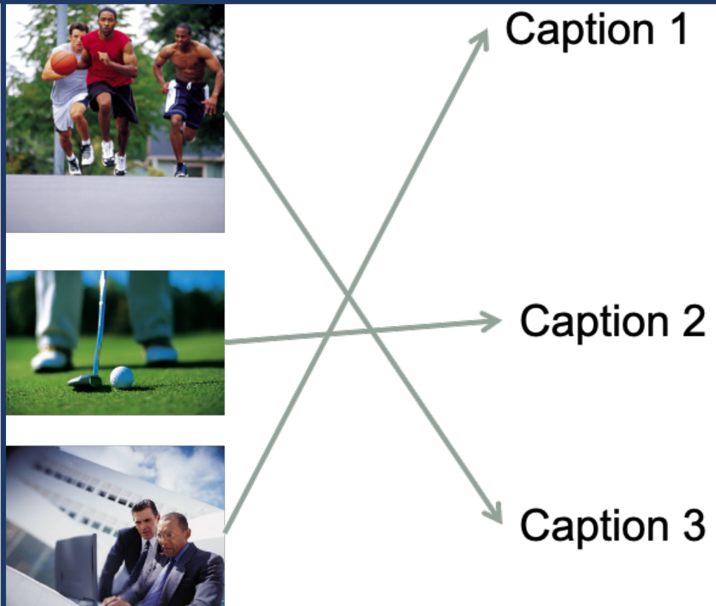
National Institute of
Standards and Technology
U.S. Department of Commerce

Goals and Motivation

- Measure how well an automatic system can describe a video in natural language.
- Measure how well an automatic system can match high-level textual descriptions to low-level computer vision features.
- Transfer successful image captioning technology to the video domain.
- Real world applications
  - Video summarization
  - Supporting search and browsing
  - Accessibility - video description to the blind
  - Video event prediction

Subtasks

This is a system generated caption.



- **Description Generation (Core):**

  Automatically generate a text description for a given video.

- **Matching & Ranking (Optional):**

  Return for each video a ranked list of the most likely text description from each of the five sets.

Note: Images were selected from Google Images with Creative Commons license.

Testing Dataset

- VTT tasks from 2016 to 2019 used the Twitter Vines dataset.
  - Videos were ~6 sec long
  - Quality control issues
  - Links distributed instead of videos, leading to problem of removed links.
- Mixed up things a little with addition of Flickr videos in 2019.
- New dataset: V3C
  - The Vimeo Creative Commons Collection (V3C) is divided into 3 partitions.
  - Total duration: 3800+ hours.
  - V3C2 duration: 1300+ hours. Divided into more than 1.4M segments. Only segments between 3 to 10 sec selected for this task.
  - Videos distributed directly to participants.

Testing Dataset

- Manual selection of videos.
  - We watched 8000+ videos.
  - Selected 1700 videos for annotation.
- Selection criteria mainly concerned with diversity in videos.
- The V3C dataset removes some previous concerns:
  - Videos with multiple, unrelated segments that are not coherent.
  - Offensive videos.

Annotation Process

- A total of 9 assessors annotated the videos.
- Each video was annotated by 5 different assessors.
- Assessors were provided with annotation guidelines by NIST.
- For each video, assessors were asked to combine 4 facets if applicable:
  - Who is the video showing (objects, persons, animals, …etc) ?
  - What are the objects and beings doing (actions, states, events, …etc)?
  - Where (locale, site, place, geographic, ...etc) ?
  - When (time of day, season, ...etc) ?

Annotation Process

- Assessors were provided training for the task.
- Their work was monitored, and feedback provided.
- NIST personnel were available for any questions or confusion.
- Our annotation process differentiates our dataset from other datasets.
  - Arguably better/more detailed descriptions than crowd-sourced datasets.

Annotation – Observations

NIST

- Average sentence length for each assessor:

Annotator	Avg. Length	# Videos
1	16.60	825
2	16.65	875
3	17.67	1700
4	19.62	825
5	21.22	875
6	22.61	875
7	22.71	875
8	24.14	825
9	25.81	825

Avg. sentence length: 20.46 words

- Additional questions:

Please rate how difficult it was to describe the video.

Very Easy (1) Easy (2) Medium (3) Hard (4) Very Hard (5)

How likely is it that other assessors will write similar descriptions for the video?

Not Likely (1) Somewhat Likely (2) Very Likely (3)

Q1 Avg Score: 2.53 (Scale of 5)

Q2 Avg Score: 2.24 (Scale of 3)

Correlation between difficulty scores: -0.61

Participants

Teams	Matching & Ranking	Description Generation
IMFD_IMPRESEE		✓
KSLAB		✓
KU_ISPL		✓
MMCUniAugsburg		✓
PICSOM		✓
RUC_AIM3	✓	✓

- 6 teams participated
  - 19 Description Generation Runs
  - 4 Matching and Ranking Runs

Description Generation

- Up to 4 runs in the *Description Generation* subtask.
- Metrics used for evaluation:
  - CIDEr (Consensus-based Image Description Evaluation)
  - SPICE (Semantic Propositional Image Caption Evaluation)
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering)
  - BLEU (BiLingual Evaluation Understudy)
  - STS (Semantic Textual Similarity)
  - DA (Direct Assessment), which is a crowdsourced rating of captions using Amazon Mechanical Turk (AMT)

Run Types

Training Data Types:

'**I**': Only image captioning datasets

'**V**': Only video captioning datasets

'**B**': Both image and video captioning datasets

Features Used:

'**V**': Visual features only

'**A**': Both audio and visual features

Submissions - Run Types

1 ‘VV’ (Video Data/Visual Feats)

Teams: 3

Runs: 9

2 ‘IV’ (Image Data/Visual Feats)

Teams: 1

Runs: 2

3 ‘BV’ (I+V Data/Visual Feats)

Teams: 1

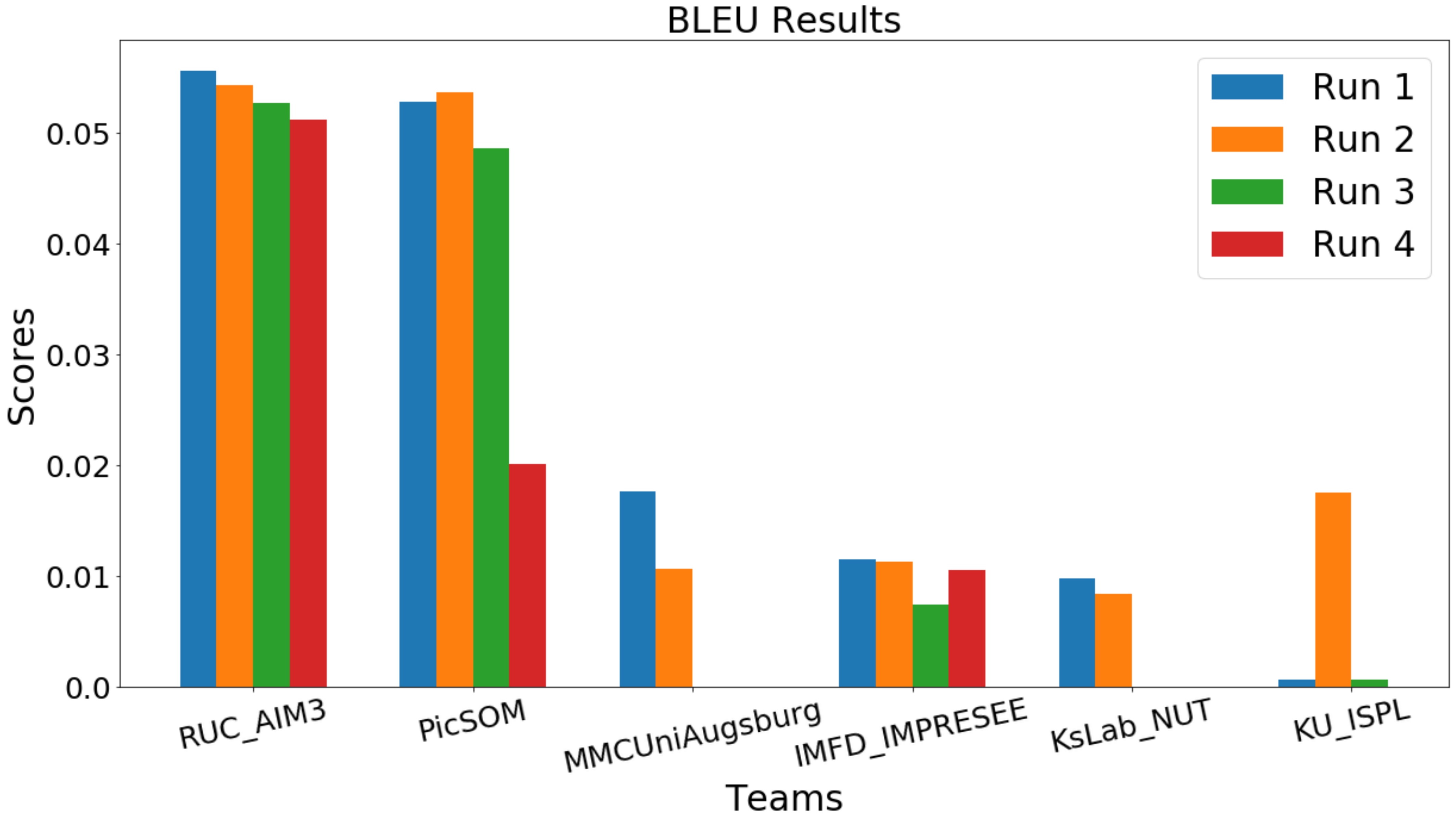
Runs: 4

4 ‘VA’ (Video Data/V+A Feats)

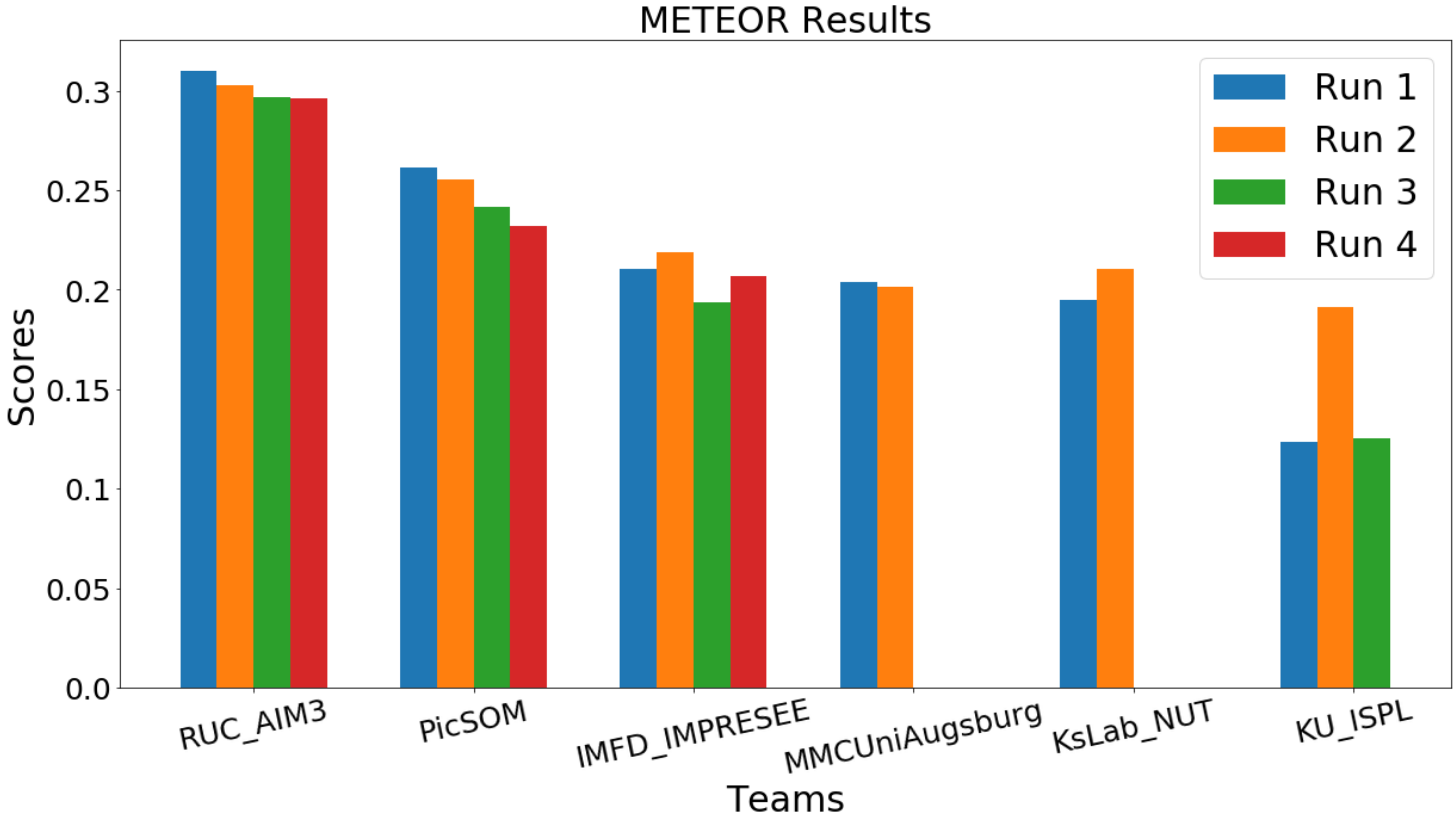
Teams: 1

Runs: 4

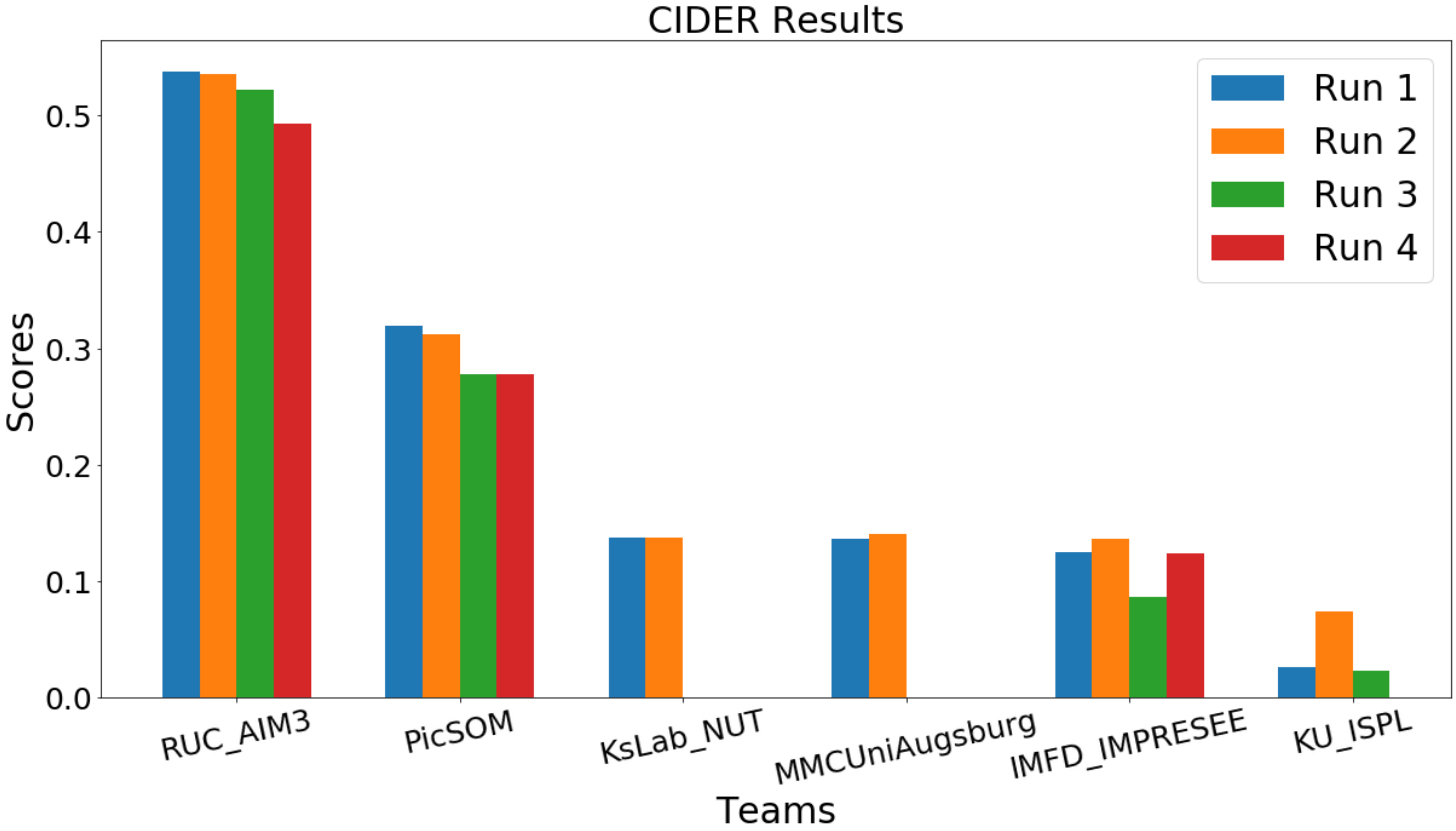
BLEU Results



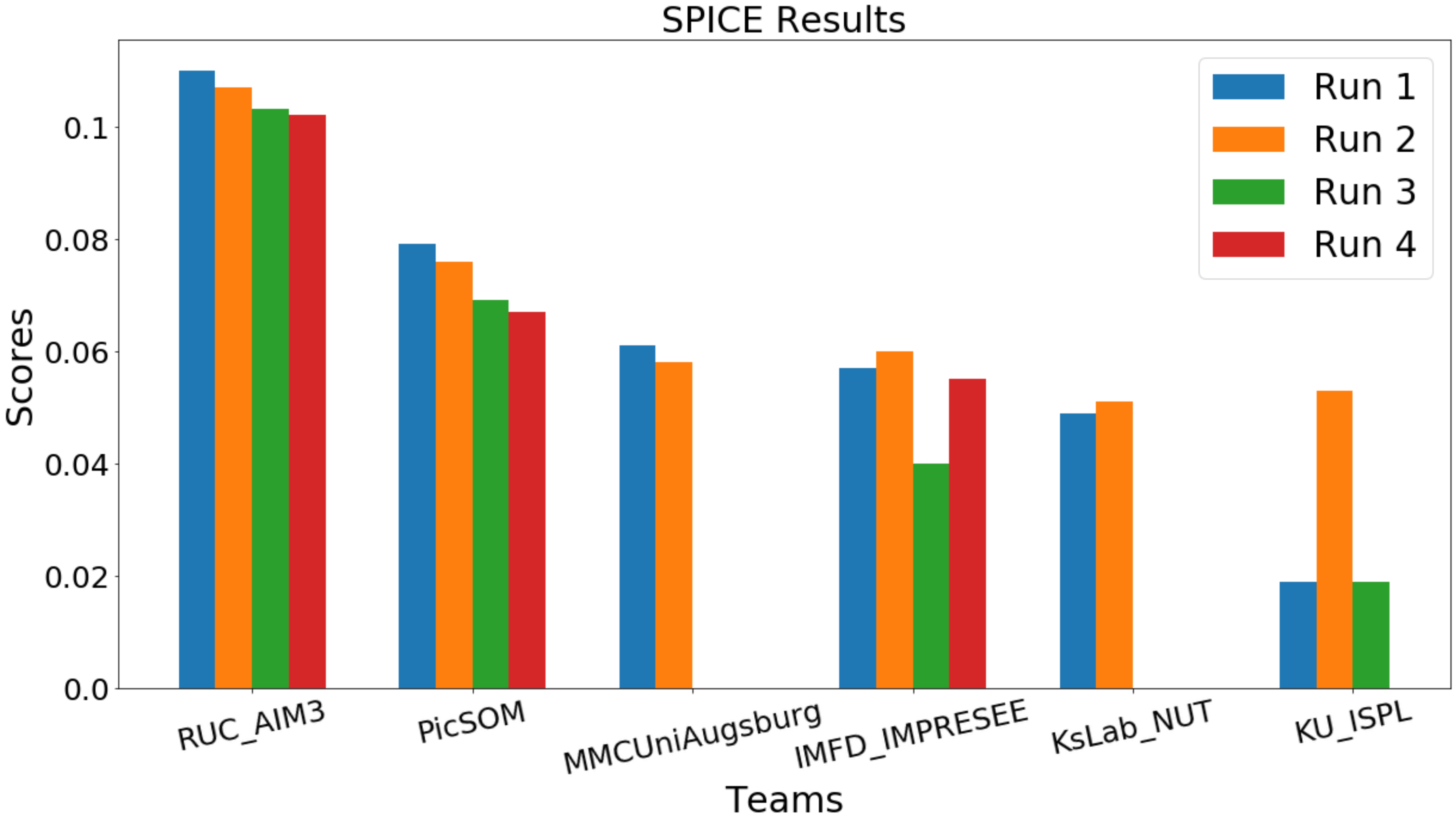
METEOR Results



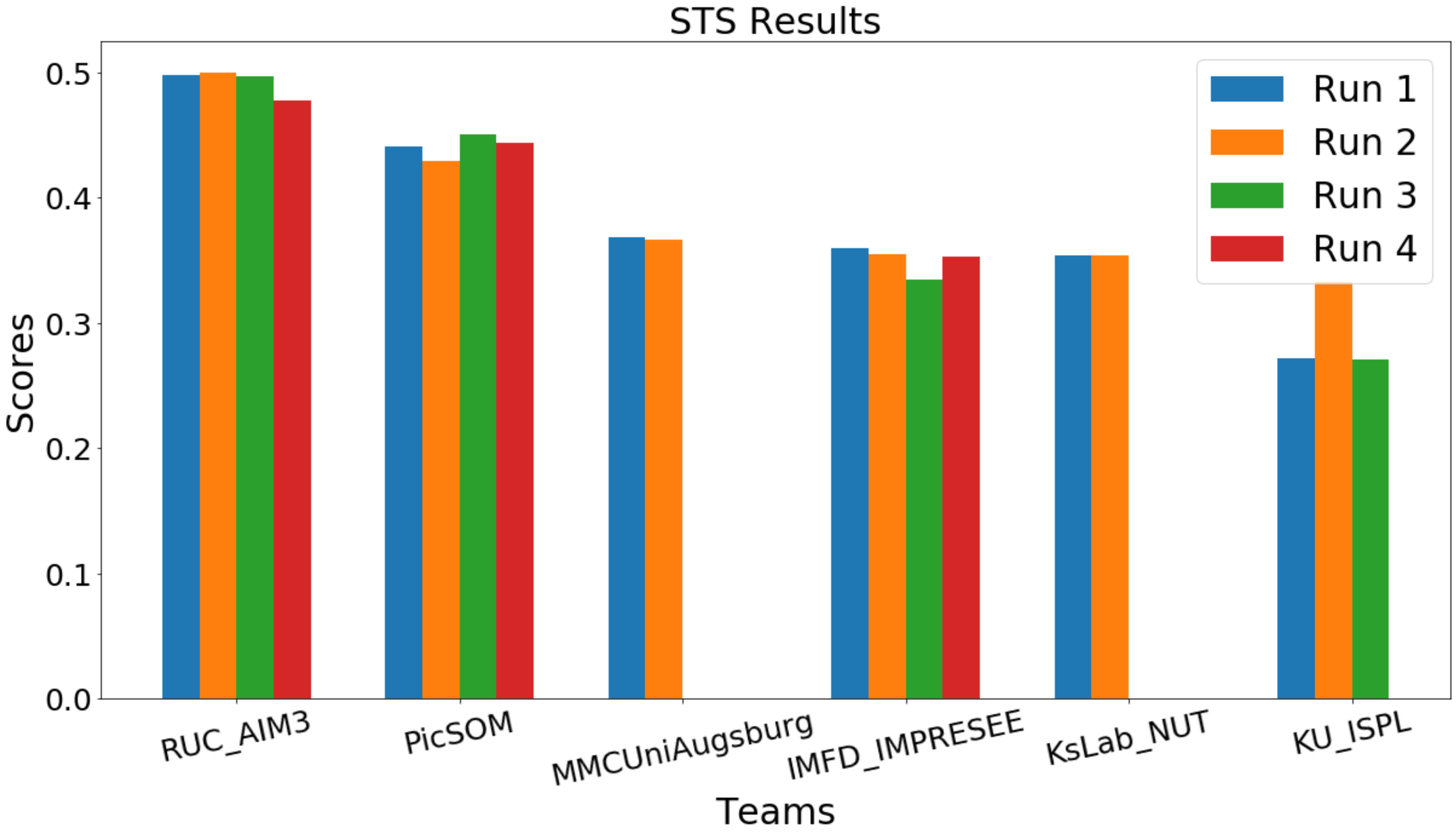
CIDER Results



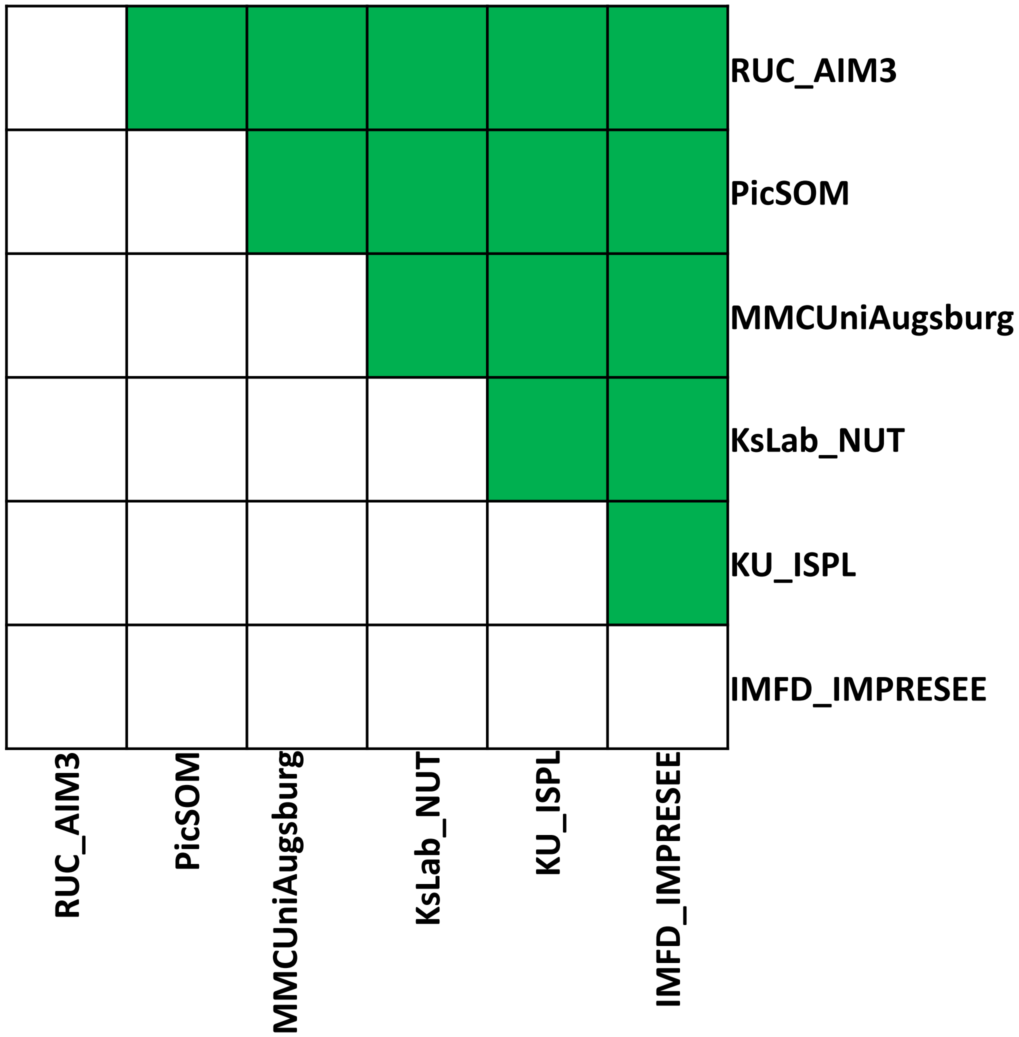
SPICE Results



Average STS Results



Significance Test - CIDEr



- Green squares indicate a significant “win” for the row over the column using the CIDEr metric.
- Significance calculated at $p<0.05$

Correlation of Run Scores – Automated Metrics

NIST

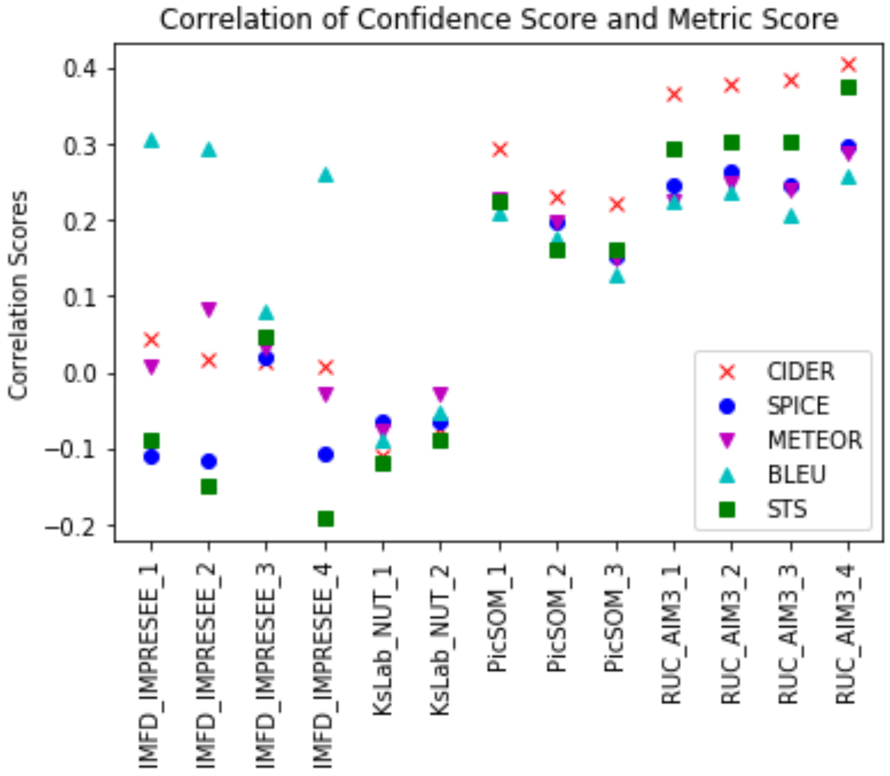
	CIDER_Score	CIDER-D_Score	SPICE_Score	METEOR_Score	BLEU_Score	Average_STS
CIDER_Score	1	0.992	0.959	0.948	0.911	0.961
CIDER-D_Score	0.992	1	0.953	0.945	0.929	0.942
SPICE_Score	0.959	0.953	1	0.986	0.889	0.963
METEOR_Score	0.948	0.945	0.986	1	0.893	0.969
BLEU_Score	0.911	0.929	0.889	0.893	1	0.914
STS	0.961	0.942	0.963	0.969	0.914	1

Correlation – Individual Video Scores

	CIDER_Score	CIDER-D_Score	SPICE_Score	METEOR_Score	BLEU_Score	Average_STS
CIDER_Score	1	0.908	0.588	0.654	0.524	0.535
CIDER-D_Score	0.908	1	0.6	0.652	0.508	0.622
SPICE_Score	0.588	0.6	1	0.69	0.543	0.637
METEOR_Score	0.654	0.652	0.69	1	0.562	0.682
BLEU_Score	0.524	0.508	0.543	0.562	1	0.458
STS	0.535	0.622	0.637	0.682	0.458	1

Confidence Scores

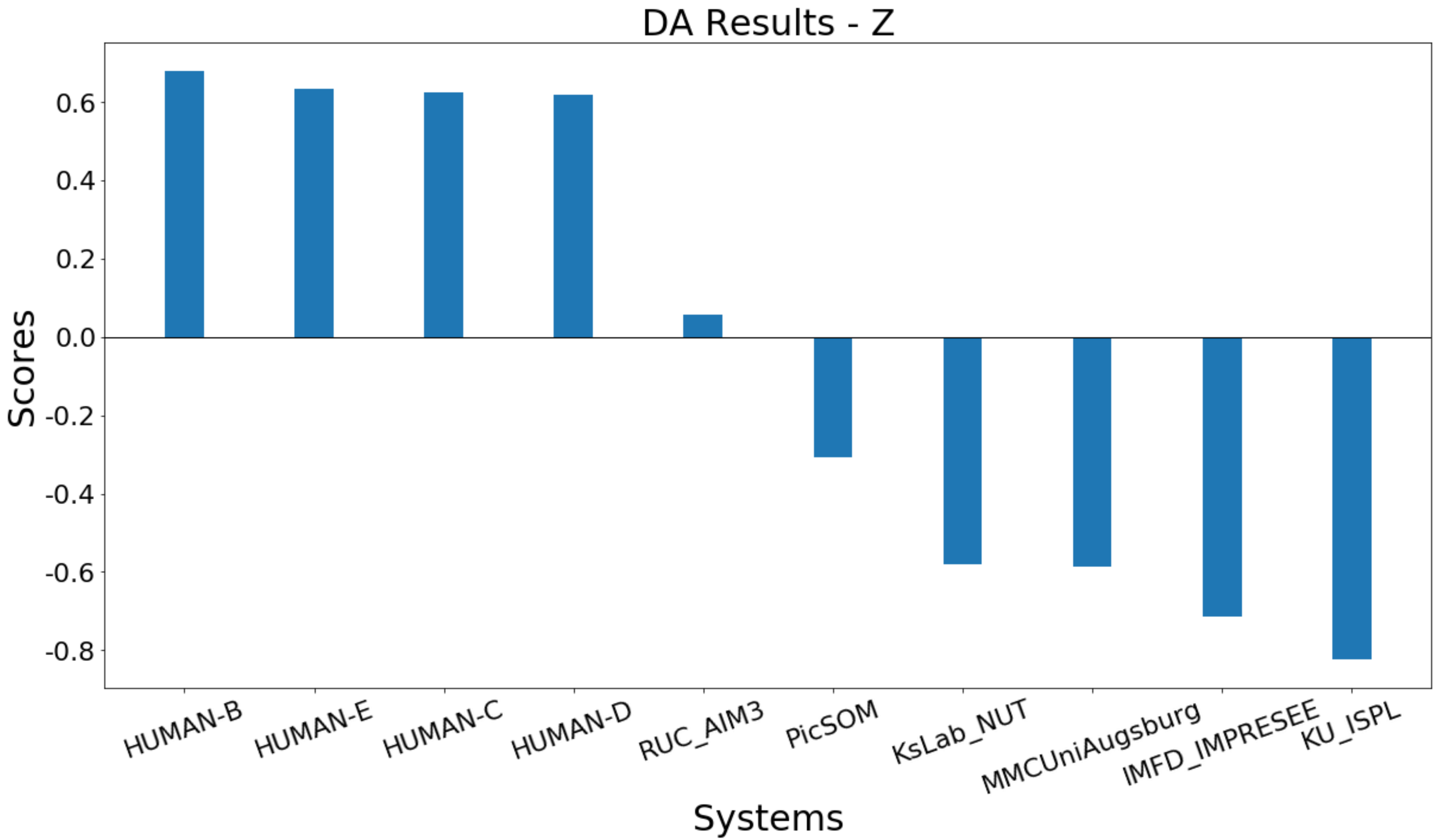
- Teams were asked to provide confidence scores for the generated sentences.
- Correlation was calculated between these confidence scores and evaluation metric scores for all runs.



Direct Assessment

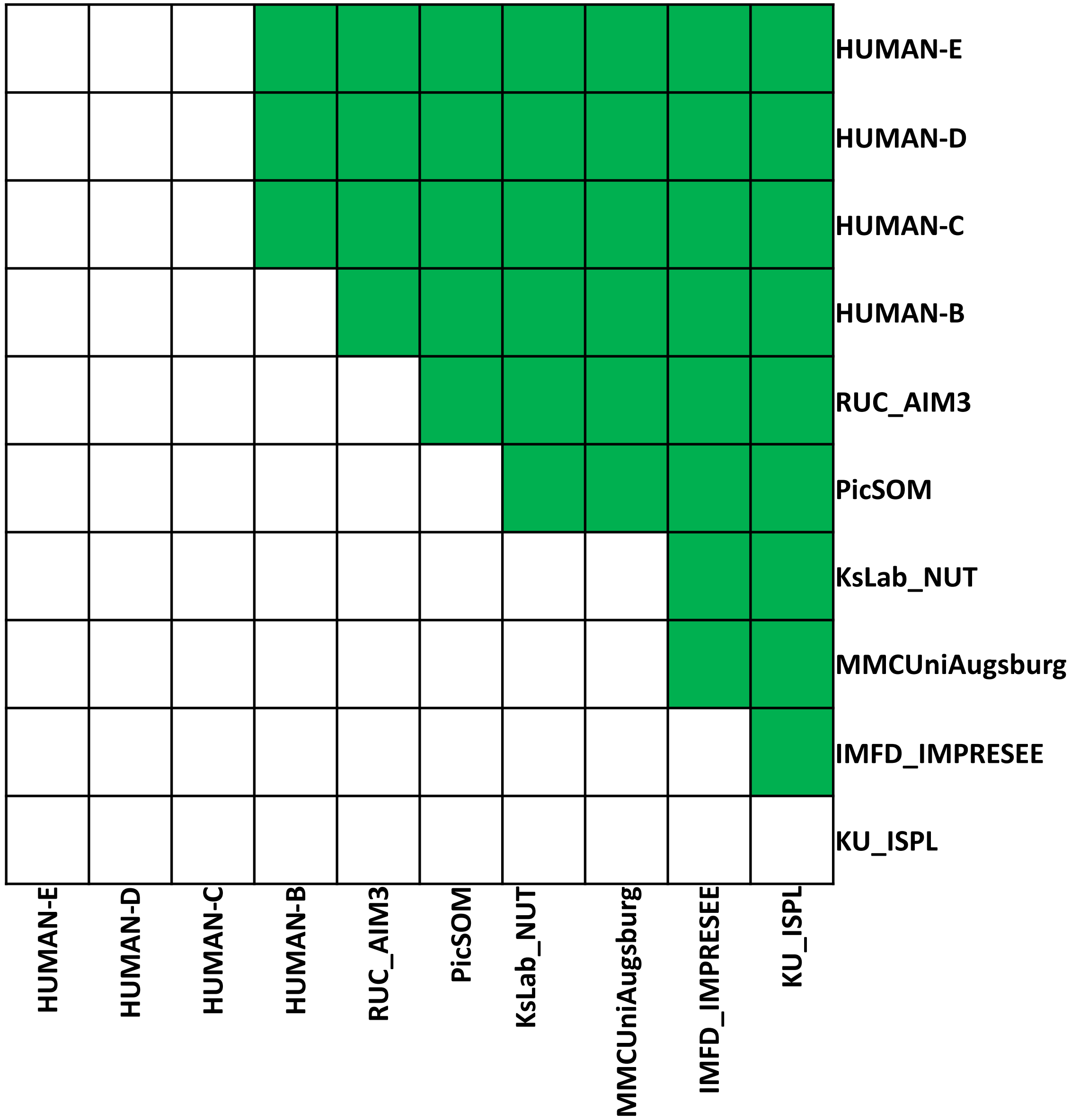
- DA uses crowdsourcing to evaluate how well a caption describes a video.
- Human evaluators rate captions on a scale of 0 to 100.
- DA conducted on only primary runs for each team.
- The DA score is reported as follows:
  - Z score is standardized per individual AMT worker’s mean and standard deviation score. The average Z score is then reported for each run.

DA Results - Z


DA Results - Z
Scores
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
HUMAN-B
HUMAN-E
HUMAN-C
HUMAN-D
RUC_AIM3
PicSOM
KsLab_NUT
MMCUniAugsburg
IMFD_IMPRESEE
KU_ISPL
Systems

DA Result - Significance

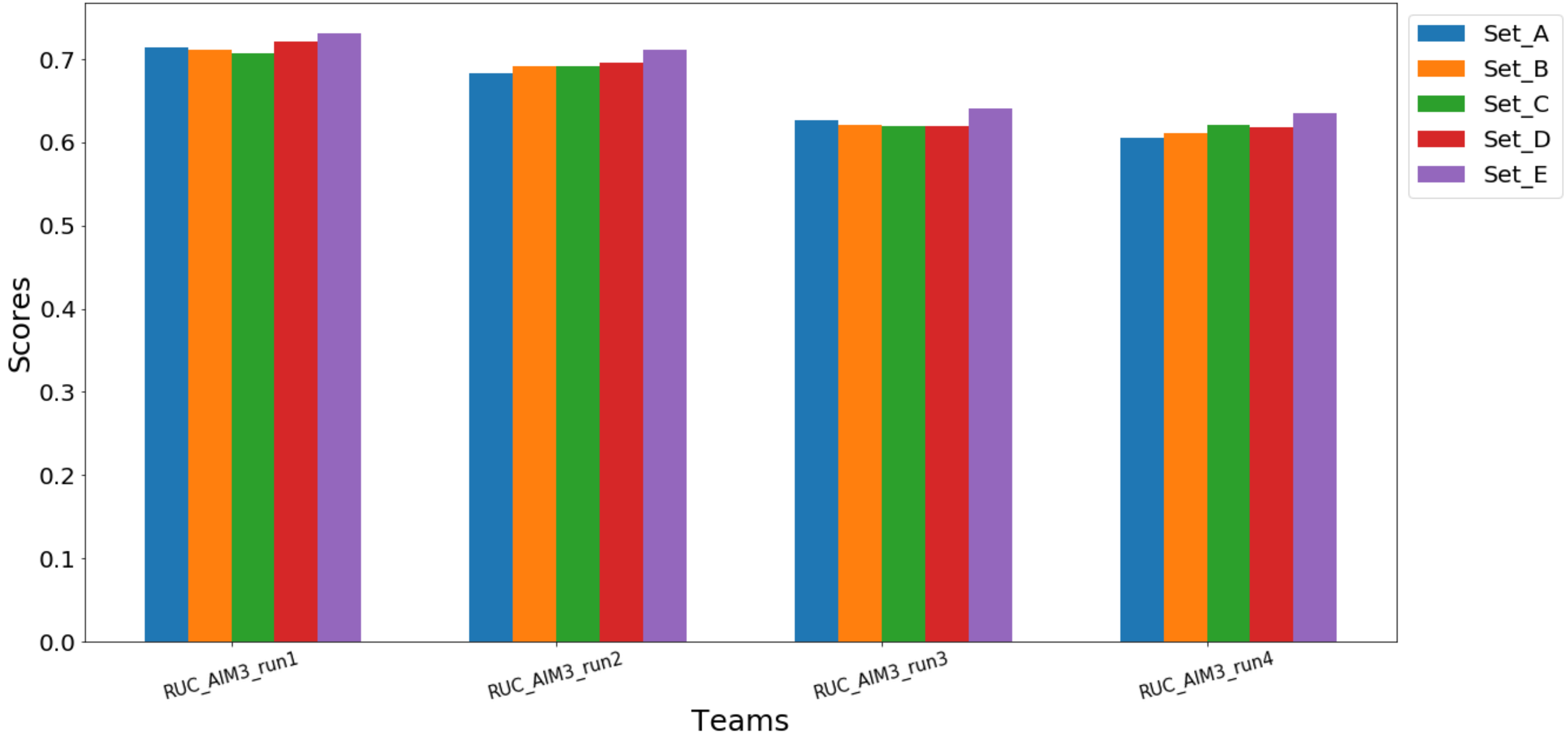
- Green squares indicate a significant “win” for the row over the column.
- No system yet reaches human performance.
- Amongst systems, RUC-AIM3 outperforms the rest, with significant wins. PicSOM is firmly in the second place.



Matching and Ranking

- This subtask was designated optional in 2019.
- Only 1 team (4 runs) submitted in 2020.
- Training was done using video datasets and both audio and visual features were used (‘VA’).
- Mean inverted rank used for evaluation.

Matching and Ranking Results



Matching and Ranking

- We included (obviously) fake sentences to check how they would be ranked. None of these sentences corresponded to any videos in the dataset.
- These fake sentences included:
  - Grammatically correct sentences that made no logical sense.
  - Grammatically incorrect sentences (e.g. random words just strung together).
- Median rank of fake sentences: 461 (Out of 1720)
- 13.5% of fake sentences ranked in top 100.
- 53% of fake sentences ranked in top 500.

High Level Overview of Some Approaches

KsLab_NUT

- Keyframes are extracted from the video
  - First and last frames + 3 frames with largest changes in features.
  - Image features extracted by a GoogLeNet. ImageNet dataset used for pre-training.
- Encoder-decoder method used to caption each frame.
  - Neural Image Captioning (NIC) Model.
  - MS COCO used for pre-training.
- Caption aggregation using extractive methods.
  - BERTSUM and LexRank used.
- Proposal to use abstractive methods in the future to improve scores.

KU_ISPL

- Different methods for each run.
- SA-LSTM used as baseline method (Run 1).
- Transformer and LSTM connected for runs 2 and 3.
- Attention mechanism used.
- Only TRECVID VTT data used for training.

MMCUniAugsburg

- Model based on Transformer architecture [1].
  - Modified to take videos as input by adding an image embedding layer and positional encoding.
  - Three datasets used for training:
    - Auto-captions on GIF
    - TRECVID-VTT
    - MSR-VTT
- Systems pretrained on merged datasets and fine tuned on TRECVID-VTT.
- Found significant improvement over traditional image captioning pipelines.

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, pp. 5998–6008, 2017.

Conclusion and Future Work

- This year we used a new video source – V3C2
- Lots of training sets are available.
- Need to increase visibility of the task. Dataset consolidated and made available to allow new teams to participate. (https://ir.nist.gov/tv_vtt_data/)
- The task will be renewed.
    - Upcoming changes will be discussed at the end of the session.

Thank you!