TRECVID 2021: 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

Description Generation System

Cars are racing on the road.

Description Generation:

Automatically generate a text description for a given video.

A group of men are playing _____.

Fill-in-the-Blank System

basketball

Fill-in-the-Blanks:

For a video and a corresponding sentence with missing word(s), find the most appropriate word(s) to fill in the blank.

Test Dataset (IRB # ITL-17-0025)

- 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.
- Dataset from 2020 onwards: 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.

Test Dataset

- Manual selection of videos.
  - We watched 9000+ videos.
  - Selected 1977 videos for annotation.
  - Subset of 300 videos will be used to measure system progress over 3 years.
- 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 10 assessors annotated the videos.
- Each video was annotated by 5 different assessors to get 5 captions.
- 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.

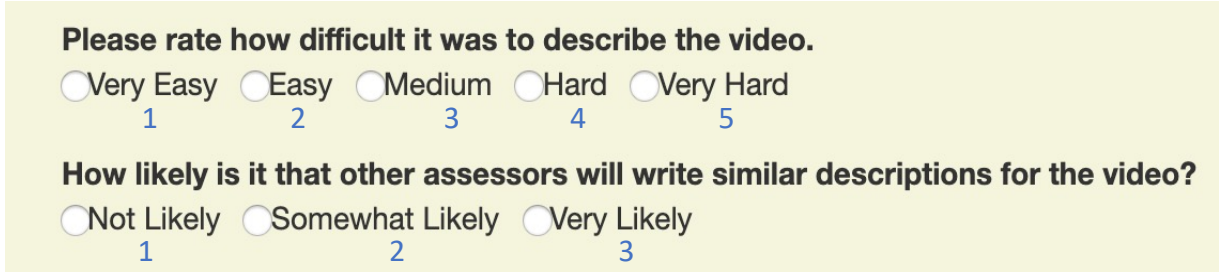
Annotation – Observations

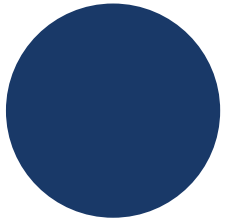
- Average sentence length (words per sentence) for each assessor:

Annotator	Avg. Length	# Videos
1	17.27	867
2	18.77	867
3	18.97	810
4	19.14	810
5	19.50	834
6	20.00	810
7	20.33	843
8	25.42	810
9	27.78	867
10	32.24	867

Avg. sentence length: 21.99 words

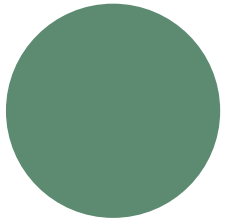
- Additional questions:





Q1 Avg Score: 2.71 (Scale of 5)

Q2 Avg Score: 2.15 (Scale of 3)



Correlation between difficulty scores: -0.59

Participants

Teams	Fill-in-the-Blanks	Description Generation
KSLAB		✓
MMCUniAugsburg		✓
RUC_AIM3	✓	✓
RUCMM	✓	✓
UEC		✓

- 5 teams participated
  - 15 Description Generation Runs
  - 3 Fill-in-the-Blanks Runs

Description Generation

- Up to 4 runs in the *Description Generation* subtask.
- Metrics used for evaluation:
  - CIDEr (Consensus-based Image Description Evaluation) [1]
  - SPICE (Semantic Propositional Image Caption Evaluation) [2]
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering) [3]
  - BLEU (BiLingual Evaluation Understudy) [4]
  - STS (Semantic Textual Similarity) [5]
  - DA (Direct Assessment), which is a crowdsourced rating of captions using Amazon Mechanical Turk (AMT) [6]

Run Types

Training Data Types:

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

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

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

Features Used:

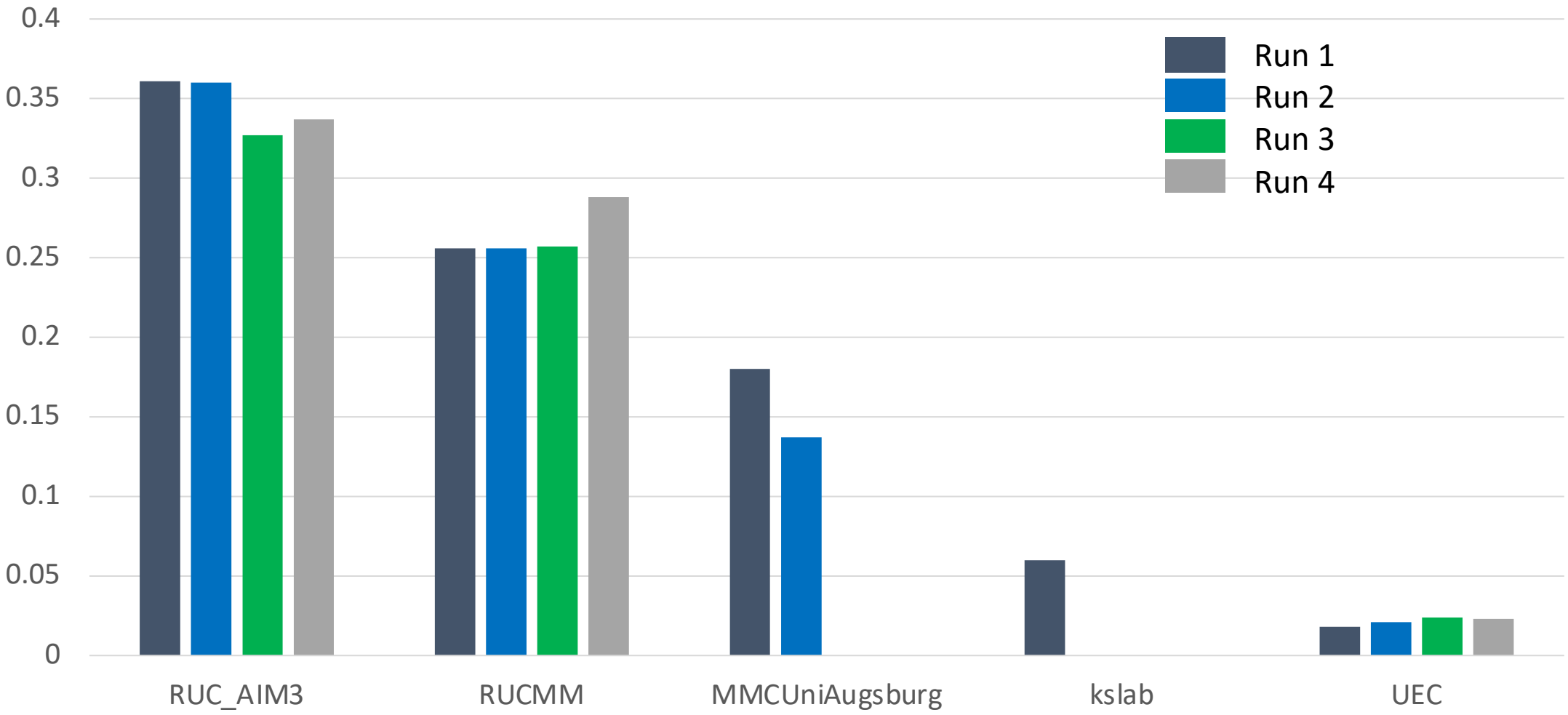
'V': Visual features only

'A': Both audio and visual features

Submissions - Run Types

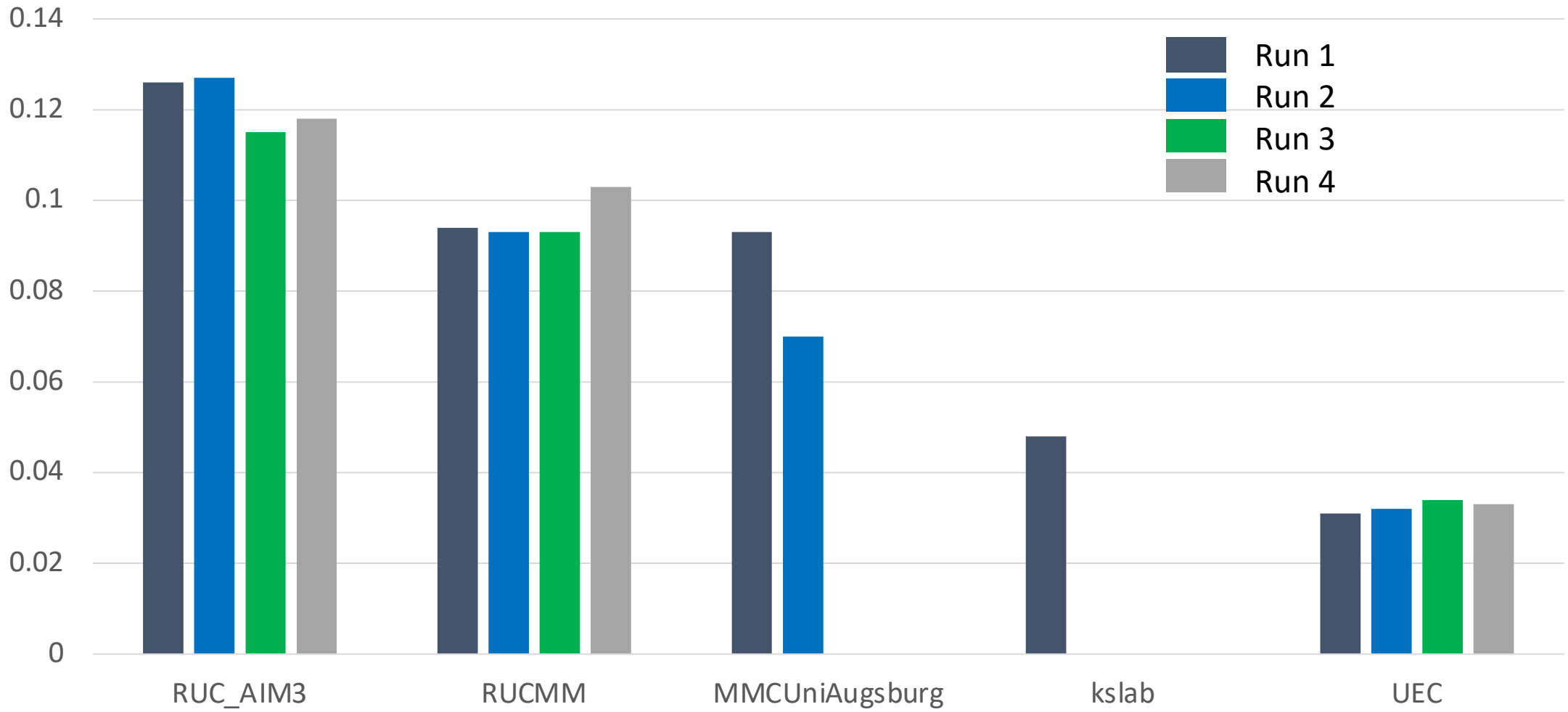
1 ‘VV’ (Video Data/Visual Feats) • RUC_AIM3 • RUCMM	**2 ‘IV’ (Image Data/Visual Feats)** • KsLab
3 ‘BV’ (I+V Data/Visual Feats) • UEC	**4 ‘VA’ (Video Data/V+A Feats)** • MMCUniAugsburg

CIDER-D Results

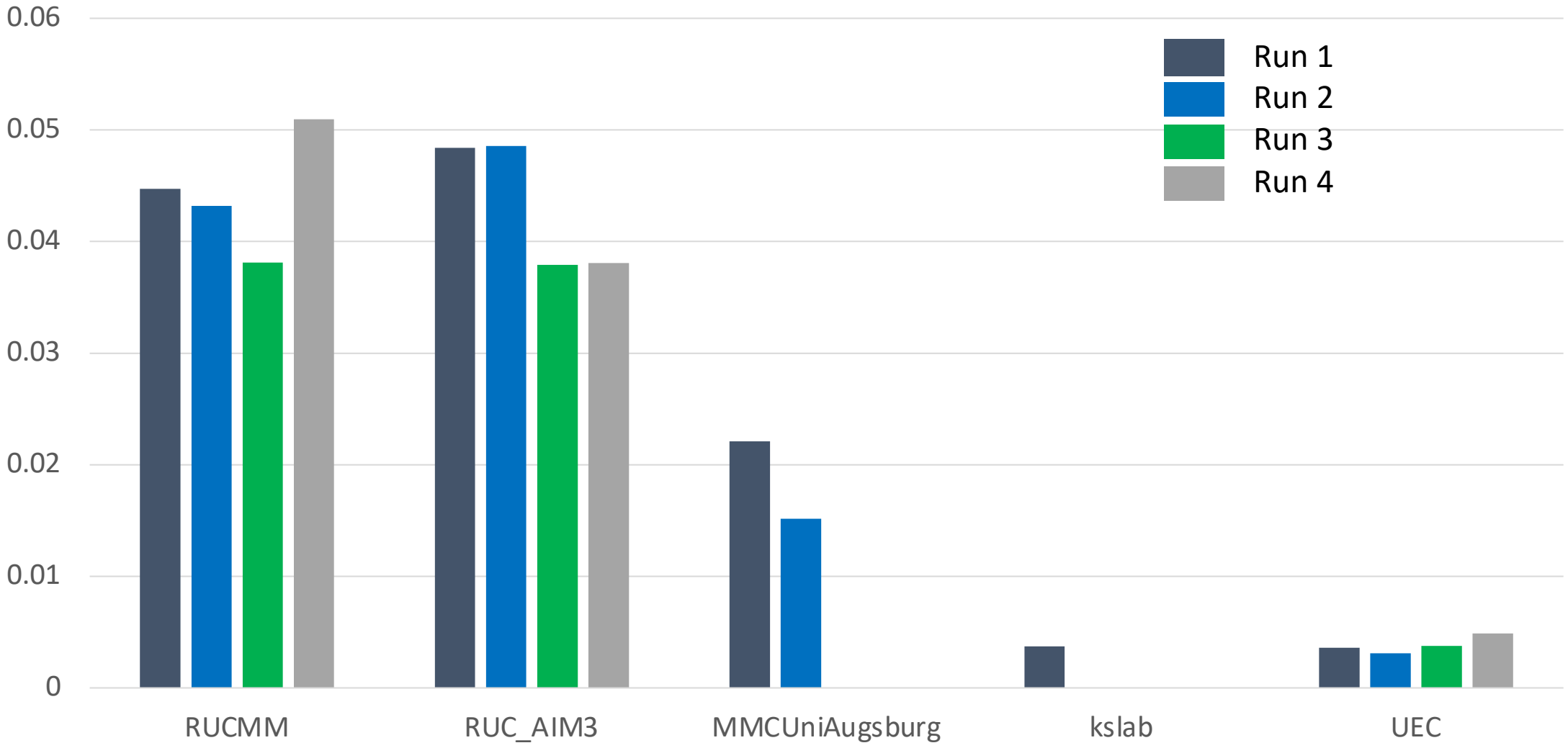


Note: For all metric scores, higher value is better.

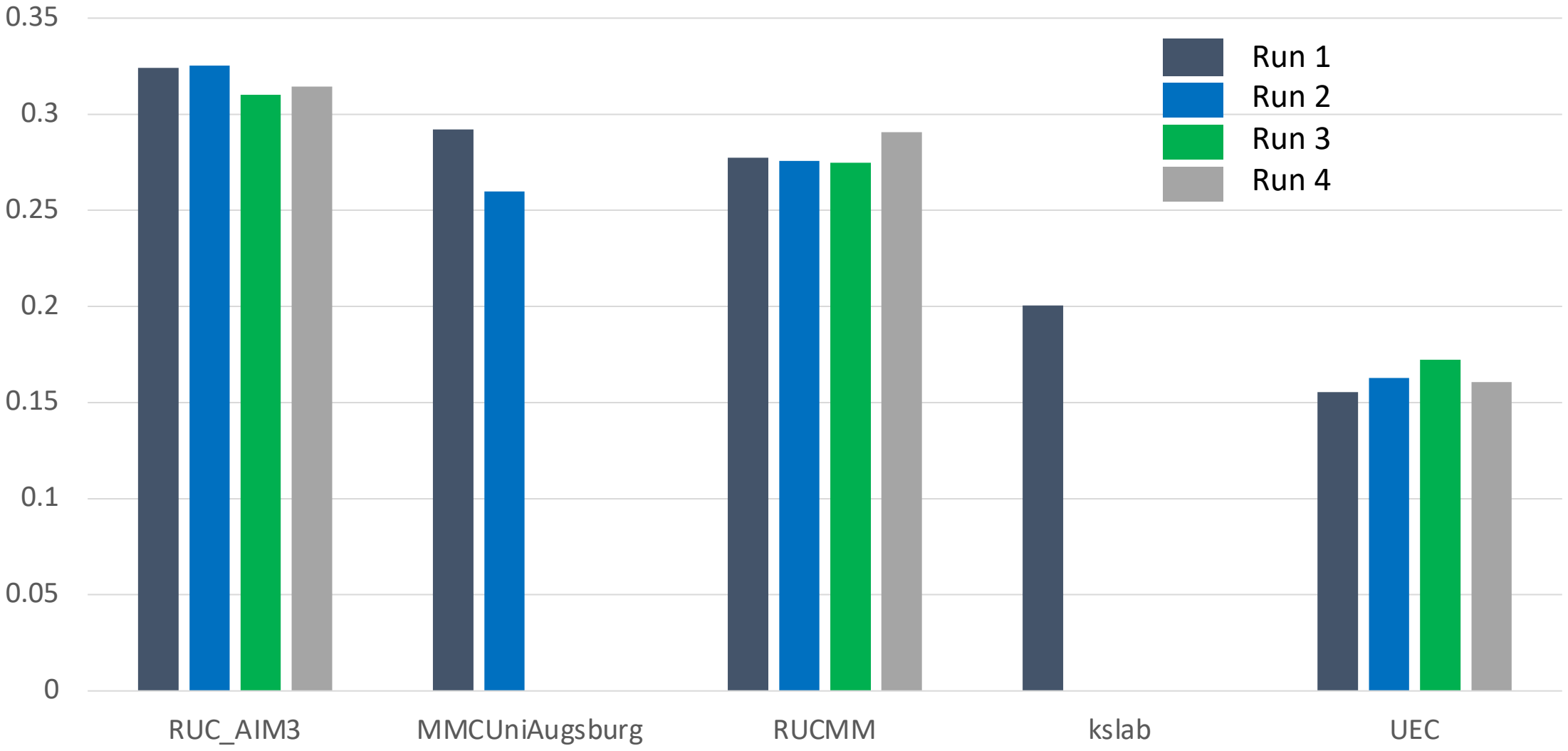
SPICE Results



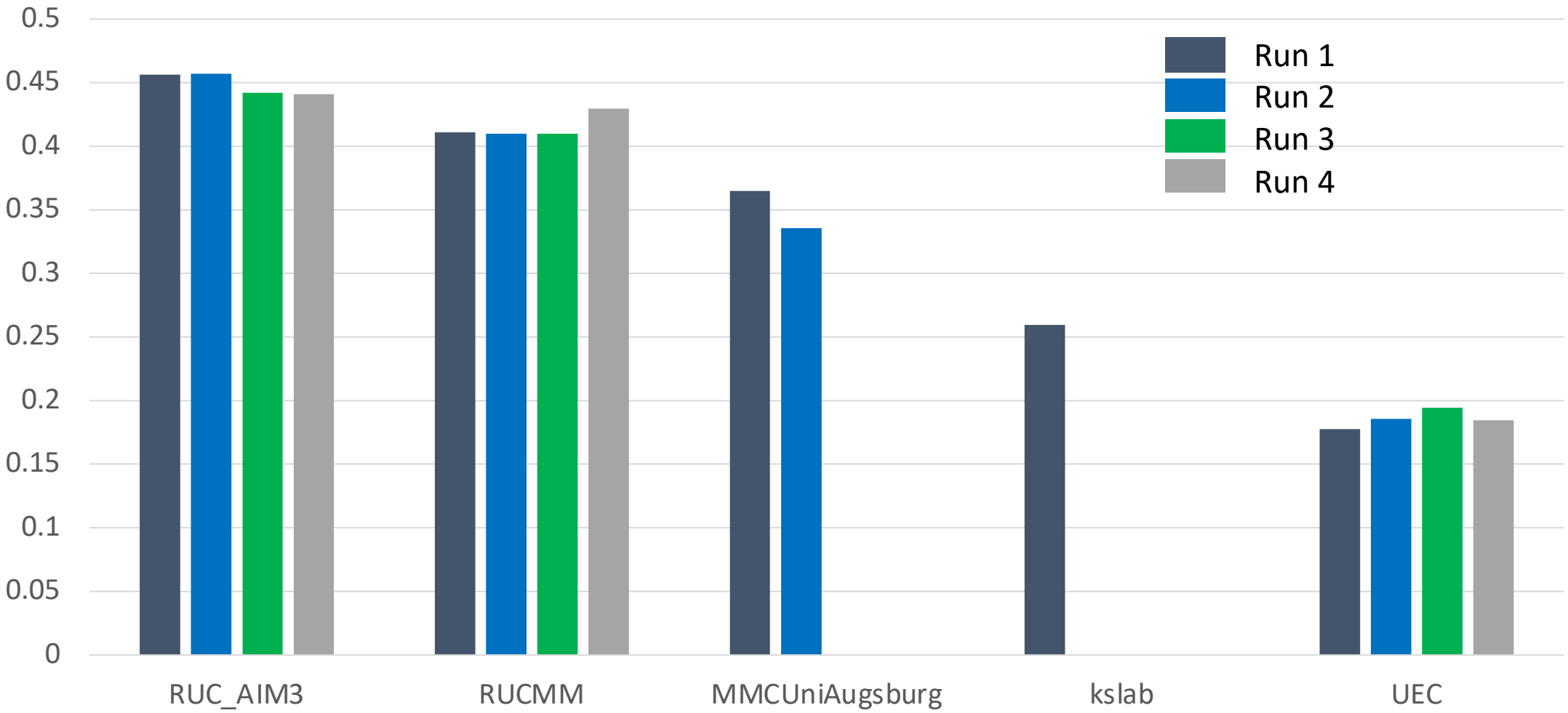
BLEU Results



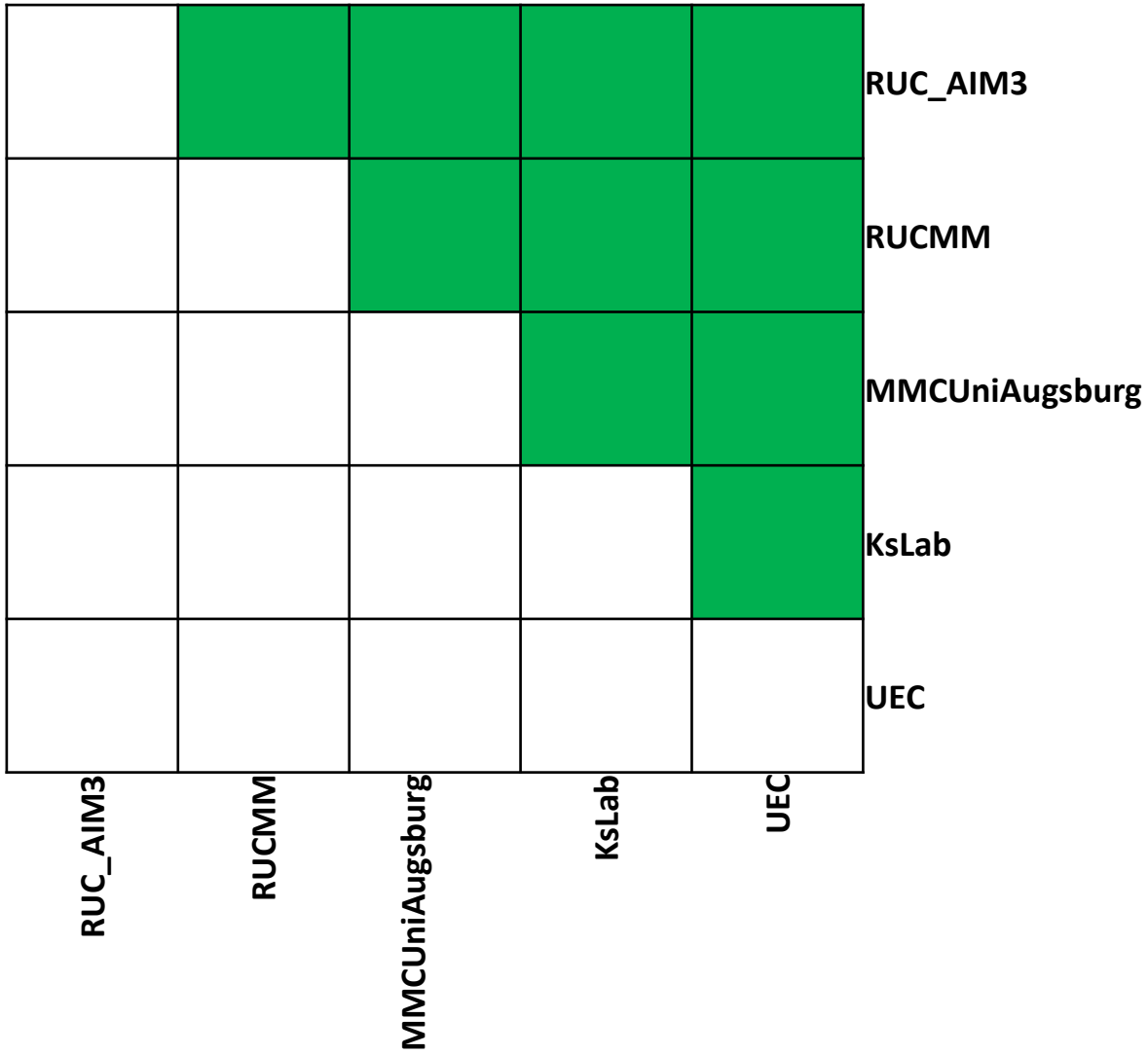
METEOR Results



STS Results


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Run 1
Run 2
Run 3
Run 4
RUC_AIM3
RUCMM
MMCUniAugsburg
kslab
UEC

Significance Test - CIDEr



- Green squares indicate a significant “win” for the row over the column using the CIDEr metric.
- Significance calculated at $p<0.05$.
- CIDEr-D, SPICE, BLEU, STS significance tests show the same results.

Correlation of Run Scores – Automated Metrics

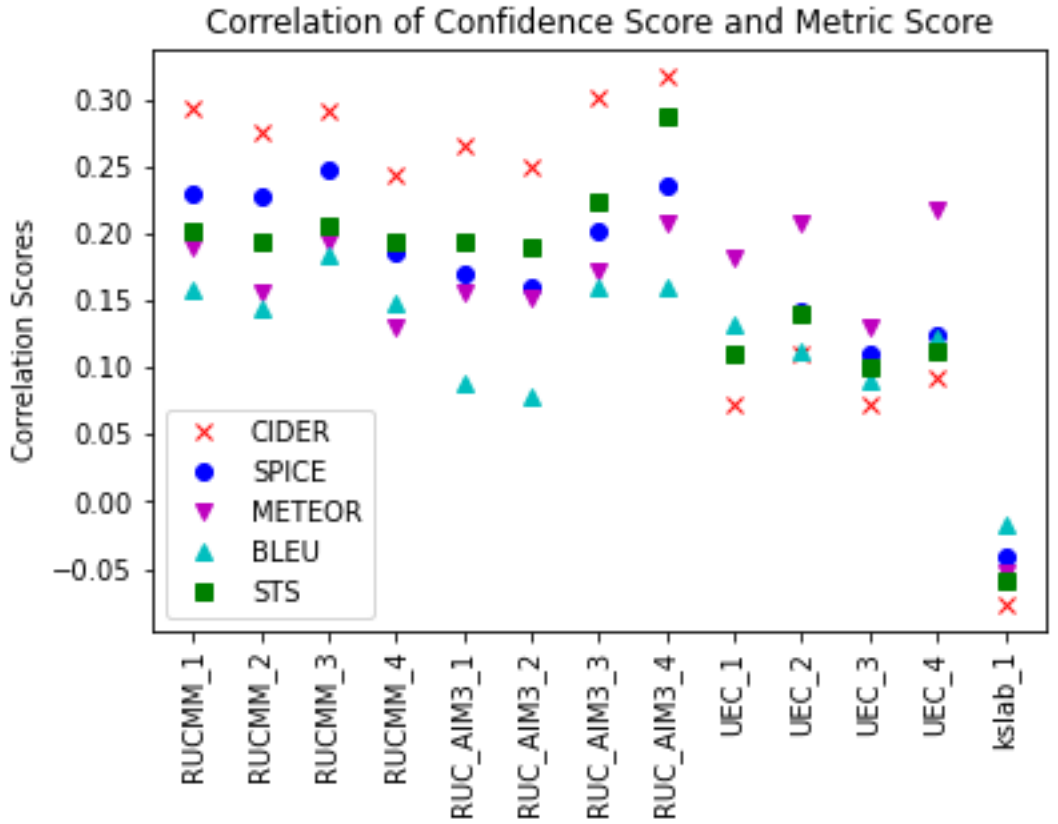
	CIDER_Score	CIDER-D_Score	SPICE_Score	METEOR_Score	BLEU_Score	STS
CIDER_Score	1	0.997	0.984	0.96	0.96	0.986
CIDER-D_Score	0.997	1	0.99	0.964	0.956	0.98
SPICE_Score	0.984	0.99	1	0.988	0.929	0.982
METEOR_Score	0.96	0.964	0.988	1	0.897	0.984
BLEU_Score	0.96	0.956	0.929	0.897	1	0.947
STS	0.986	0.98	0.982	0.984	0.947	1

Correlation – Individual Video Scores

	CIDER_Score	CIDER-D_Score	SPICE_Score	METEOR_Score	BLEU_Score	STS
CIDER_Score	1	0.906	0.673	0.701	0.633	0.691
CIDER-D_Score	0.906	1	0.65	0.695	0.629	0.607
SPICE_Score	0.673	0.65	1	0.734	0.623	0.711
METEOR_Score	0.701	0.695	0.734	1	0.646	0.724
BLEU_Score	0.633	0.629	0.623	0.646	1	0.533
STS	0.691	0.607	0.711	0.724	0.533	1

Confidence Scores

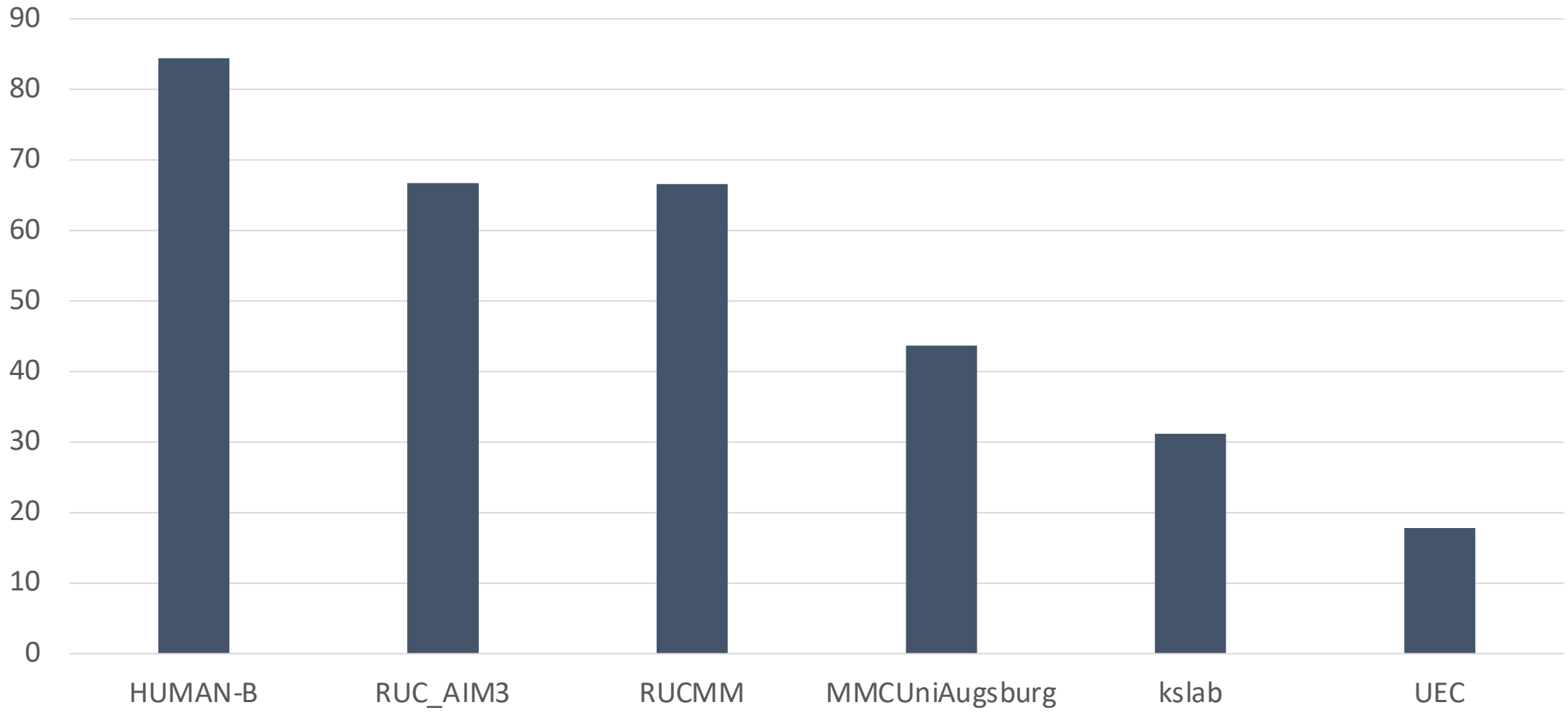
- Teams were asked to provide confidence scores for the generated sentences.
- Correlation was calculated between these confidence scores and evaluation metric scores for all runs.



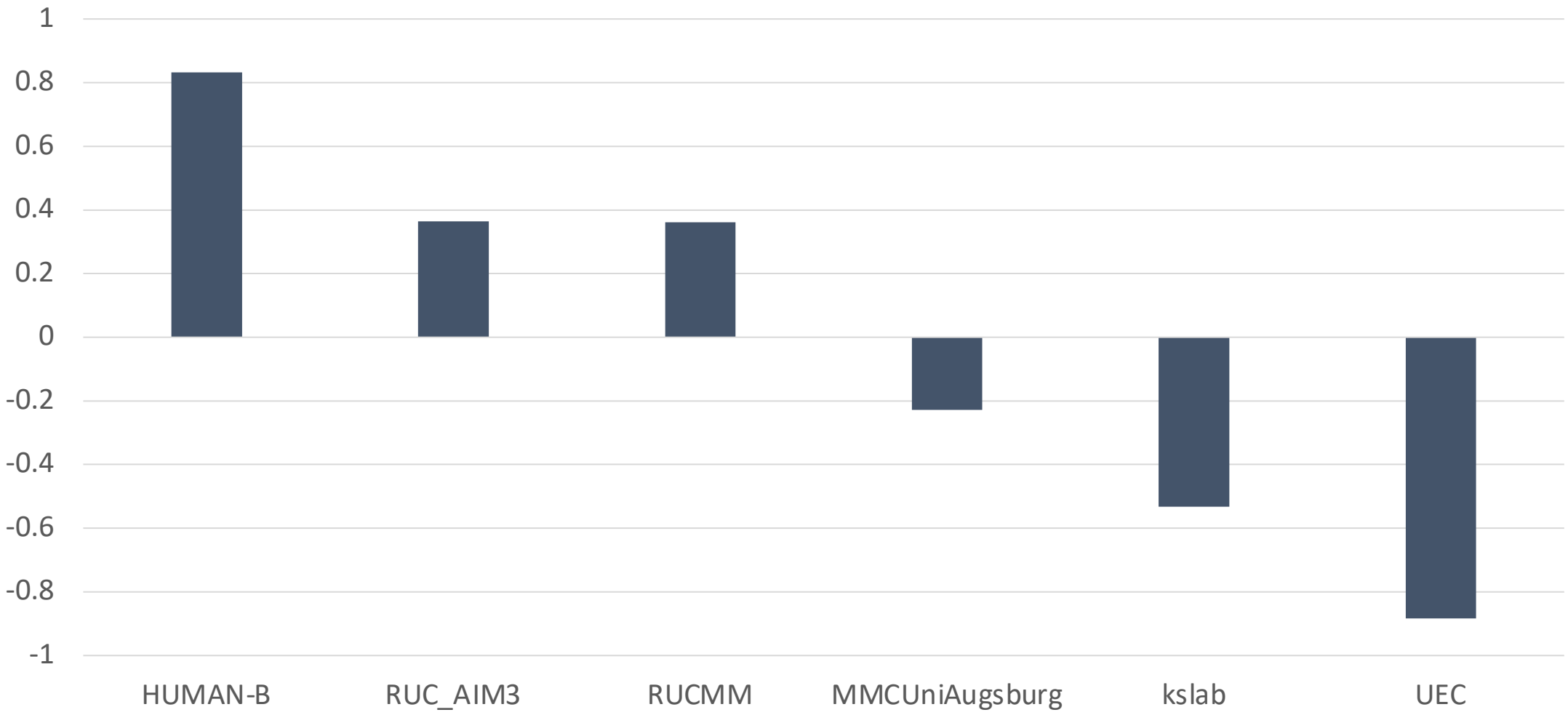
Direct Assessment (DA)

- 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:
  - Raw score is the average score for each run over all videos. It ranges between 0 and 100.
  - 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 - Raw

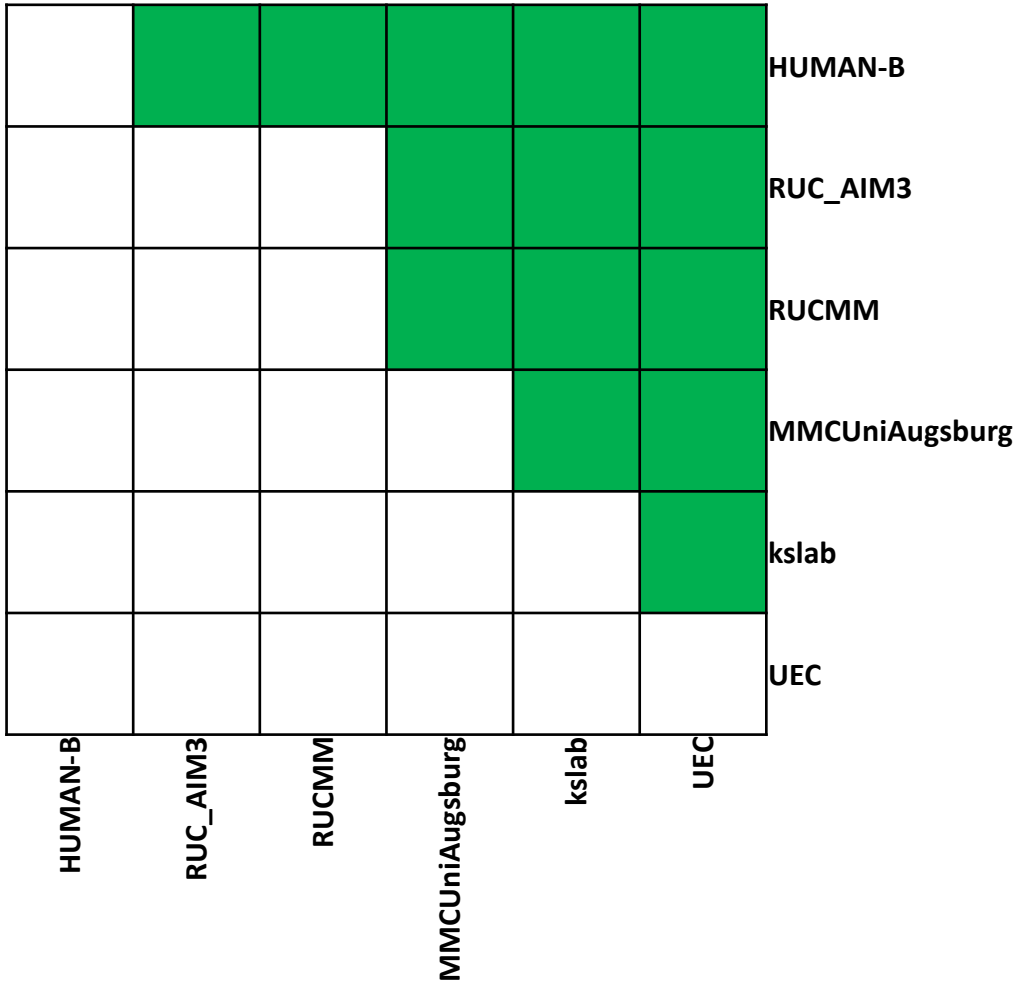


DA Results - Z


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
HUMAN-B
RUC_AIM3
RUCMM
MMCUniAugsburg
kslab
UEC

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 and RUCMM lead the others.



Easy Video Example

GT:

1. A woman in a black dress is playing the harp.
2. A young woman plays the harp next to a sparkling curtain.
3. An Asian girl plays a harp in a room.
4. A young woman is plucking along ta harp.
5. A dark haired Asian woman, wearing a black outfit and seated at a harp in front of a stand used for holding sheet music, gently plays the multi colored strings in a white room with a purple curtain

Submissions:

1: a man holding a surfboard and a cat in the background.
2. a woman is sitting down and playing a song on a harp .
3. a woman is playing a song on a harp in a room . and
4. a young woman is playing a song on a harp in a room
5. a young woman in a blue dress is playing a harp in a room
6. A young woman is playing a harp in a room
7. A woman in a black dress and a man are sitting at a table with a

Hard Video Example

GT:

1. A massive fireworks display, with many fireworks suspended by a series of overhead wires, is ignited in an outdoor setting
2. Ground explosions, that create small clouds of smoke, are taking place on an area with cables and hanging short cables, outdoors.
3. Steam is coming out of the ground near an electrical fence.
4. Firecrackers explode inside a wire fence outside creating a lot of smoke.
5. In a field during the day, there is a barbed wire fence and with massive explosions of fireworks creating a lot of smoke.

Submissions:

1. a group of birds are standing on a window sill .
2. a man in a mask is putting off a fire in a field in a field .
3. a fire is burning on a street at night . and
4. a person is hitting a large piece of metal with a chain in a cage
5. a man is standing in front of a gate while two men are using a fire extinguisher to clear a large amount of rain
6. a group of people are racing around a race track in a rain
7. A group of people are standing on a platform and they are spraying a fire
8. A person is using a hose to put out a fire in a cage

Fill in the Blanks

- Fill-in-the-Blanks subtask introduced this year.
- Participants are provided with a video and a corresponding sentence with word(s) missing. The goal is to predict the best words to complete the sentence.
- Up to 2 runs per team allowed.
- Two teams participated in the pilot with a total of 3 runs.
- Manual evaluation used for this subtask.
  - AMT workers shown video, sentence, and system output.
  - They score the word(s) on a scale of 100.

Fill in the Blanks - Example

- Sentence:

Male and female university athletes chant together on _____ to show that they are united.

- Ground Truth Answer:
  - a sports field

- System Answers:
  - a sports field
  - huddle

Fill in the Blanks - Evaluation

Teams	AVERAGE-Z
HUMAN	0.420
RUC_AIM3_RUN2	0.173
RUC_AIM3_RUN1	0.130
RUCMM	-0.102

- Clusters are separated by the orange lines.
- Lower ranked clusters are significantly outperformed by higher ranked clusters.
- Wilcoxon rank sum test used with $p<0.05$.

Fill in the Blanks - Evaluation

- Character n-gram F-score used for automatic evaluation to compare with the manual evaluation. [7]
- Only a single ground truth per sentence available.
- CHRF scores with 4-gram and 6-gram shown.

Teams	Scores C-4	C-6
RUC_AIM3_RUN1	44.06	36.91
RUC_AIM3_RUN2	41.89	34.98
RUCMM	9.94	7.00

High Level Overview of Some Approaches

MMCUniAugsburg

- Model based on Transformer architecture [8].
- Image and audio embedding layers used along with positional encoding.
- Both 2D and 3D visual features (extracted using I3D) used.
- Train on VATEX and 90% TRECVID-VTT data.
- Models are fine tuned with self critical sequence training that optimizes CIDEr and BLEU-4 metrics.

UEC

- Fine tune the image captioning model of [9].
- Use pretrained ResNet-101 model to extract visual features.
- Captioning model uses attention mechanism.
- The model is trained on COCO and TRECVID datasets.

Conclusion and Future Work

- This was the second year using the V3C2 test data.
- Lots of training sets are available.
- Matching and ranking subtask completely phased out.
- Fill-in-the-Blank subtask was introduced.
- We have also introduced a 3-year progress video dataset.
  - 300 videos were selected as progress videos.
  - The results of algorithms over 3 years will be compared on the progress dataset to measure progress. We hope teams will be resubmitting next year.
  - The ground truth of these videos will not be made public till 2023.

References

[1] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." European conference on computer vision. Springer, Cham, 2016.

[3] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005.

[4] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[5] Han, L., Kashyap, A., Finin, T.,Mayfield, J., and Weese, J. (2013).UMBCEBIQUITY-CORE: Semantic Textual Similarity Systems. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics, volume 1, pages 44–52

[6] Graham, Yvette, George Awad, and Alan Smeaton. "Evaluation of automatic video captioning using direct assessment." PloS one 13.9 (2018): e0202789.

[7] Popović, Maja. "chrF: character n-gram F-score for automatic MT evaluation." *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 2015.

[8] 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.

[9] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. arXiv preprint arXiv:1803.04376, 2018.

Thank you!