TRECVID 2018

Video to Text Description

Asad A. Butt
NIST

George Awad
NIST; Dakota Consulting, Inc

Alan Smeaton
Dublin City University

NIST
National Institute of Standards and Technology

Goals and Motivations

- ✓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

TASKS

- Systems are asked to submit results for two subtasks:
  1. Matching & Ranking:

     Return for each URL a ranked list of the most likely text description from each of the five sets.
  2. Description Generation:

     Automatically generate a text description for each URL.

Video Dataset

- Crawled 50k+ Twitter Vine video URLs.
- Max video duration == 6 sec.
- A subset of 2000 URLs (quasi) randomly selected, divided amongst 10 assessors.
  - Significant preprocessing to remove unsuitable videos.
- Final dataset included 1903 URLs due to removal of videos from Vine.

Steps to Remove Redundancy

- Before selecting the dataset, we clustered videos based on visual similarity.
  - Used a tool called SOTU [1], which used Visual Bag of Words to cluster videos with 60% similarity for at least 3 frames.
  - Resulted in the removal of duplicate videos, as well as those which were very visually similar (e.g. soccer games), resulting in a more diverse set of videos.

[1] Zhao, Wan-Lei and Ngo Chong-Wah. "SOTU in Action." (2012).

Dataset Cleaning

- Dataset Creation Process: Manually went through large collection of videos.
  - Used list of commonly appearing videos from last year to select a diverse set of videos.
  - Removed videos with multiple, unrelated segments that are hard to describe.
  - Removed any animated (or otherwise unsuitable) videos.
- Resulted in a much cleaner dataset.

Annotation Process

- Each video was annotated by 5 assessors.
  - Annotation guidelines by NIST:
    - For each video, annotators were asked to combine 4 facets *if applicable*:
      - Who is the video describing (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 – Observations

1. Different assessors provide varying amount of detail when describing videos. Some assessors had very long sentences to incorporate all information, while others gave a brief description.
2. Assessors interpret scenes according to cultural or pop cultural references, not universally recognized.
3. Specifying the time of the day was often not possible for indoor videos.
4. Given the removal of videos with multiple disjointed scenes, assessors were better able to provide descriptions.

Sample Captions of 5 Assessors

1. Orange car #1 on gray day drives around curve in road race test.
2. Orange car drives on wet road curve with observers.
3. An orange car with black roof, is driving around a curve on the road, while a person, wearing grey is observing it.
4. The orange car is driving on the road and going around a curve.
5. Advertisement for automobile mountain race showing the orange number one car navigating a curve on the mountain during the race in the evening; an individual is observing the vehicle dressed in jeans and cold weather coat.

1. A woman lets go of a brown ball attached to overhead wire that comes back and hits her in the face.
2. In a room, a bowling ball on a string swings and its a woman with a white shirt on in the face.
3. During a demonstration a white woman with black hair wearing a white top and holding a ball tether to a line from above as the demonstrator tells her to let go of the ball which returns on its tether and hits the woman in the face.
4. A man in blue holds a ball on a cord and lets it swing, and it comes back and hits a woman in white in the face.
5. A young girl, before an audience of students, allows a pendulum to swing from her face and all are surprised when it returns to strike her.

2018 Participants (12 teams finished)

	Matching & Ranking (26 Runs)	Description Generation (24 Runs)
INF	✓	✓
KSLAB	✓	✓
KU_ISPL	✓	✓
MMSys_CCMIP	✓	✓
NTU_ROSE	✓	✓
PicSOM		✓
UPCer		✓
UTS_CETC_D2DCRC_CAI	✓	✓
EURECOM	✓	
ORAND	✓	
RUCMM	✓	
UCR_VCG	✓	

Sub-task 1: Matching & Ranking

Person reading newspaper outdoors at daytime

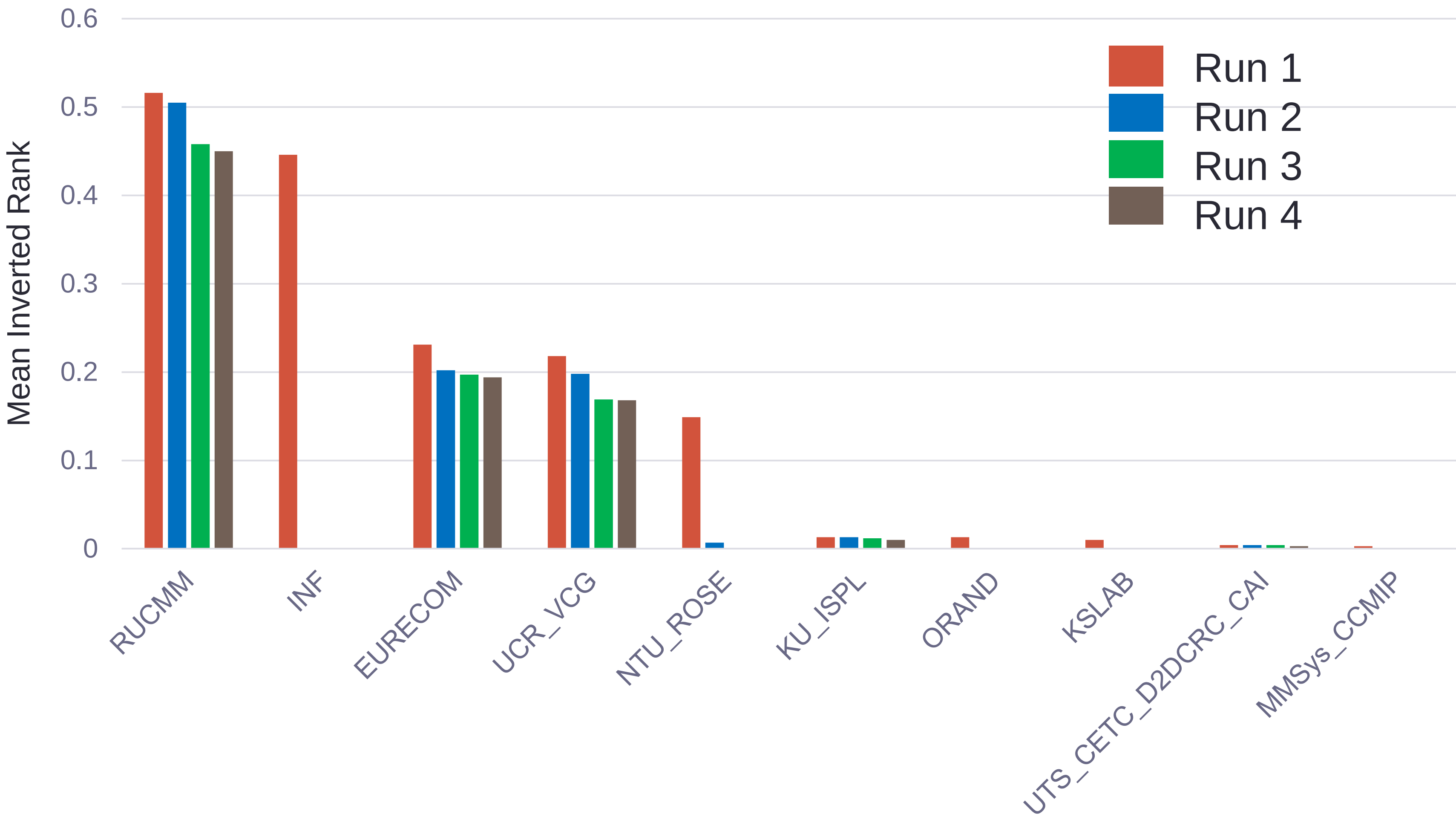
Person playing golf outdoors in the field

Three men running in the street at daytime

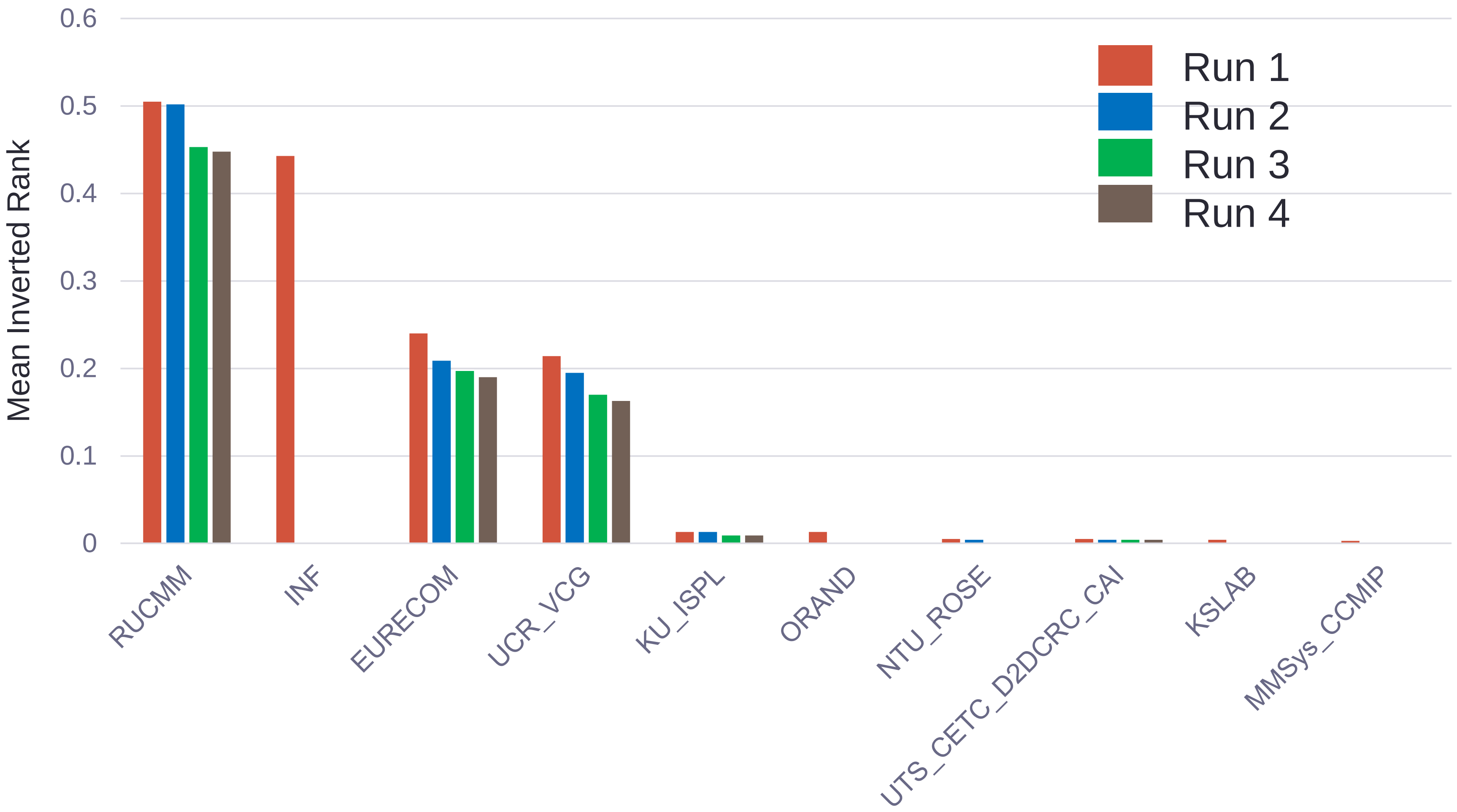
Two men looking at laptop in an office

- Up to 4 runs per site were allowed in the *Matching & Ranking* subtask.
- Mean inverted rank used for evaluation.
- Five sets of descriptions used.

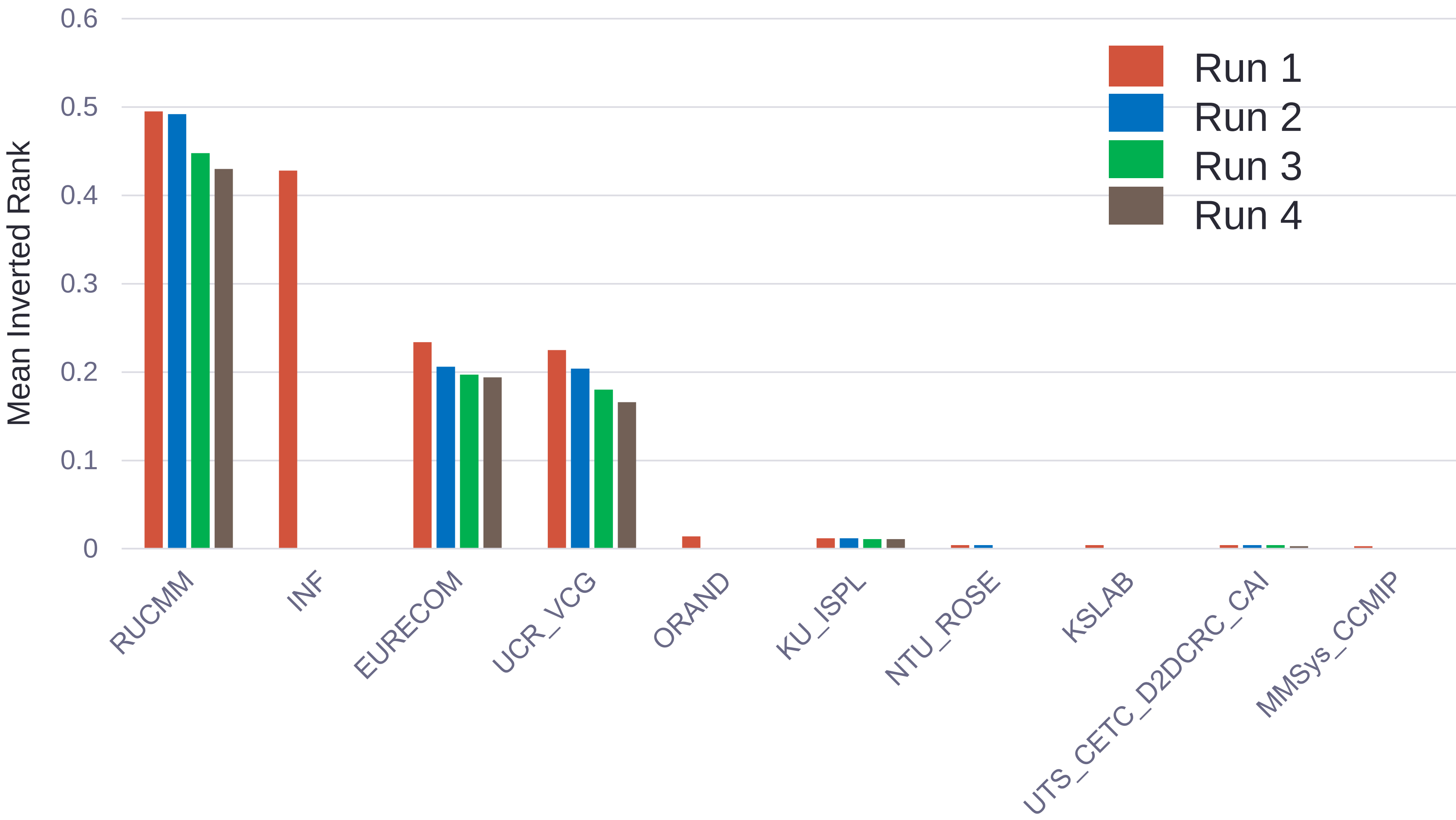
Matching & Ranking Results – Set A



Matching & Ranking Results – Set B

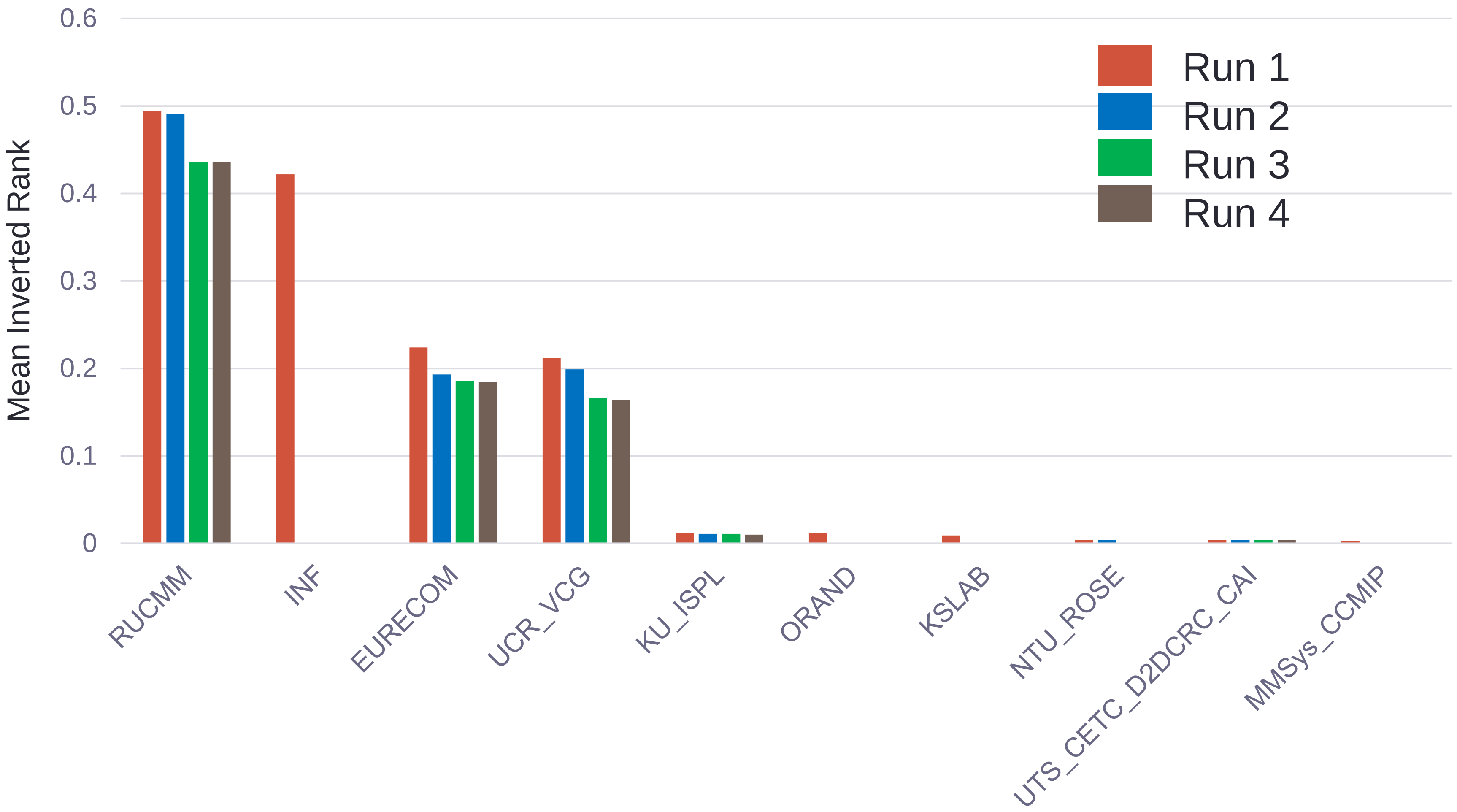


Matching & Ranking Results – Set C



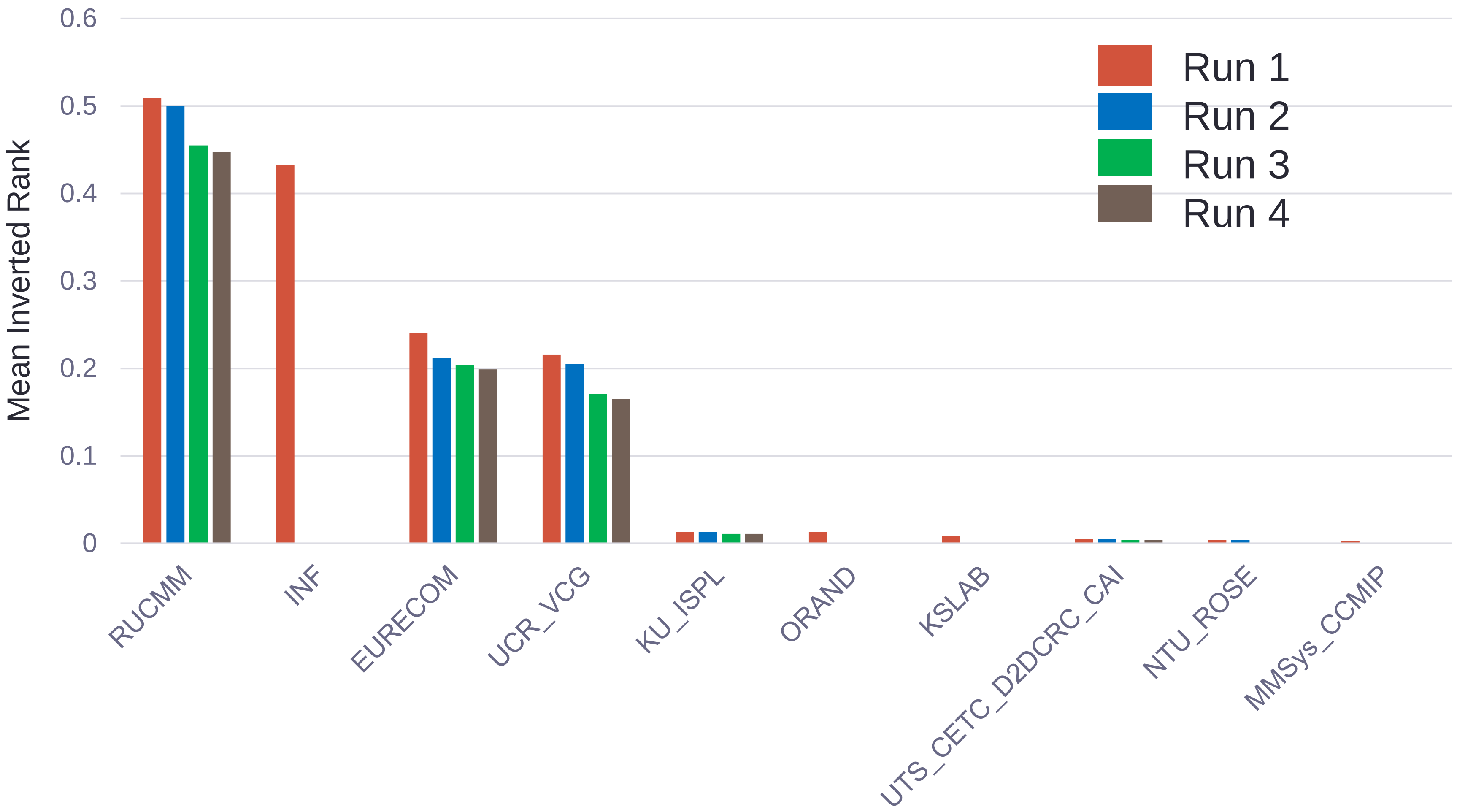
NIST National Institute of Standards and Technology

Matching & Ranking Results – Set D



NIST
National Institute of Standards and Technology

Matching & Ranking Results – Set E



Systems Rankings for each Set

A	B	C	D	E
RUCMM	RUCMM	RUCMM	RUCMM	RUCMM
INF	INF	INF	INF	INF
EURECOM	EURECOM	EURECOM	EURECOM	EURECOM
UCR_VCG	UCR_VCG	UCR_VCG	UCR_VCG	UCR_VCG
NTU_ROSE	KU_ISPL	ORAND	KU_ISPL	KU_ISPL
KU_ISPL	ORAND	KU_ISPL	ORAND	ORAND
ORAND	NTU_ROSE	NTU_ROSE	KSLAB	KSLAB
KSLAB	UTS_CETC_D2DCR C_CAI	KSLAB	NTU_ROSE	UTS_CETC_D2DCR C_CAI
UTS_CETC_D2DCR C_CAI	KSLAB	UTS_CETC_D2DCR C_CAI	UTS_CETC_D2DCR C_CAI	NTU_ROSE
MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP

Not much difference between these runs.

Top 3 Results

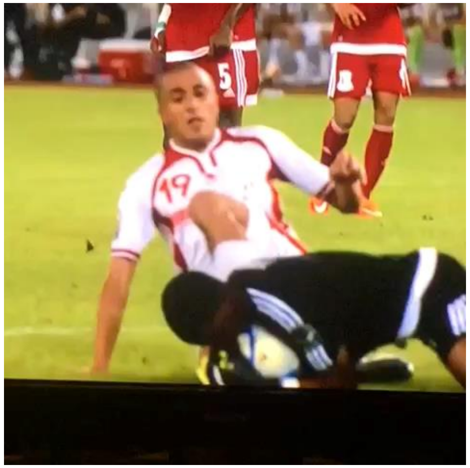
#1874

#1681

#598

Bottom 3 Results

#1029



#958

#1825

Sub-task 2: Description Generation

Given a video

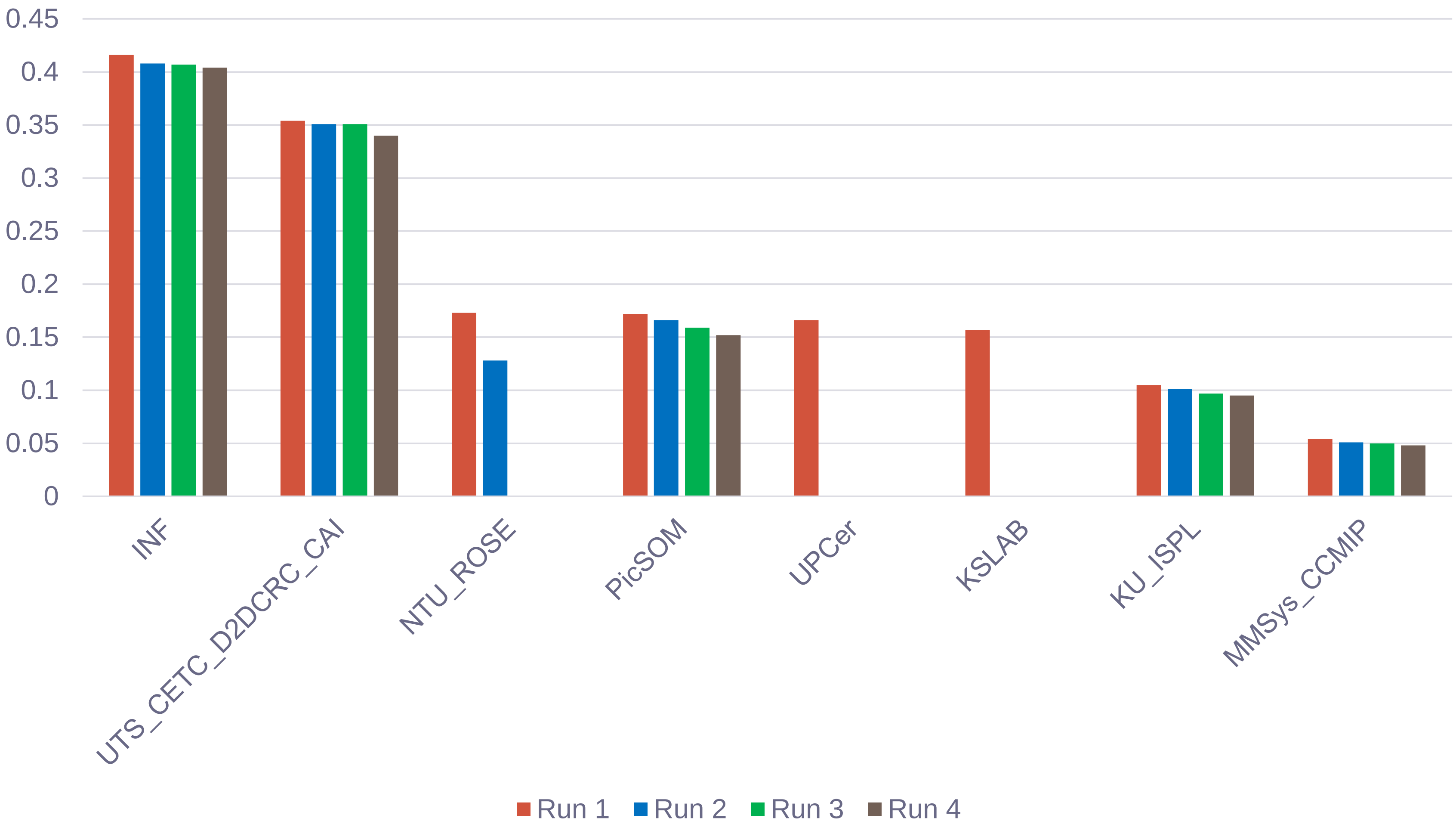
Generate a textual description

Who ? What ? Where ? When ?

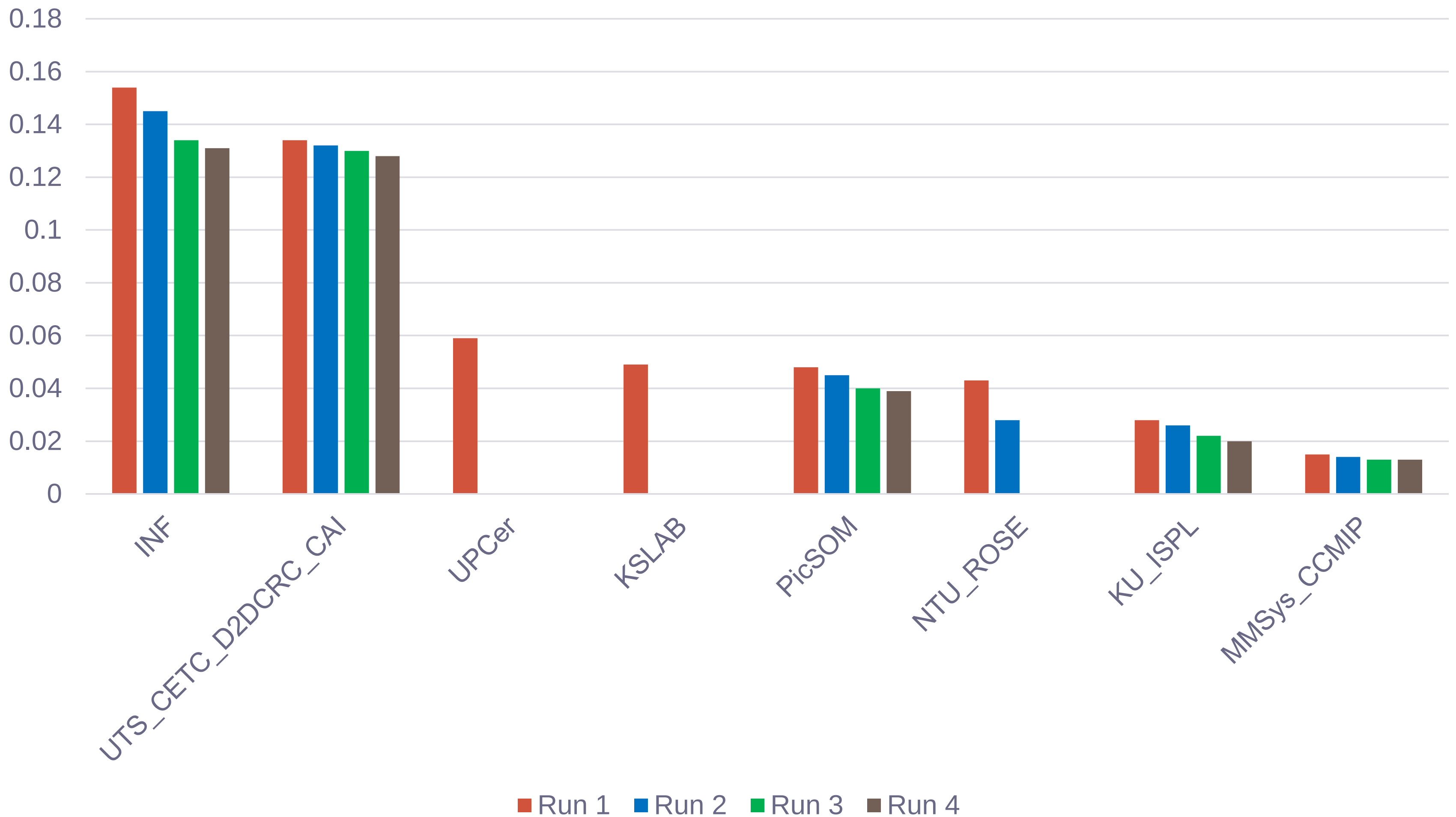
"a dog is licking its nose"

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

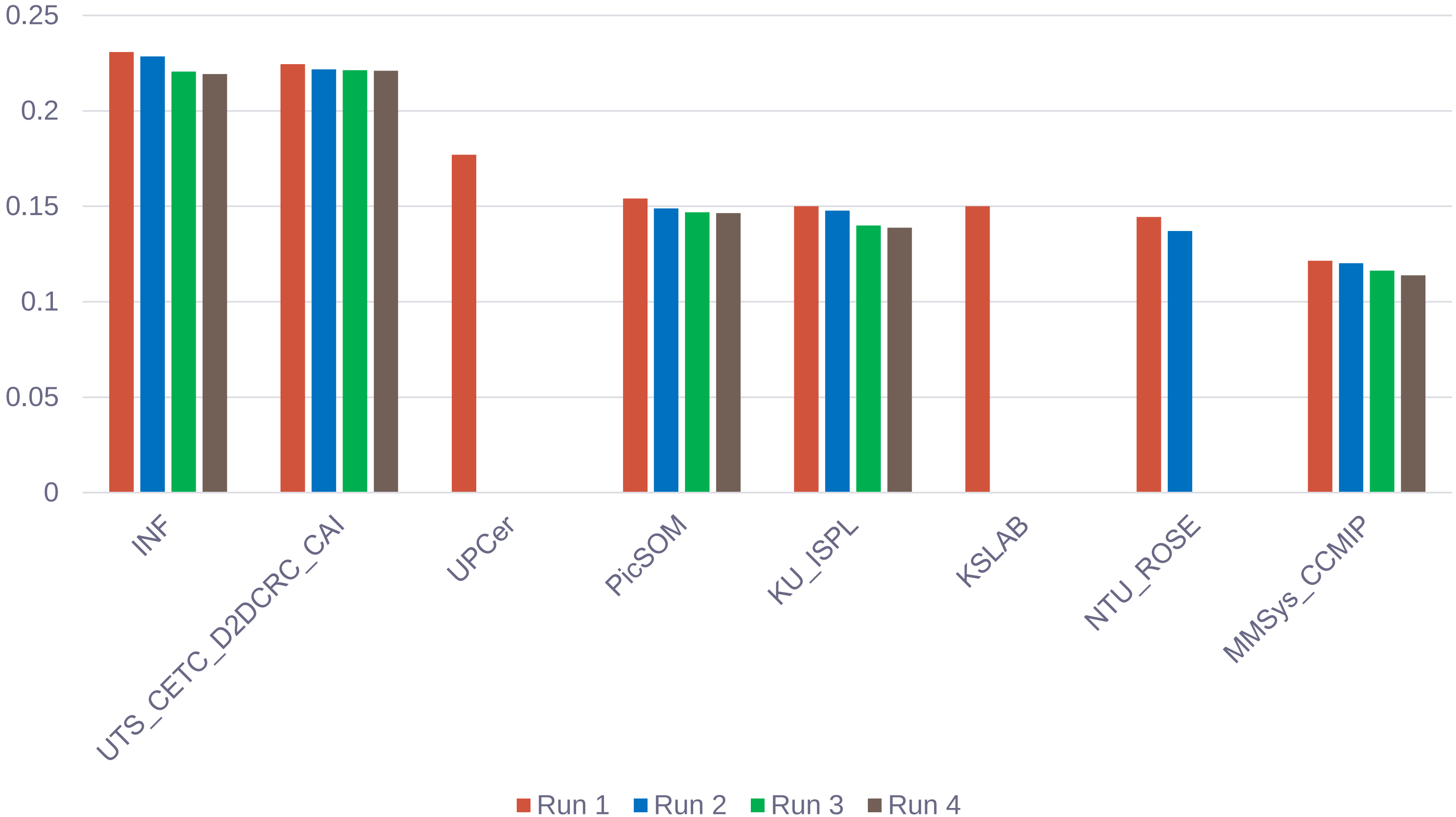
CIDEr Results



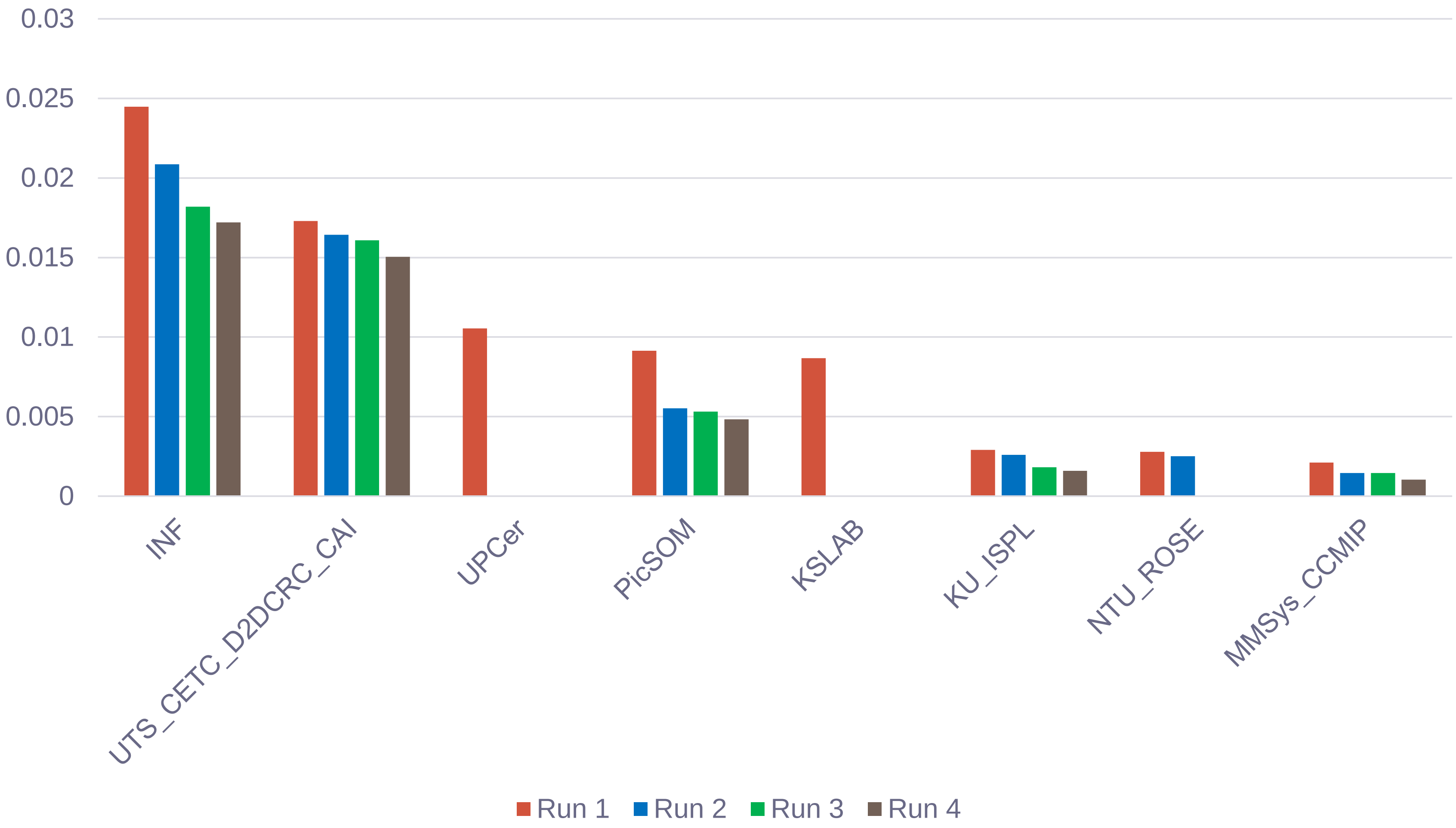
CIDEr-D Results



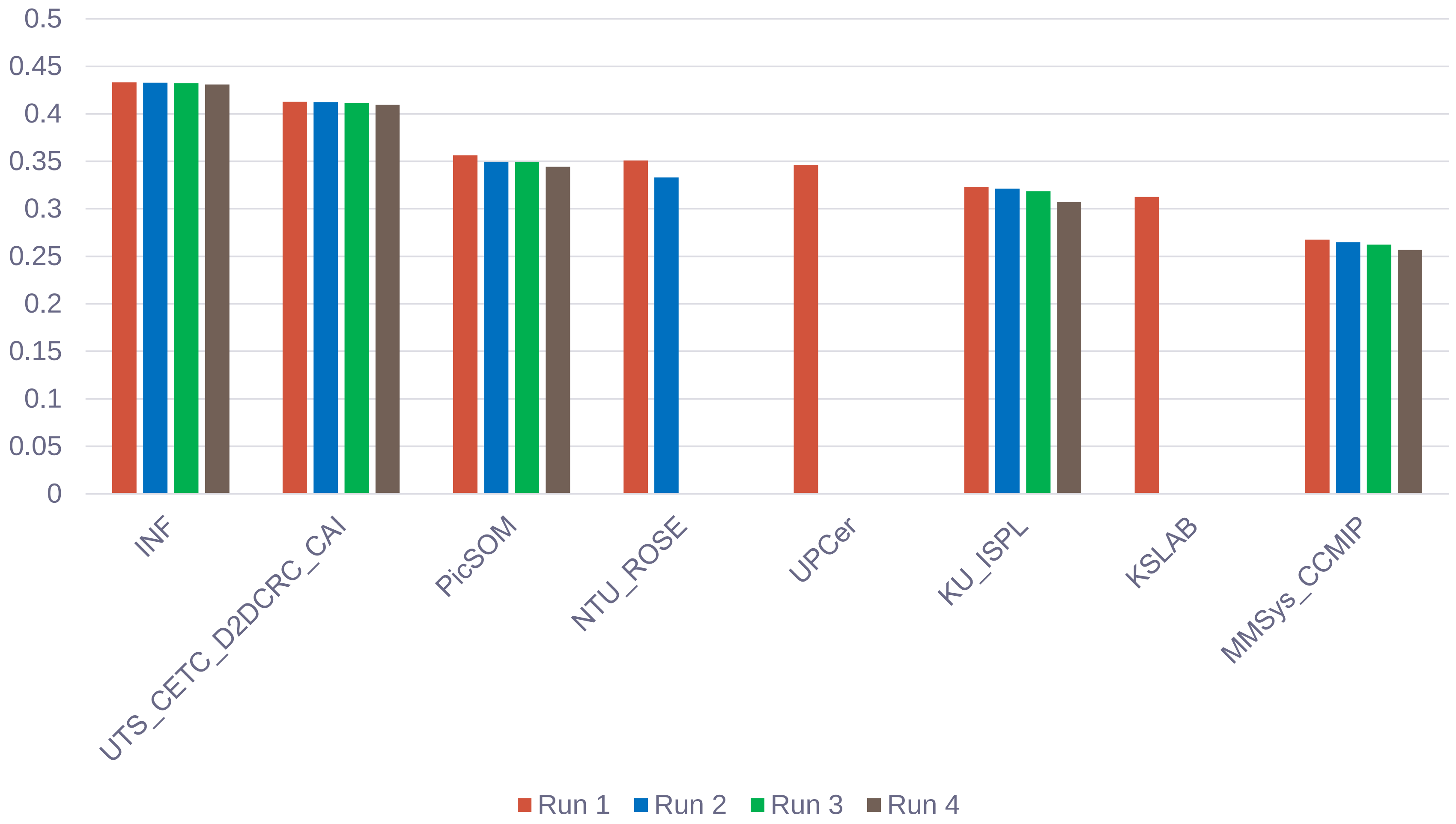
METEOR Results



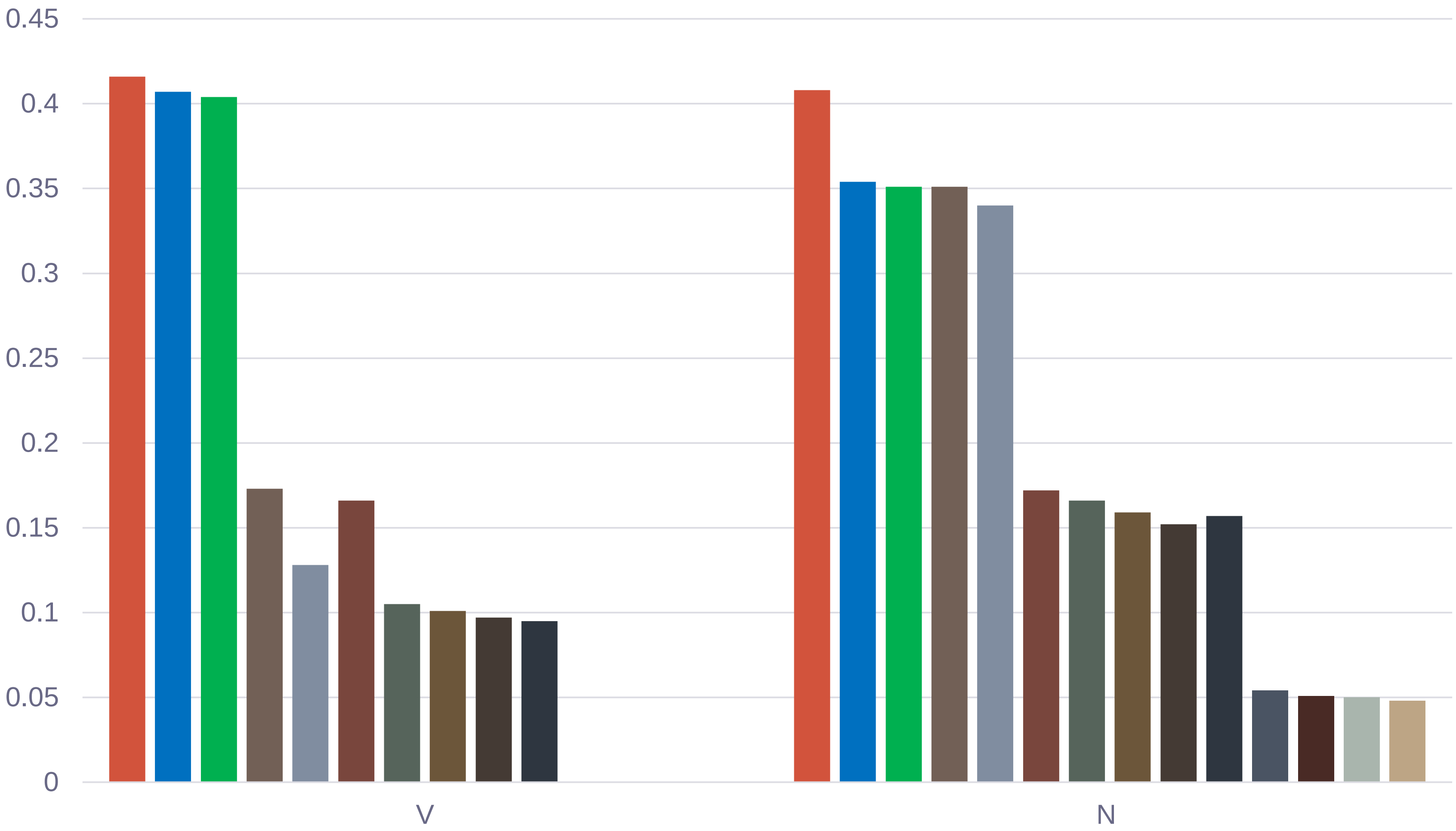
BLEU Results



STS Results


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
INF
UTS_CETC_D2DCRC_CAI
PicSOM
NTU_ROSE
UPCer
KU_ISPL
KSLAB
MMSys_CCMIP
Run 1
Run 2
Run 3
Run 4

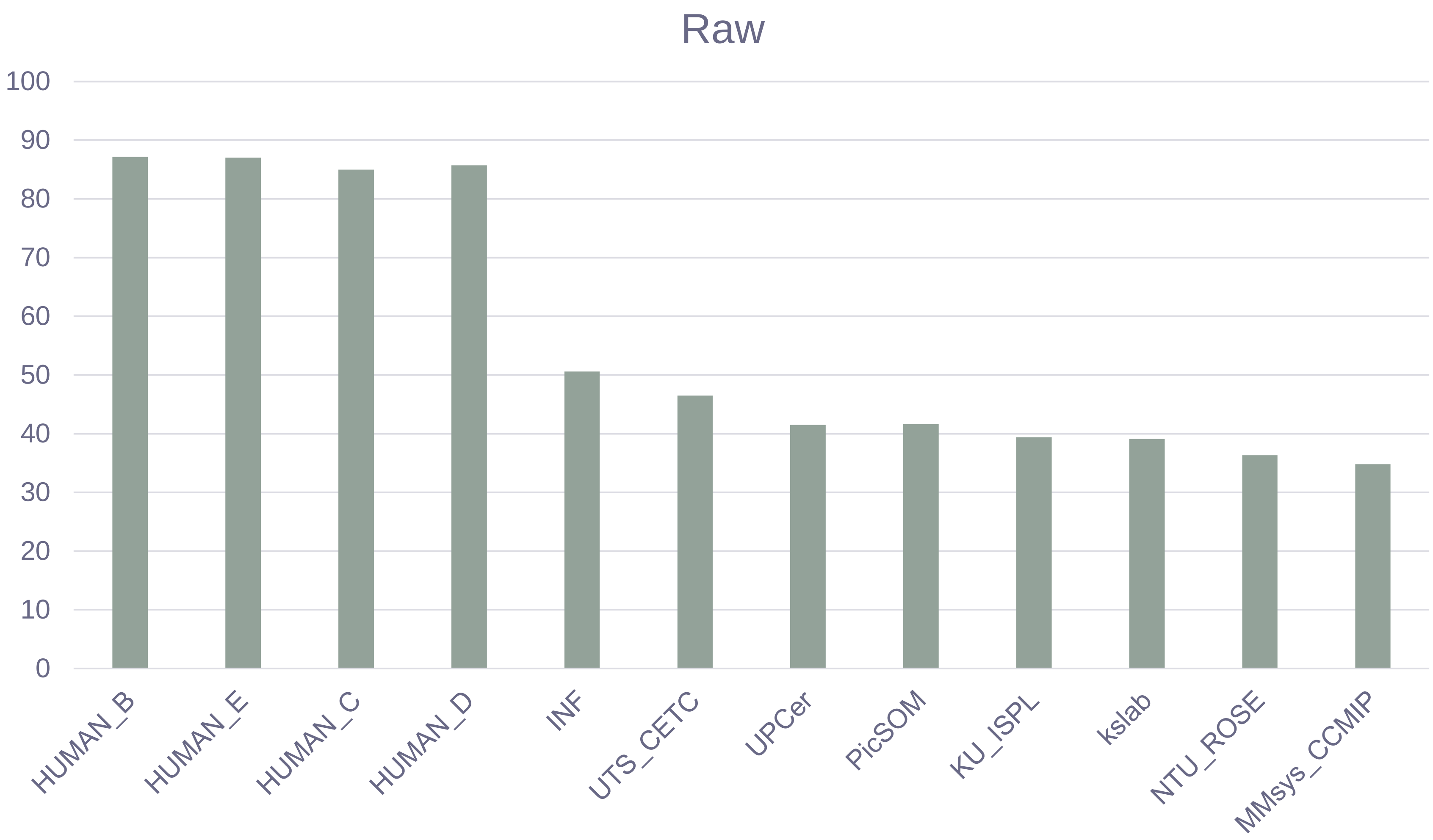
CIDEr Results – Run Type



Direct Assessment (DA)

- Measures …
  - **RAW**: Average DA score [0..100] for each system (non-standardised) – micro-averaged per caption then overall average
  - **Z**: Average DA score per system after standardisation per individual AMT worker's mean and std. dev. score.

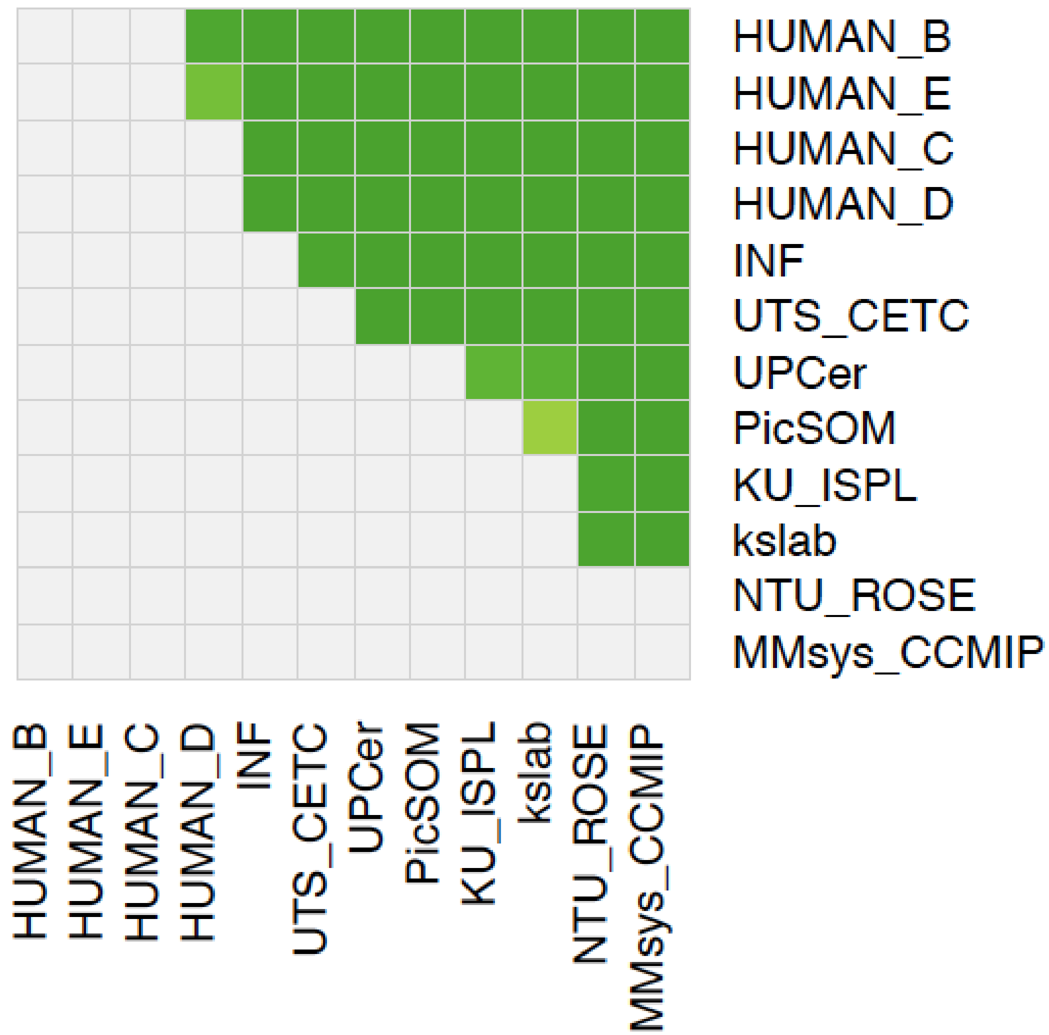
DA results - Raw



DA results - Z


Z
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
HUMAN_B
HUMAN_E
HUMAN_C
HUMAN_D
INF
UTS_CETC
UPCer
PicSOM
KU_ISPL
kslab
NTU_ROSE
MMsys_CCMIP

What DA Results Tell Us ..



1. Green squares indicate a significant “win” for the row over the column.
2. No system yet reaches human performance.
3. Humans B and E statistically perform better than Human D.
4. Amongst systems, INF outperforms the rest.

Systems Rankings for each Metric

CIDEr	CIDEr-D	METEOR	BLEU	STS	DA
INF	INF	INF	INF	INF	INF
UTS_CETC_D2 DCRC_CAI	UTS_CETC_D2 DCRC_CAI	UTS_CETC_D2 DCRC_CAI	UTS_CETC_D2 DCRC_CAI	UTS_CETC_D2 DCRC_CAI	UTS_CETC_D2 DCRC_CAI
NTU_ROSE	UPCer	UPCer	UPCer	PicSOM	UPCer
PicSOM	KSLAB	PicSOM	PicSOM	NTU_ROSE	PicSOM
UPCer	PicSOM	KU_ISPL	KSLAB	UPCer	KU_ISPL
KSLAB	NTU_ROSE	KSLAB	KU_ISPL	KU_ISPL	KSLAB
KU_ISPL	KU_ISPL	NTU_ROSE	NTU_ROSE	KSLAB	NTU_ROSE
MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP	MMSys_CCMIP

Observations

- The task continues to evolve as the number of annotations per video were standardized to 5 (compare to last year's task).
- Tried to remove redundancy and create a diverse set with little or no ambiguity for matching sub-task.
- Steps were taken to ensure that a cleaner dataset was used for the task.

Participants

- Teams that will present today:
  - RUCMM
  - KU_ISPL
  - INF

- Very high level bullets on approaches by other teams.

UTS_CETC_D2DCRC

- Widely used LSTM based sequence to sequence model.
- Focus on improving generalization ability of the model.
  - Different training strategies used.
- Several combinations of spatial and temporal features are ensembled together.
- Simple model structure preferred to help generalization ability.

- Training data: MSVD, MSR-VTT 2016, TGIF, VTT 2016, VTT 2017

PicSOM

Description Generation

- LSTM recurrent neural networks used to generate descriptions using multi-modal features.
- Visual features include image and video features and trajectory features.
- Audio features also used.
- Training datasets used: MS COCO, MSR-VTT, TGIF, MSVD.
- Significant improvement by expanding MSR-VTT training dataset with MS COCO.

KSLAB

- The main idea is to extract representations from only key frames.
- Key frames are detected for different types of events.
- The method uses a CNN encoder and LSTM decoder.
- Model trained using MS COCO dataset.

NTU_ROSE

- Matching & Ranking
  - Trained 2 different models on MS COCO dataset.
  - Image based retrieval methods found suitable.
- Description Generation
  - Training dataset: MSR-VTT and MSVD.
  - CST-captioning (Consensus-based Sequence Training) used as baseline and adapted.
  - Both visual and audio features used.
  - Model trained on MSR-VTT performed better, probably because it generates longer sentences than one trained on MSVD.

MMSys

- Matching & Ranking
  - Wikipedia and Pascal Sentence datasets used for training.
  - Used pre-trained cross-modal retrieval method for matching task.
- Description Generation
  - MSR-VTT dataset used for training.
  - Extract 1 fps per video and used pre-trained Inception-ResNetV2 to extract features.
  - Used sen2vec for text features.
  - Model trained on frame and text features.

EURECOM

Matching & Ranking

- Improved approach of best team of 2017 (DL-61-86).
  - Feature vectors derived from frames extracted at 2 fps using final layer of ResNet-152.
  - Contextualized features obtained and combined through soft attention mechanism.
  - Resulting vector v fed into two fully connected layers using RELU activation.
- Vector v concatenated with vector from last layer of an RGB-I3D.
- Instead of using Res-Net152 trained on ImageNet, it is also finetuned on MSCOCO.

UCR_VCG

Matching & Ranking

- MS-COCO dataset used for training.
- Keyframes extracted from videos – representative frames
- A joint image-text embedding approach used to match videos to descriptions.

Conclusion

- Good number of participation. Task will be renewed.
- This year we had more annotations per video.
- A cleaner dataset created.
- Direct Assessment was used for a second year running. This year we included multiple human responses. The results are interesting.
- Lots of available training sets, some overlap ... MSR-VTT, MS-COCO, ImageNet, YouTube2Text, MSVD, TRECVid2016-2017 VTT, TGIF
- Some teams used audio features in addition to visual features.

Discussion

- Is there value in the caption ranking sub-task? Should it be continued, especially with some teams participating only in this subtask?
- Is the inclusion of run type (N or V) valuable?
  - Other possible run types? Video datasets only vs. video + image captioning training datasets.
- Possibilities for a new dataset?
- Are more teams planning to use audio features? What about motion from video?
- What did individual teams learn?