TRECVID 2017

Video to Text Description

Alan Smeaton
Dublin City University

George Awad
NIST; Dakota Consulting, Inc

Asad A. Butt
NIST

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 four 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 1,880 URLs randomly selected, divided amongst 10 assessors.
- Each video was annotated by at least 2 assessors, and at most 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) ?

Video Dataset

- Matching & Ranking Task
  - 4 groups created based on number of descriptions (2, 3, 4, or 5).

Group	# of Videos in Set
G2 (2x)	1,613
G3 (3x)	795
G4 (4x)	388
G5 (5x)	159

- Description Generation Task
  - All 1,880 videos were used

Runs Submitted

Subtask	Group	Runs Submitted
Matching and Ranking	G2	68
	G3	90
	G4	124
	G5	155
Description Generation	-	43

Annotation Process – Observations

1. Some complex scenes contain a lot of information to describe.
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. There may be some similar videos, resulting in similar descriptions. This was minimized by redundancy removal.

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.

- Some videos have very similar descriptions making matching and ranking difficult.
  - Based on often used keywords, we clustered the entire dataset into 800 clusters out of more than 5000 text descriptions.
  - Clusters were inspected manually to remove videos with very similar descriptions to avoid confusion for the systems.
  - This resulted in fewer videos for the matching and ranking dataset (1,613) compared to the description generation dataset (1,880).

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

Sample Captions of 5 Assessors

1. Many people hold long trampoline and person does double somersault.
2. A group of men hoist a man into the air and he does a flip.
3. Group of young men holding a portable trampoline/mat and when they raise it man on top of trampoline flips and somersaults into the air and lands on his feet.
4. Man thrown in air, manages at least five head over heels in high somersault.
5. One trampoline athlete demonstrates perfectly.

1. Basketball player misses shot, goes out of bounds, and teammate makes basket and physically hangs onto basket for a time.
2. A basketball player hangs on the basket, at basketball play.
3. A basketball player is barreling towards the basket when he is sideswiped by and opponent looses control of the ball; his teammate recovers the basketball, scores for two points and swings from the basketball rim.
4. A player scored a point in a basketball game.
5. Basketball game in progress; black jersey player makes basket and hangs on rim.

Run Submissions & Evaluation Metrics

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

- Up to 4 runs in the *Description Generation* subtask.
- Machine Translation metrics including
  - BLEU (BiLingual Evaluation Understudy) [2]
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering) [3]
  - CIDEr (Consensus-based Image Description Evaluation) metric was also used for *Description Generation* [4]
- The "Semantic Textual Similarity" metric (STS) was used again following last year's trial [5].
- A measure called "Direct Assessment" which is a crowdsourced rating of caption rankings using Amazon Mechanical Turk (AMT)

[2] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
[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*. Vol. 29. 2005.
[4] 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.
[5] Han, Lushan, et al. "UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems." * *SEM@ NAACL-HLT*. 2013.

BLEU and METEOR

- BLEU [0..1] used in MT (Machine Translation) to evaluate quality of text. It approximate human judgement at a corpus level.
- Measures the fraction of N-grams (up to 4-gram) in common between source and target.
- N-gram matches for a high N (e.g., 4) rarely occur at sentence-level, so poor performance of BLEU@*N* especially when comparing only individual sentences, better comparing paragraphs or higher.
- Often we see B@1, B@2, B@3, B@4 … we do B@4.
- Heavily influenced by number of references available.

METEOR

- METEOR Computes unigram precision and recall, extending exact word matches to include similar words based on WordNet synonyms and stemmed tokens
- Based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision

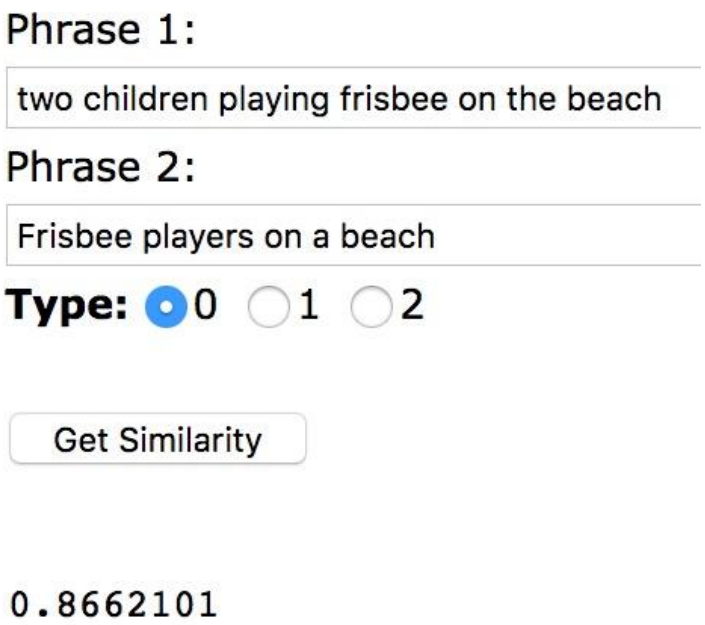
CIDEr

- CIDEr computes the TF-IDF (term frequency inverse document frequency) for each n-gram.
- N-grams of lengths from 1 to 4 are used to compute the CIDEr score.
- The metric is shown to agree with human judgment when comparing two different system descriptions with a reference sentence.

- This is an active area of research … there are no universally agreed metric(s).

UMBC STS measure [0..1]

- We're exploring STS – based on distributional similarity and Latent Semantic Analysis (LSA) … complemented with semantic relations extracted from WordNet



Direct Assessment

- Brings human assessment (AMT) into the evaluation by crowdsourcing how well a caption describes a video.
- Automatically degraded the quality of some manual captions to then rate the quality of the human assessors and take into account in the evaluation, distinguishing genuine assessors from those gaming the system
- A variation on what is used in the main benchmark in MT the Workshop on Statistical Machine Translation (WMT)
- Human evaluator is required to rate a caption on a 0..100 score
- Re-ran this on VTT 2016 submissions, twice, with 0.99 correlation on scores and rankings, showing consistency

2017 Participants (16 teams finished)

	Matching & Ranking (437 Runs)	Description Generation (43 Runs)
ARETE	✓	✓
CCNY		✓
DCU		✓
KU_ISPL	✓	✓
Mediamill	✓	✓
NII Hitachi UIT	✓	✓
RUC CMU	✓	✓
SDNU MMSys	✓	✓
TJU		✓
UPCer	✓	✓
UTS CAI	✓	✓
VIREO	✓	✓
INF		✓
KBVR	✓	
DL	✓	
CMU_BOSCH	✓	

Sub-task 1: Matching & Ranking

Person reading newspaper outdoors at daytime

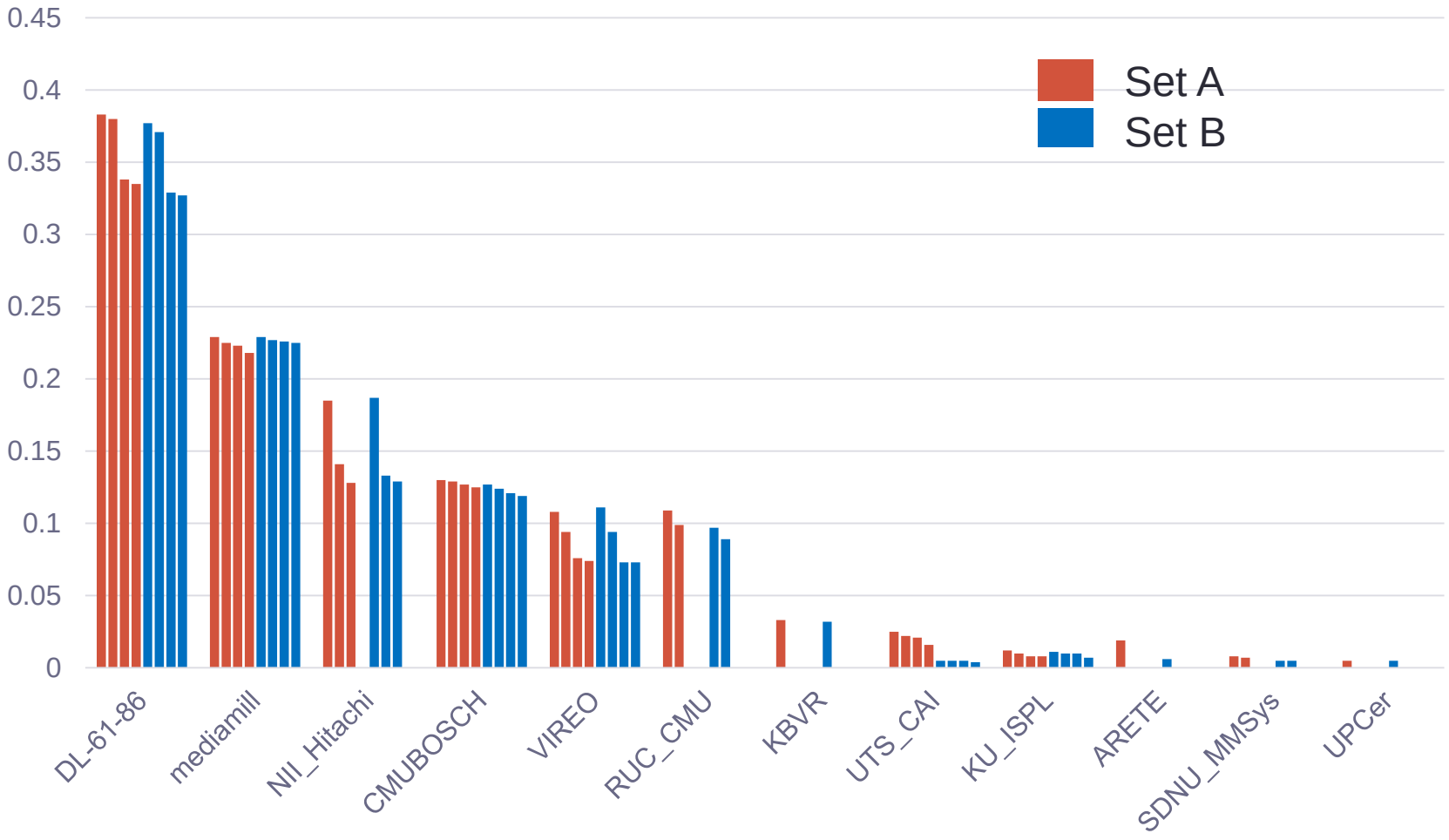
Person playing golf outdoors in the field

Three men running in the street at daytime

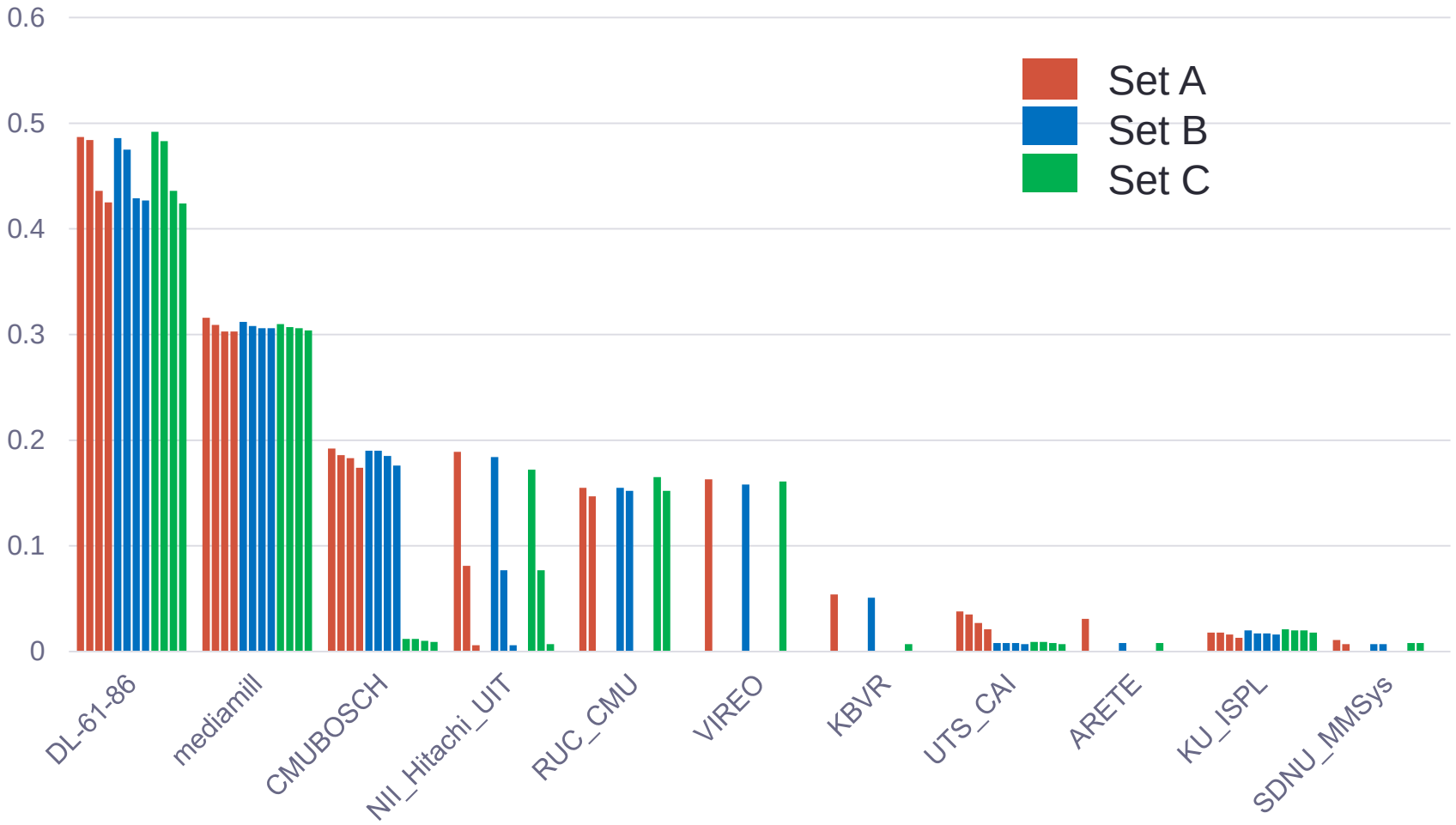
Two men looking at laptop in an office

Multiple runs for each group.

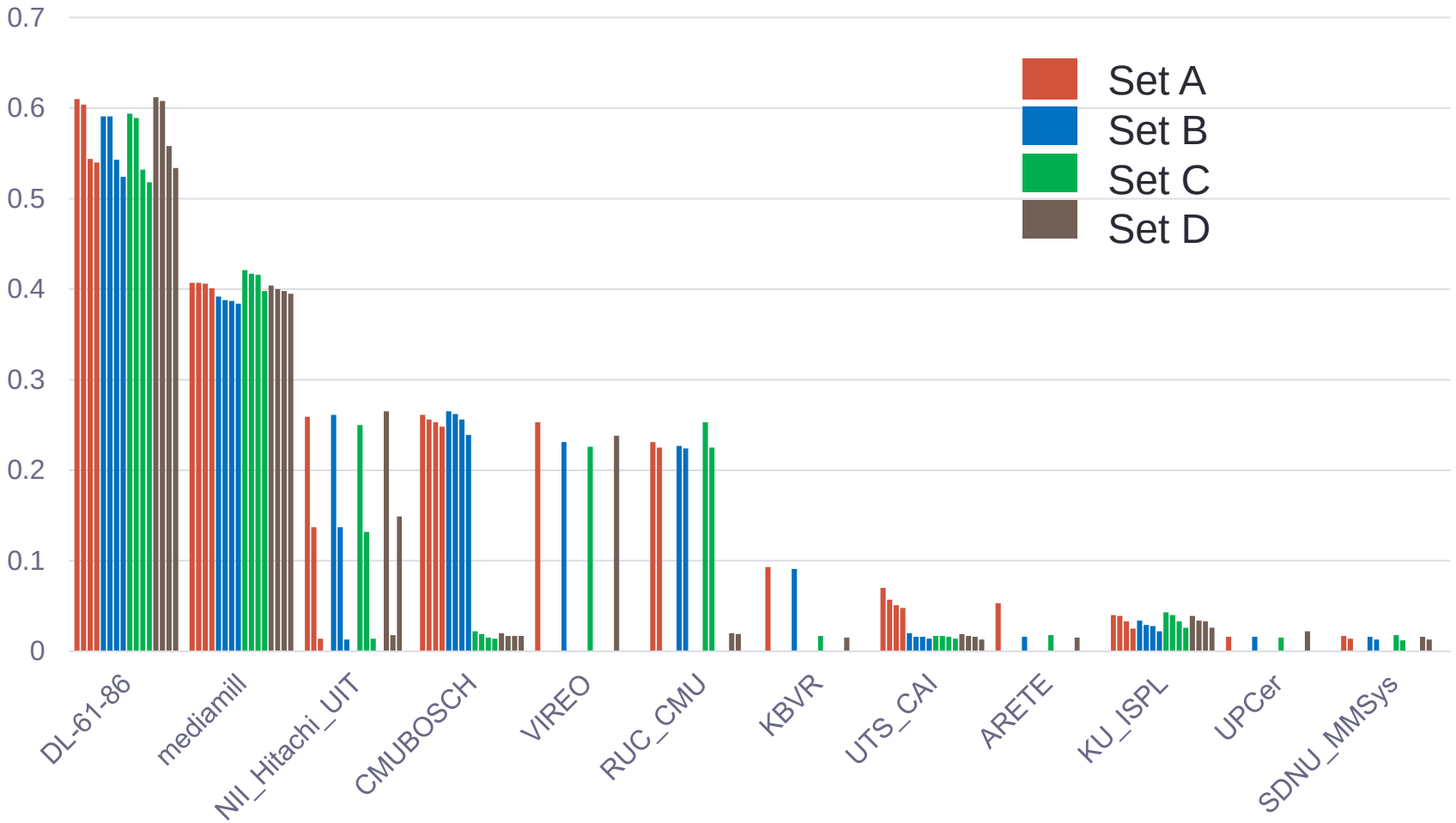
Matching & Ranking Results – Group 2x



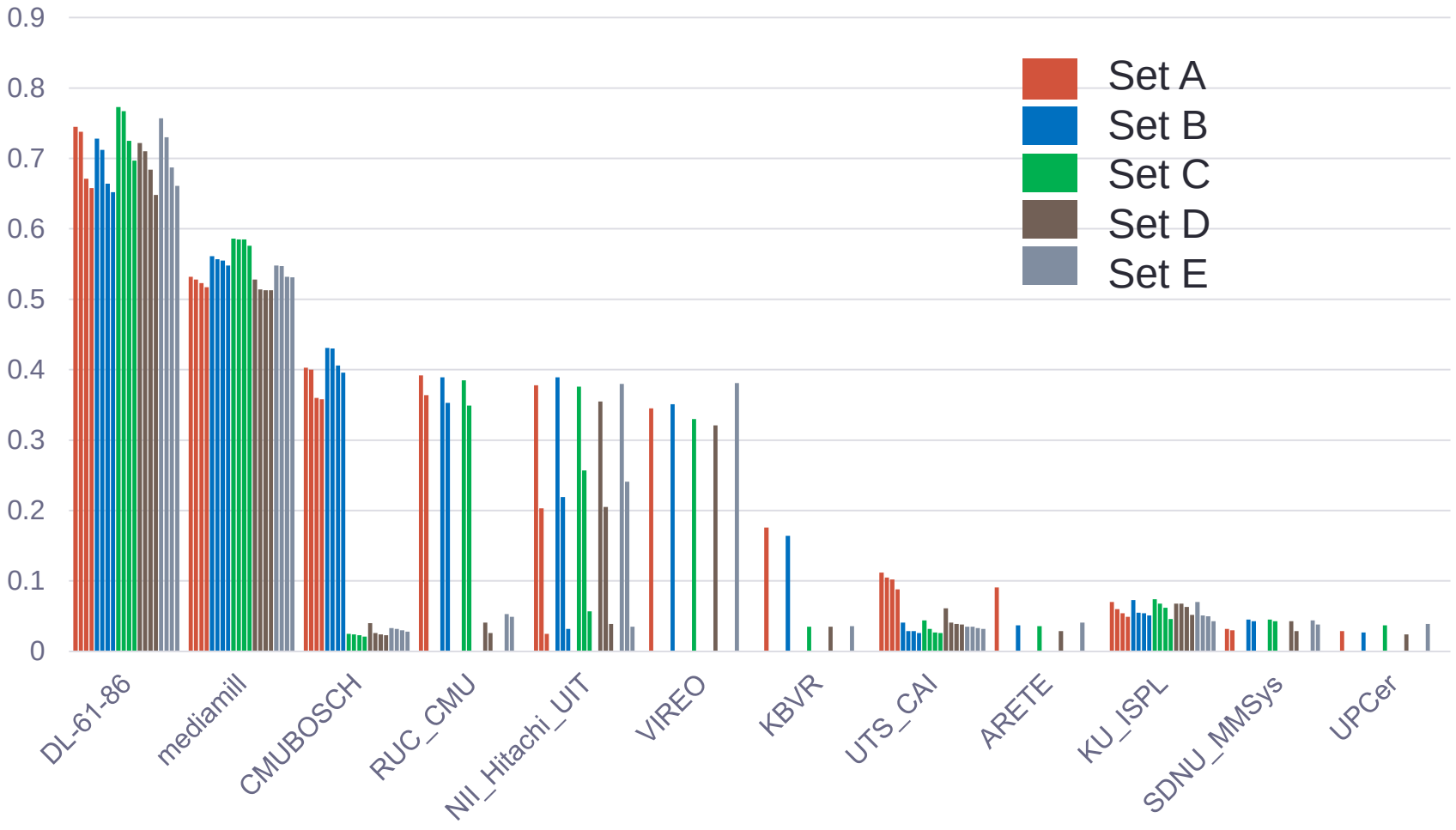
Matching & Ranking Results – Group 3x



Matching & Ranking Results – Group 4x



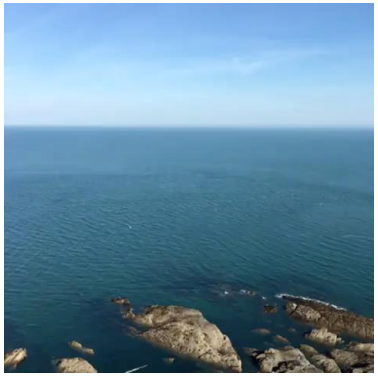
Matching & Ranking Results – Group 5x



Systems Rankings for each Group

G2	G3	G4	G5
DL-61-86	DL-61-86	DL-61-86	DL-61-86
mediamill	mediamill	mediamill	mediamill
NII_Hitachi	CMUBOSCH	NII_Hitachi_UIT	CMUBOSCH
CMUBOSCH	NII_Hitachi_UIT	CMUBOSCH	RUC_CMU
VIREO	RUC_CMU	VIREO	NII_Hitachi_UIT
RUC_CMU	VIREO	RUC_CMU	VIREO
KBVR	KBVR	KBVR	KBVR
UTS_CAI	UTS_CAI	UTS_CAI	UTS_CAI
KU_ISPL	ARETE	ARETE	ARETE
ARETE	KU_ISPL	KU_ISPL	KU_ISPL
SDNU_MMSys	SDNU_MMSys	UPCer	SDNU_MMSys
UPCer		SDNU_MMSys	UPCer

Top 3 Results – G2



#1489
The ocean view from a cliff.

#599
A young woman licking an ice cream, and talking on the beach at day time.

#603
Trash truck picks up trash can, dumping contents on the street instead of into the truck

Top 3 Results – G3

#599
A young woman licking an ice cream, and talking on the beach at day time.

#1503
Car on wet road spins in complete circle and drives on

#1695
A guy bikes with his front wheel up, along other bikers on the road.

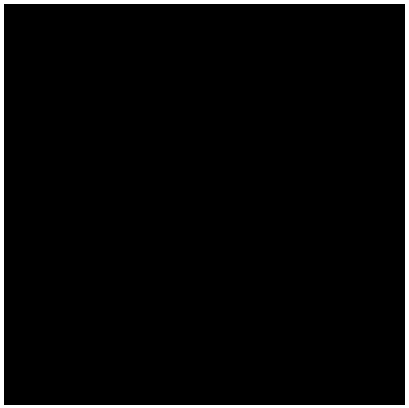
Top 3 Results – G4

#990
A baseball player hitting an opposite field homerun during a game.

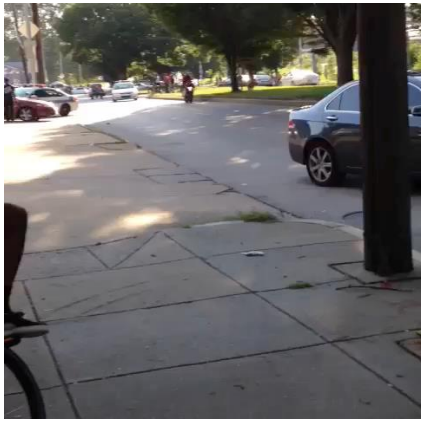
#1599
white cat with collar sniffs plate on table with purple placemat

#1503
Car on wet road spins in complete circle and drives on

Top 3 Results – G5



#168
a man in a skate board ran across an intersection and a van hit him



#1178
A police car is chasing a tricyclist, in the street, daytime.

#1183
Child holds container of watermelon bits and talks.

Bottom 3 Results – G2

#1136
Two young guys are facing each other and move their fingers to each other..

#920
Video shows baby covered in some brown lotion or mud-like substance; switches to basketball player in jersey bouncing ball.

#522
Donald Trump giving a speech.

Bottom 3 Results – G3

#1455
An Asian male is hugging an Asian woman, laying on his back on a stage banging his arms and feet, and then laying down on a playing field banging his hands on the ground

#920
Video shows baby covered in some brown lotion or mud-like substance; switches to basketball player in jersey bouncing ball.

#522
Donald Trump giving a speech.

Bottom 3 Results – G4

#522
Donald Trump giving a speech.

#1613
Large inflated mascot dog on game field sideline, "swallows" cheerleader.

#646
Girl in cafeteria setting makes playful noises and gestures while two people sprawl on seating.

Bottom 3 Results – G5

#1613
Large inflated mascot dog on game field sideline, "swallows" cheerleader.

#646
Girl in cafeteria setting makes playful noises and gestures while two people sprawl on seating.

#1661
blond person dances, knocks over yellow pole inside transit vehicle.

Sub-task 2: Description Generation

Given a video

Generate a textual description

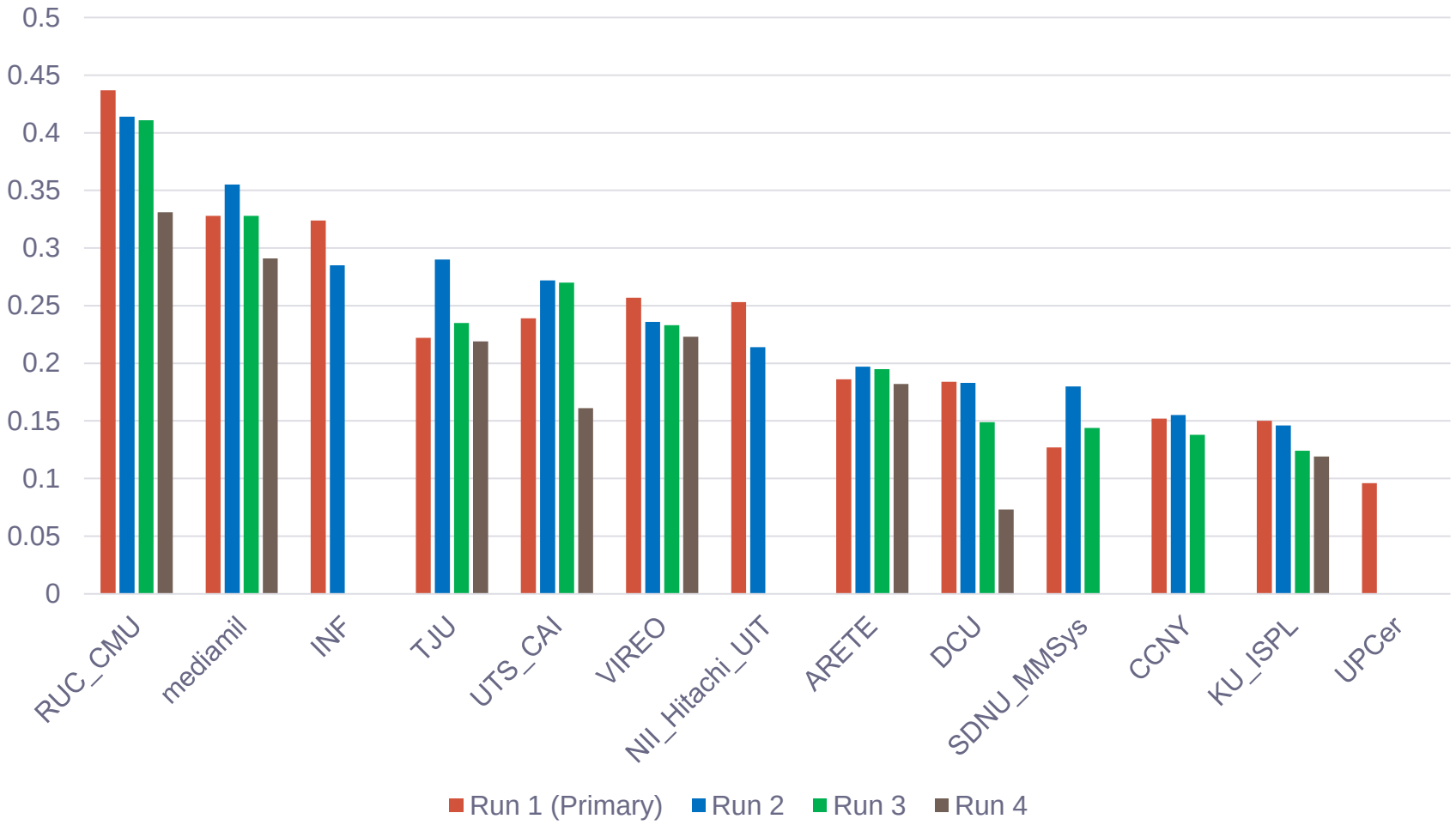
Who ? What ? Where ? When ?

"a dog is licking its nose"

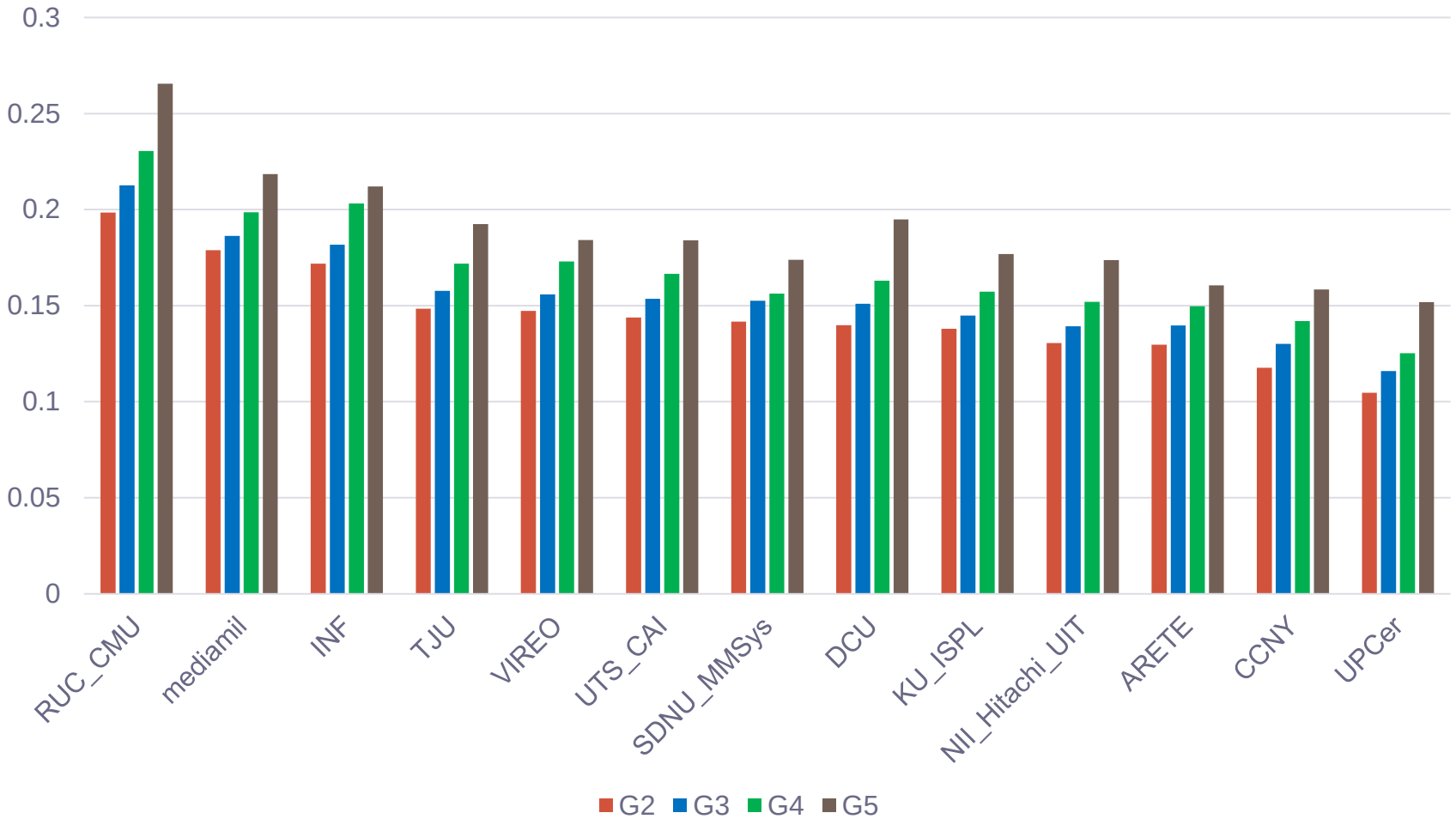
Metrics

- Popular MT measures : BLEU, METEOR, CIDEr
- Semantic textual similarity measure (STS)
- Direct Annotation using Mechanical Turk to rate captions
- All runs and GT were normalized (lowercase, punctuations, stop words, stemming) before evaluation by metrics (except STS)
- Each site asked to nominate one run as "primary"

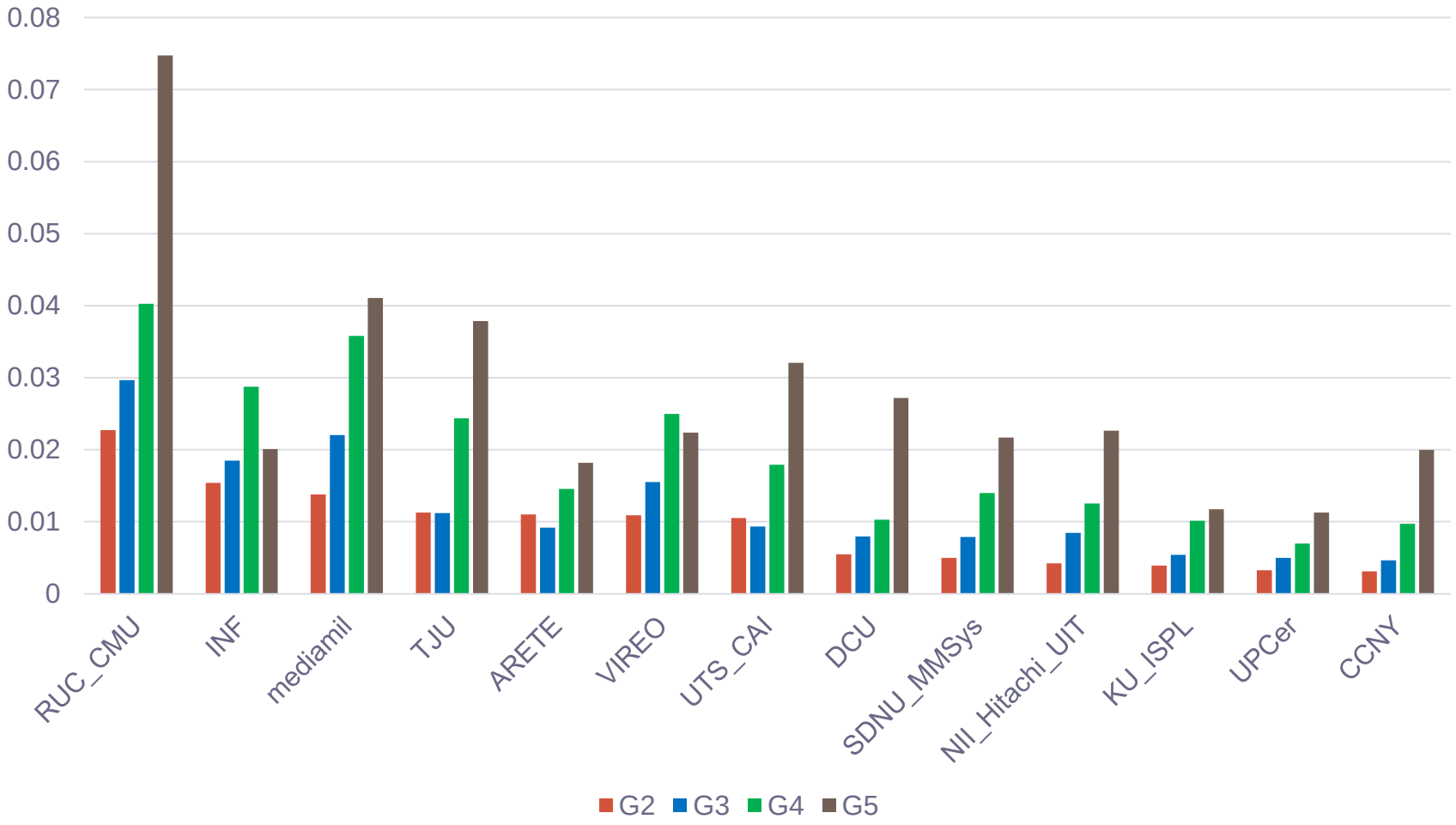
CIDEr Results



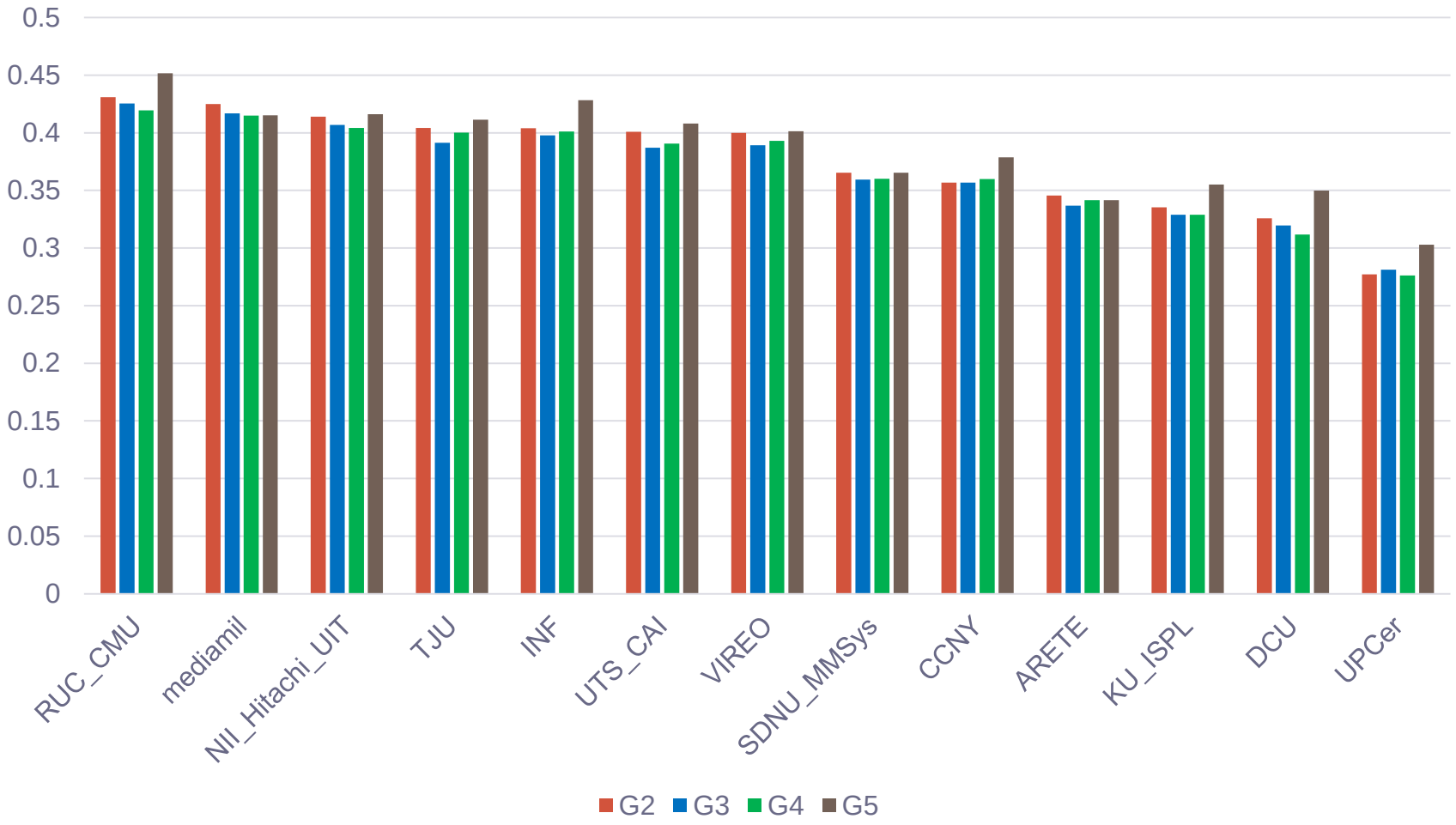
METEOR Results – Best Runs



BLEU Results – Best Runs

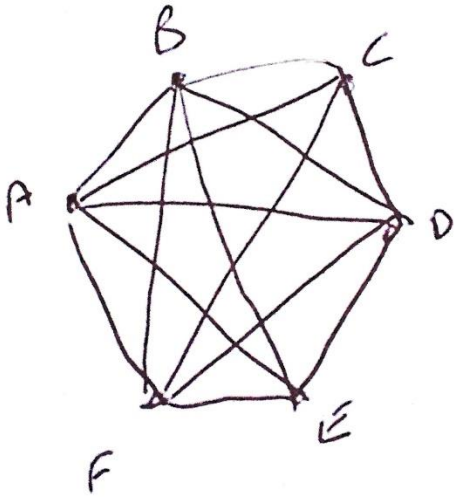


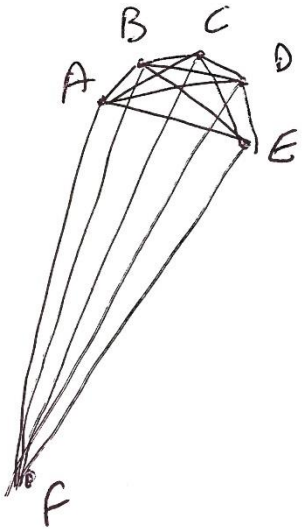
STS Results – Best Runs



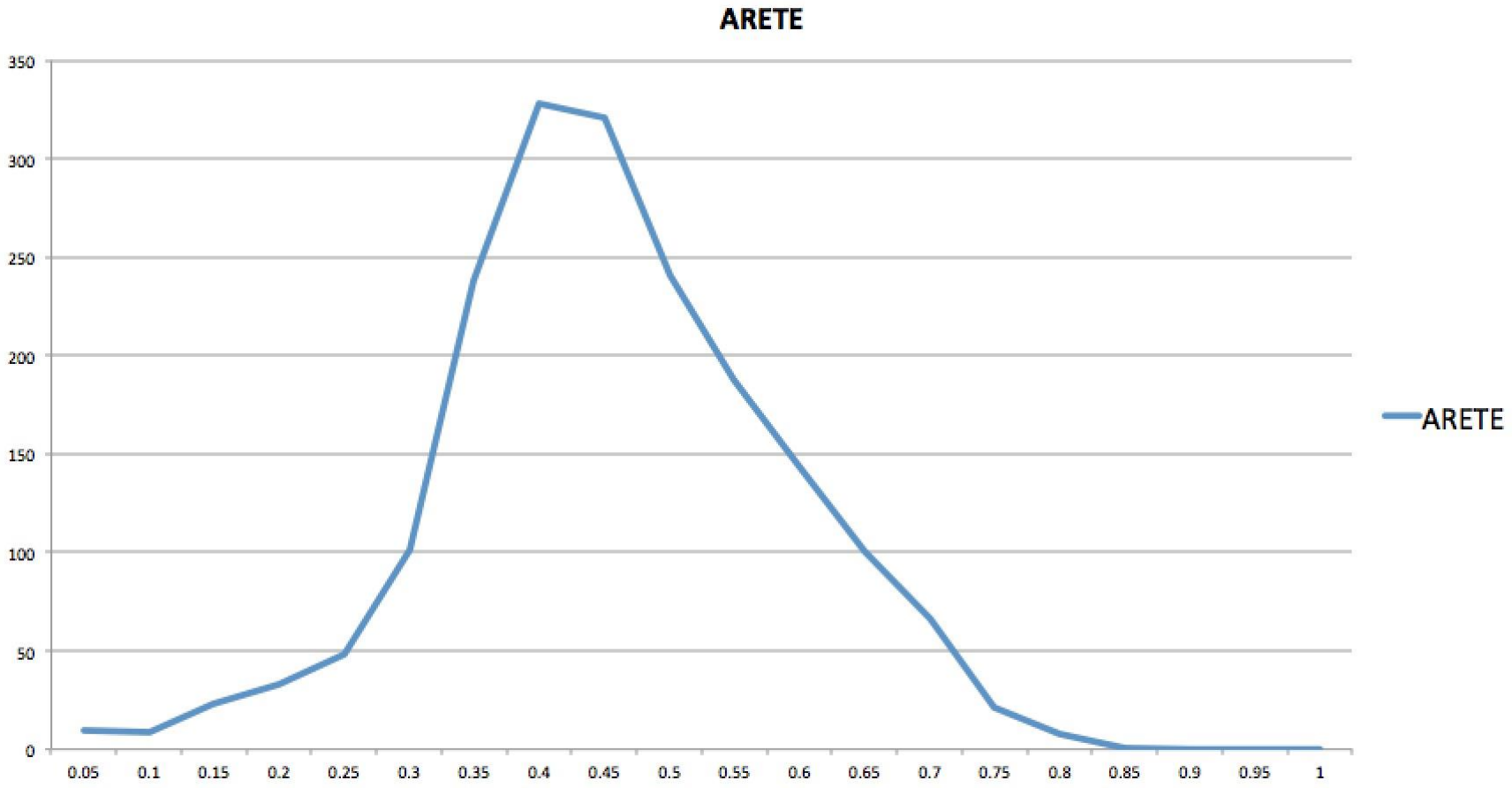
STS Results - Analysis

- With STS API we can measure, for each video, pairwise similarity among all captions (primary run only) for 13 systems + 1 manual (171,080 pairwise comparisons - thanks UMBC)
- Ideally all systems very similar but the more “outlier-ish” a system, across all 1,880 videos (lower averaged STS value), says something

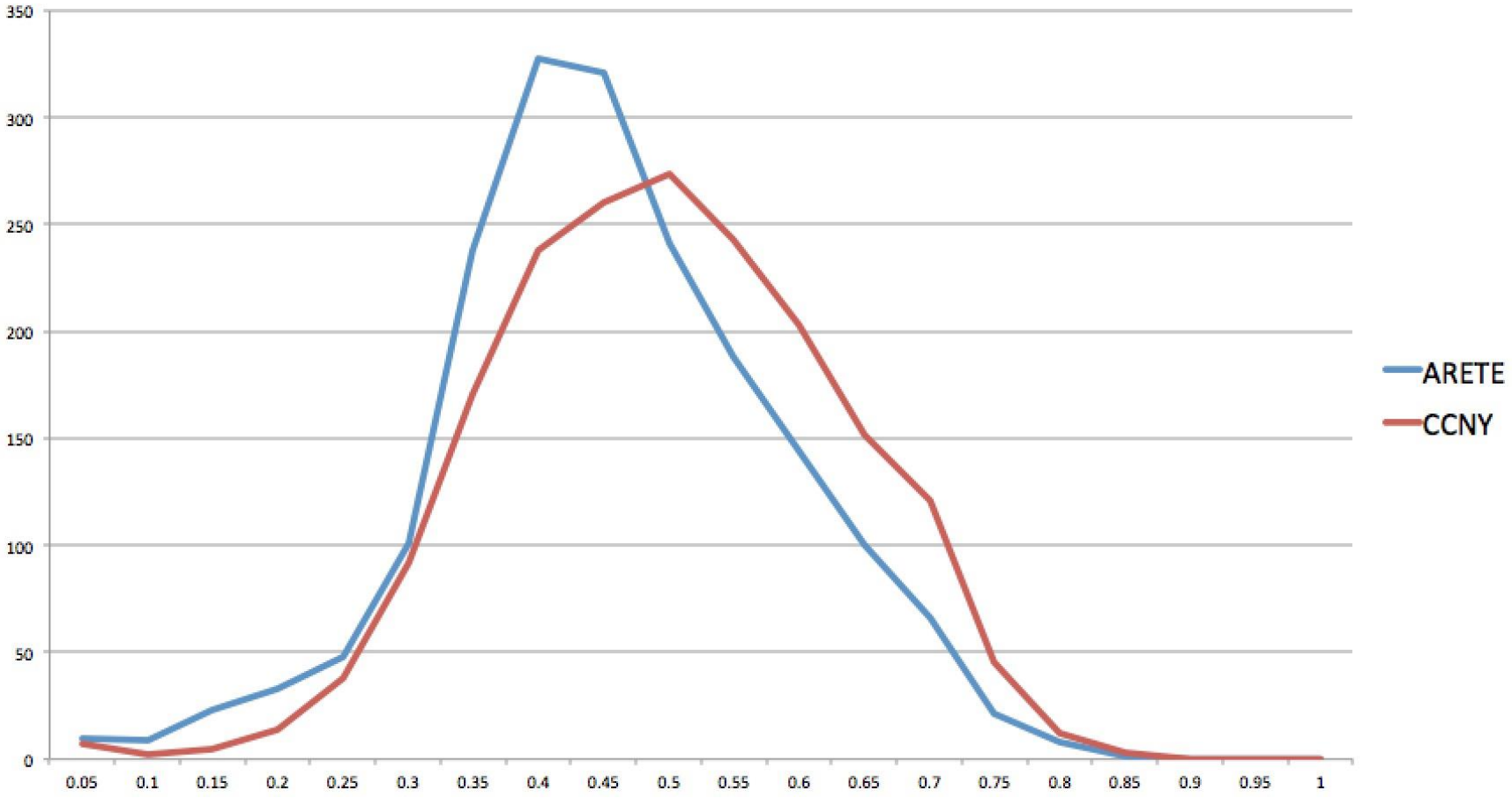




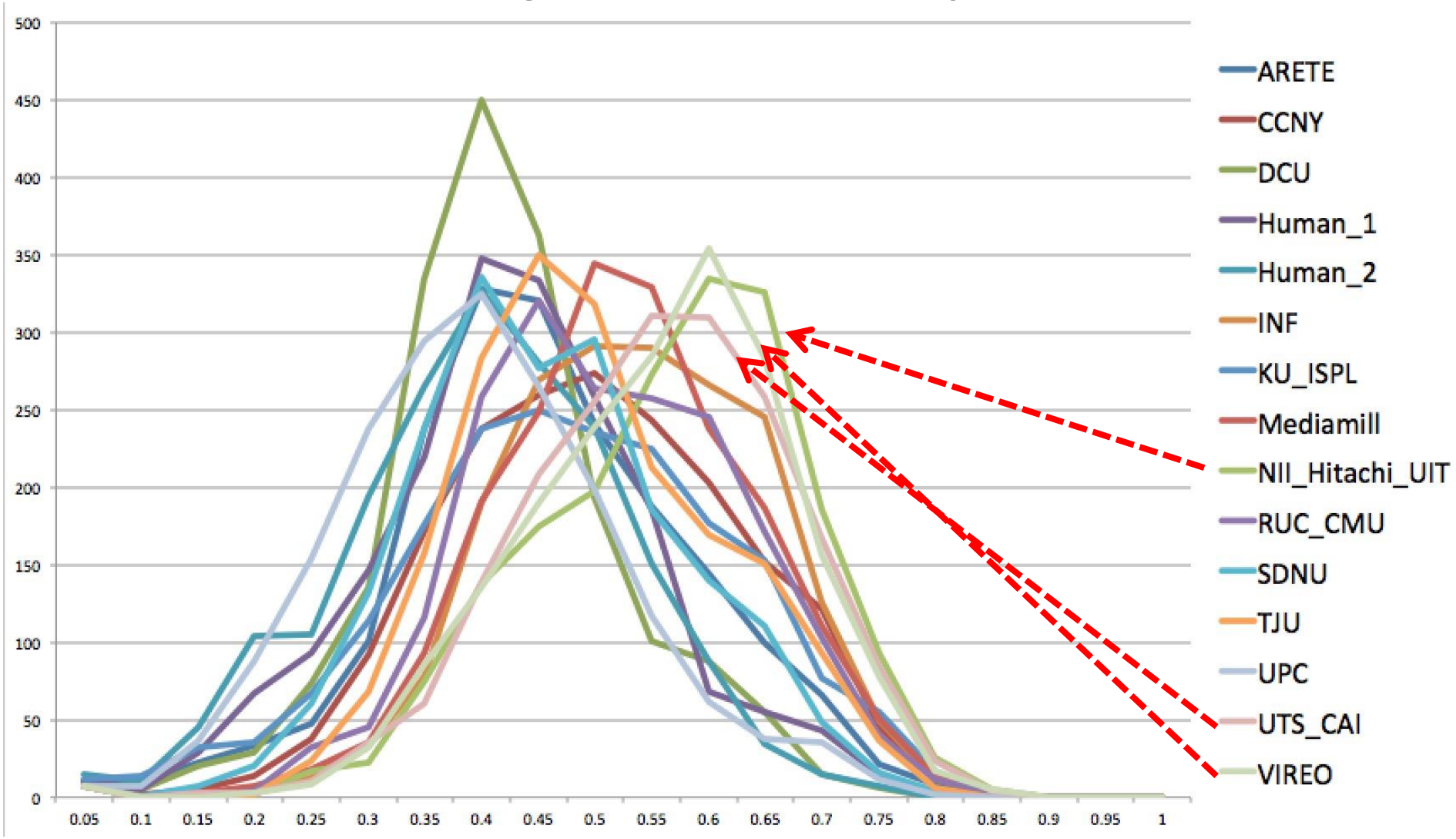
- Take ARETE … for each 1,880 videos, compute STS(ARETE, SystemN), for each of N other systems (+human), put value into 1 of 20 buckets and plot



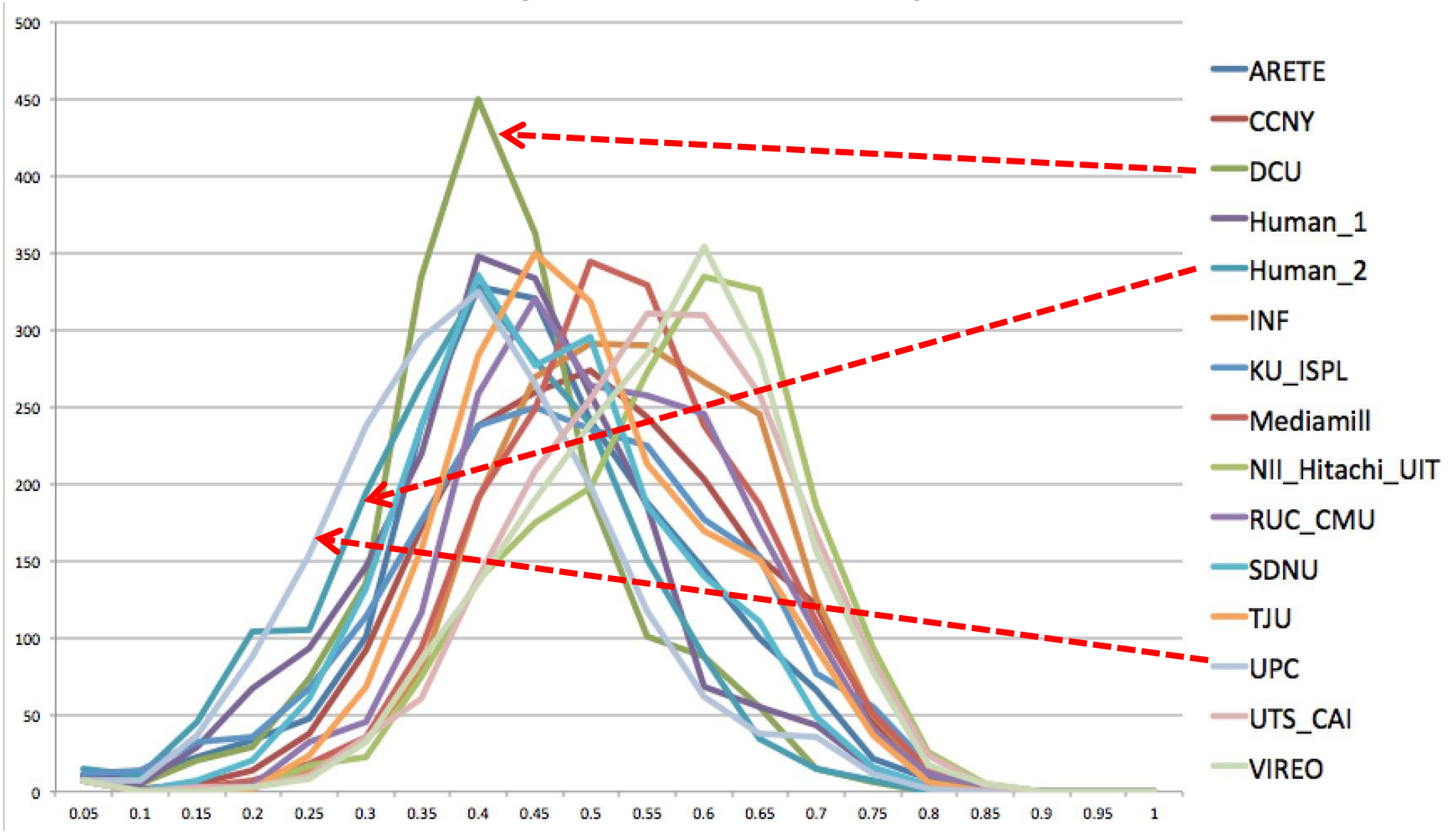
- Now compare ARETE with, say, CCNY … CCNY has higher similarities with “the rest”, is more “with the crowd”
- With a big crowd, is there crowd wisdom, or over-fitting ?



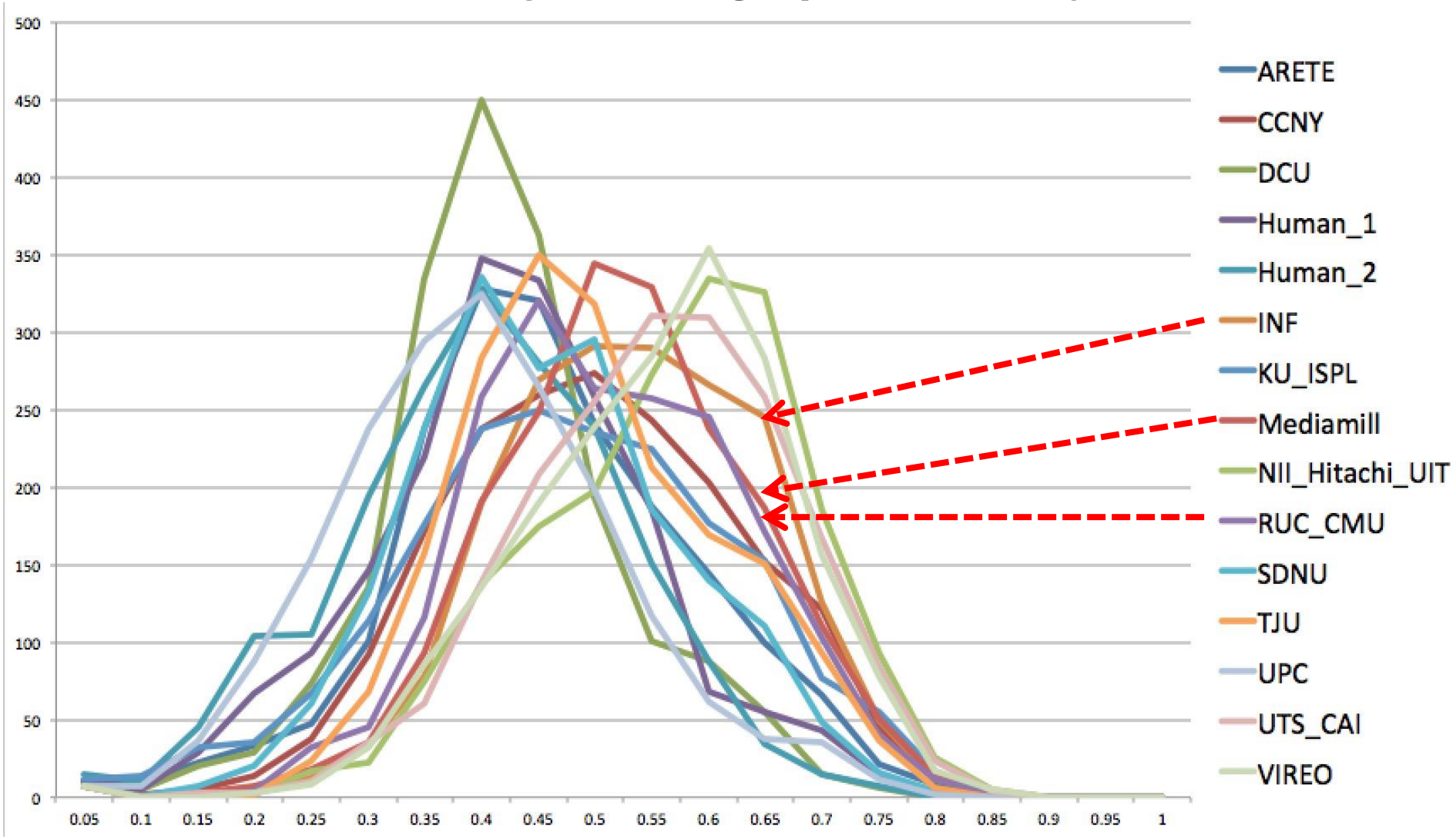
- Now every system v every other, across all videos
- There is an ordering – the **“popular”** systems



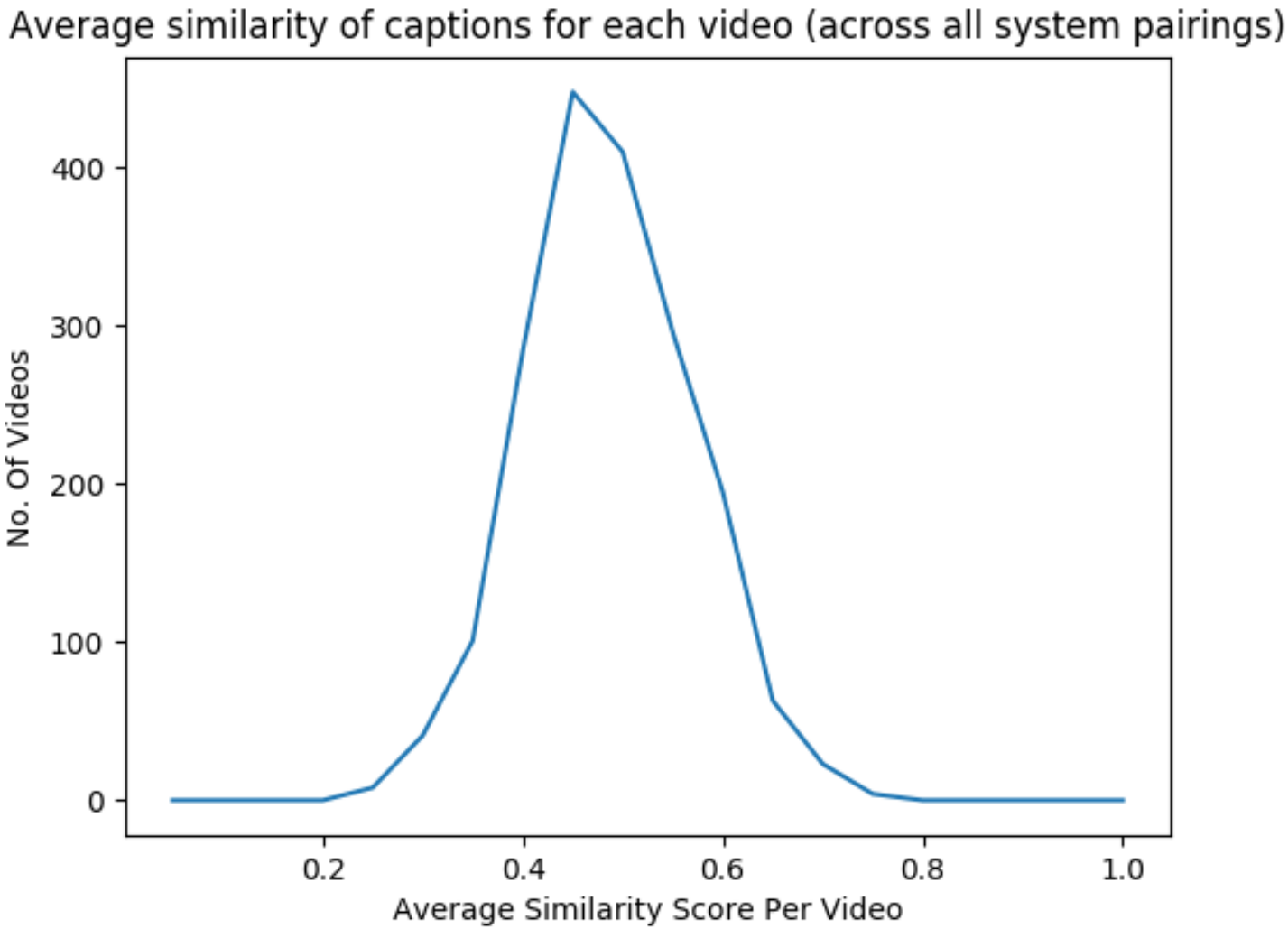
- Now every system v every other, across all videos
- There is an ordering – the **“outlier”** systems



- Now every system vs. every other, across all videos
- There is an ordering – the **high performer** systems



- If there's an ordering of **systems** in terms of "popularity" / "outlier", is there an ordering of **videos** in terms of agreeability among captions



Top 4 Agreeable Videos (Highest Score)

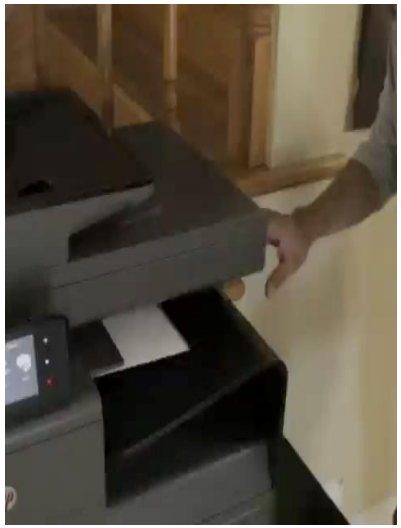
1002

1457

85

370

4 Least Agreeable Videos !



1249

1734

1262

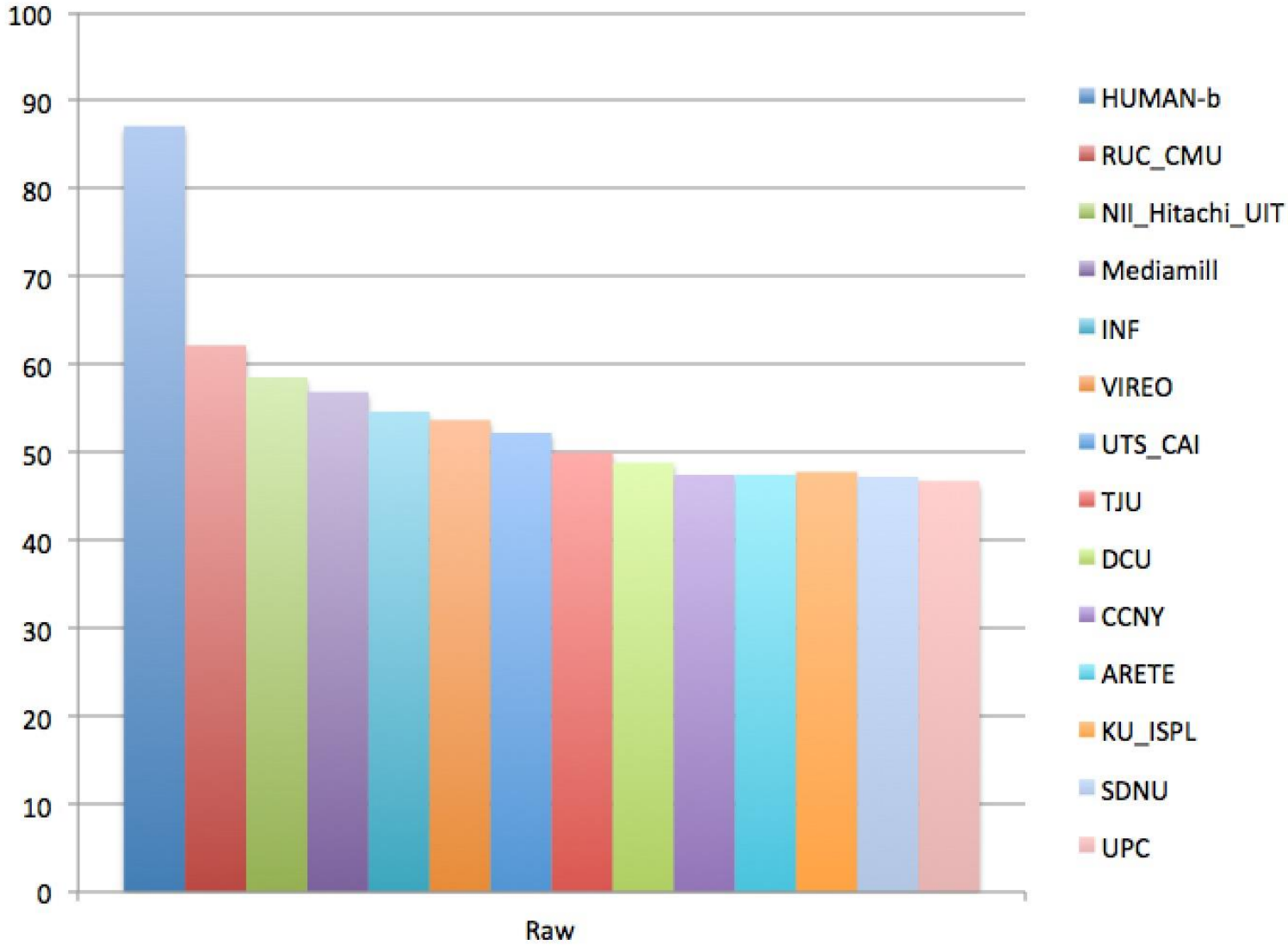
190

Direct Assessment (DA)

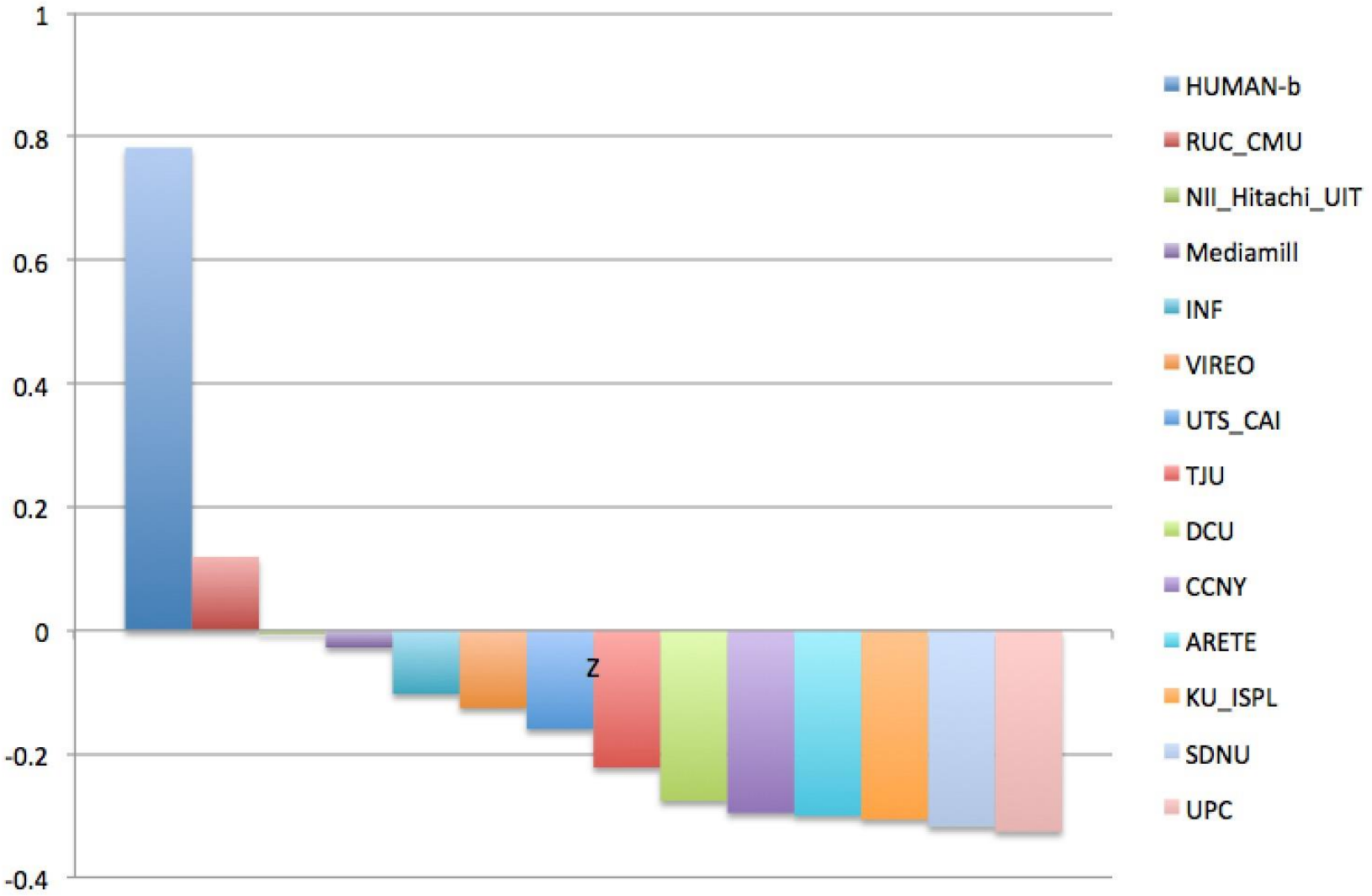
- Cost-effective in terms of $, used +100 assessors, each assessor was rated, video scoring divided into HITs of 100x, all completed by 12 Sept !

- 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.
  - **N**: Number of caption scores combined to compute Raw and Z

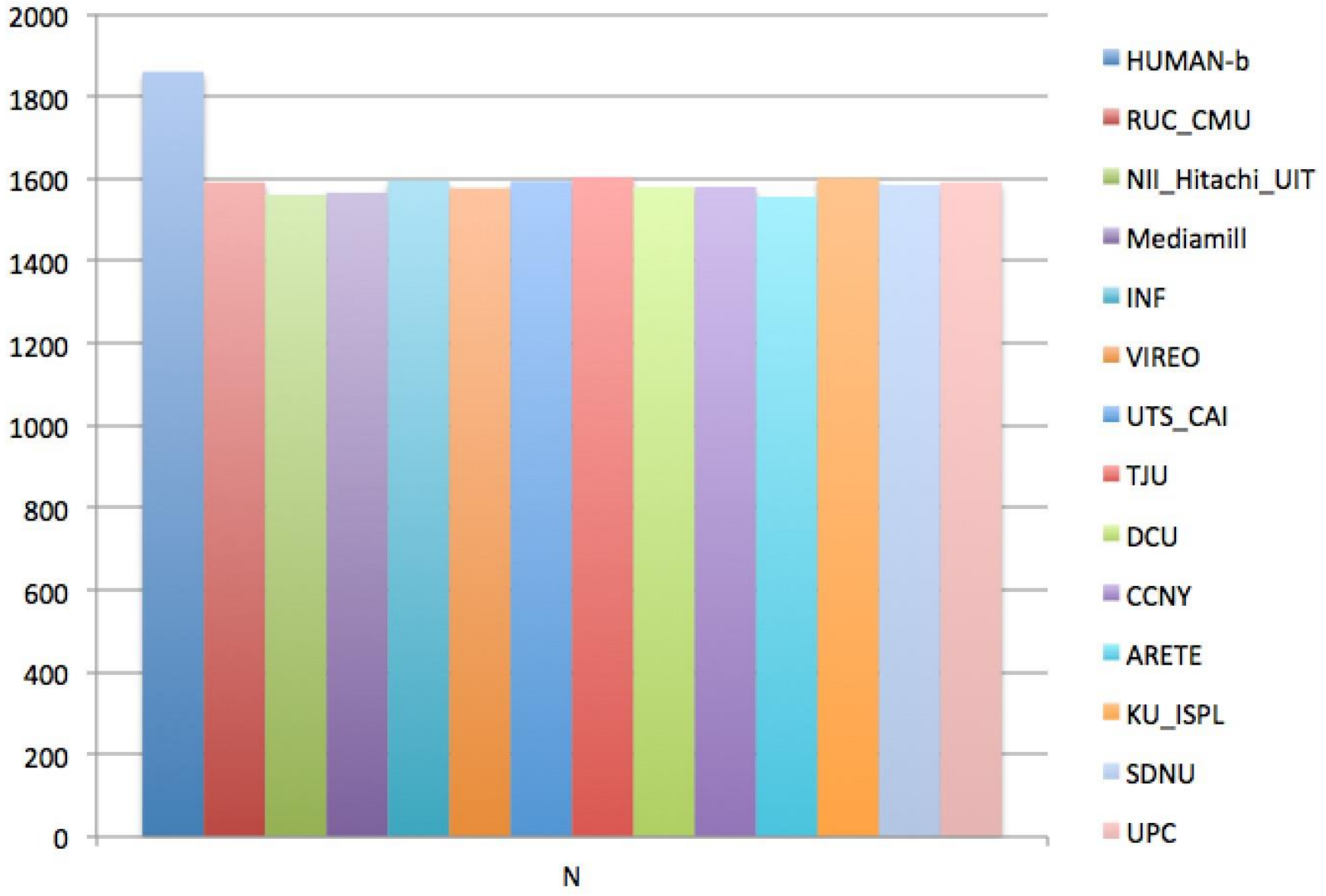
DA results - Raw



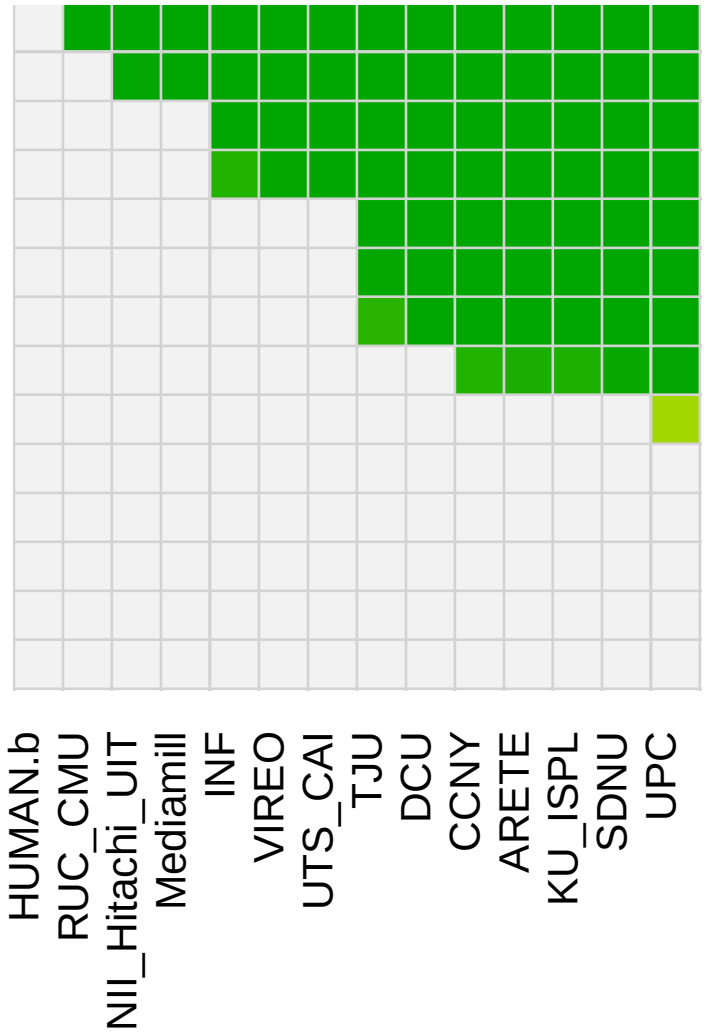
DA results - Z


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Z
HUMAN-b
RUC_CMU
NII_Hitachi_UIT
Mediamill
INF
VIREO
UTS_CAI
TJU
DCU
CCNY
ARETE
KU_ISPL
SDNU
UPC

DA results - N


2000
1800
1600
1400
1200
1000
800
600
400
200
0
N
HUMAN-b
RUC_CMU
NII_Hitachi_UIT
Mediamill
INF
VIREO
UTS_CAI
TJU
DCU
CCNY
ARETE
KU_ISPL
SDNU
UPC

What DA Results Tell Us ..



1. No system yet reaches human performance (Human-B beats everyone)

2. According to DA there is a clear best-performer over other systems, but still 25% less than Human

3. Green squares indicate a significant “win” for the row over the column.

Systems Rankings for each Metric

CIDEr	METEOR	BLEU	STS	DA
RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU
mediamil	mediamil	mediamil	INF	NII_Hitachi_UIT
INF	INF	TJU	mediamil	mediamil
TJU	DCU	UTS_CAI	NII_Hitachi_UIT	INF
UTS_CAI	TJU	INF	TJU	VIREO
VIREO	VIREO	DCU	UTS_CAI	UTS_CAI
NII_Hitachi_UIT	UTS_CAI	VIREO	VIREO	TJU
ARETE	KU_ISPL	NII_Hitachi_UIT	CCNY	DCU
DCU	SDNU_MMSys	SDNU_MMSys	SDNU_MMSys	CCNY
SDNU_MMSys	NII_Hitachi_UIT	CCNY	KU_ISPL	ARETE
CCNY	ARETE	ARETE	DCU	KU_ISPL
KU_ISPL	CCNY	KU_ISPL	ARETE	SDNU_MMSys
UPCer	UPCer	UPCer	UPCer	UPCer

An example from run submissions – 8 unique examples

1. A woman holding a microphone
2. A woman is dancing
3. A woman wearing a hat is singing into a microphone
4. A woman sings on a stage
5. A girl is singing on a stage
6. A woman is singing a song
7. A woman is singing a song on stage in a beauty salon
8. A woman is talking to a man

Observations

- Task evolved from last year owing to different number of manual descriptions, more participants
- In future, we may standardize the number of annotations per video for uniform evaluation.
- Tried to remove redundancy and create a diverse set with little or no ambiguity for matching sub-task.
- For the description sub-task, in general higher number of descriptions results in higher scores, due to the possibility of higher number of word matches.
- There seems to be general agreement between the metrics, with STS being the exception in some cases, and that’s not a bad thing

Participants

- Very high level bullets on what approaches participants took – more details in posters and presentations

1. Areté Associates, Arlington VA

- Used Venugopalan et al.'s ICCV 2015 Sequence to Sequence - Video to Text (S2VT) model
- Trained on Microsoft, MPII-movie, Montreal-VAD and TRECVID VTT2016 datasets

- Matching & Ranking
  - generate caption for each video then re-rank by METEOR

- Description Generation
  - all runs are variants on S2VT model but trained on different training data

2. *CCNY, KBVR, UPCer, Mediamill*

- Did not submit any papers

3. Dublin City University

- Description Generation runs only, trained on MS-COCO
- Two runs extracted keyframes, NeuralTalk2 to generate captions for each keyframe, then combine them
- Third automatically generated saliance-based crops for each keyframe, rank crops for aesthetic appeal, NeuralTalk2 used on top-10 aesthetic crops to generate captions, combined into 1 caption.
- Fourth run, a trained end-to-end system, used a stack of two Long-Short Term Memory (LSTM) cells together with a pre-trained convolutional neural network (CNN), like ARETE, used Venugopalan et al.'s ICCV 2015 Sequence to Sequence - Video to Text (S2VT) model

4. KU-ISPL

- Intelligent Systems Processing Laboratory, Korea University

- Used a stacked LSTM model with inputs being various mid-level deep learning and multi-object detection features, plus audio.
- Research question was on improving training data rather than models, and runs combined different features,
- Word2Vec used to encode sentences
- A variety of trained datasets including MSVD (Microsoft YouTube clips), MPII-MD (Max Planck Institute), MVAD (Montreal Institute for Learning), MSR-VTT (MSR Video to Language ACM Challenge), TRECVid2016-VTT

6. NII Hitachi UIT

- National Institute of Informatics (Japan) + University of Information Technology (Vietnam) + UPC (Spain) + Hitachi (Japan) + 2 others
- Building on work of MediaMill and VisualWord2Vec

- Combine multiple features extracted from
  - frames (VGG, ResNet, C3D)
  - spatial-temporal volumes
  - audio (MFCC) segments.

- Trained on MSR-VTT and their MANet method described at MULTMEDIA 2017

7. RUC CMU

- Is this the same as INF ?
- Renmin University of China and Carnegie Mellon University
- Alex Hauptmann et al.

8. Shandong Normal University, China

- Cross-modal retrieval method learns mapping matrices and projects different modality features into a common latent space where similarity can be measured directly

- Used MSR-VTT for training, extracted one keyframe per second and the Inception V3 NCC to identify keyframe features

9. TJU

- Tianjin University, China and National University of Singapore

- Builds on LSTM for sentence generation with one attention layer

- Research question is whether additional data can boost video captioning, so added MSVD and MSR-VTT datasets

11. UTS CAI

- University of Technology Sydney, Australia

- 3 sub-modules for
  - feature extraction (ResNet and C3D),
  - feature aggregation (Recurrent NNs, specifically HRNN and MVRM)
  - sentence generation

- Runs trained on MSR VTT only, MSVD + TRECVid2016, and MSR VTT + TRECVid2016, MS-COCO

12. VIREO

- Video Retrieval Group, City University of Hong Kong

- A spatio-temporal attention network to learn inter-modality correspondence without explicit concept detectors.

- Select the most salient parts of videos in both spatial and temporal dimensions.
  - no attention model
  - spatial attention model
  - temporal attention model

- LSTM used to generate captions

13. INF(ormedia)

Matching and Ranking

- Build models that have better discriminative ability as the model needs to distinguish between different captions give the video

Description Generation

- Focus on generalization of caption generation, two research questions with results:

1. Which one is more promising for better generalization on unseen datasets, high quality training dataset or more robust model - its the dataset
2. Can get more stable generalization ability by ensembling more different models

Used MS-COCO for training

15. DL-61-86

University of Sydney, Australia and Zhejiang University, China

Matching & Ranking task

- Based on Word2VisualVec, improved by replacing the average pooling on the textual input with the multi-scale sentence vectorization and a newly devised Spatial Enhanced Representation (SER).
- Best run is the ensemble of ten models which are variants of Word2VisualVec and SER but all runs are very good
- Trained on MSR-VTT (2016)

16. CMU_BOSCH

- CMU LTU + U Calif Riverside + Bosch Associates

- Adopted the joint visual semantic embedding approach for image-text retrieval and applied to video-text retrieval task using key-frames.
- Trained on MS-COCO and TRECVid VTT
- Focus on KF identification, extracted 4x KFs per video
- Matching & Ranking task submissions

Observations

- Good number of participation, again, growing, will renew this task
- Results *appear* to be better than last year.
- Is there value in the caption ranking sub-task ?
- We used CIDEr as well as BLEU and METEOR and STS
- STS as a metric has some questions, making us ask what makes more sense? MT metrics or semantic similarity ? Which metric measures real system performance in a realistic application ?
- Direct Annotation was introduced – its very clever
- Lots of available training sets, some overlap ... MSR-VTT (MSR Video to Language ACM MM Challenge), MS-COCO, Place2, ImageNet, YouTube2Text, MSVD (Microsoft YouTube clips), TRECVid2016 VTT, MPII-MD (Max Planck Institute), MVAD (Montreal Institute for Learning),
- What did individual teams learn ?