TRECVID 2016

Video to Text Description
NEW *Showcase / Pilot Task(s)*

Alan Smeaton
Dublin City University

Marc Ritter
Technical University Chemnitz

George Awad
NIST; Dakota Consulting, Inc

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

TASK

- Given a set of :
  - ➢2000 URLs of Twitter vine videos.
  - ➢2 sets (A and B) of text descriptions for each of 2000 videos.

- 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 set of A and of B.
  2. Description Generation:

     Automatically generate a text description for each URL.

Video Dataset

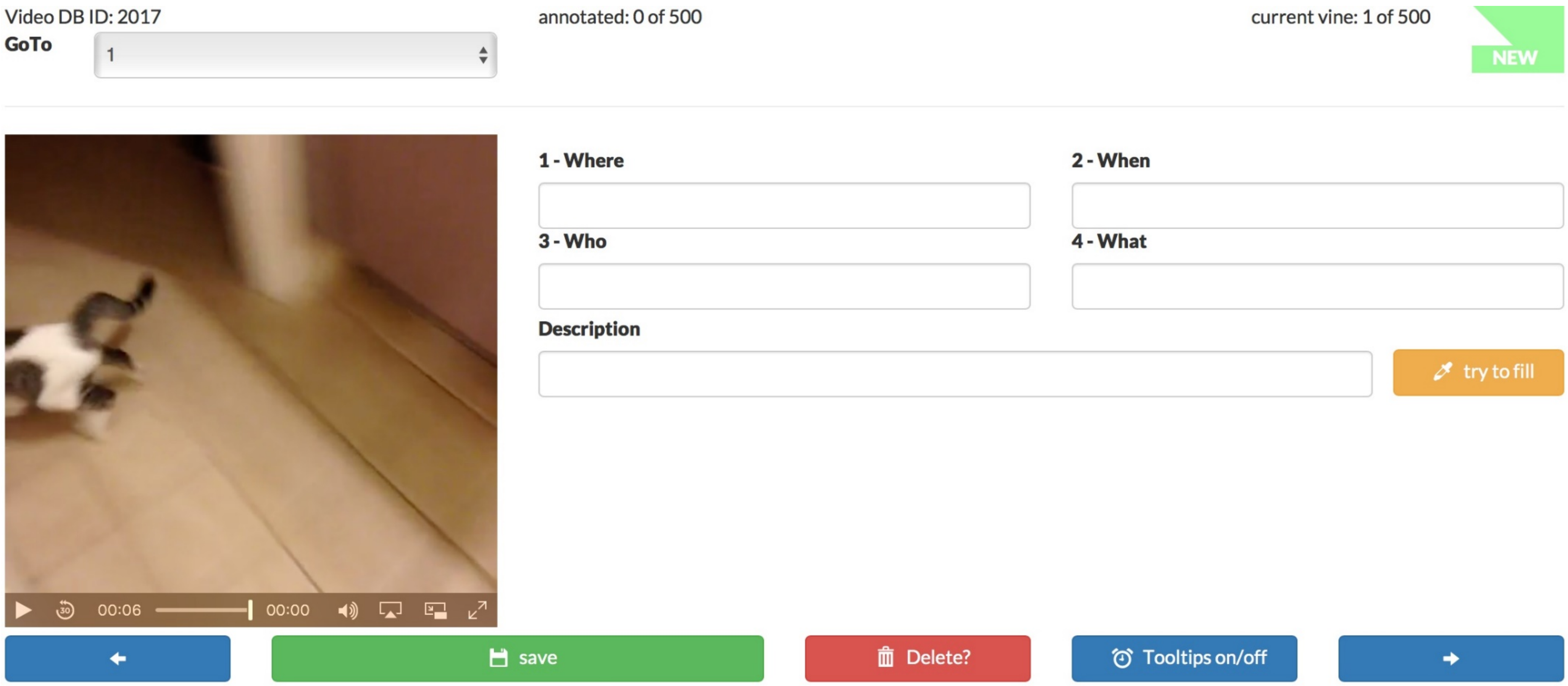
- Crawled 30k+ Twitter vine video URLs.
- Max video duration == 6 sec.
- A subset of 2000 URLs randomly selected.
- Marc Ritter's TUC Chemnitz group supported manual annotations:
  - Each video annotated by 2 persons (A and B).
  - In total 4000 textual descriptions (*1 sentence each*) were produced.
  - 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) ?

NIST
National Institute of Standards and Technology

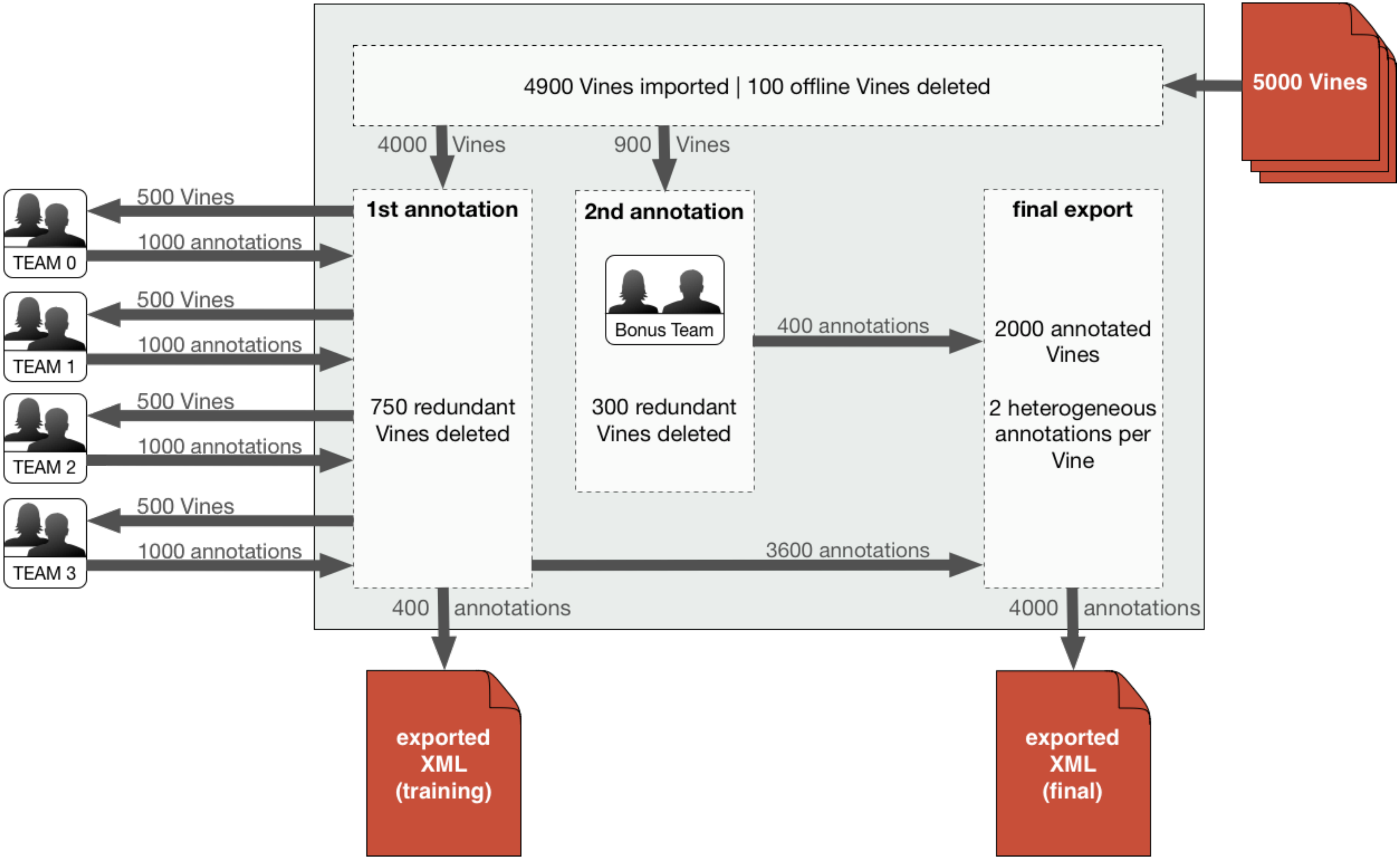
Annotation Process Obstacles

- Bad video quality
- A lot of simple scenes/events with repeating plain descriptions
- A lot of complex scenes containing too many events to be described
- Clips sometimes appear too short for a convenient description
- Audio track relevant for description but has not been used to avoid semantic distractions
- Non-English Text overlays/subtitles hard to understand
- Cultural differences in reception of events/scene content
- Finding a neutral scene description appears as a challenging task
- Well-known people in videos may have influenced (inappropriately) the description of scenes
- Specifying time of day (frequently) impossible for indoor-shots
- Description quality suffers from long annotation hours
- Some offline vines were detected
- A lot of vines with redundant or even identical content

Annotation UI Overview



Annotation Process



Annotation Statistics

UID	# annotations	Ø (sec)	(sec)	(sec)	# time (hh:mm:ss)
0	700	62.16	239.00	40.00	12:06:12
1	500	84.00	455.00	13.00	11:40:04
2	500	56.84	499.00	09.00	07:53:38
3	500	81.12	491.00	12.00	11:16:00
4	500	234.62	499.00	33.00	32:35:09
5	500	165.38	493.00	30.00	22:58:12
6	500	57.06	333.00	10.00	07:55:32
7	500	64.11	495.00	12.00	08:54:15
8	200	82.14	552.00	68.00	04:33:47
total	**4400**	**98.60**	**552.00**	**09.00**	**119:52:49**

Samples of captions

A	B
a dog jumping onto a couch	a dog runs against a couch indoors at daytime
in the daytime, a driver let the steering wheel of car and slip on the slide above his car in the street	on a car on a street the driver climb out of his moving car and use the slide on cargo area of the car
an asian woman turns her head	an asian young woman is yelling at another one that poses to the camera
a woman sings outdoors	a woman walks through a floor at daytime
a person floating in a wind tunnel	a person dances in the air in a wind tunnel

Run Submissions & Evaluation Metrics

- Up to 4 runs per set (for A and for B) were allowed in the *Matching & Ranking* subtask.
- Up to 4 runs in the *Description Generation* subtask.
- Mean inverted rank measured the *Matching & Ranking* subtask.
- Machine Translation metrics including BLEU (BiLingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) were used to score the *Description Generation* subtask.
- An experimental "Semantic Textual Similarity" metric (STS) was also tested.

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

- This is an active area … CIDEr (Consensus-Based Image Description Evaluation) is another recent metric … 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

Phrase 1:

two children playing frisbee on the beach

Phrase 2:

Frisbee players on a beach

Type: (•) 0 ( ) 1 ( ) 2

Get Similarity

0.8662101

Phrase 1:

two children playing frisbee on the beach

Phrase 2:

A child running on the sand

Type: (•) 0 ( ) 1 ( ) 2

Get Similarity

0.44439912

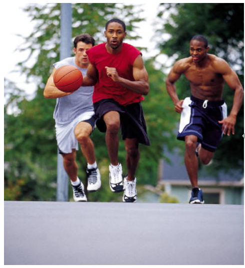
Participants (7 out of 11 teams finished)

	Matching & Ranking	Description Generation
DCU	✓	✓
INF(ormedia)	✓	✓
Mediamill (AMS)	✓	✓
NII (Japan + Vietnam)	✓	✓
Sheffield_UETLahore	✓	✓
VIREO (CUHK)	✓	
Etter Solutions	✓	

Total of 46 runs

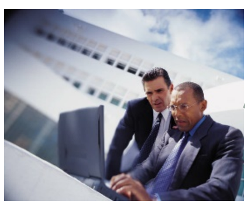
Total of 16 runs

Task 1: Matching & Ranking



Person reading newspaper outdoors at daytime

Person playing golf outdoors in the field



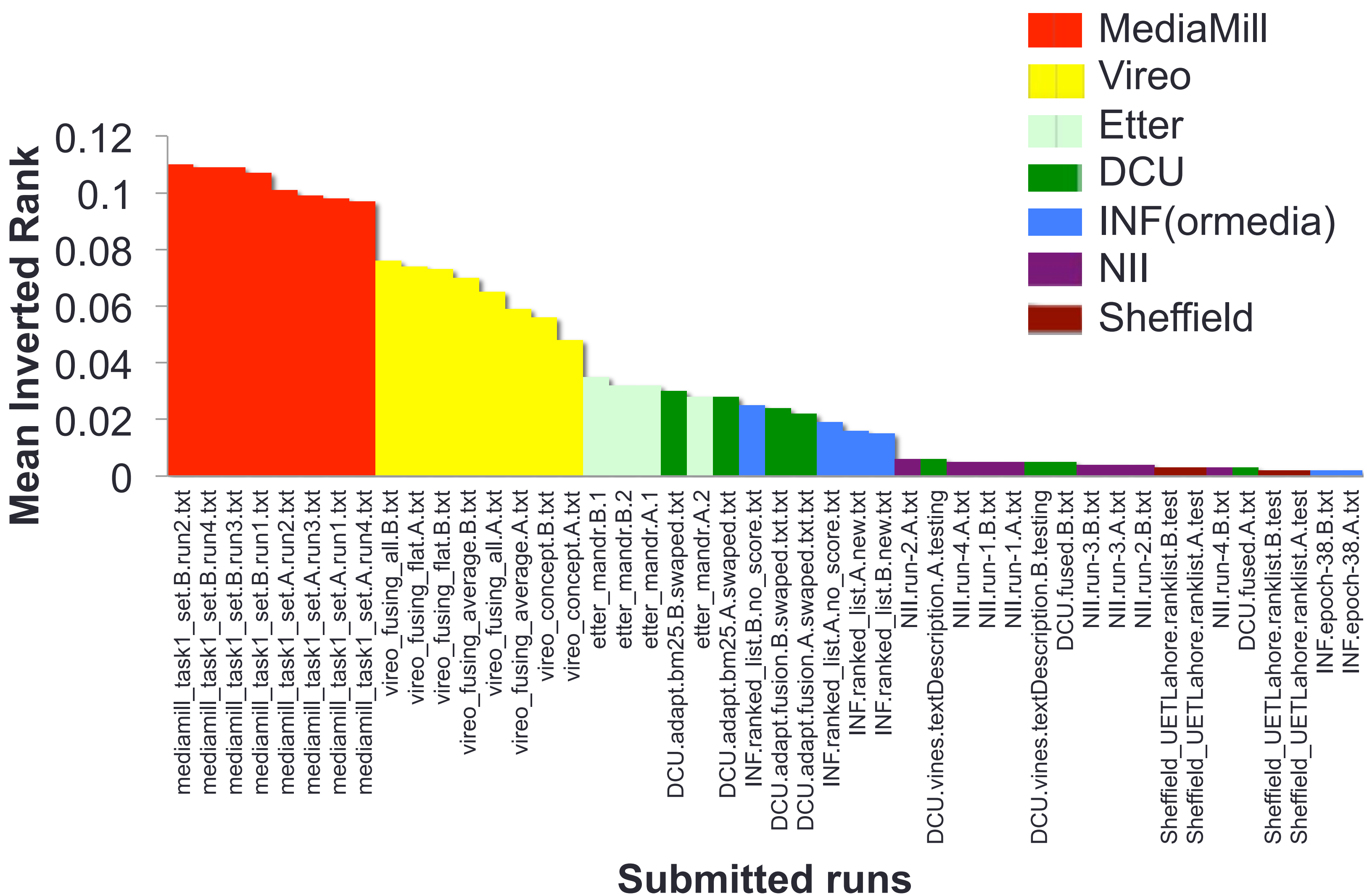
Three men running in the street at daytime

Two men looking at laptop in an office

x 2000

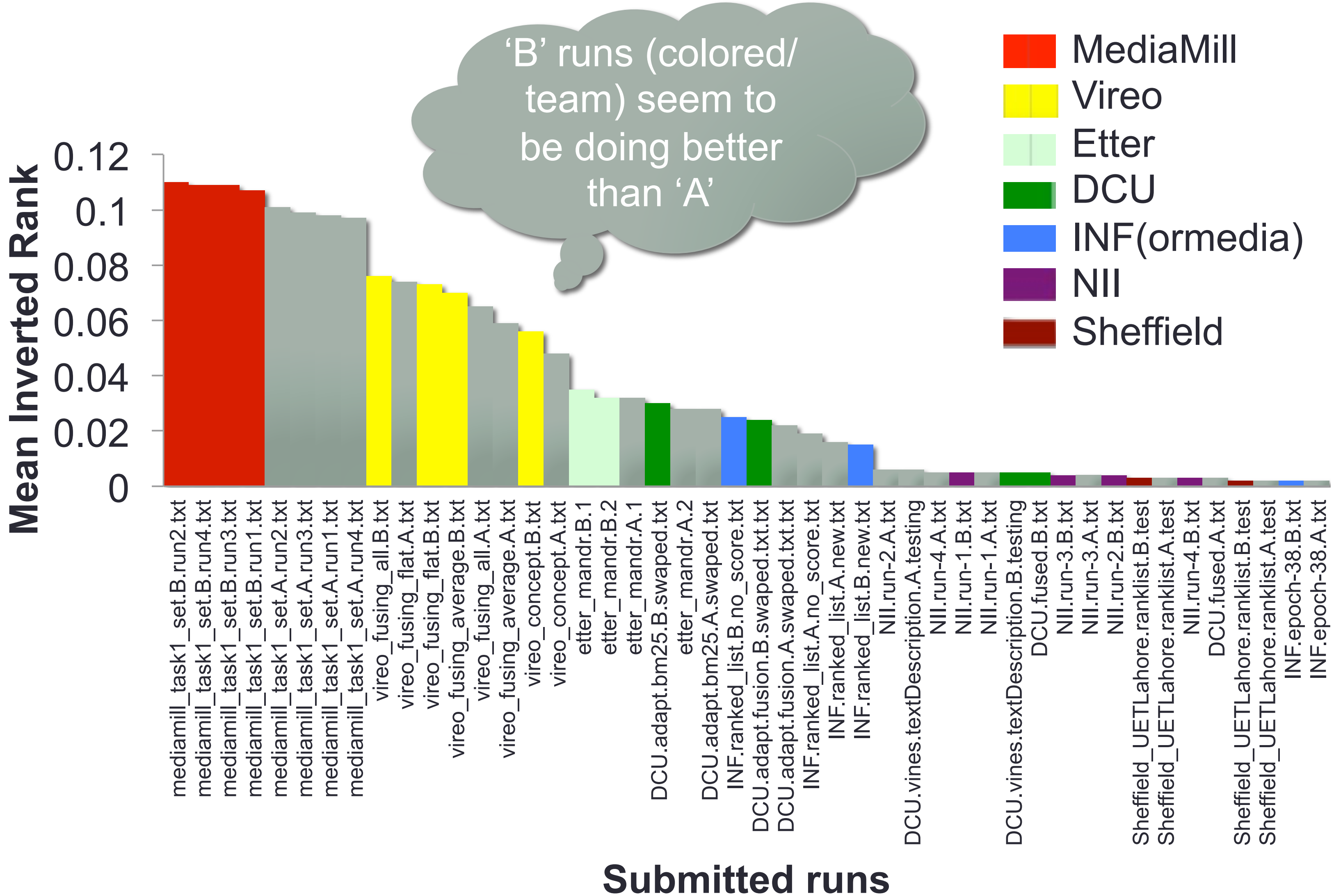
x 2000 type A … and ... X 2000 type B

Matching & Ranking results by run

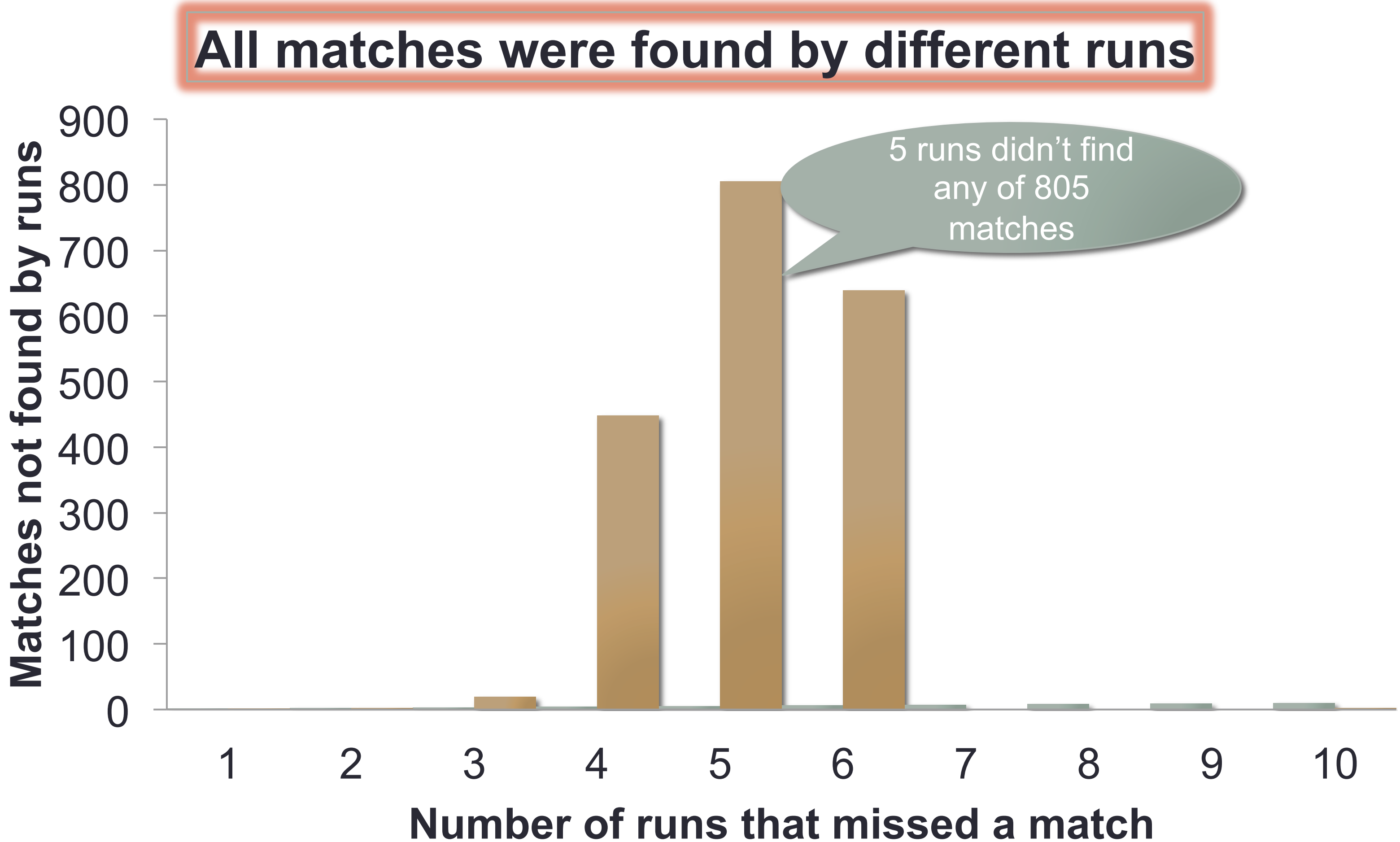


NIST
National Institute of Standards and Technology

Matching & Ranking results by run



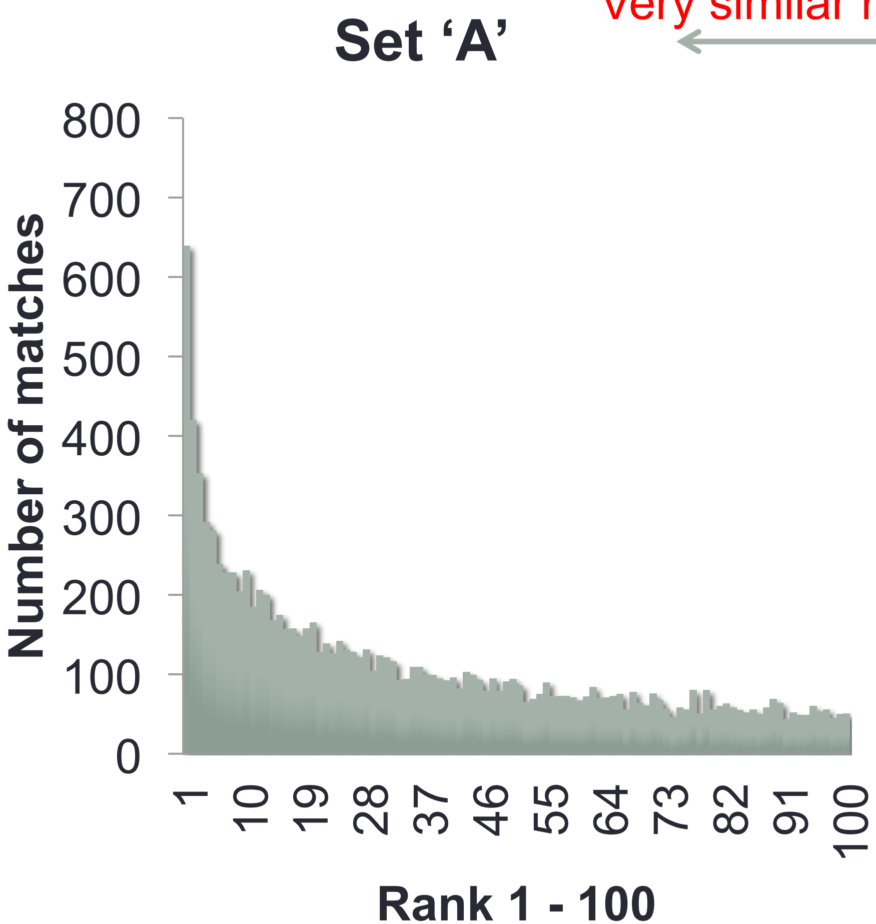
Runs vs. matches

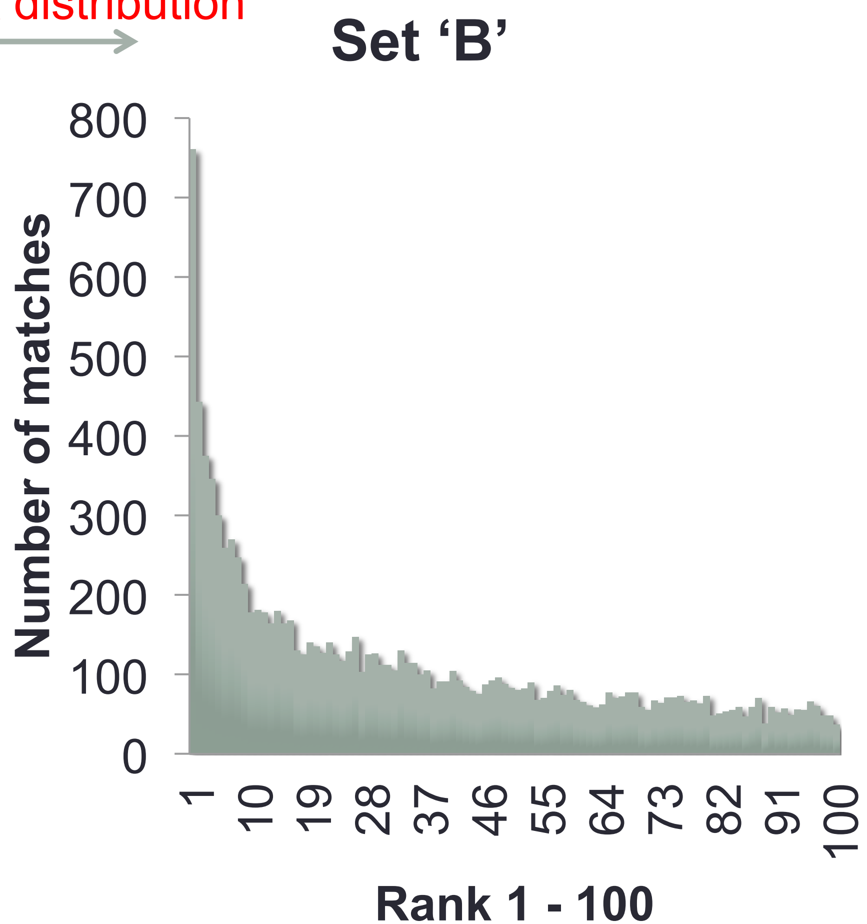


NIST National Institute of Standards and Technology

Matched ranks frequency across all runs

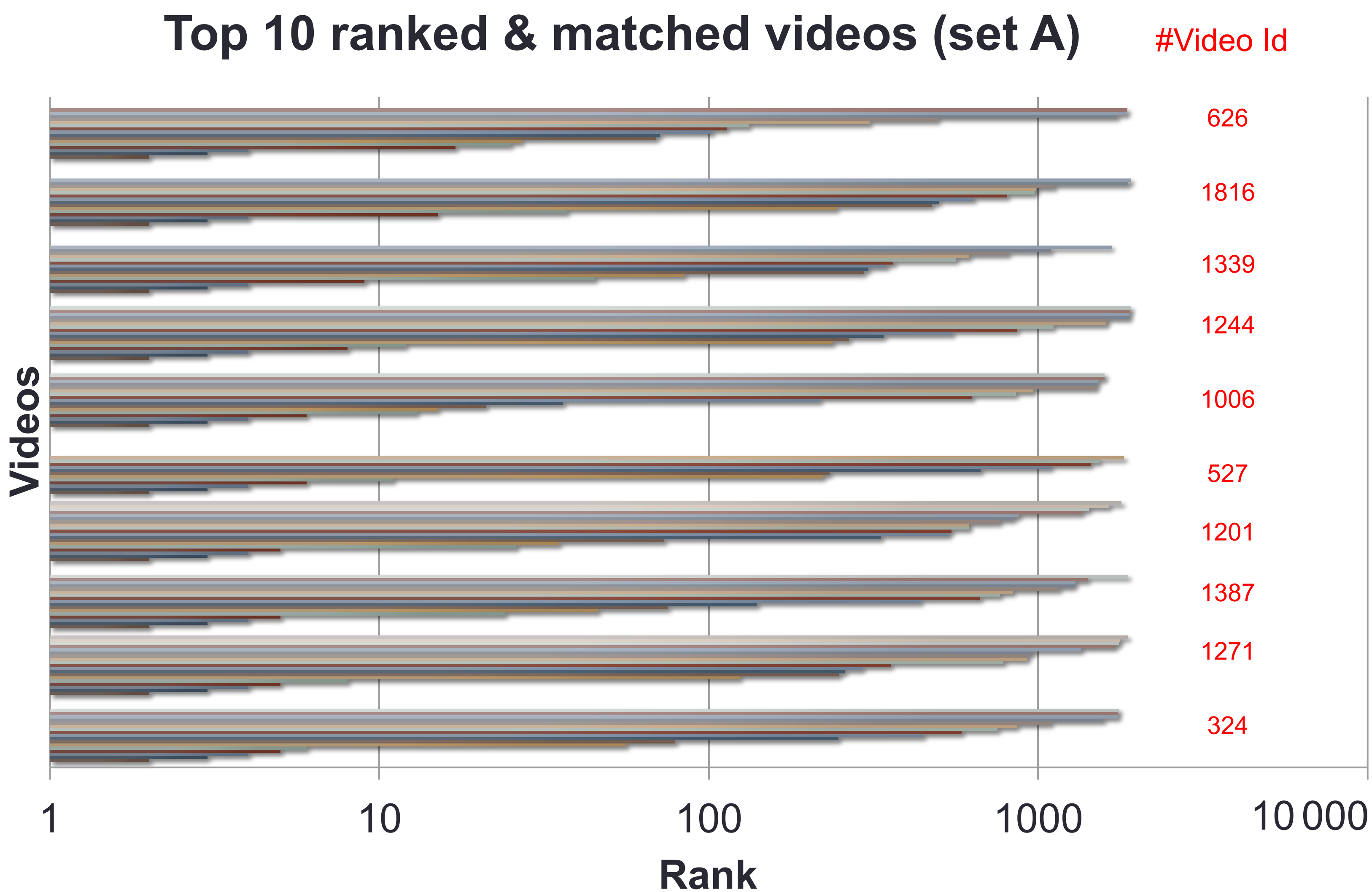
Very similar rank distribution



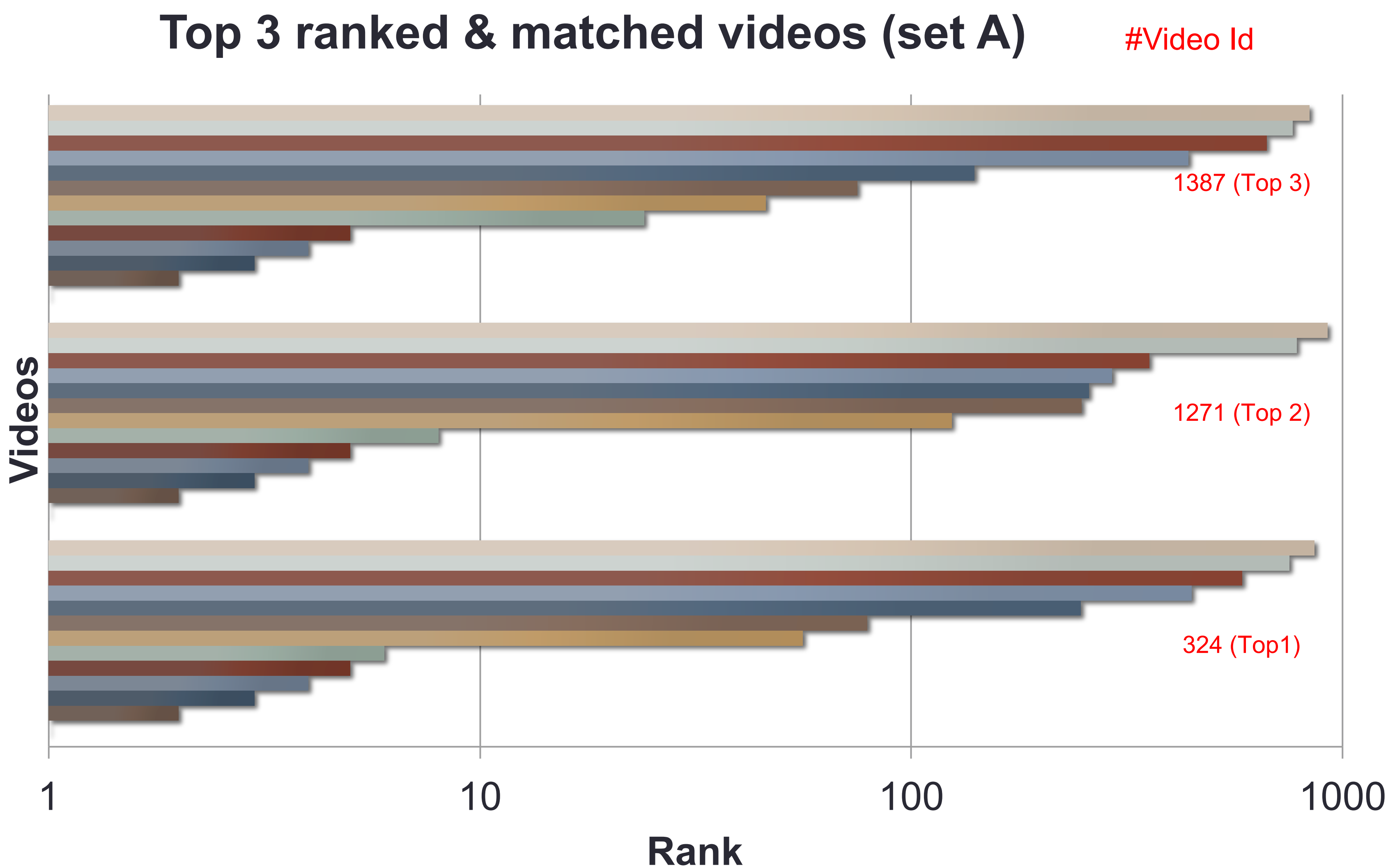


NIST
National Institute of Standards and Technology

Videos vs. Ranks



Videos vs. Ranks

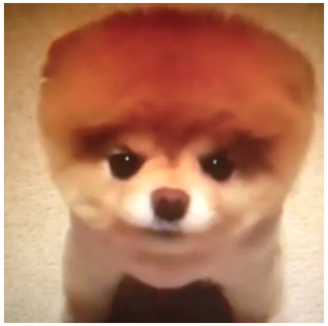


NIST
National Institute of Standards and Technology

Samples of top 3 results (set A)

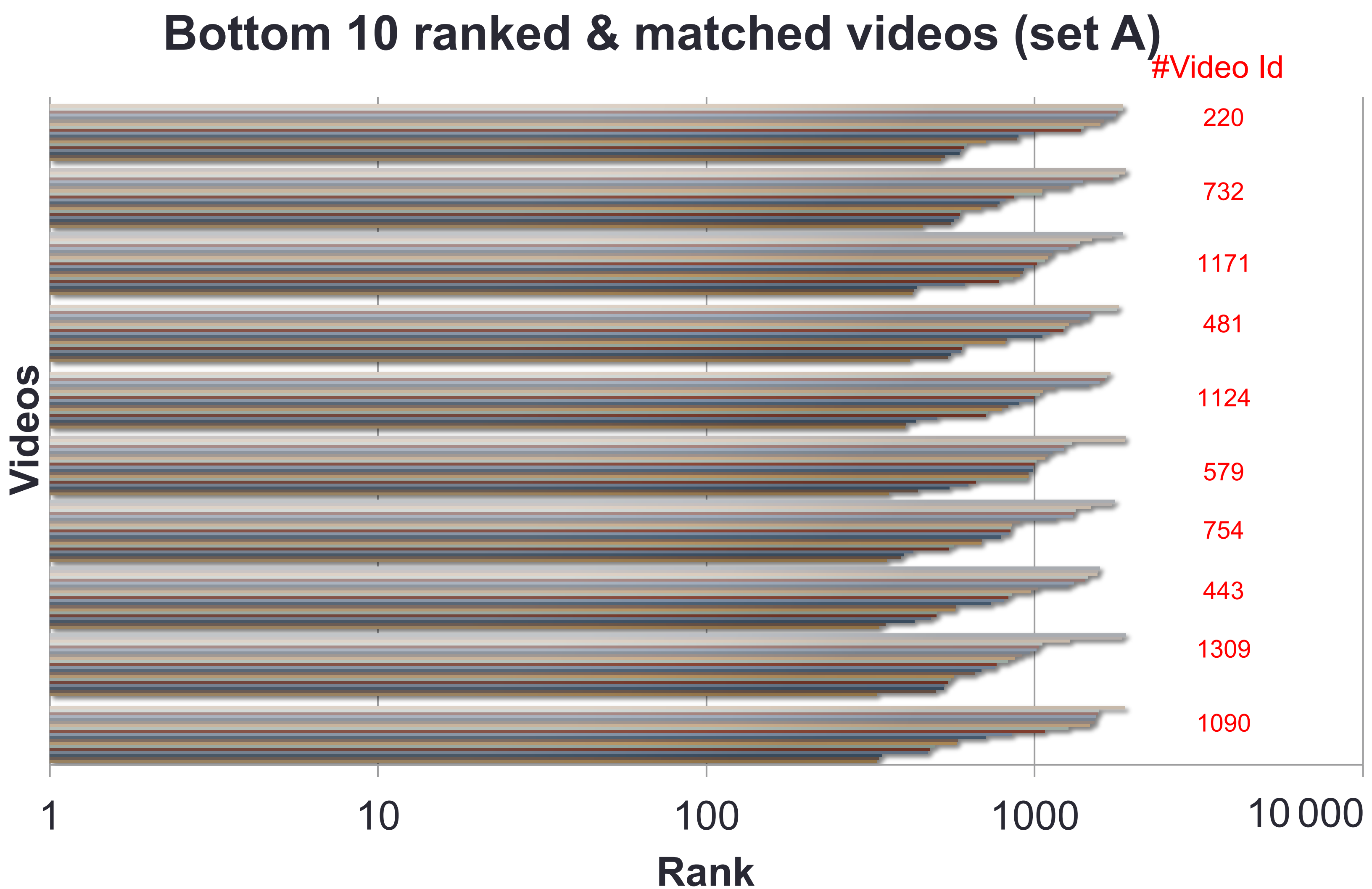
#1271
a woman and a man are kissing each other

#1387
a dog imitating a baby by crawling on the floor in a living room



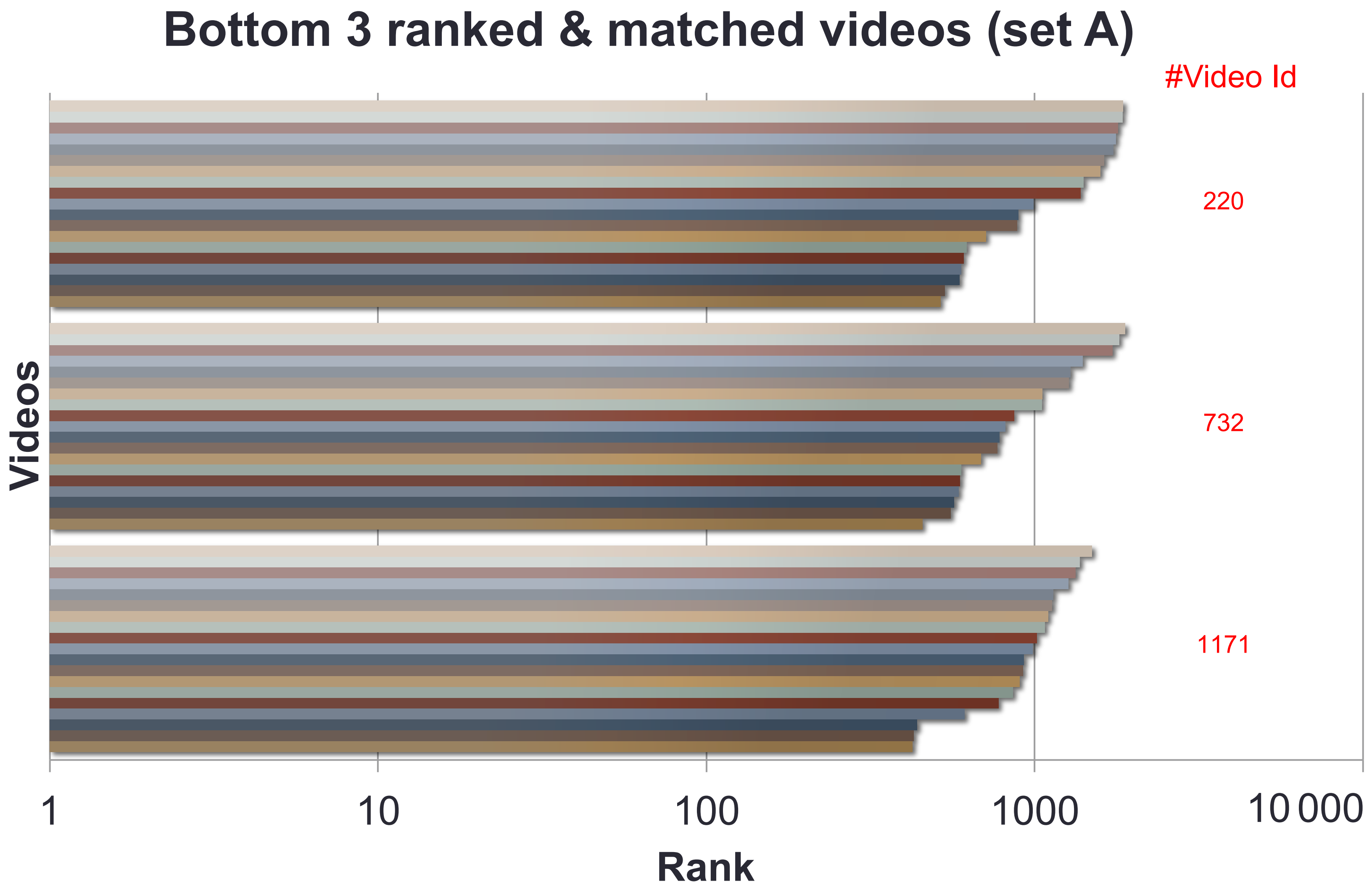
#324
a dog is licking its nose

Videos vs. Ranks



NIST
National Institute of Standards and Technology

Videos vs. Ranks



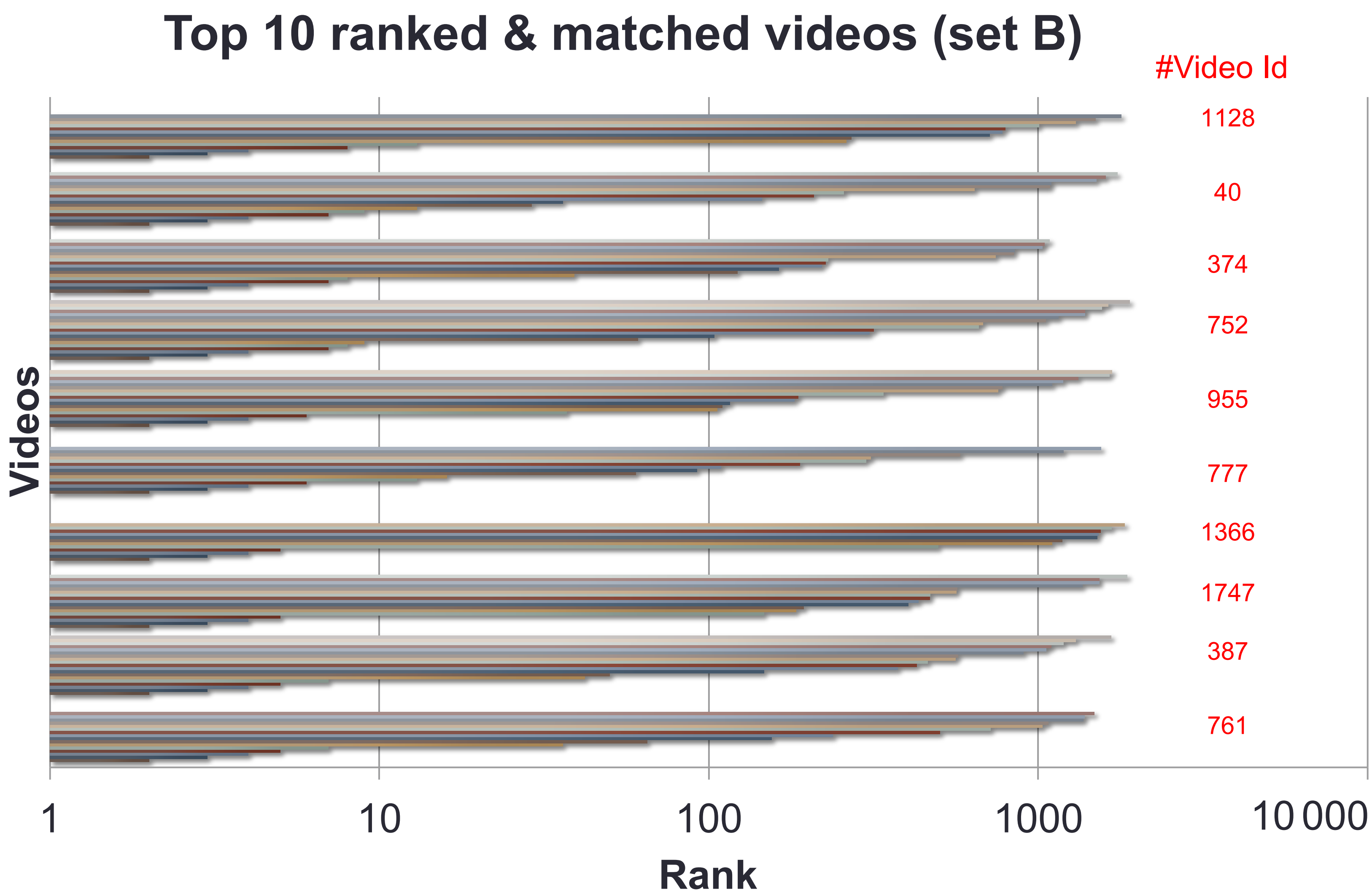
Samples of bottom 3 results (set A)

#1171
3 balls hover in front of a man

#220
2 soccer players are playing rock-paper-scissors on a soccer field

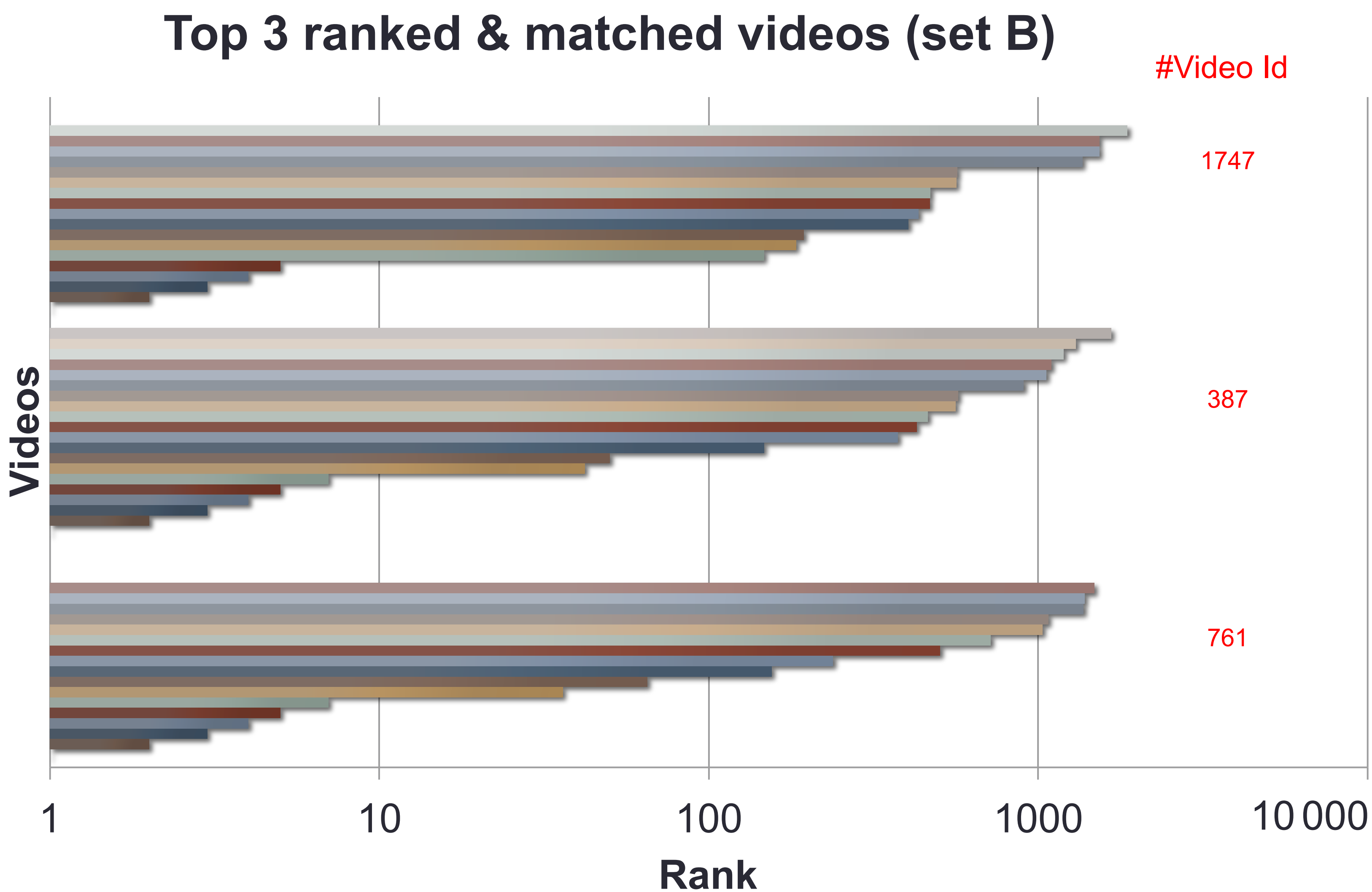
#732
a person wearing a costume and holding a chainsaw

Videos vs. Ranks

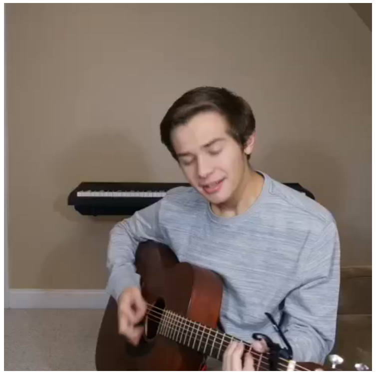


NIST National Institute of Standards and Technology

Videos vs. Ranks



Samples of top 3 results (set B)

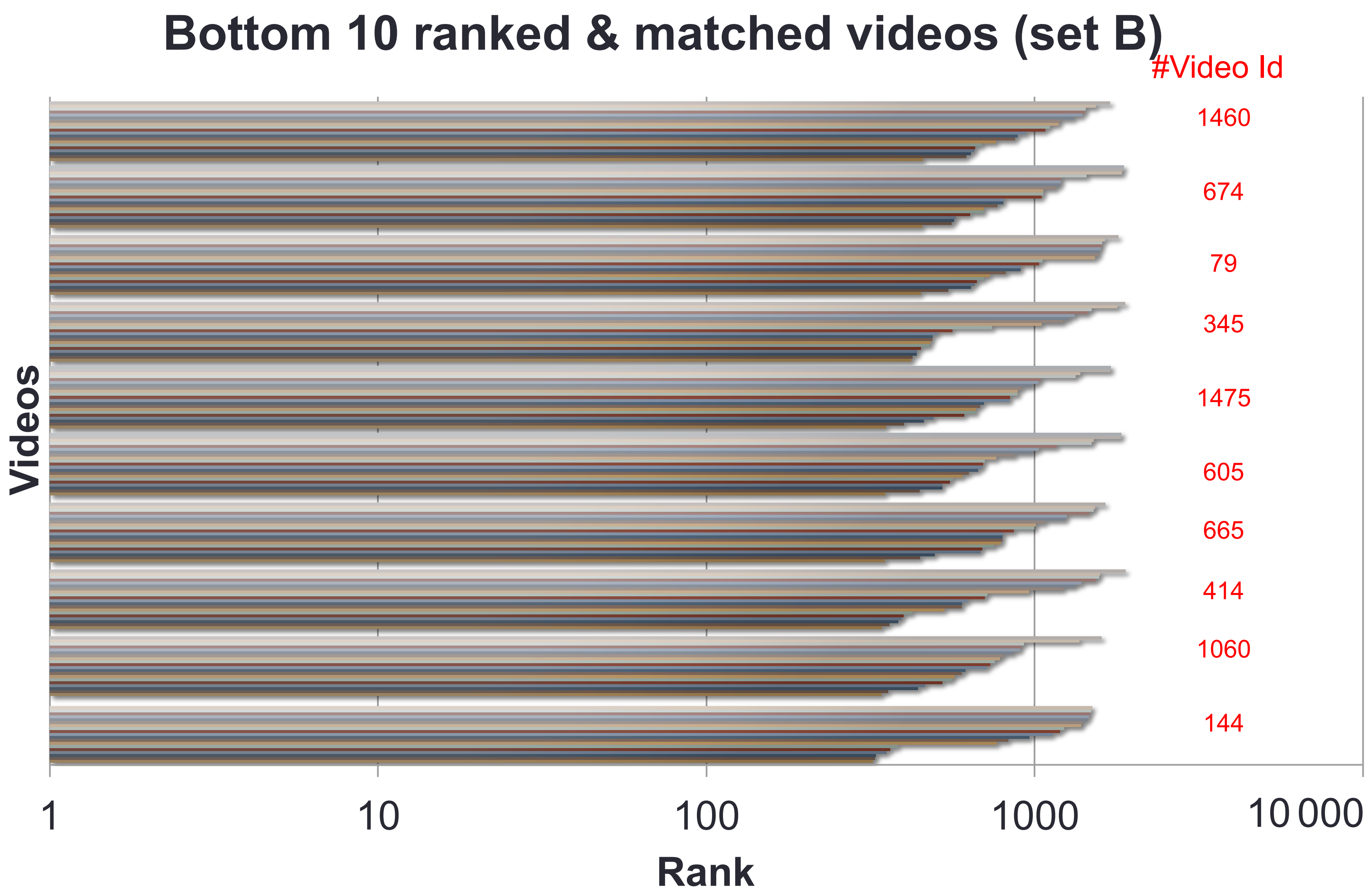


#761
White guy playing the guitar in a room

#387
An Asian young man sitting is eating something yellow

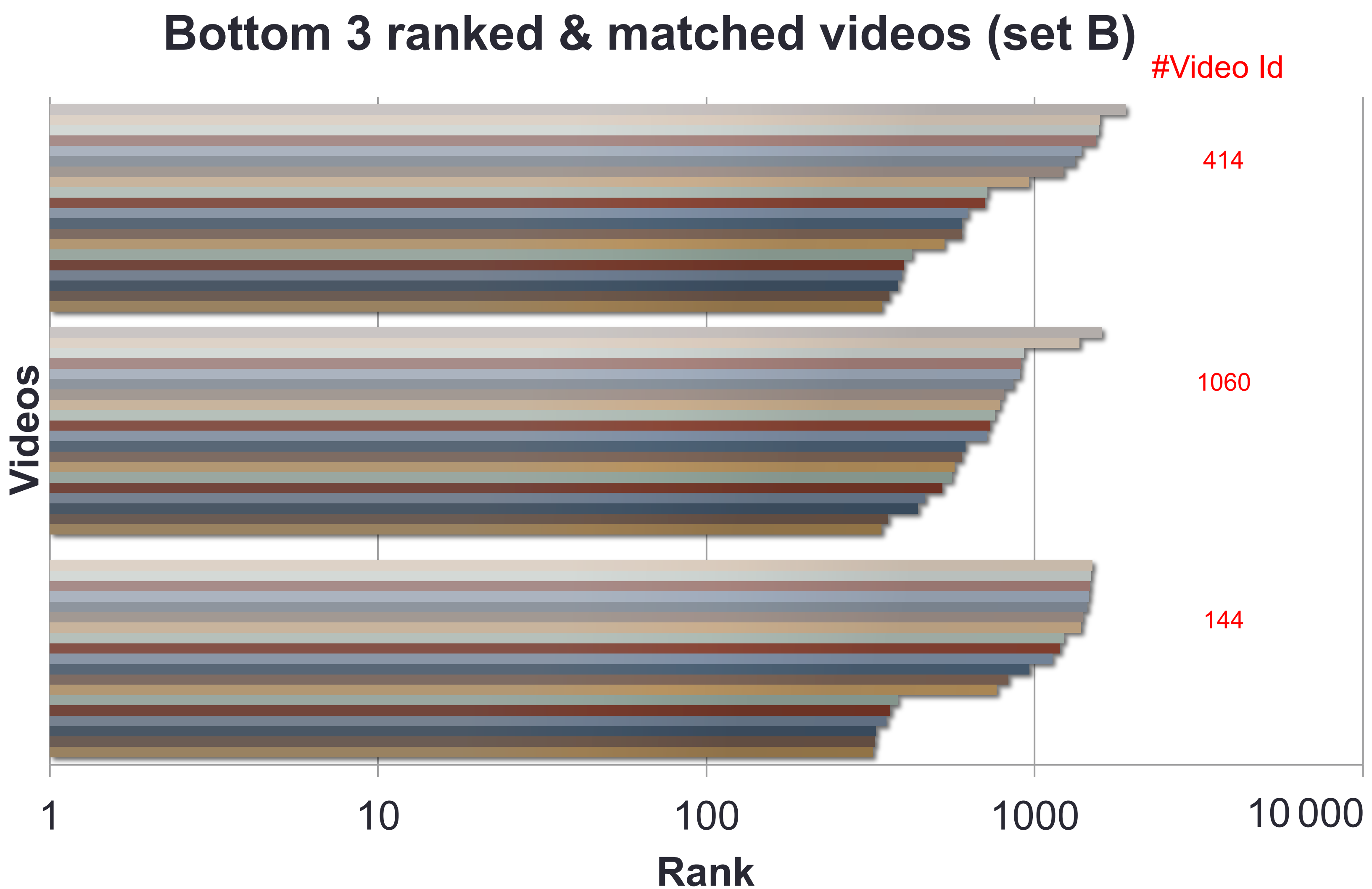
#1747
a man sitting in a room is giving baby something
to drink and it starts laughing

Videos vs. Ranks



NIST National Institute of Standards and Technology

Videos vs. Ranks



Samples of bottom 3 results (set B)

#144
A man touches his chin in a tv show

#1060
A man piggybacking another man outdoors

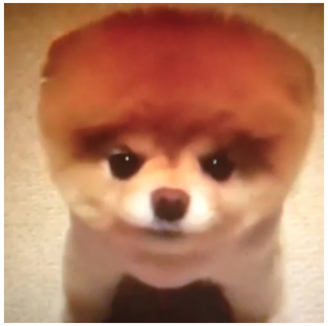
#414
a woman is following a man walking on the street at daytime trying to talk with him

Lessons Learned ?

- Can we say something about A vs B
- At the top end we’re not so bad … best results can find the correct caption in almost top 1% of ranking

Task 2: Description Generation

Given a video



Generate a textual description

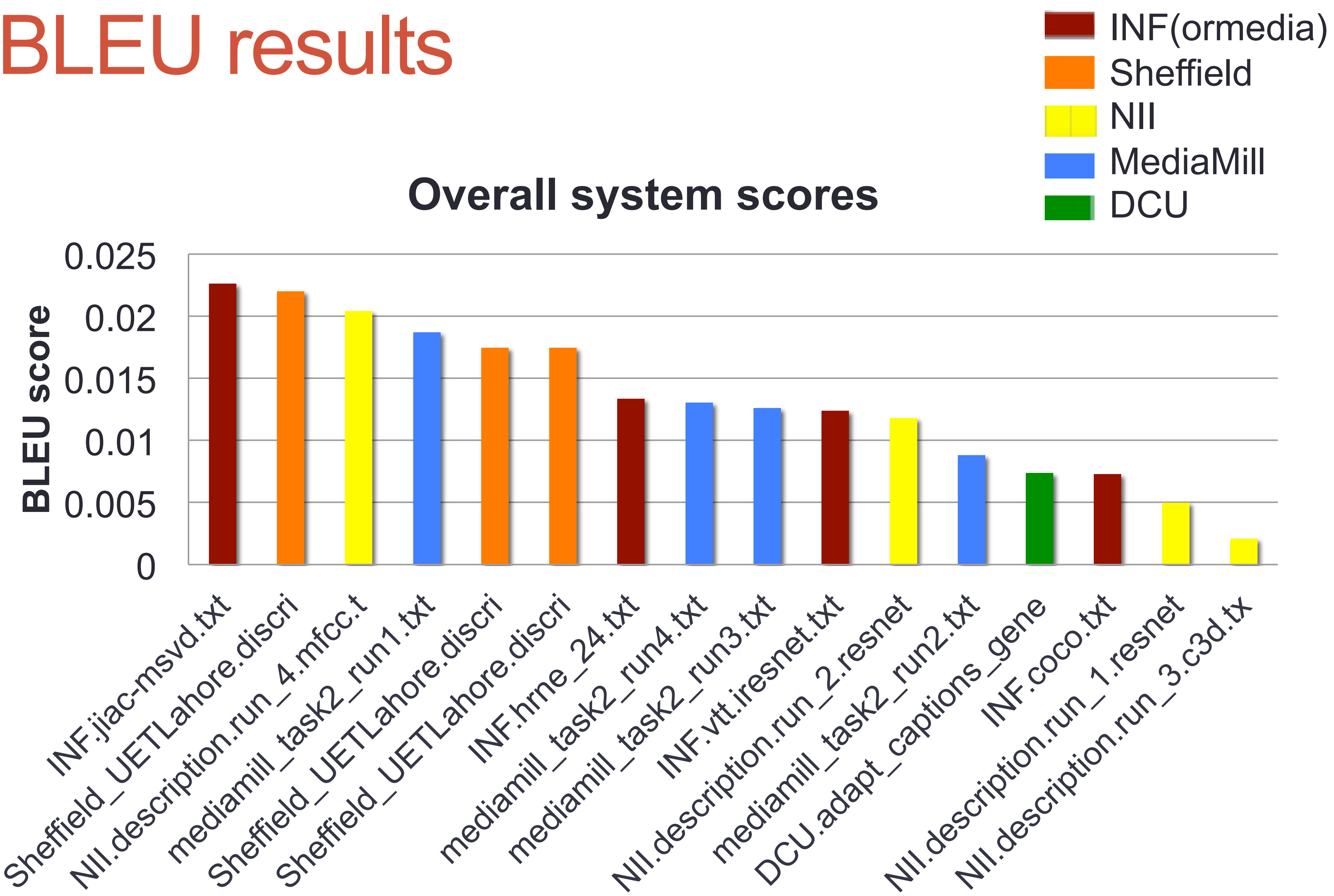
Who ? What ? Where ? When ?

“a dog is licking its nose”

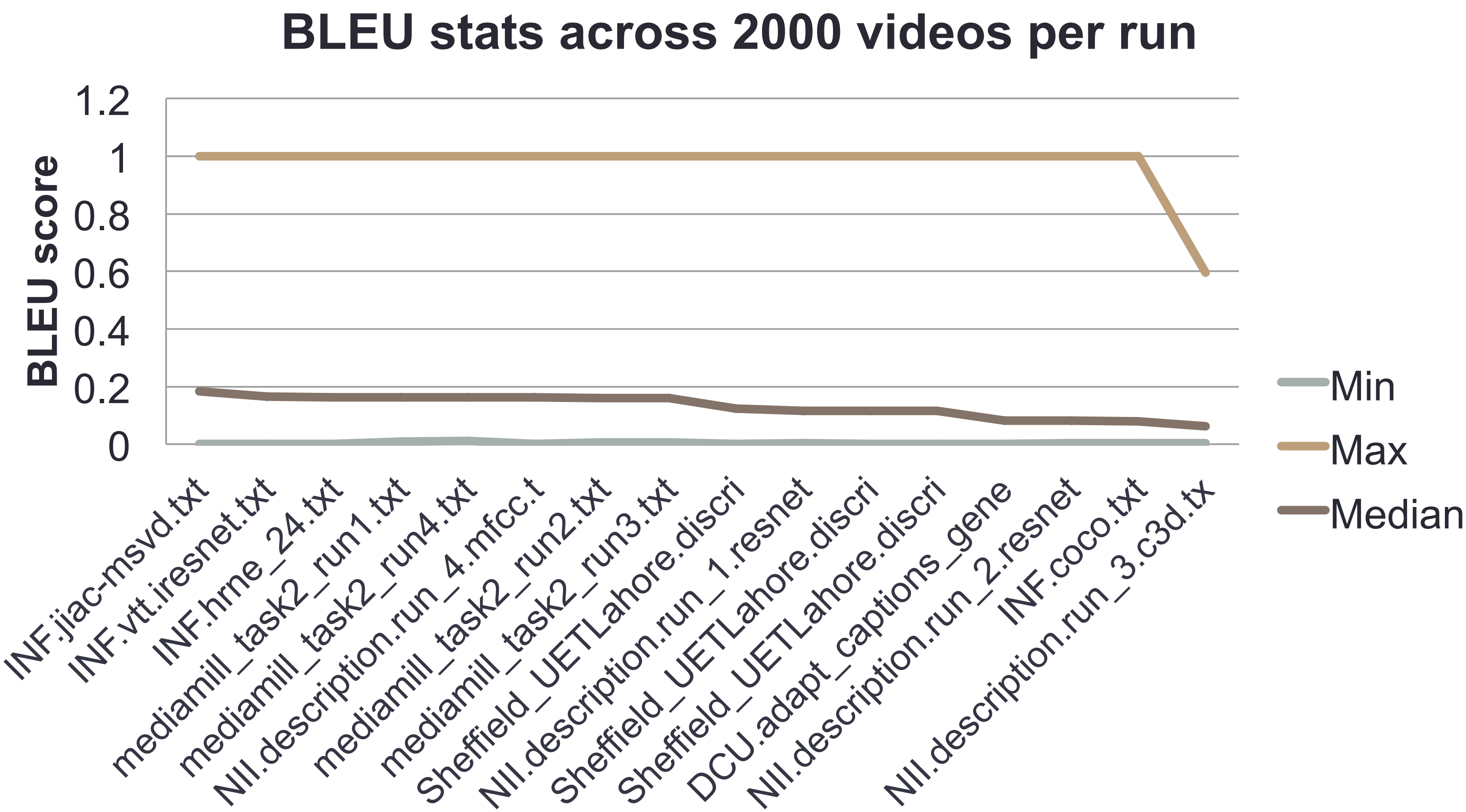
Metrics

- Popular MT measures : BLEU , METEOR
- Semantic textual similarity measure (STS).
- All runs and GT were normalized (lowercase, punctuations, stop words, stemming) before evaluation by MT metrics (except STS)

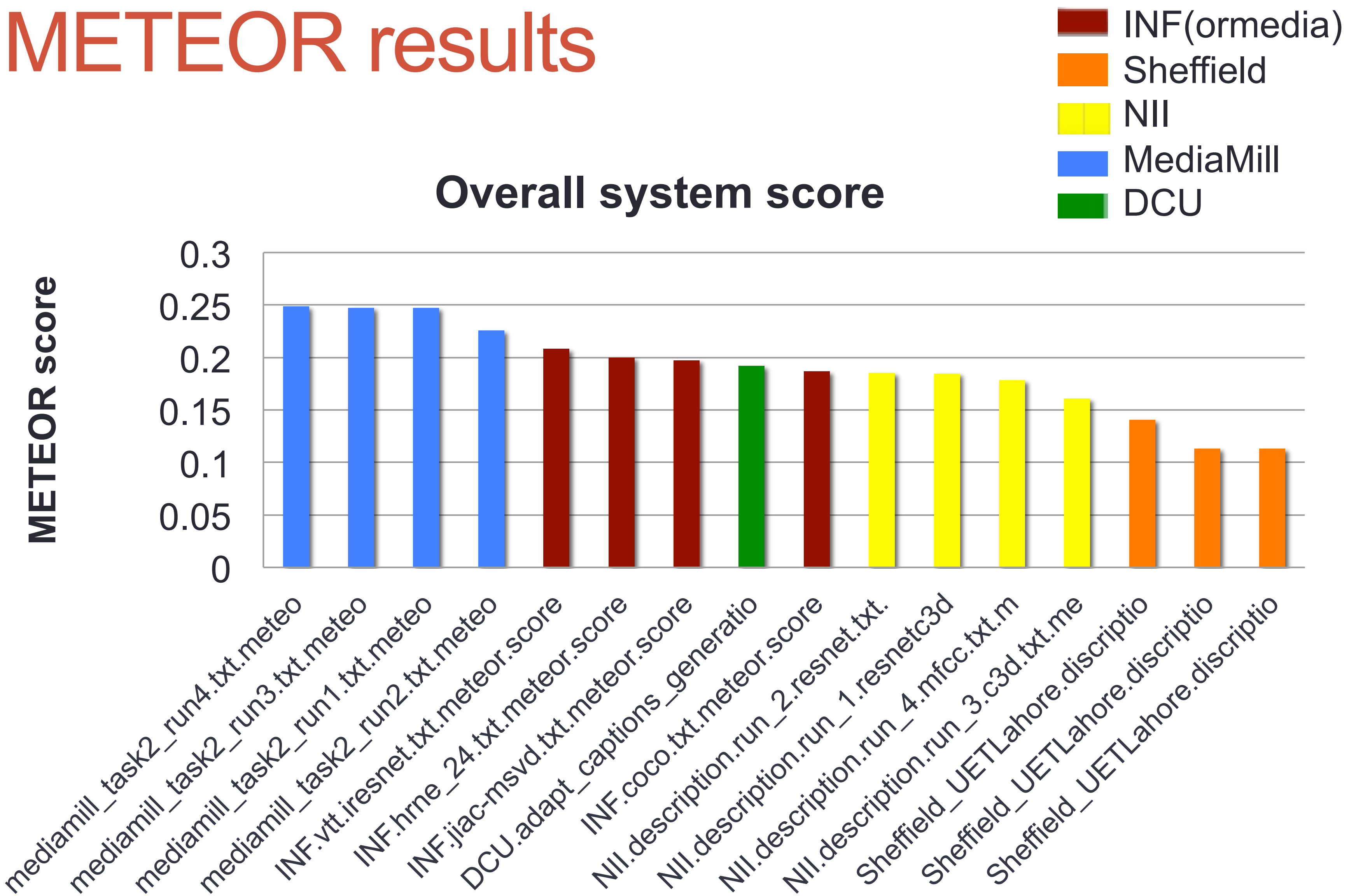
BLEU results



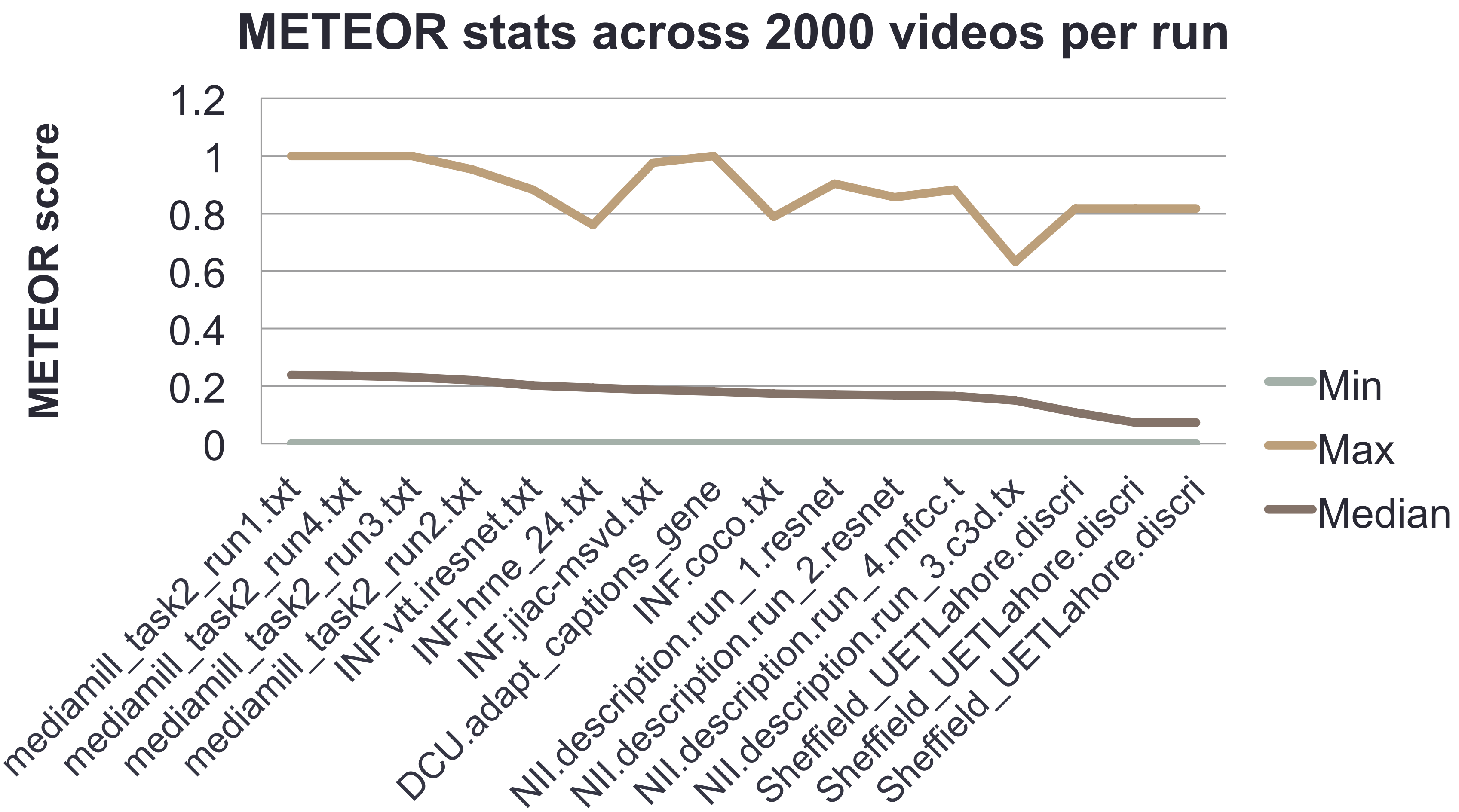
BLEU stats sorted by median value



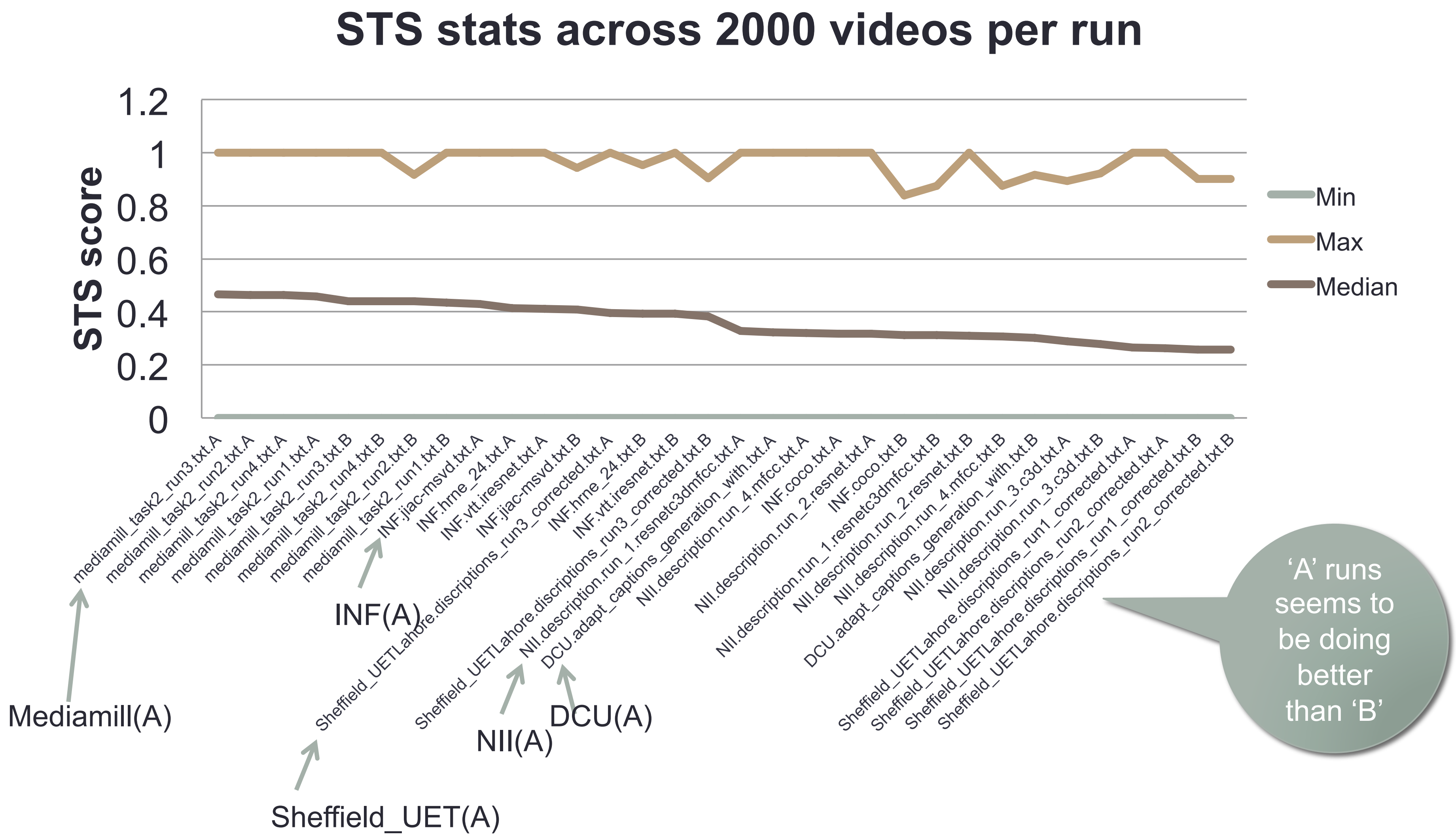
METEOR results



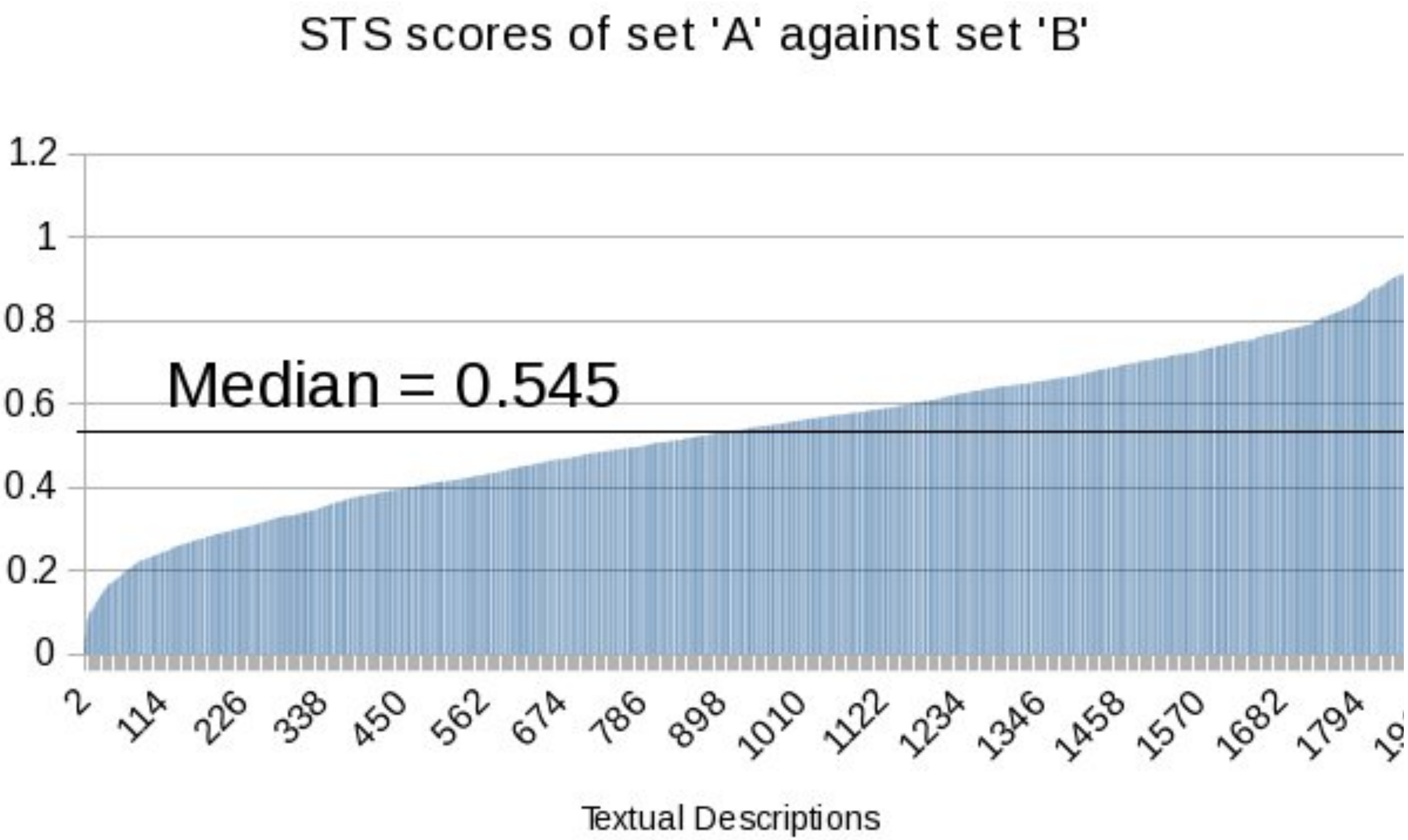
METEOR stats sorted by median value



Semantic Textual Similarity (STS) sorted by median value



STS(A, B) Sorted by STS value



An example from run submissions – 7 unique examples

1. a girl is playing with a baby
2. a little girl is playing with a dog
3. a man is playing with a woman in a room
4. a woman is playing with a baby
5. a man is playing a video game and singing
6. a man is talking to a car
7. A toddler and a dog

Participants

- High level descriptions of what groups did from their papers … more details on posters

Participant: DCU

Task A: Caption Matching

- Preprocess 10 frames/video to detect 1,000 objects (VGG-16 CNN from ImageNet), 94 crowd behaviour concepts (WWW dataset), locations (Place2 dataset on VGG16)
- 4 runs, baseline BM25, Word2vec, and fusion

Task B: Caption Generation

- Train on MS-COCO using NeuralTalk2, a RNN
- One caption per keyframe, captions then fused

Participant: Informedia

Focus on generalization ability of caption models, ignoring Who, What, Where, When facets

Trained 4 caption models on 3 datasets (MS-COCO, MS-VD, MSR-VTT), achieving sota on those models based on VGGNet concepts and Hierarchical Recurrent Neural Encoder for temporal aspects

Task B: Caption Generation

- Results explore transfer models to TRECVid-VTT

Participant: MediaMill

Task A: Caption Matching

Task B: Caption Generation

Participant: NII

Task A: Caption Matching

- 3DCNN for video representation trained on MSR-VTT + 1970 YouTube2Text + 1M captioned images
- 4 run variants submitted, concluding the approach did not generalise well on test set and suffers from over-fitting

Task B: Caption Generation

- Trained on 6500 videos from MSR-VTT dataset
- Confirmed that multimodal feature fusion works best, with audio features surprisingly good

Participant: Sheffield / Lahore

Task A: Caption Matching

Did some run

Task B: Caption Generation

- Identified a variety of high level concepts for frames
- Detect and recognize faces, age and gender, emotion, objects, (human) actions
- Varied the frequency of frames for each type of recognition
- Runs based on combinations of feature types

Participant: VIREO (CUHK)

Adopted their zero-example MED system in reverse

Used a concept bank of 2000 concepts trained on MSR-VTT, Flickr30k, MS-COCO and TGIF datasets

Task A: Caption Matching

- 4(+4) runs testing traditional concept-based approach vs attention-based deep models, finding deep models perform better, motion features dominate performance

Participant: Etter Solutions

Task A: Caption Matching

- Focused on concepts for Who, What, When, Where
- Used a subset of ImageNet plus scene categories from the Places database
- Applied concepts to 1 fps (frame per second) with sliding window, mapped this to “document” vector, and calculated similarity score

Observations

- Good participation, good finishing %, ‘B’ runs did better than ‘A’ in matching & ranking while ‘A’ did better than ‘B’ in the semantic similarity.
- METEOR scores are higher than BLEU, we should have used CIDEr also (some participants did)
- 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 ?
- Lots of available training sets, some overlap ... MSR-VTT, MS-COCO, Place2, ImageNet, YouTube2Text, MS-VD .. Some trained with AMT (MSR-VTT-10k has 10,000 videos, 41.2 hours and 20 annotations each !)
- What did individual teams learn ?
- Do we need more reference (GT) sets ? (good for MT metrics)
- Should we run again as pilot ? How many videos to annotate, how many annotations on each?
- Only some systems applied the 4-facet description in their submissions ?

Observations

- There are other video-to-caption challenges like ACM MULTIMEDIA 2016 Grand Challenges
- Images from YFCC100N with captions in a caption-matching/prediction task for 36 884 test images. Majority of participants used CNNs and RNNs
- Video MSR VTT with 41.2h, 10 000 clips each with x20 AMT captions … evaluation measures BLEU, METEOR, CIDEr and ROUGE-L ... GC results do not get aggregated and disssipate at the ACM MM Conference, so hard to gauge.