Findings of the Second Shared Task on Multimodal Translation and Multilingual Image Description

Desmond Elliott[*], Stella Frank[*], Loïc Barrault[†], Fethi Bougares[†], Lucia Specia[◦]

[*]University of Edinburgh, [†]University of Le Mans, [◦]University of Sheffield

Key Idea: visual context can improve translation

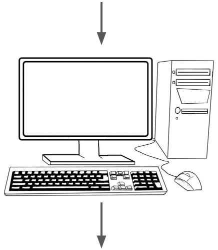
A wall divided the city

Eine Wand teilte die Stadt

Credit: Stella Frank (WMT 2016)

Key Idea: visual context can improve translation

A wall divided the city



~~Eine Wand teilte die Stadt~~

Credit: Stella Frank (WMT 2016)

Key Idea: visual context can improve translation

A wall divided the city

Eine **Mauer** teilte die Stadt

Credit: Stella Frank (WMT 2016)

Multimodality improves semantic classes

Source: A woman wearing a **hat** is making bread.

No Image: Eine Frau mit einer Mütze macht Brot.

Credit: Specia et al. (2016)

Multimodality improves semantic classes

Source: A woman wearing a **hat** is making bread.

No Image: Eine Frau mit einer Mütze macht Brot. ✗

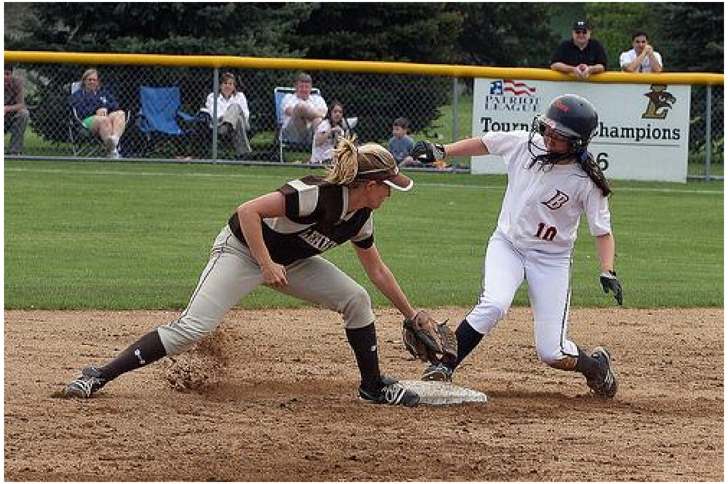
With Image: Eine Frau mit einem Hut macht Brot. ✔

Credit: Specia et al. (2016)

Multimodality improves gender marking

Source: A **baseball player** in a black shirt just tagged **a player** in a white shirt.

No Image: Ein Baseballspieler in einem schwarzen Shirt fängt einen Spieler in einem weißen Shirt.



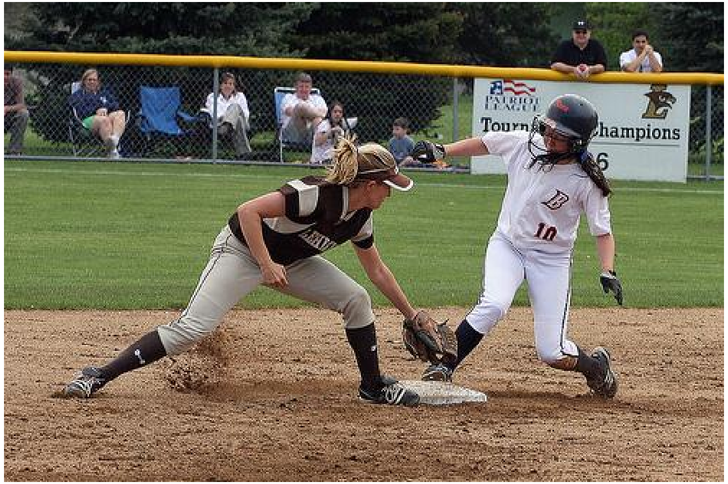
Credit: Specia et al. (2016)

Multimodality improves gender marking

Source: A **baseball player** in a black shirt just tagged **a player** in a white shirt.

With Image: Eine Baseballspielerin in einem schwarzen Shirt fängt eine Spielerin in einem Weißen Shirt.

✔



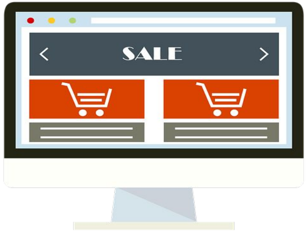
Credit: Specia et al. (2016)

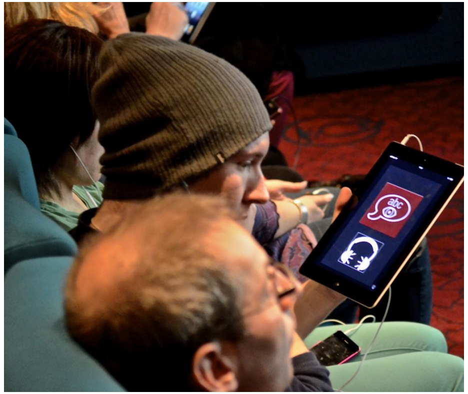
Use Cases for Multimodal Translation

- Localised alt-text generation across the Web
- Richer e-commerce experiences
- Audio described movies for more languages

The Danish flag flying against a cloudy sky

Det danske flag vajende mod en blå himmel





Task 1: Multimodal Machine Translation

Q: What can **images** bring to translation?

A bird flies
over the water

Model

Ein Vogel fliegt
über das Wasser

Task 2: Multilingual Image Description

- Source-target-image parallel data is **rare**
- More realistic:
  - unannotated images
  - monolingually described images
- We need models that can tolerate absent data

Task 2: Multilingual Image Description

- Q: What can **multilinguality** bring to image description?

Evaluation: only image

Model

Ein Vogel fliegt
über das Wasser

Task 2: Multilingual Image Description

- Q: What can **multilinguality** bring to image description?

Training: with source language and image

Model

Ein Vogel fliegt über das Wasser

A bird flies over the water

Data

Multi30K Dataset

31,000 Images

31,000 Professional Translations

155,000 Crowdsourced Descriptions

Elliott et al. (2016)

Translated Sentences

A brown dog is running after the black dog.

Ein brauner Hund rennt dem schwarzen Hund hinterher

Independent Descriptions

A brown dog is running after the black dog.

Ein schwarzer und ein brauner Hund rennen auf steinigem Boden aufeinander zu

New Data: Multi30K French

- Multi30K is now 4-way aligned

- 31,000 Images
  - En descriptions
  - De professional translations
  - Fr crowdsourced translations

En: A group of people are eating noodles.

De: Eine Gruppe von Leuten isst Nudeln.

Fr: Un groupe de gens mangent des nouilles.

New Data: Multi30K 2017 test

- Harvest 12K CC-licensed images from the Flickr30K photo groups
- Filter down to 2,071 new images
- Fewer near-duplicate images

Group	Task 1	Task 2
Strangers!	150	154
Wild Child	83	83
Dogs in Action	78	92
Action Photography	238	259
Flickr Social Club	241	263
Everything Outdoor	206	214
Outdoor Activities	4	6

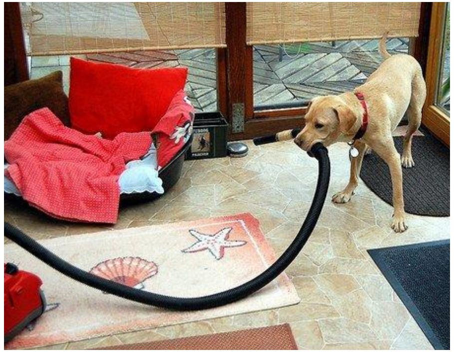
Fewer Near-Duplicates

- Less of this ...



Fewer Near-Duplicates

- More of this …



New Data: Ambiguous COCO (teaser)

- 461 images from the VerSe dataset (Gella et al., 2016)
- English verb sense ambiguity
- Covering 56 ambiguous verbs
  - Shake - 3 images (least)
  - Reach - 26 images (most)

Example of ambiguity: “to pass”

.. red train is passing over ..

Example of ambiguity: “to pass”

.. red train is passing over ..

.. on a motorcycle passing ..

Example of ambiguity: “to pass”

.. red train is passing over ..

.. on a motorcycle passing ..

Ein roter Zug fährt auf einer Brücke über das Wasser

German

Ein Mann auf einem Motorrad fährt an einem anderen Fahrzeug vorbei

Example of ambiguity: “to pass”

.. red train is passing over ..

.. on a motorcycle passing ..

Un train rouge traverse l'eau sur un pont.

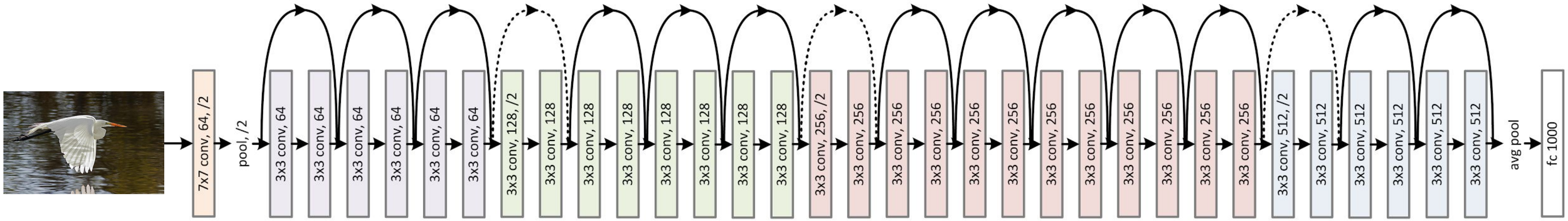
French

Un homme sur une moto **dépasse** un autre véhicule.

Provided Image Representation

Intermediate layers from ResNet-50 Convolutional Neural Network (He et al., 2016) trained on ImageNet for object recognition task:

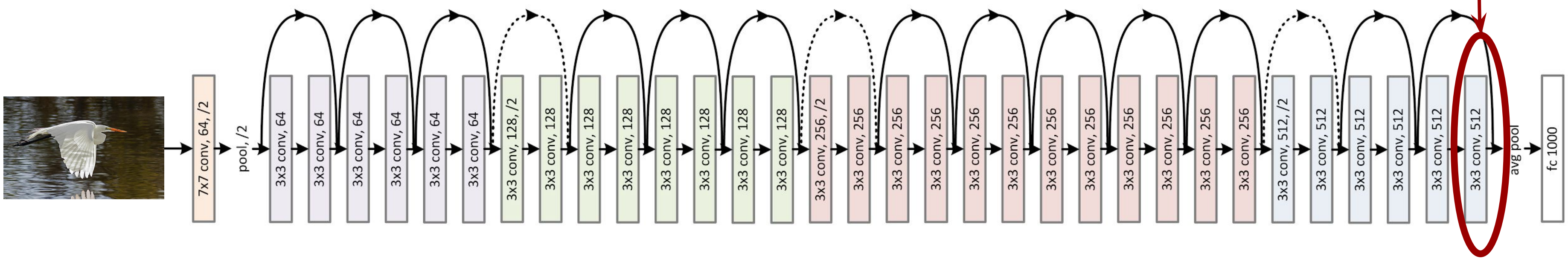
- `res4_relu`: last convolutional layer (14x14x1024D tensor)
- `avgpool`: pooled output of the final convolutional layer (2048D vector)



Provided Image Representation

Intermediate layers from ResNet-50 Convolutional Neural Network (He et al., 2016) trained on ImageNet for object recognition task:

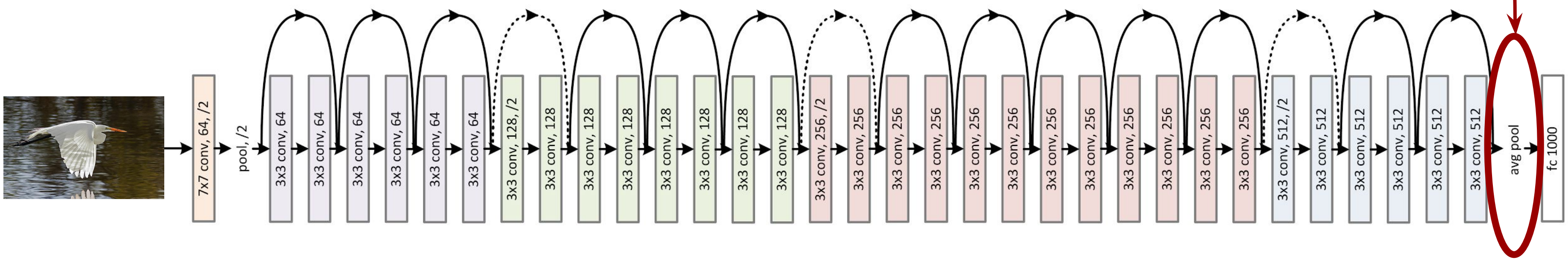
- `res4_relu`: last convolutional layer (14x14x1024D tensor)
- `avgpool`: pooled output of the final convolutional layer (2048D vector)



Provided Image Representation

Intermediate layers from ResNet-50 Convolutional Neural Network (He et al., 2016) trained on ImageNet for object recognition task:

- `res4_relu`: last convolutional layer (14x14x1024D tensor)
- `avgpool`: pooled output of the final convolutional layer (2048D vector)



Datasets overview

	Training set		Development set	
	Images	Sentences	Images	Sentences
Translation	29,000	29,000	1,014	1,014
Description	29,000	145,000	1,014	5,070

Datasets overview

	Training set		Development set	
	Images	Sentences	Images	Sentences
Translation	29,000	29,000	1,014	1,014
Description	29,000	145,000	1,014	5,070

	2017 test	
	Images	Sentences
Translation	1,000	1,000
Description	1,071	5,355

Datasets overview

	Training set		Development set	
	Images	Sentences	Images	Sentences
Translation	29,000	29,000	1,014	1,014
Description	29,000	145,000	1,014	5,070
	2017 test		**COCO**	
	Images	Sentences	Images	Sentences
Translation	1,000	1,000	461	461
Description	1,071	5,355	—	

Main questions for this year

1. Do multimodal systems improve on text-only systems?
    - Text-similarity and human assessments this year

Main questions for this year

1. Do multimodal systems improve on text-only systems?
    - Text-similarity and human assessments this year

2. What is the role of external data in this low resource task?
    - Participants free to use any external data this year

Results

Participants

ID	Participating team
AFRL-OHIOSTATE	Air Force Research Laboratory & Ohio State University (Duselis et al., 2017)
CMU	Carnegie Melon University (Jaffe, 2017)
CUNI	Univerzita Karlova v Praze (Helcl and Libovický, 2017)
DCU-ADAPT	Dublin City University (Calixto et al., 2017a)
LIUMCVC	Laboratoire d'Informatique de l'Université du Maine & Universitat Autonoma de Barcelona Computer Vision Center (Caglayan et al., 2017a)
NICT	National Institute of Information and Communications Technology & Nara Institute of Science and Technology (Zhang et al., 2017)
OREGONSTATE	Oregon State University (Ma et al., 2017)
SHEF	University of Sheffield (Madhyastha et al., 2017)
UvA-TiCC	Universiteit van Amsterdam & Tilburg University (Elliott and Kádár, 2017)

General Trends (1/3)

- More ResNet-50 `avgpool` features; less `res4_relu`

- Exceptions
  - SHEF: ImageNet 1000-class softmax distribution
  - UvA-TiCC: GoogLeNet v3 `avgpool`

General Trends (2/3)

- Most submissions
  - encoder / decoder feature initialisation, or
  - double-attention mechanisms

- Exceptions
  - AFRL-OHIOSTATE: retrieval approach
  - LIUMCVC: condition the target embeddings on image
  - UvA-TiCC: image representation prediction

General Trends (3/3)

- Most submissions used Constrained data

- Exceptions:
  - CUNI: parallel text
  - UvA-TiCC: monolingual image data & parallel text

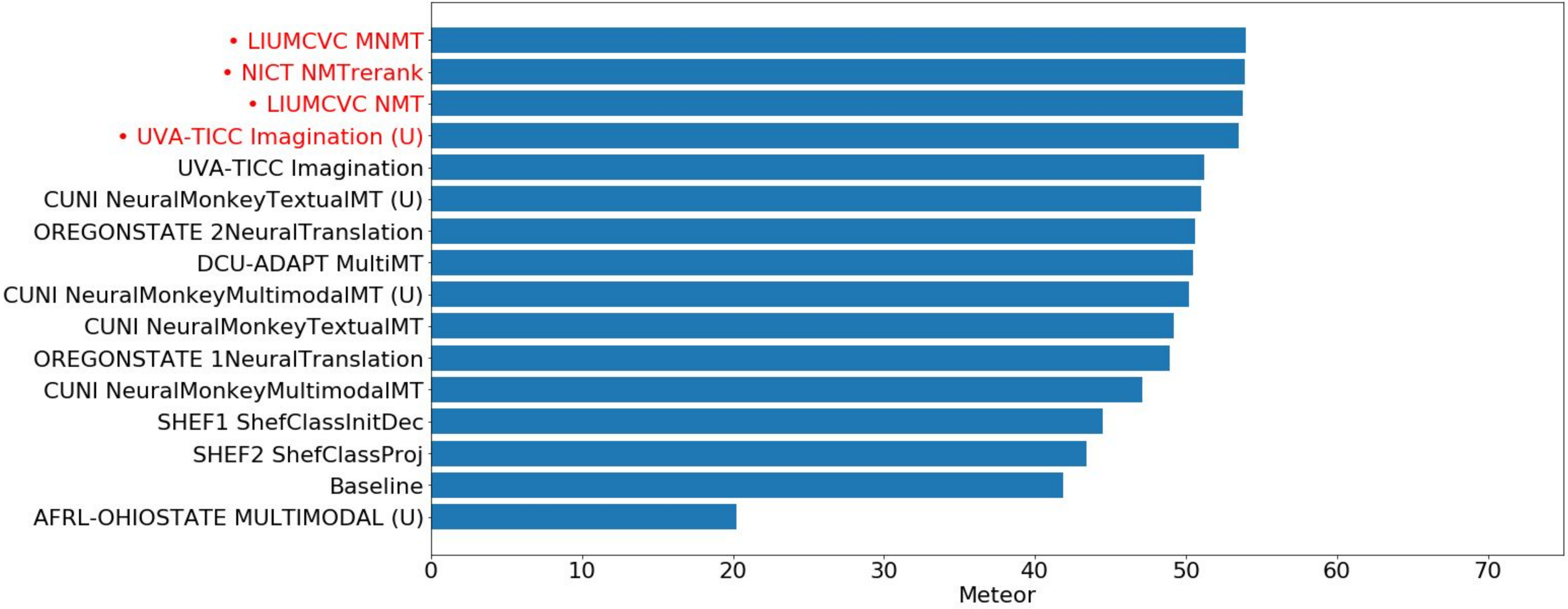
Task 1 Evaluation

- Meteor 1.5 (Denkowski et al., 2014)
- Direct Assessment (Graham et al., 2017)

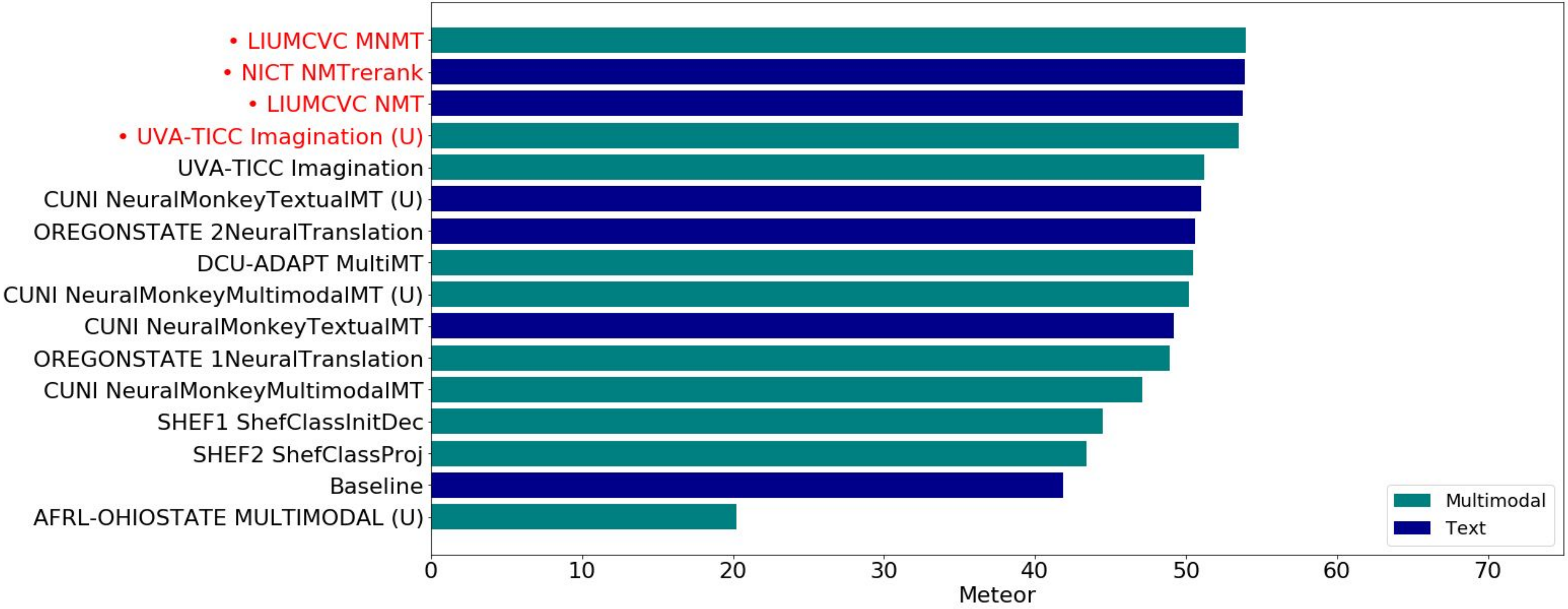
Baselines

- Text-only Nematus (Sennrich et al., 2017)
  - Train on only the 29K En-De/Fr pairs

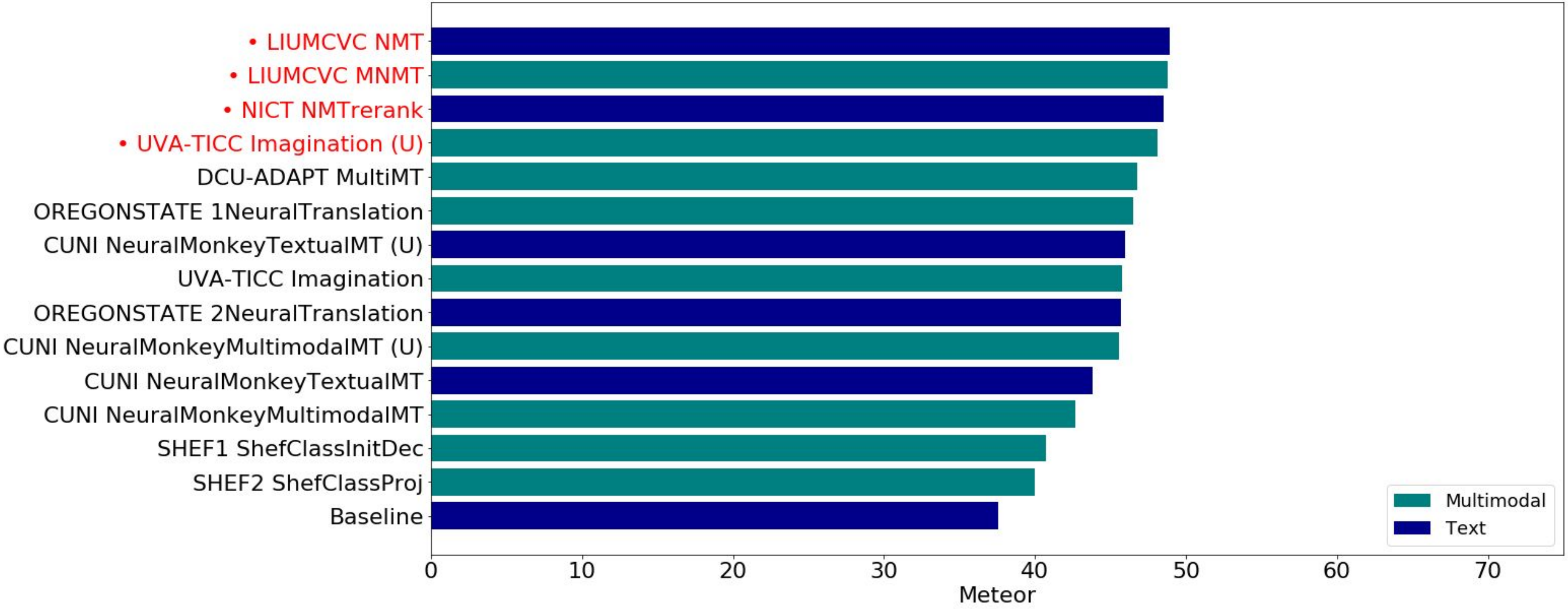
En-De Multi30K 2017



En-De Multi30K 2017



En-De Ambiguous COCO



Direct Assessment interface

0/10 blocks, 8 items left in block | MultiModalTask #28:Segment #265 | English → German (deutsch)

— Corresponding image

A graffiti covered wall depicting astronauts flying a magic carpet.

— Source text

ein mit graffiti bedeckter wand fliegt über einen zauber teppich .

— Candidate translation

— How accurately does the above candidate text convey the original semantics of the reference text? Slider ranges from Not a all (left) to Perfectly (right).

Reset

Submit

En-De Multi30K 2017 Human (n=3,485)

#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

Multimodal
Text

En-De Multi30K 2017 Human (n=3,485)

#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

Visual context helped

Multimodal
Text

En-De Multi30K 2017 Human (n=3,485)

External resources helped

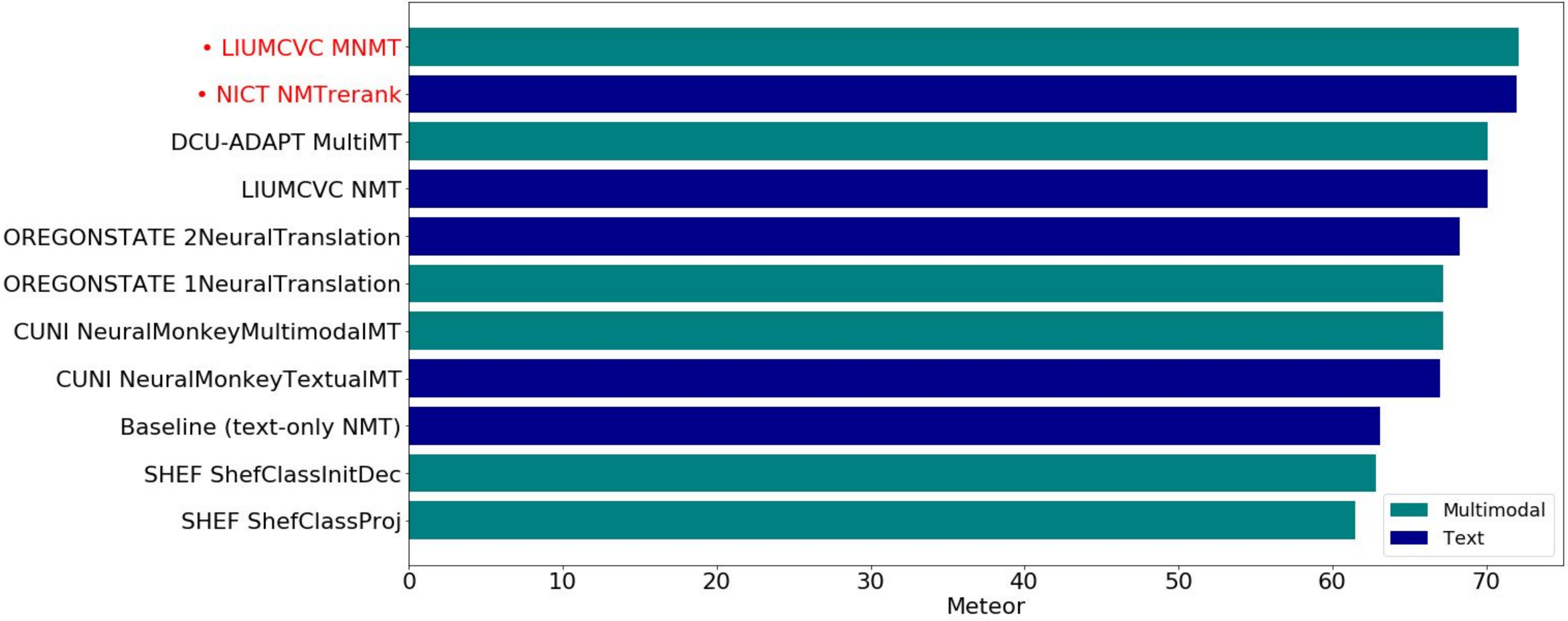
#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

Visual context helped

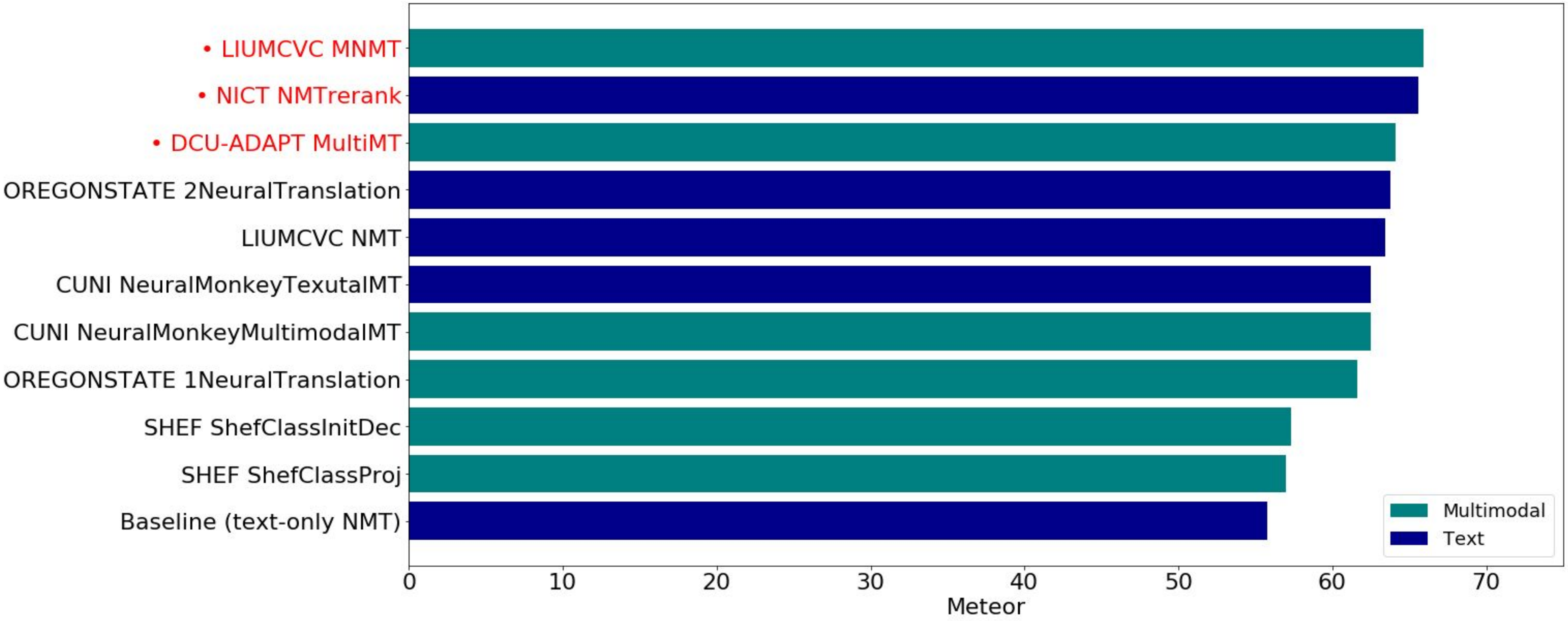
Multimodal

Text

En-Fr Multi30K 2017



En-Fr Ambiguous COCO



En-Fr Multi30K 2017 Human (n=2,521)

#	Raw	z	System
1	79.4	0.446	NICT_NMTrerank_C
	74.2	0.307	CUNI_NeuralMonkeyMultimodalMT_C
	74.1	0.3	DCU-ADAPT_MultiMT_C
4	71.2	0.22	LIUMCVC_MNMT_C
	65.4	0.056	OREGONSTATE_2NeuralTranslation_C
	61.9	-0.041	CUNI_NeuralMonkeyTextualMT_C
	60.8	-0.078	OREGONSTATE_1NeuralTranslation_C
	60.5	-0.079	LIUMCVC_NMT_C
9	54.7	-0.254	SHEF_ShefClassInitDec_C
	54.0	-0.282	SHEF_ShefClassProj_C
11	44.1	-0.539	Baseline (text-only NMT)

Multimodal
Text

En-Fr Multi30K 2017 Human (n=2,521)

#	Raw	z	System
1	79.4	0.446	NICT_NMTrerank_C
	74.2	0.307	CUNI_NeuralMonkeyMultimodalMT_C
	74.1	0.3	DCU-ADAPT_MultiMT_C
4	71.2	0.22	LIUMCVC_MNMT_C
	65.4	0.056	OREGONSTATE_2NeuralTranslation_C
	61.9	-0.041	CUNI_NeuralMonkeyTextualMT_C
	60.8	-0.078	OREGONSTATE_1NeuralTranslation_C
	60.5	-0.079	LIUMCVC_NMT_C
9	54.7	-0.254	SHEF_ShefClassInitDec_C
	54.0	-0.282	SHEF_ShefClassProj_C
11	44.1	-0.539	Baseline (text-only NMT)

Visual context helped

Multimodal

Text

En-Fr Multi30K 2017 Human (n=2,521)

#	Raw	z	System
1	79.4	0.446	NICT_NMTrerank_C
	74.2	0.307	CUNI_NeuralMonkeyMultimodalMT_C
	74.1	0.3	DCU-ADAPT_MultiMT_C
4	71.2	0.22	LIUMCVC_MNMT_C
	65.4	0.056	OREGONSTATE_2NeuralTranslation_C
	61.9	-0.041	CUNI_NeuralMonkeyTextualMT_C
	60.8	-0.078	OREGONSTATE_1NeuralTranslation_C
	60.5	-0.079	LIUMCVC_NMT_C
9	54.7	-0.254	SHEF_ShefClassInitDec_C
	54.0	-0.282	SHEF_ShefClassProj_C
11	44.1	-0.539	Baseline (text-only NMT)

Visual context hurt

Visual context helped

Multimodal

Text

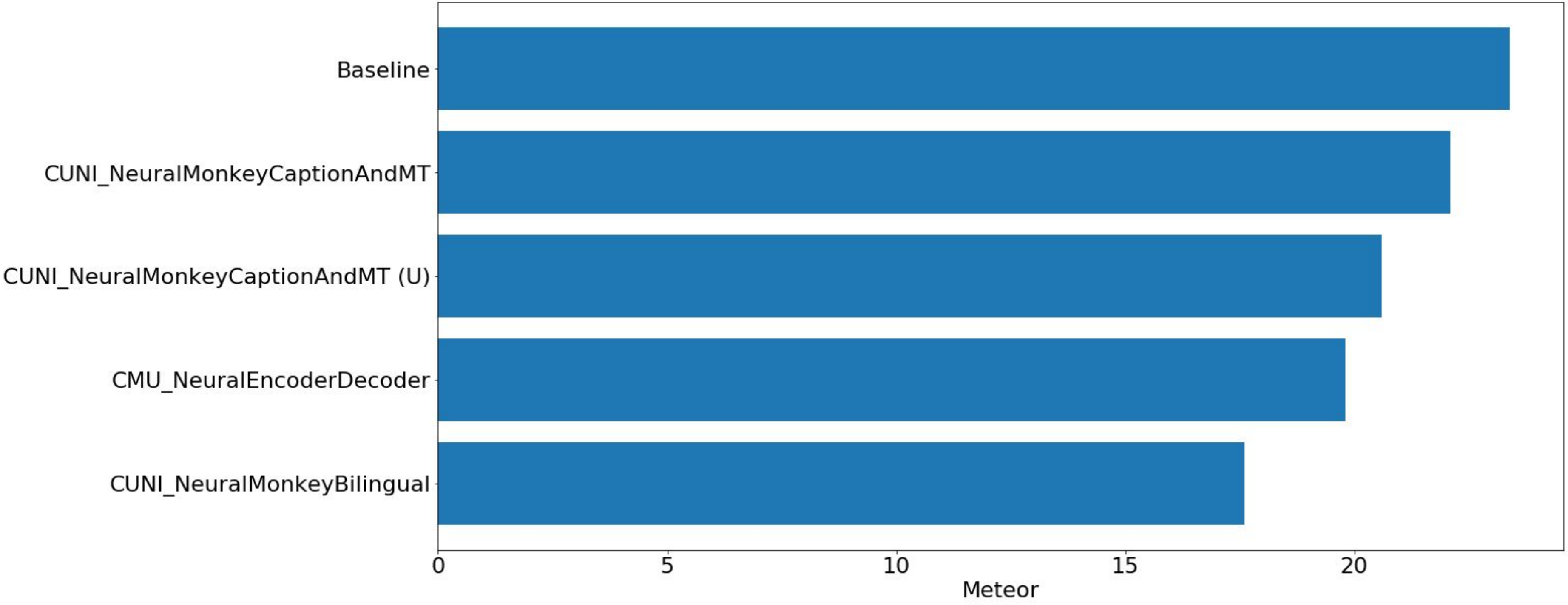
Task 2 Evaluation

- Meteor 1.5 (Denkowski et al., 2014)
  - Multiple independently collected reference descriptions

Baseline

- Attention-based image description (Xu et al., 2015)
  - Train on only the 155K Image-German data

Task 2: En-De Multi30K 2017

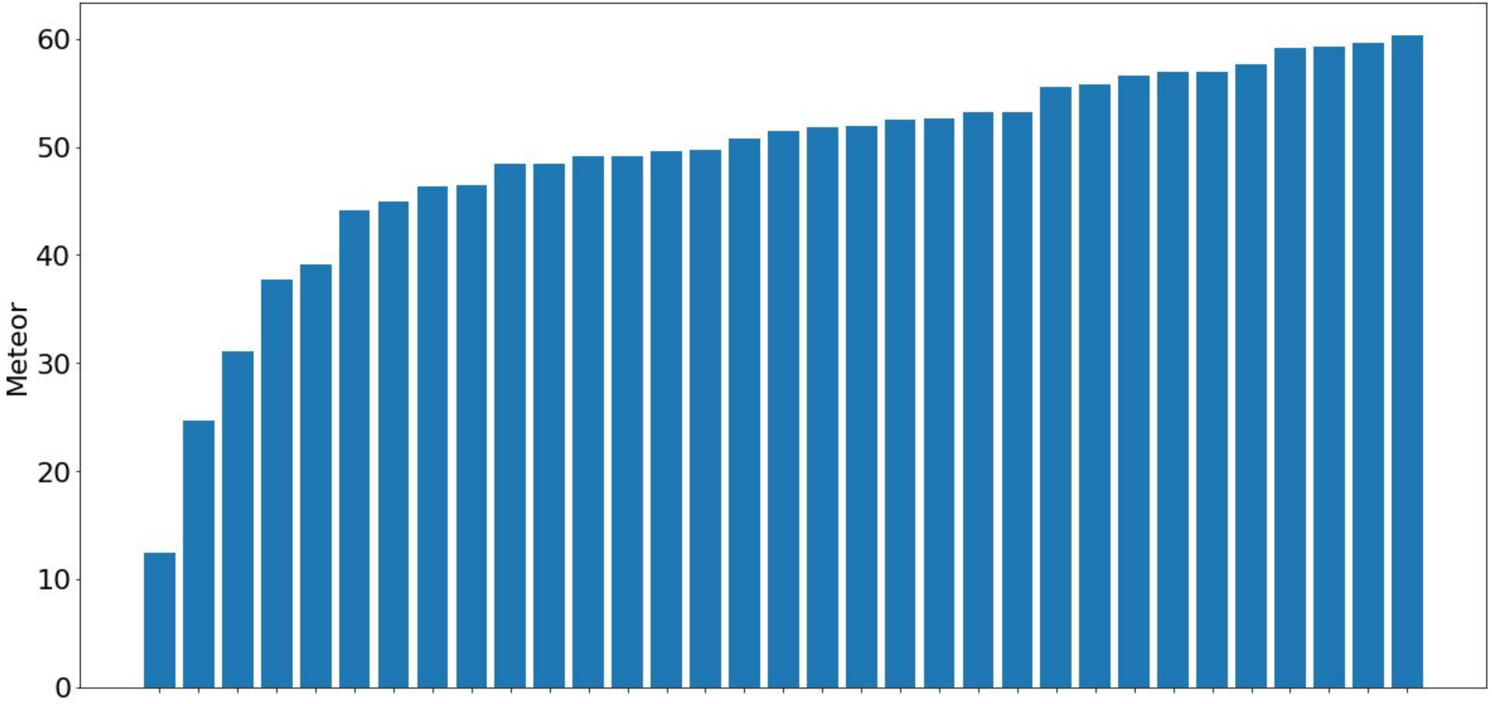


Conclusions

- Text-similarity metrics are masking real progress
  - Direct Assessment shows that multimodal > text-only

- Extra parallel text improves multimodal translation
- Ambiguous COCO is more challenging than Multi30K
- Multilingual Image Description is very challenging

Reality check: Multi30K En-De Test 2016



Reality check: Multi30K En-De Test 2016