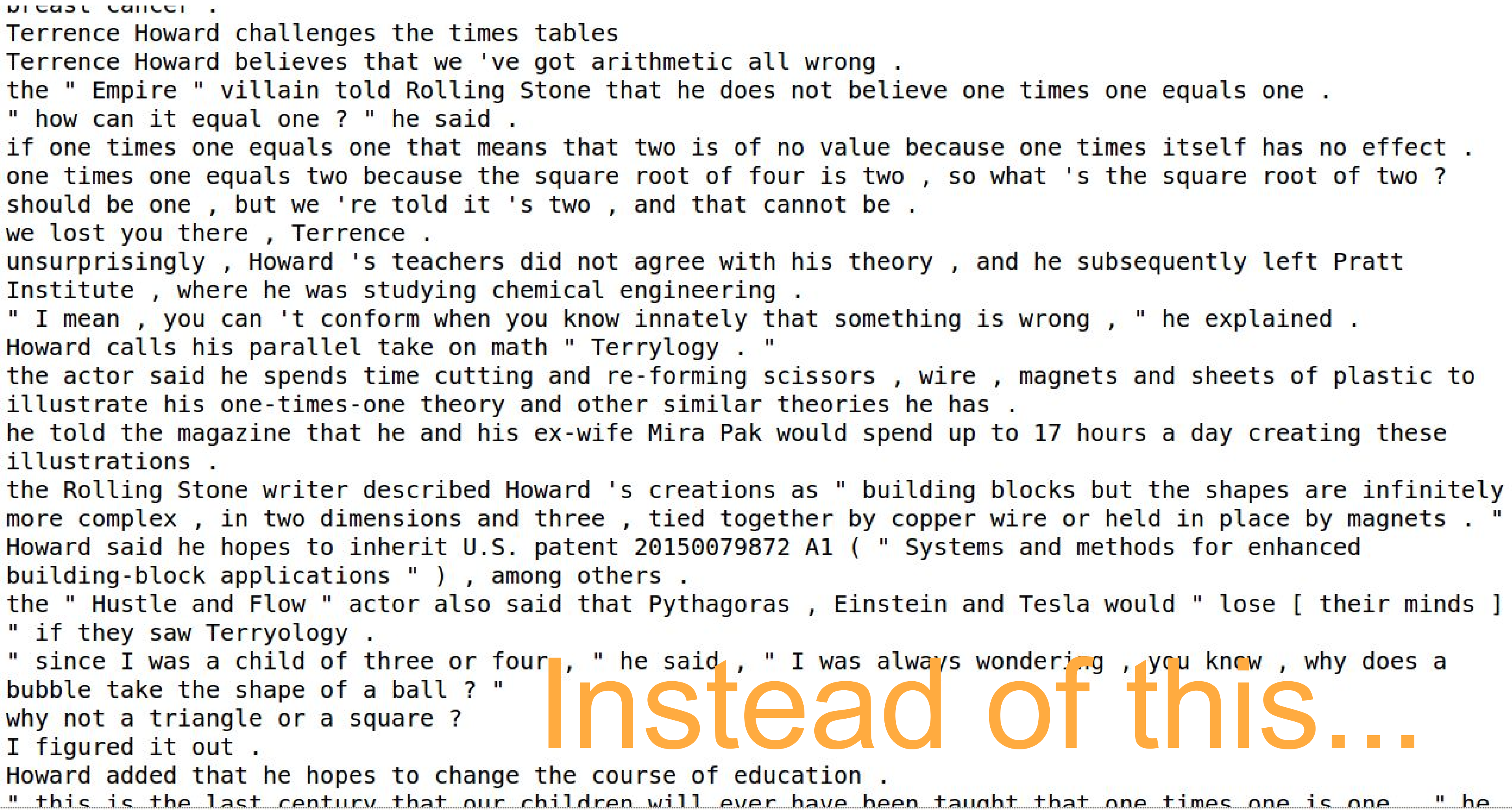
# First shared task on Multimodal Machine Translation and Crosslingual Image Description

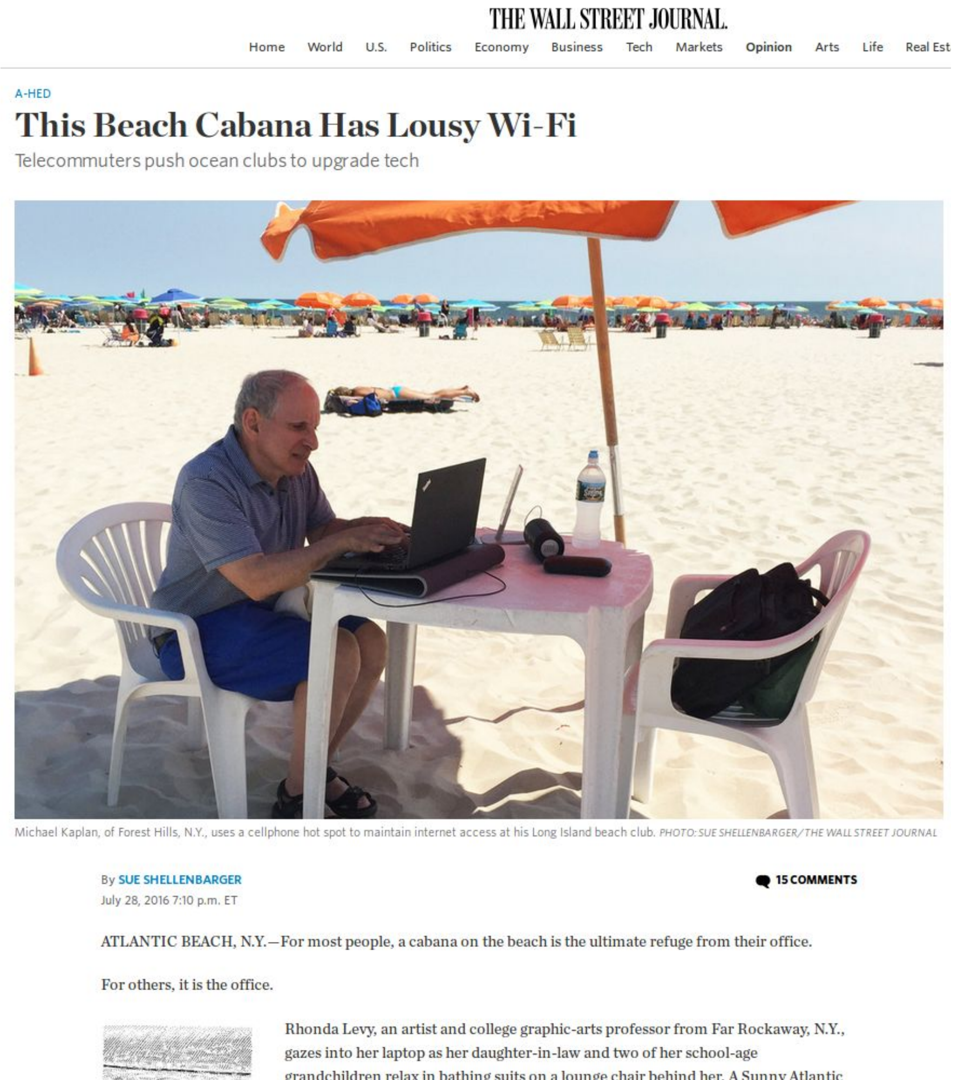
Lucia Specia, Stella Frank, Khalil Sima'an and Desmond Elliott



```
breast cancer .
Terrence Howard challenges the times tables
Terrence Howard believes that we 've got arithmetic all wrong .
the " Empire " villain told Rolling Stone that he does not believe one times one equals one .
" how can it equal one ? " he said .
if one times one equals one that means that two is of no value because one times itself has no effect .
one times one equals two because the square root of four is two , so what 's the square root of two ?
should be one , but we 're told it 's two , and that cannot be .
we lost you there , Terrence .
unsurprisingly , Howard 's teachers did not agree with his theory , and he subsequently left Pratt
Institute , where he was studying chemical engineering .
" I mean , you can 't conform when you know innately that something is wrong , " he explained .
Howard calls his parallel take on math " Terrylogy . "
the actor said he spends time cutting and re-forming scissors , wire , magnets and sheets of plastic to
illustrate his one-times-one theory and other similar theories he has .
he told the magazine that he and his ex-wife Mira Pak would spend up to 17 hours a day creating these
illustrations .
the Rolling Stone writer described Howard 's creations as " building blocks but the shapes are infinitely
more complex , in two dimensions and three , tied together by copper wire or held in place by magnets . "
Howard said he hopes to inherit U.S. patent 20150079872 A1 ( " Systems and methods for enhanced
building-block applications " ) , among others .
the " Hustle and Flow " actor also said that Pythagoras , Einstein and Tesla would " lose [ their minds ]
" if they saw Terryology .
" since I was a child of three or four , " he said , " I was always wondering , you know , why does a
bubble take the shape of a ball ? "
why not a triangle or a square ?
I figured it out .
Howard added that he hopes to change the course of education .
" this is the last century that our children will ever have been taught that one times one is one . " he
```

Instead of this...

… let's go back to the "Real World"



THE WALL STREET JOURNAL.

A-HED

## This Beach Cabana Has Lousy Wi-Fi

Telecommuters push ocean clubs to upgrade tech

Michael Kaplan, of Forest Hills, N.Y., uses a cellphone hot spot to maintain internet access at his Long Island beach club. *PHOTO: SUE SHELLENBARGER/THE WALL STREET JOURNAL*

By SUE SHELLENBARGER

15 COMMENTS

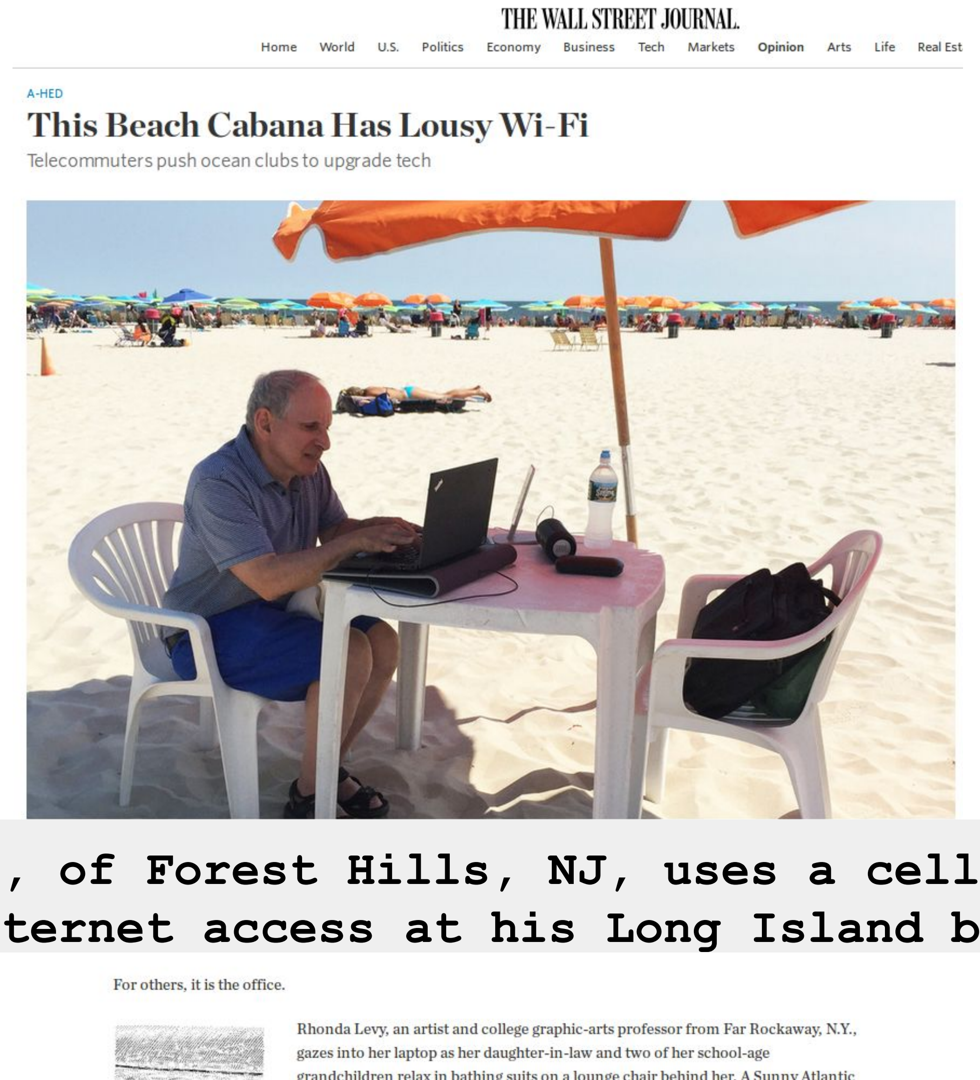
July 28, 2016 7:10 p.m. ET

ATLANTIC BEACH, N.Y.—For most people, a cabana on the beach is the ultimate refuge from their office.

For others, it is the office.

Rhonda Levy, an artist and college graphic-arts professor from Far Rockaway, N.Y., gazes into her laptop as her daughter-in-law and two of her school-age grandchildren relax in bathing suits on a lounge chair behind her. A Sunny Atlantic

...with lots of *visual* content.



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Est

A-HED

# This Beach Cabana Has Lousy Wi-Fi

Telecommuters push ocean clubs to upgrade tech

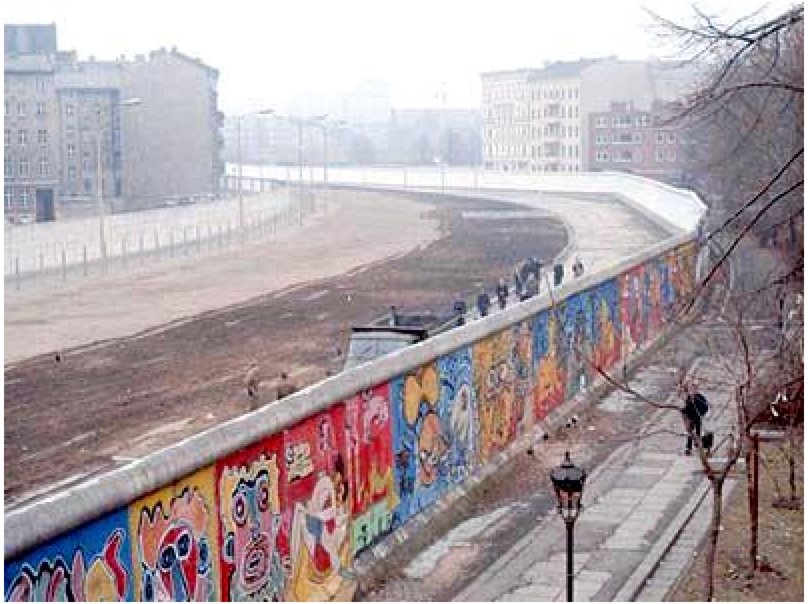
**Michael Kaglan, of Forest Hills, NJ, uses a cellphone hot spot to maintain internet access at his Long Island beach club.**

For others, it is the office.

Rhonda Levy, an artist and college graphic-arts professor from Far Rockaway, N.Y., gazes into her laptop as her daughter-in-law and two of her school-age grandchildren relax in bathing suits on a lounge chair behind her. A Sunny Atlantic

... and captions!

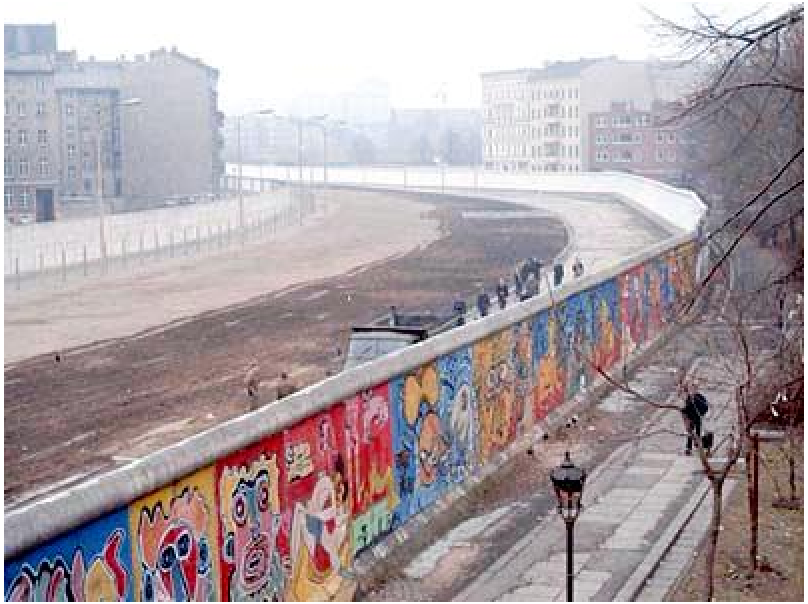
# Translations with Images



```
A wall divided the city.
```

(See also Hitschler et al. ACL 2016)

# Translations with Images



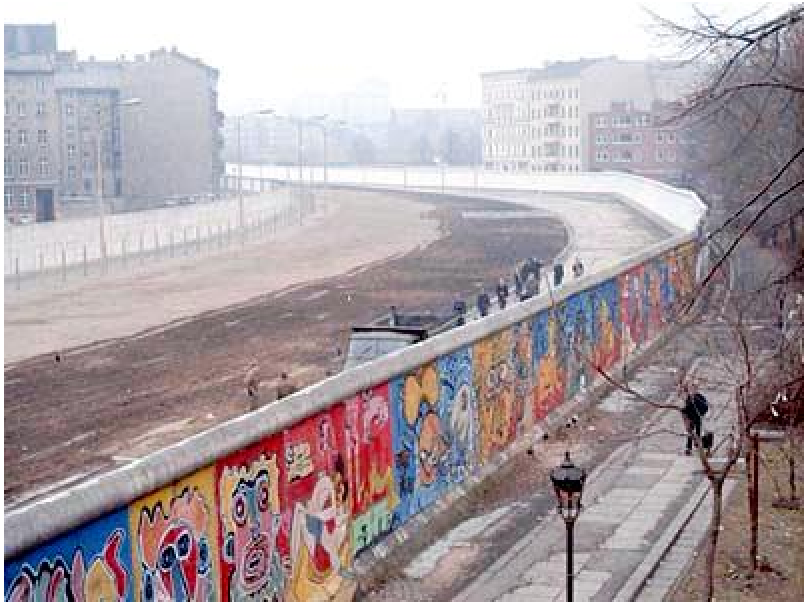
**A wall divided the city.**

**Eine Wand teilte die Stadt.**

**Eine Mauer teilte die Stadt.**

(See also Hitschler et al. ACL 2016)

# Translations with Images



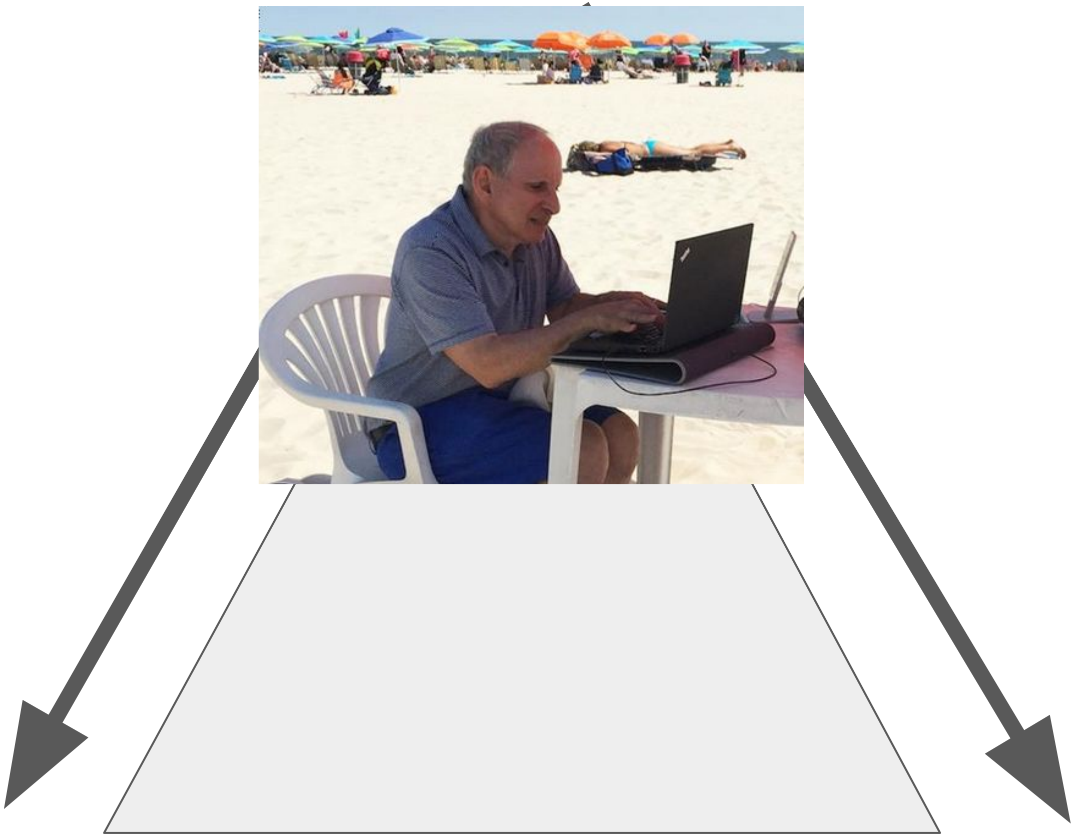
A wall divided the city.

~~Eine Wand teilte die Stadt.~~

Eine Mauer teilte die Stadt.

(See also Hitschler et al. ACL 2016)

~~Interlingua~~

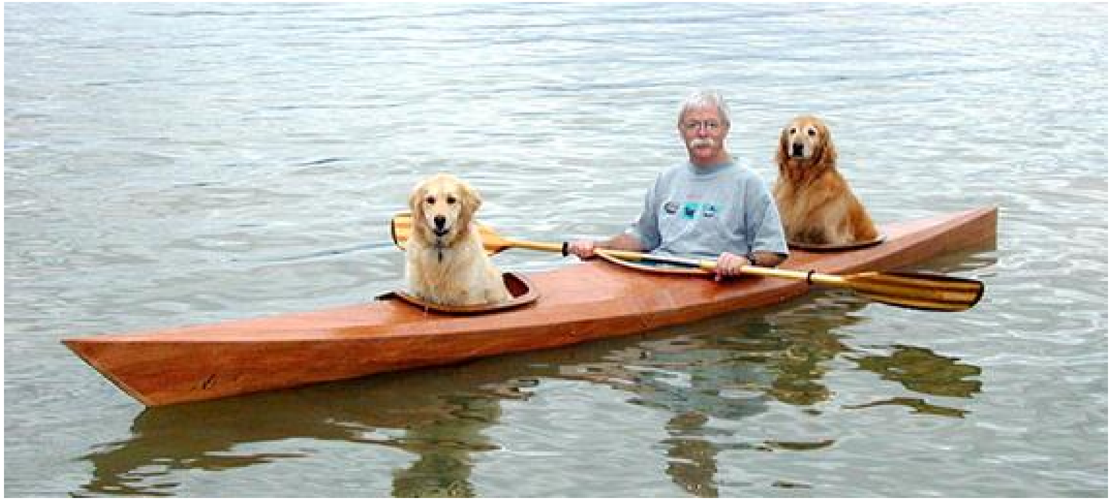


Source

Target

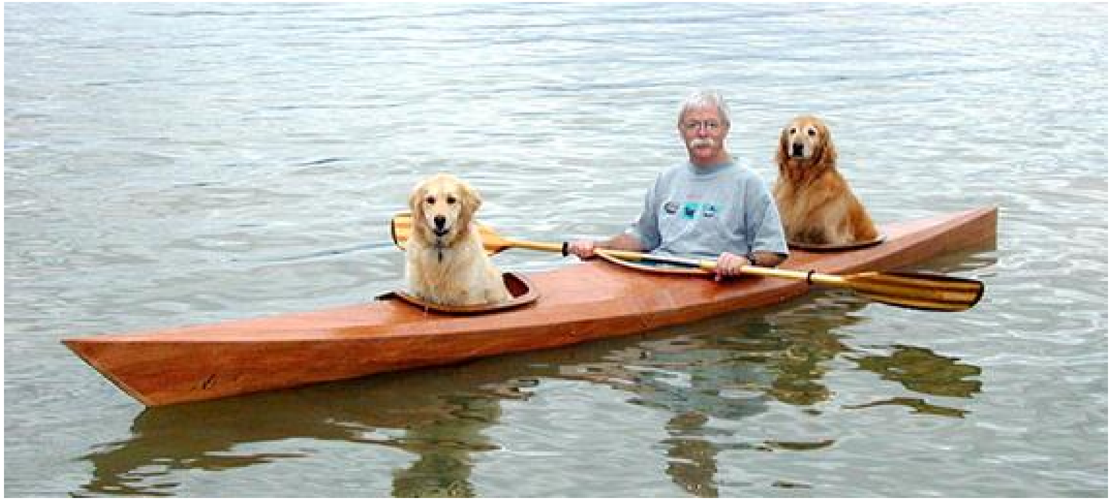
# Elsewhere in NLP: Language and Vision

**Image Description** task: generate description of image



# Elsewhere in NLP: Language and Vision

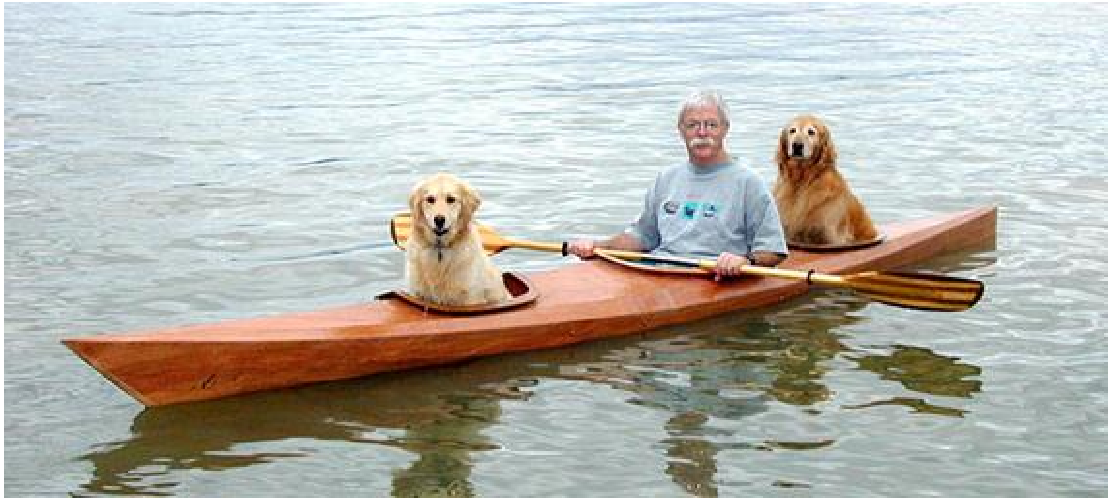
**Image Description** task: generate description of image



```
A man sitting in
a kayak with two
dogs.
```

# Elsewhere in NLP: Language and Vision

**Image Description** task: generate description of image



```
A man sitting in
a kayak with two
dogs.
```

Motivated by accessibility for visually impaired (alt-text generation)

Nearly always with only English-language datasets

# Tasks and Data

# WMT’16 shared task: subtasks

1. Multimodal Machine Translation
   What can images bring to translation?
2. Crosslinguistic Image Description
   What can multilinguality bring to image description?

See also: Elliott et al. arXiv 2015, Hitschler et al. ACL 2016

# Task 1: Multimodal Machine Translation

A brown dog is running after the black dog.

Input

Ein brauner Hund rennt dem schwarzen Hund hinterher.

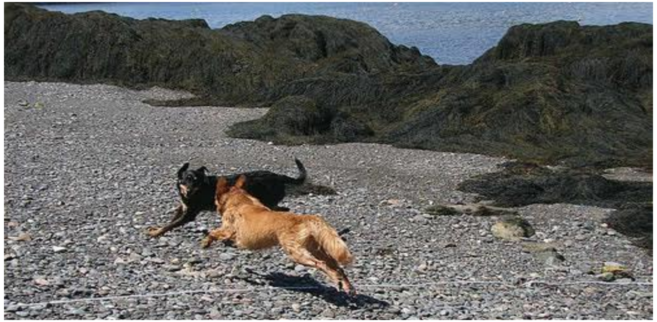
Evaluated against human translation

# Language Data for Task 1: Source Data

Flickr30K dataset (Young et al., 2014)

31,014 images from Flickr groups:
Outdoor activities, dogs in action

5 English descriptions each,
crowdsourced from U.S. workers



```
Two dogs run towards each other on
a rocky area with water in the
background.

A brown dog is running after a
black dog on a rocky shore.

Two dogs playing on a beach.

A brown dog is running after the
black dog.

Two dogs run across stones near a
body of water.
```

# Language Data for Task 1: Multimodal Translation

For each image, professionally translate one description into German.

Translator does not see image.

**Total: 31,014** parallel sentence pairs

```
Trendy girl talking on her cellphone while
gliding slowly down the street.

Ein schickes Mädchen spricht mit dem Handy
während sie langsam die Straße entlangschwebt.
```

(Data released as Multi30K, Elliott et al., 2016)

# Task 2: Crosslingual Image Description

A brown dog is running after the black dog.
Two dogs playing on a beach.
Two dogs run towards each other on a beach.
Two dogs run across stones.
A black dog and a brown dog.

Input

**Zwei Hunde spielen miteinander.**

Evaluated against German descriptions

# Language Data for Task 2: Image Description

We crowdsource 5 new German descriptions for each image.

Use (translations of) original instructions.

Much cheaper than translations!

**Total: 155,070** descriptions

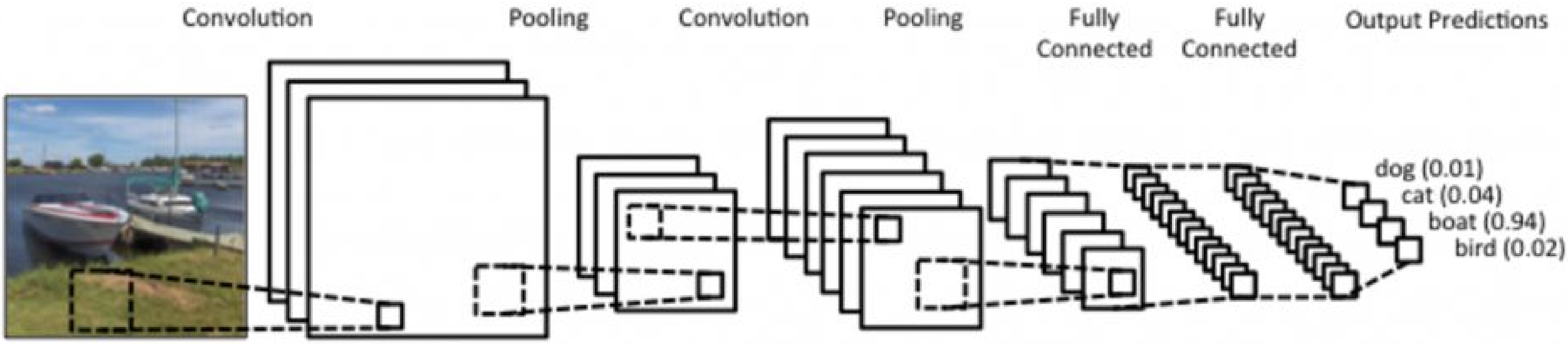
```
Two men on the scaffolding are helping to build a red brick wall.

Zwei Mauerer mauern ein Haus zusammen.
```

(Data released as Multi30K, Elliott et al., 2016)

# Image Representation

Intermediate layers from VGG19 Convolutional Neural Network trained on ImageNet for object recognition task:

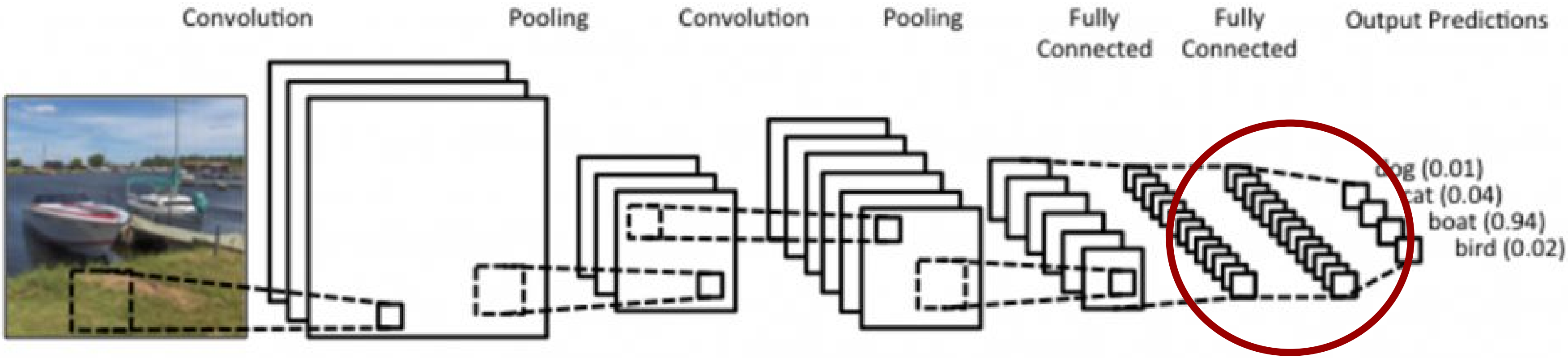
- FC7: final pre-output fully-connected layer (4096D vector)
- Conv5: last convolutional layer (14x14x512D tensor)



# Image Representation

Intermediate layers from VGG19 Convolutional Neural Network trained on ImageNet for object recognition task:

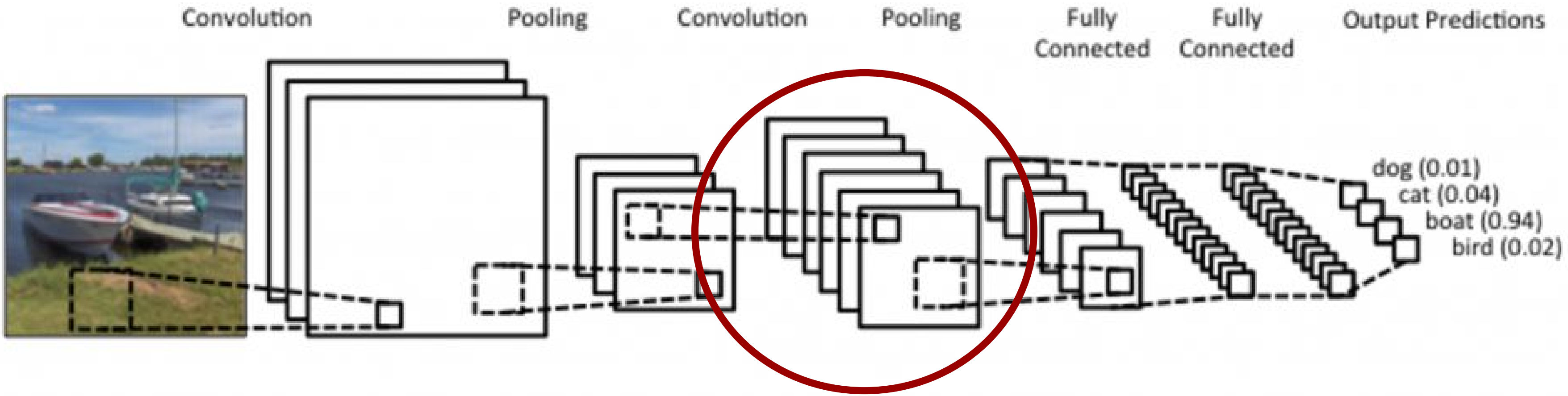
- FC7: final pre-output fully-connected layer (4096D vector)
- Conv5: last convolutional layer (14x14x512D tensor)



# Image Representation

Intermediate layers from VGG19 Convolutional Neural Network trained on ImageNet for object recognition task:

- FC7: final pre-output fully-connected layer (4096D vector)
- Conv5: last convolutional layer (14x14x512D tensor)



# Results

# Ten teams submitted 23 systems

Multimodal Translation: 16 systems (2 unconstrained)

Crosslingual Image Description: 7 systems (2 unconstrained)

# Ten teams submitted 23 systems

Multimodal Translation: 16 systems (2 unconstrained)

Crosslingual Image Description: 7 systems (2 unconstrained)

Baselines:
- Moses translations (without images)
- Neural Image Description model (GroundedTranslation, Elliott et al., 2015)

| | |
|---|---|
| CMU+NTU | Carnegie Melon University (Huang et al., 2016) |
| CUNI | Univerzita Karlova v Praze (Libovický et al., 2016) |
| DCU | Dublin City University |
| DCU-UVA | Dublin City University & Universiteit van Amsterdam (Calixto et al., 2016) |
| HUCL | Universität Heidelberg (Hitschler et al., 2016) |
| IBM-IITM-Montreal-NYU | IBM Research India, IIT Madras, Université de Montréal & New York University |
| LIUM | Laboratoire d’Informatique de l’Université du Maine (Caglayan et al., 2016) |
| LIUM-CVC | Laboratoire d’Informatique de l’Universite´ du Maine & Universitat Autonoma de Barcelona Computer Vision Center (Caglayan et al., 2016) |
| SHEF | University of Sheffield (Shah et al., 2016) |
| UPC | Universitat Politècnica de Catalunya (Rodríguez Guasch and Costa-jussà, 2016) |
| UPCb | Universitat Politècnica de Catalunya |

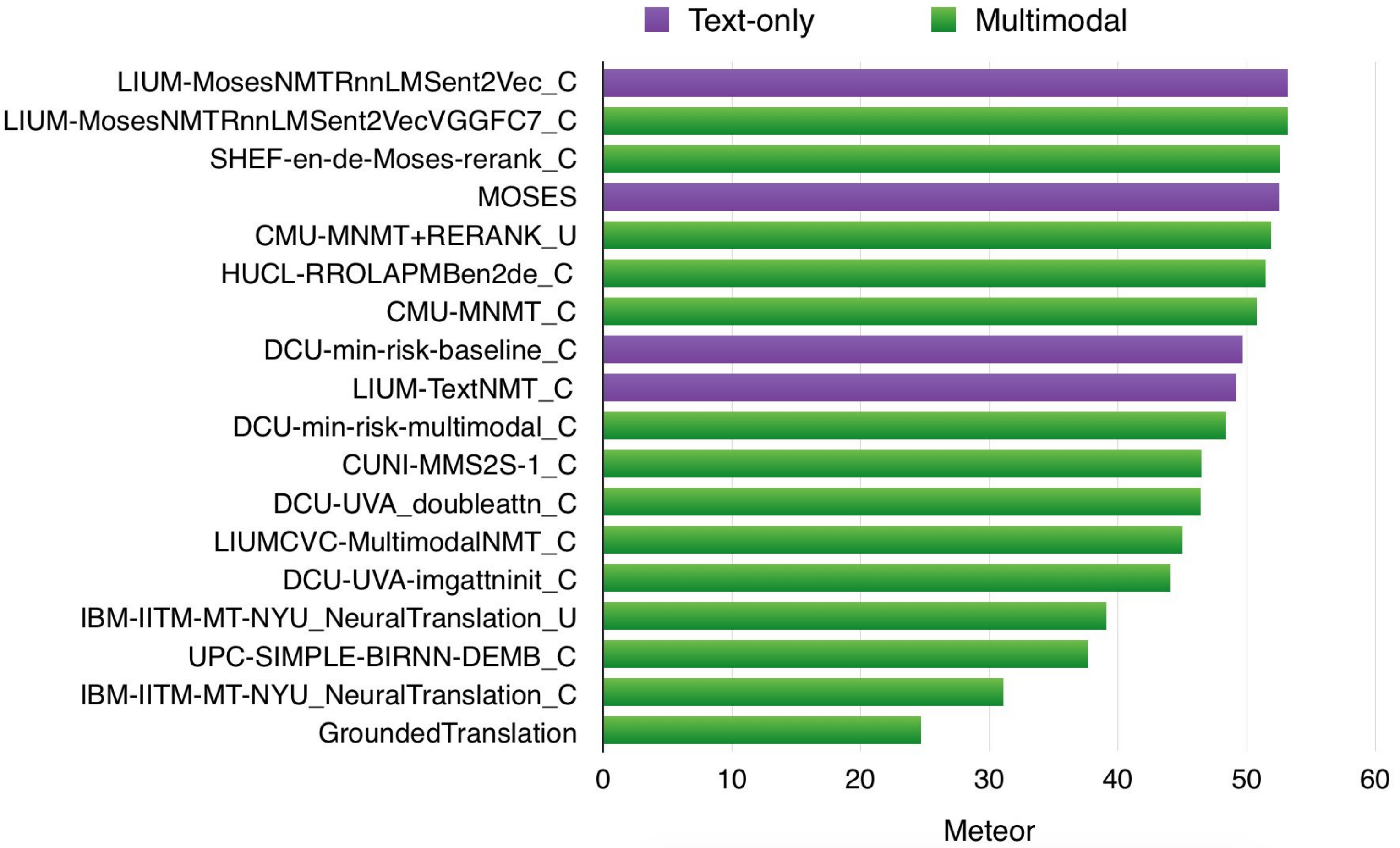
# Model architectures

| Phrase/syntax | Attention NMT | Seq-2-Seq |
|---|---|---|
| LIUM | CMU | IBM-IITM-Montreal-NYU |
| SHEF | DCU | UPC |
| HUCL | LIUM-CVC | GroundedTranslation |
| Moses | DCU-UVA | |
| | CUNI | |

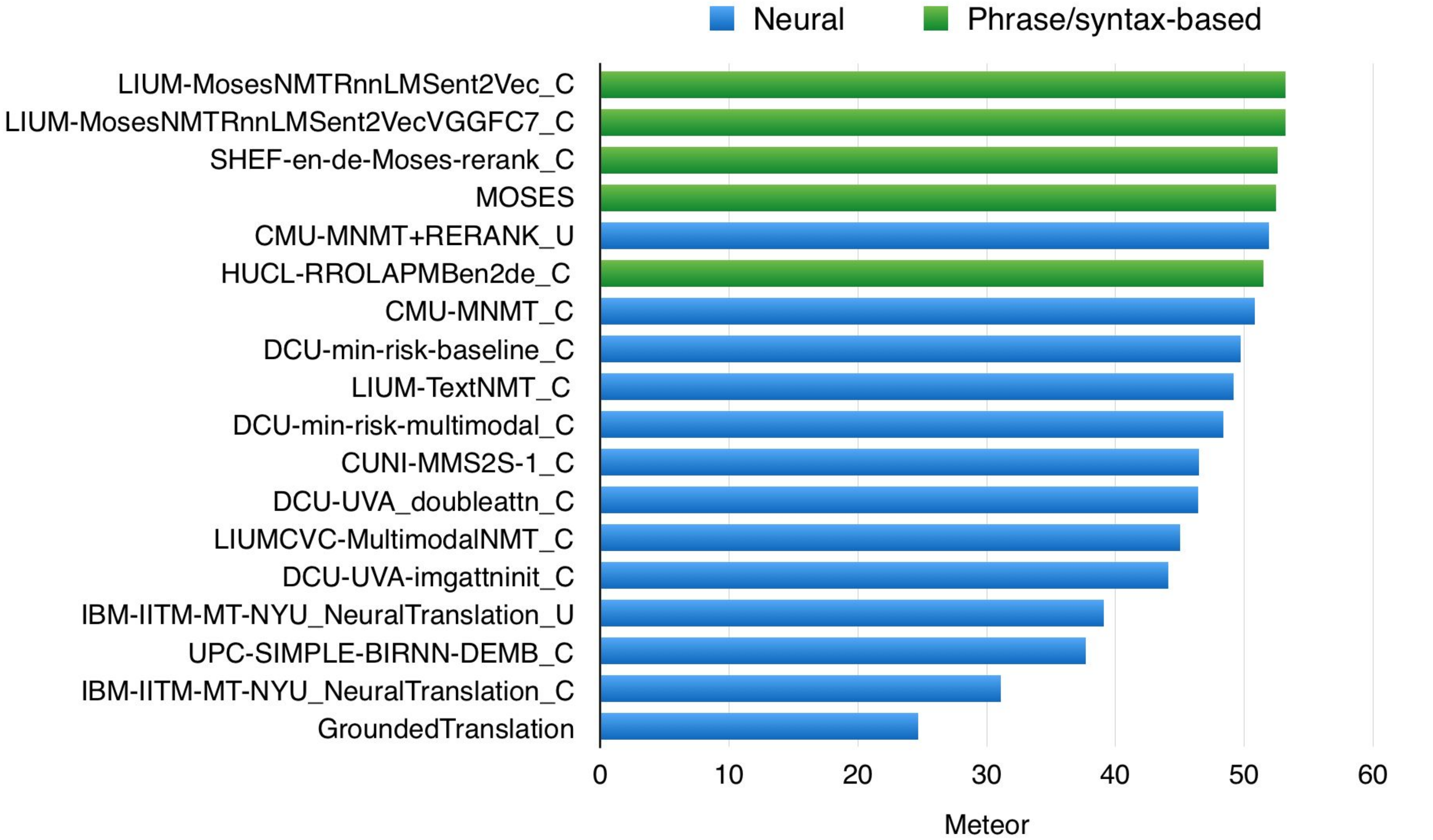
# Text-only or multimodal?

| Text-only | $FC_7$ | $CONV_5$ | Other |
|---|---|---|---|
| Moses | IBM-IITM-Montreal-NYU | CMU | SHEF |
| LIUM | DCU | DCU-UVA | LIUM |
| DCU | CUNI | | LIUM-CVC |
| | GroundedTranslation | | |
| | UPC | | |
| | HUCL | | |

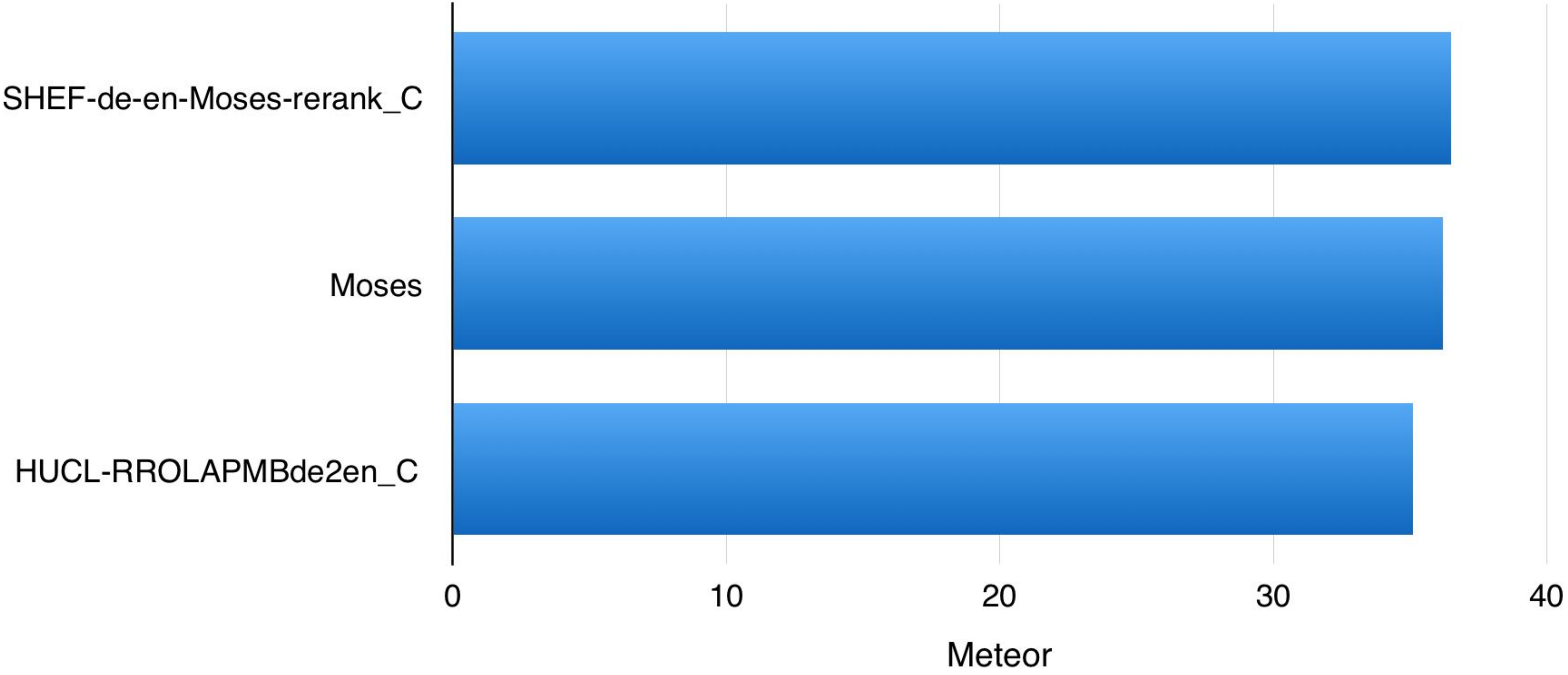
# Task 1: Multimodal Translation (En-De)



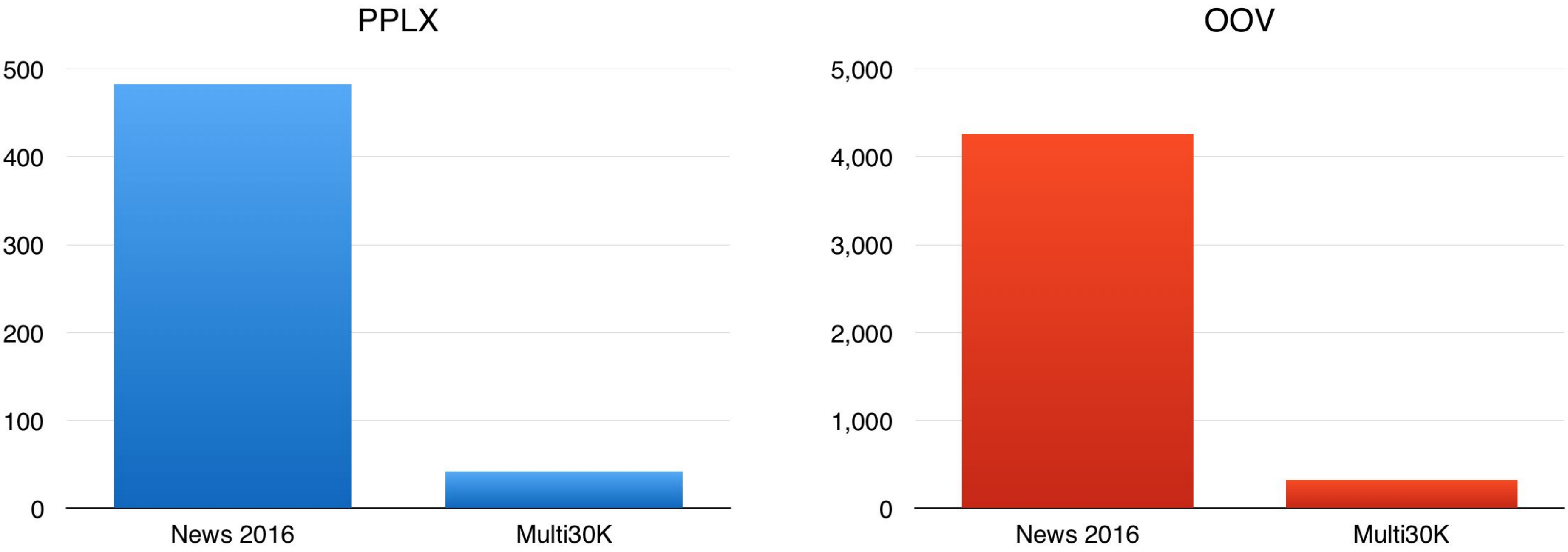
# Task 1: Multimodal Translation (En-De)



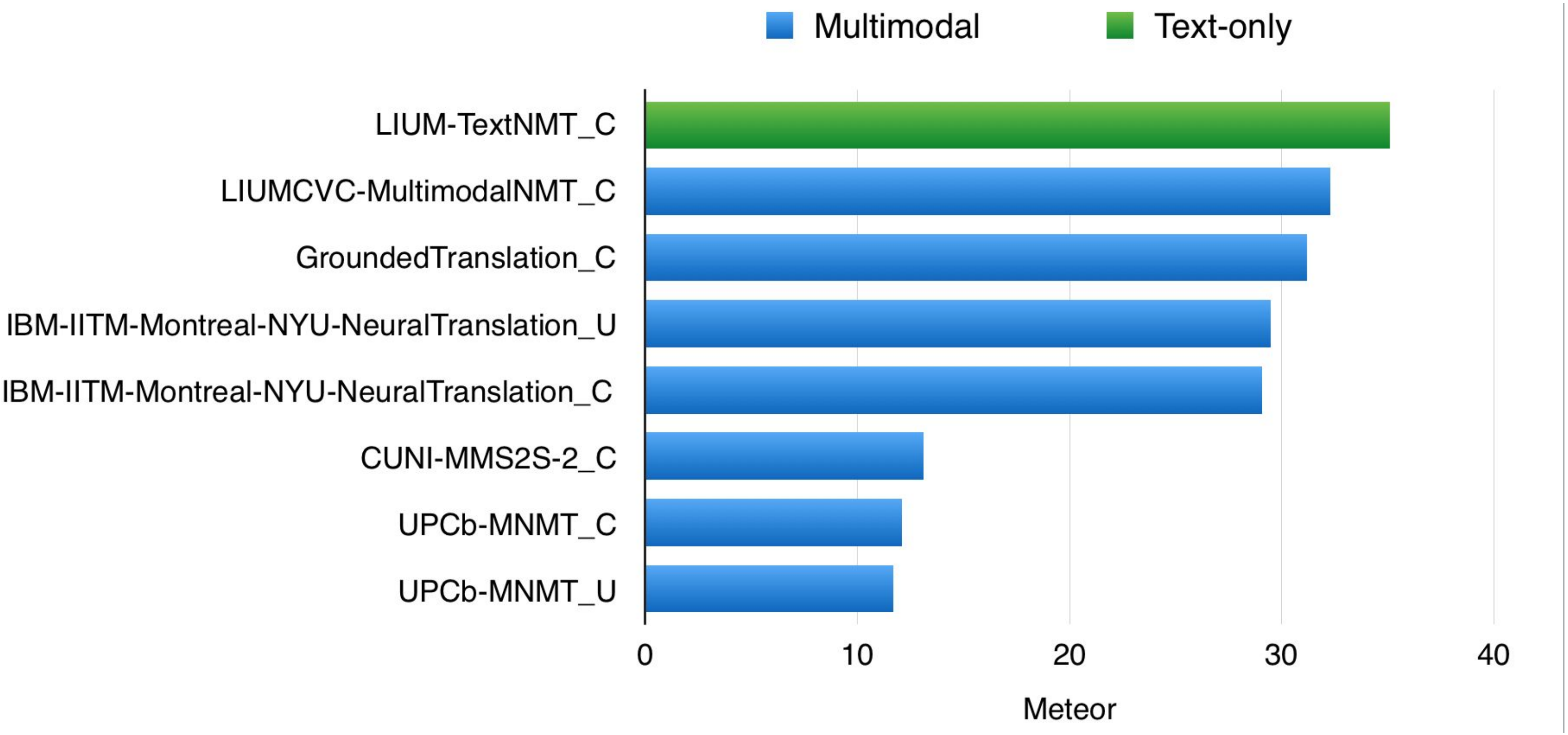
# Task 1: Multimodal Translation (De-En)



# Task 1: Data is much simpler than News 2016



# Task 2: Crosslingual Image Description (En-De)



# Conclusions

First large-scale task connecting multimodal NLP and machine translation generated lots of interest!

Text-only baselines are very strong for image description translation: when will multimodal models catch up?

# Future directions (next year?)

- Harder task:
  - Visual sense disambiguation (Gella et al., NAACL 2016)
  - Generate descriptions in both languages
- Image-aware translations as gold standard
- More languages (expanded Multi30K)
  - Dutch, Maltese, Chinese, Turkish

# Thank you!

Data available at:
www.statmt.org/wmt16/multimodal-task.html