Natural Language Understanding with Multilingual Neural Models

- or the Power of GPU's in Modern Language Technology

Jörg Tiedemann
Department of Digital Humanities
University of Helsinki
jorg.tiedemann@helsinki.fi

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

How do humans become intelligent?

reasoning

understand

argue

learn

analyze

interpret

listen

discuss

interact

read

http://www.open-pharmacy-research.ca/themes/knowledge-translation-and-exchange

How do machines become intelligent?

reasoning

argue

understand

learn

analyze

interpret

listen

discuss

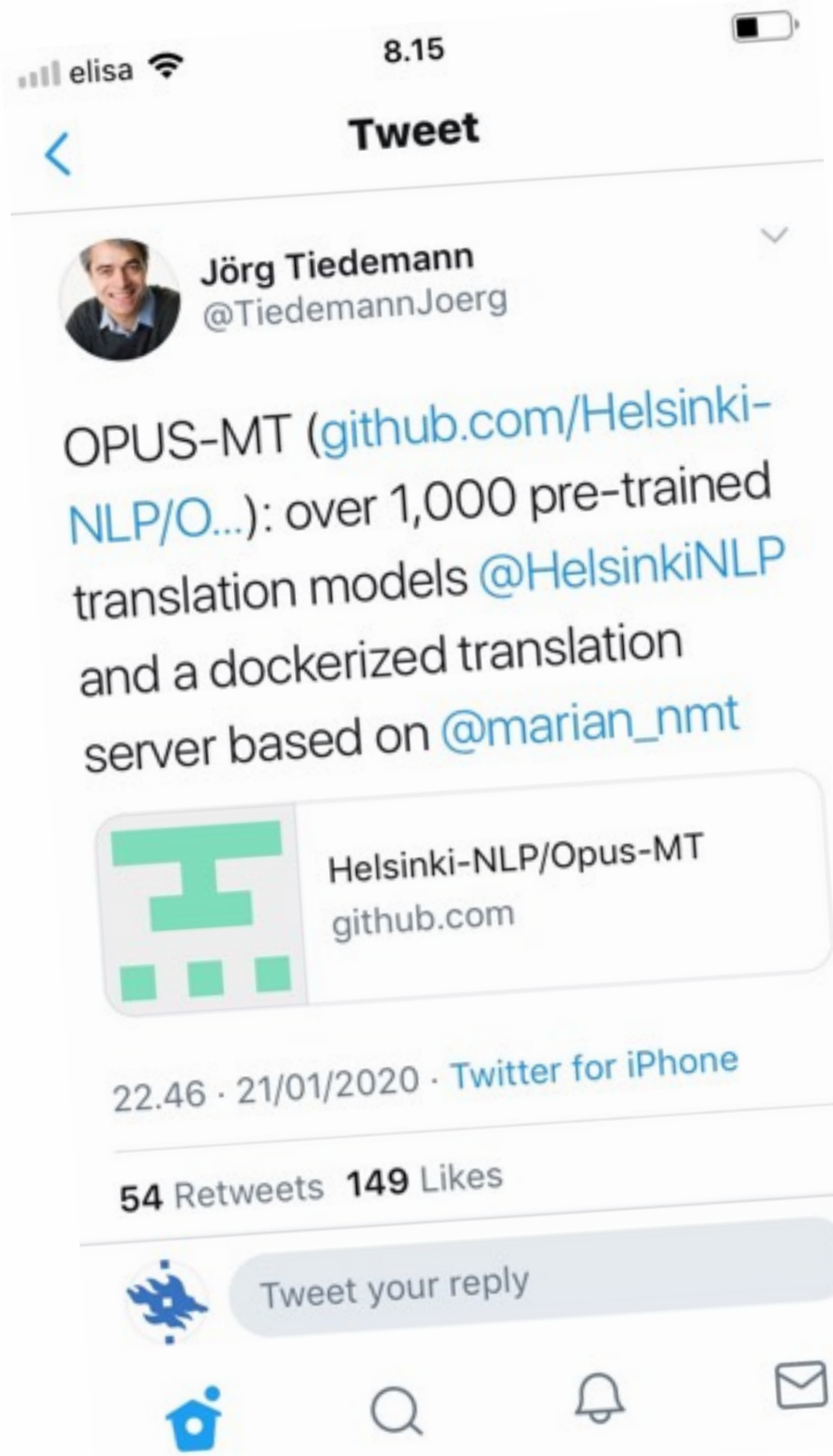
interact

read

Natural Language Processing

http://www.open-pharmacy-research.ca/themes/knowledge-translation-and-exchange

Democratize NLP ...



Democratize information & AI

04/08/1962 † 09/05/2020 https://peacemachine.net

Modern natural language processing

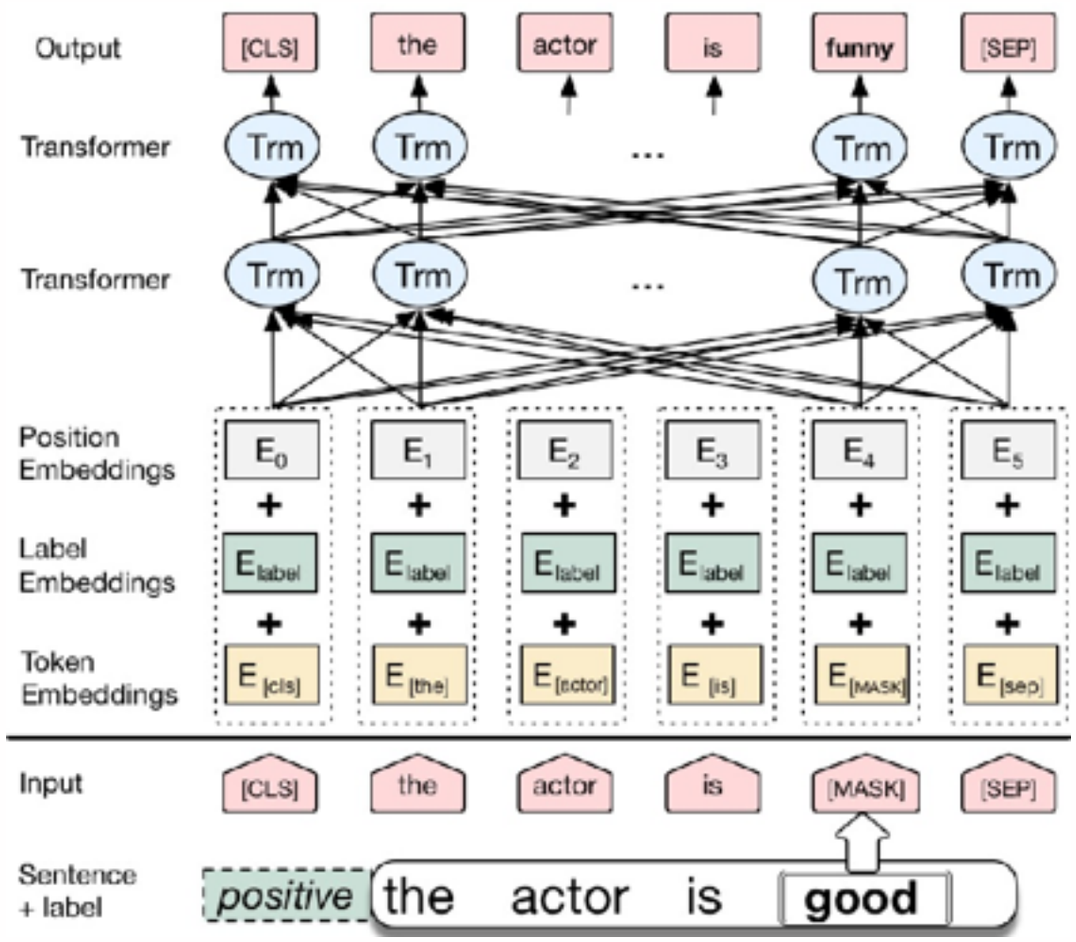
all the language data
I can possibly find

Modern natural language processing

my big fat neural network ….

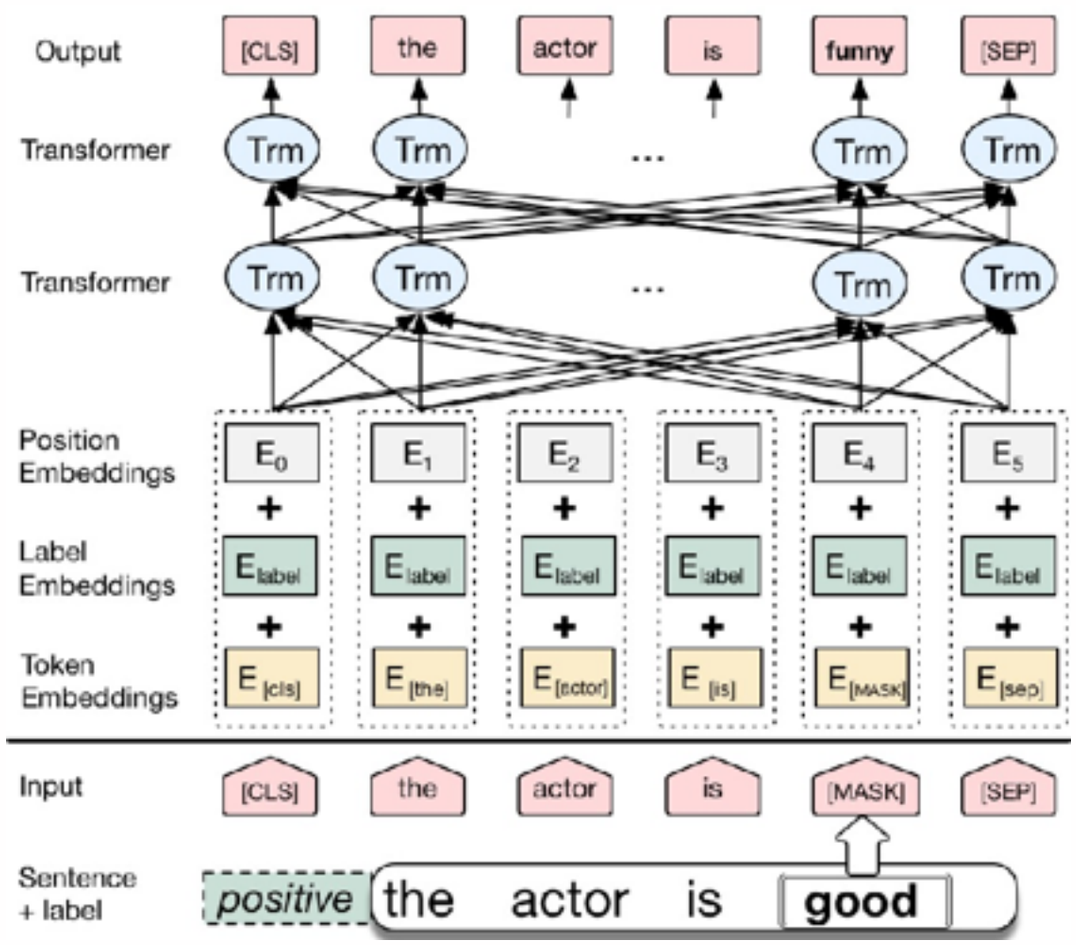
all the language data
I can possibly find

pre-training



Modern natural language processing

my big fat neural network ….



all the language data I can possibly find

pre-training

fine-tuning

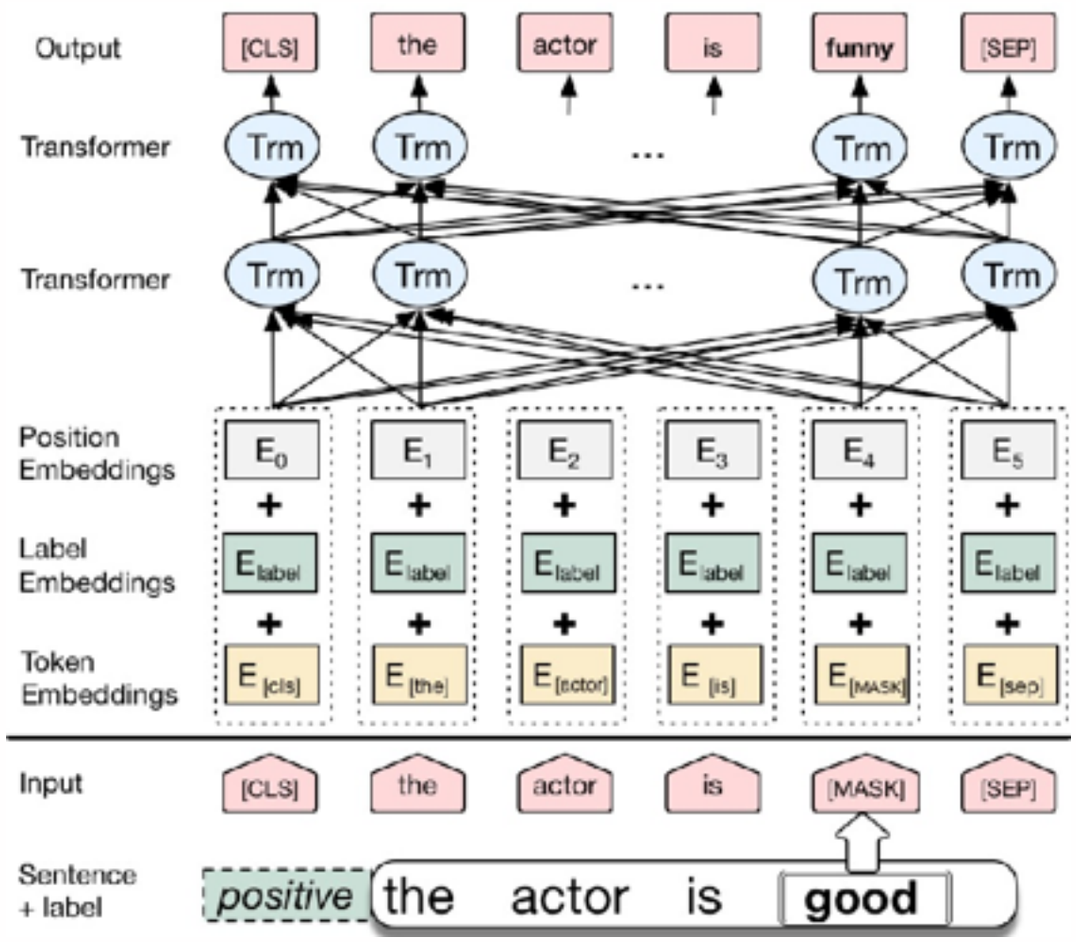
task-specific data (typically small)

Modern natural language processing

my big fat neural network ….

all the language data
I can possibly find

pre-training



fine-tuning

task-specific data
(typically small)

Requires heavy computing!

- millions of free parameters
- more and more layers
- larger and larger data sets

Some examples

Transformer language models (BERT, ELECTRA, …)

- parameters: 14M (small), 110M (base), 335M (big)
- crawled data: 420 billion words (English), 3 billion (Finnish)
- wikimedia: 4.5 billion words (English), 108 million (Finnish)
- training takes ~ 1-2 weeks (base model)

Some examples

Transformer language models (BERT, ELECTRA, …)

- parameters: 14M (small), 110M (base), 335M (big)
- crawled data: 420 billion words (English), 3 billion (Finnish)
- wikimedia: 4.5 billion words (English), 108 million (Finnish)
- training takes ~ 1-2 weeks (base model)

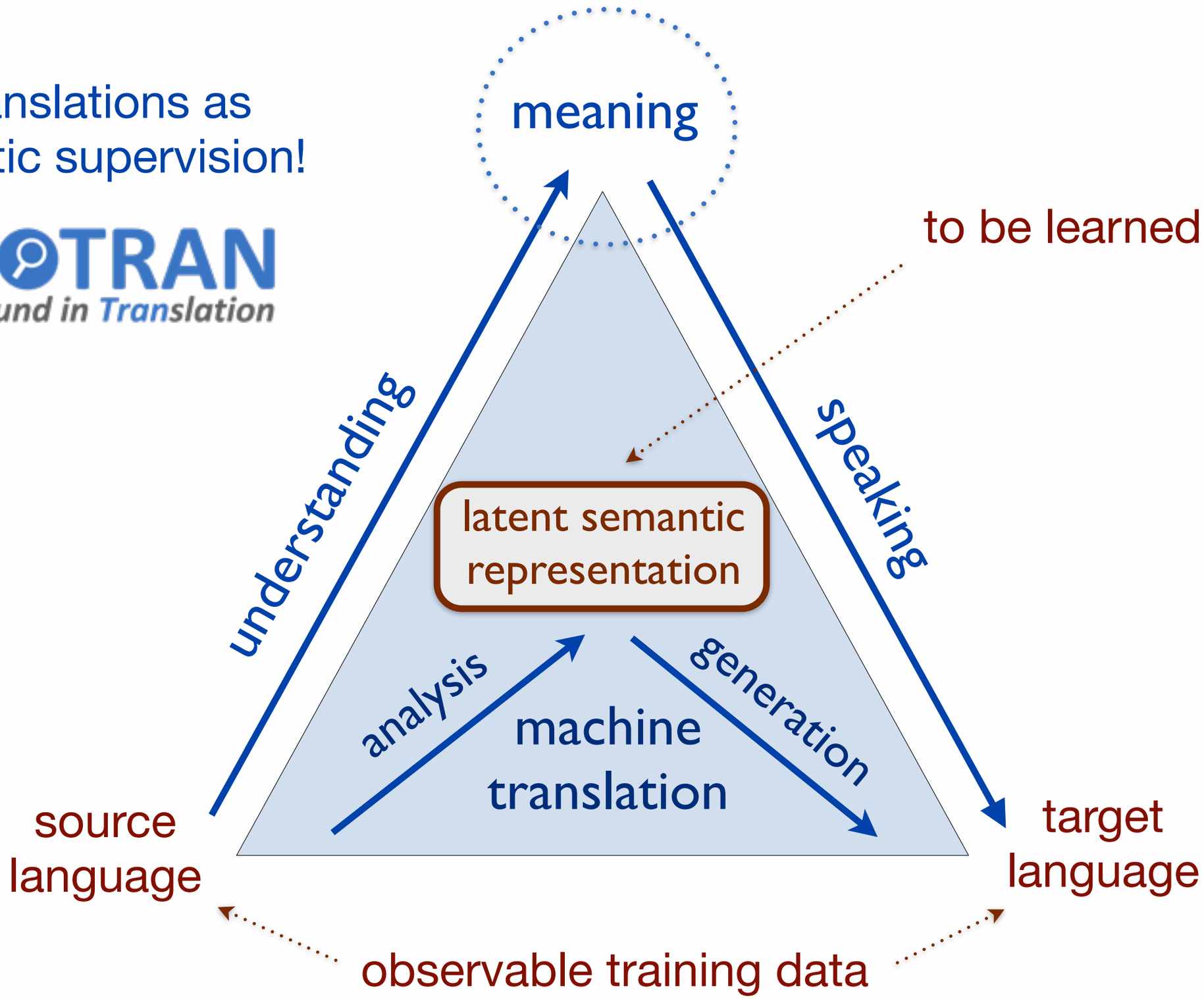
Translation model German/English/Finnish/Dutch/Swedish

- 1.46 billion sentence pair in training
- ca. 90 million parameters to be learned
- multi-GPU training on four v100 GPUs
- start to converge after ca 7 days (?) - still running …

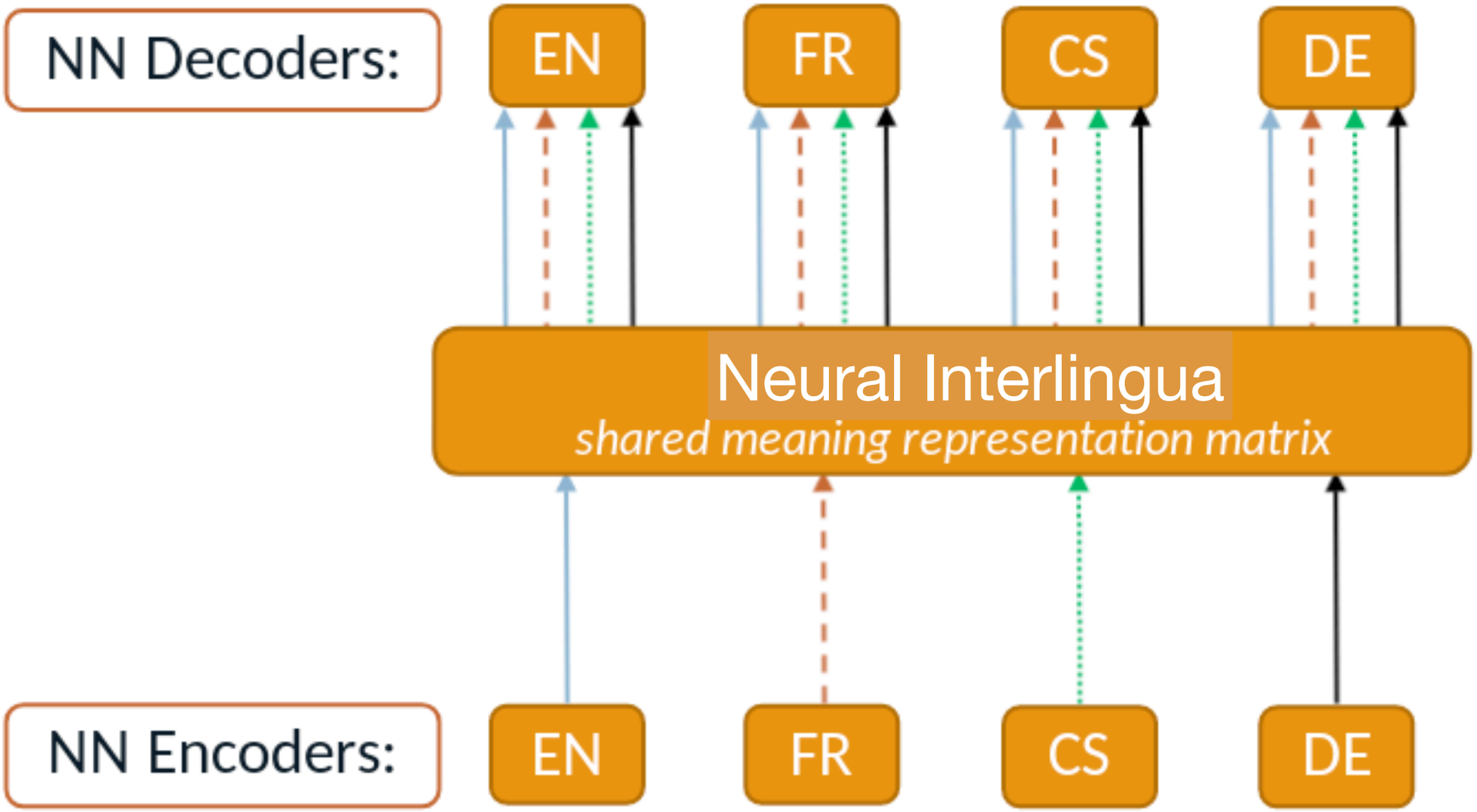
Machine translation and meaning

Use translations as semantic supervision!

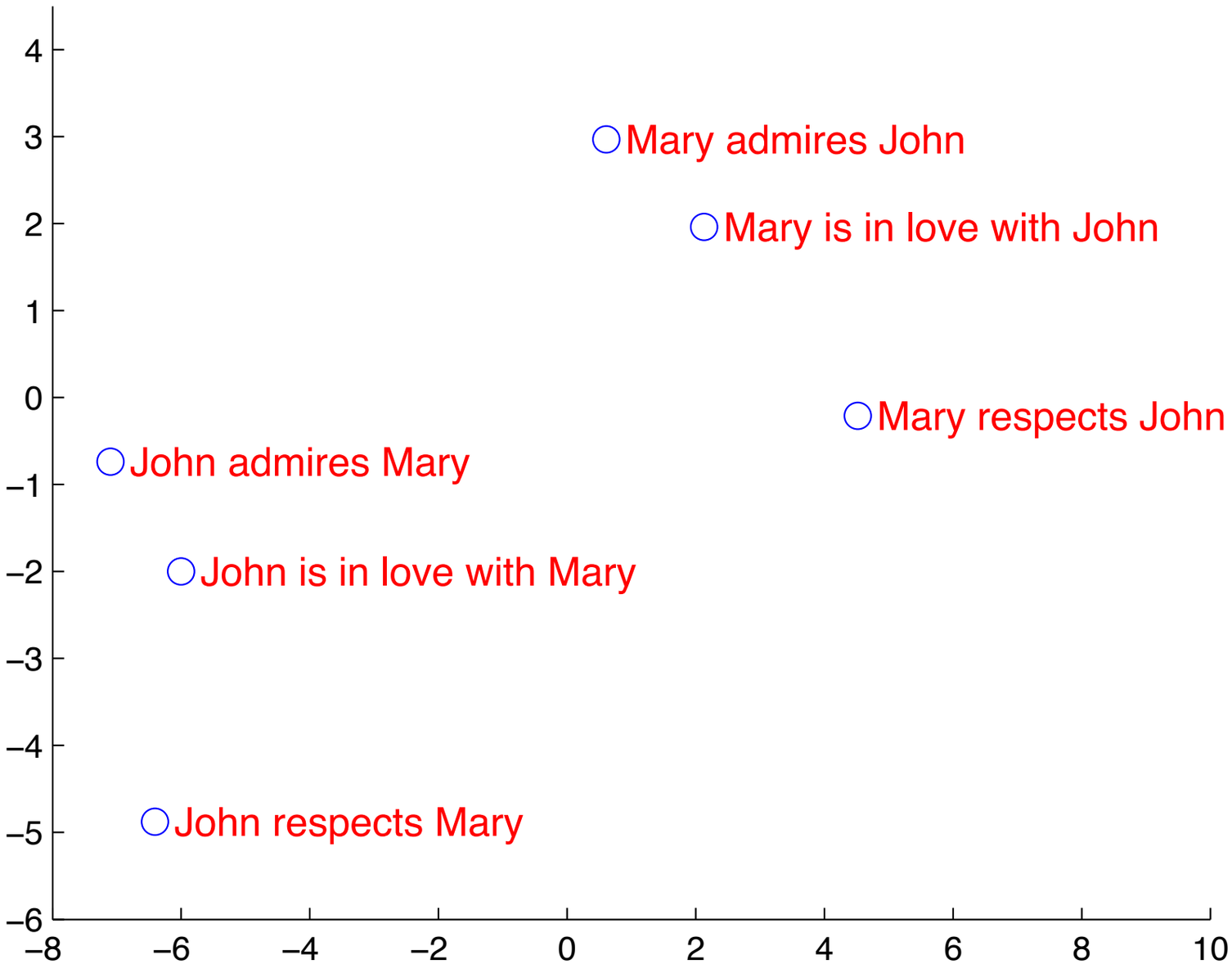
FOTRAN
Found in Translation



Multilingual machine translation



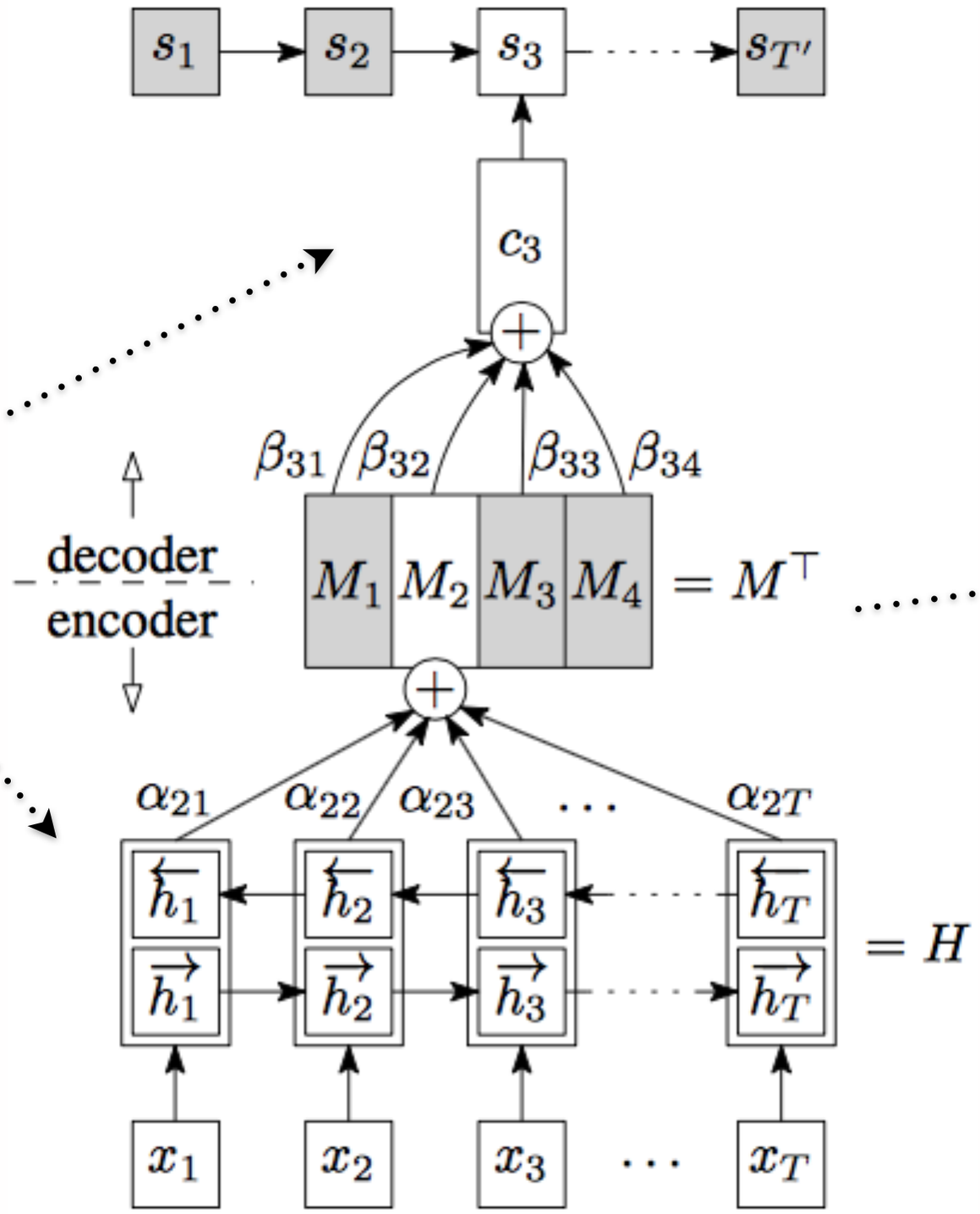
Semantic representation learning



(From Sutskever et. alet. al: "Sequence to Sequence Learning with Neural Networks")

Implementation of an “Attention Bridge”

language-specific parameters



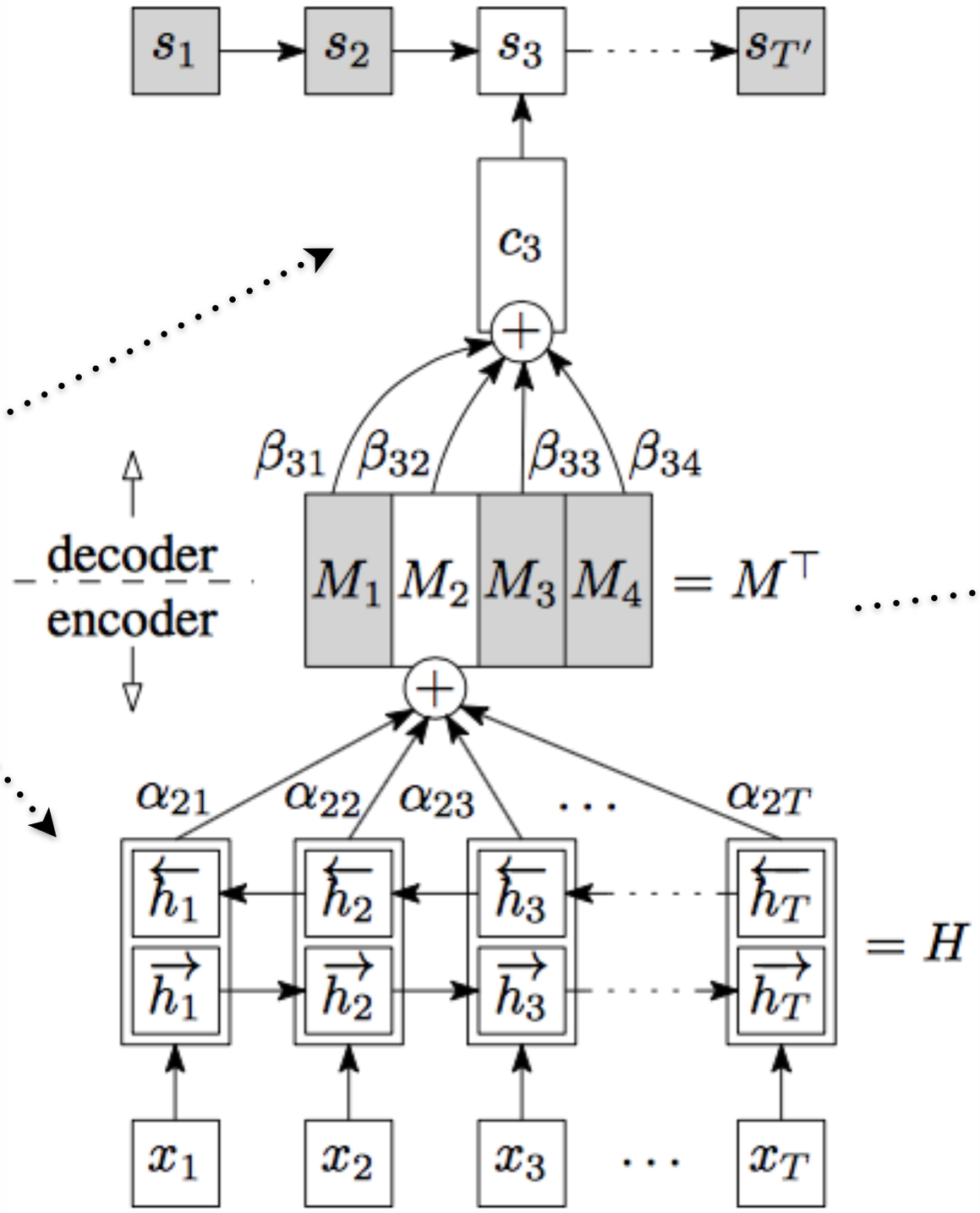
shared among all language pairs

Architecture proposed by Cífka and Bojar (2018).

Our implementation in OpenNMT-py (MTM2018)

Implementation of an “Attention Bridge”

language-specific parameters



shared among all language pairs

Rotate languages in scheduled training!

Architecture proposed by Cífka and Bojar (2018).

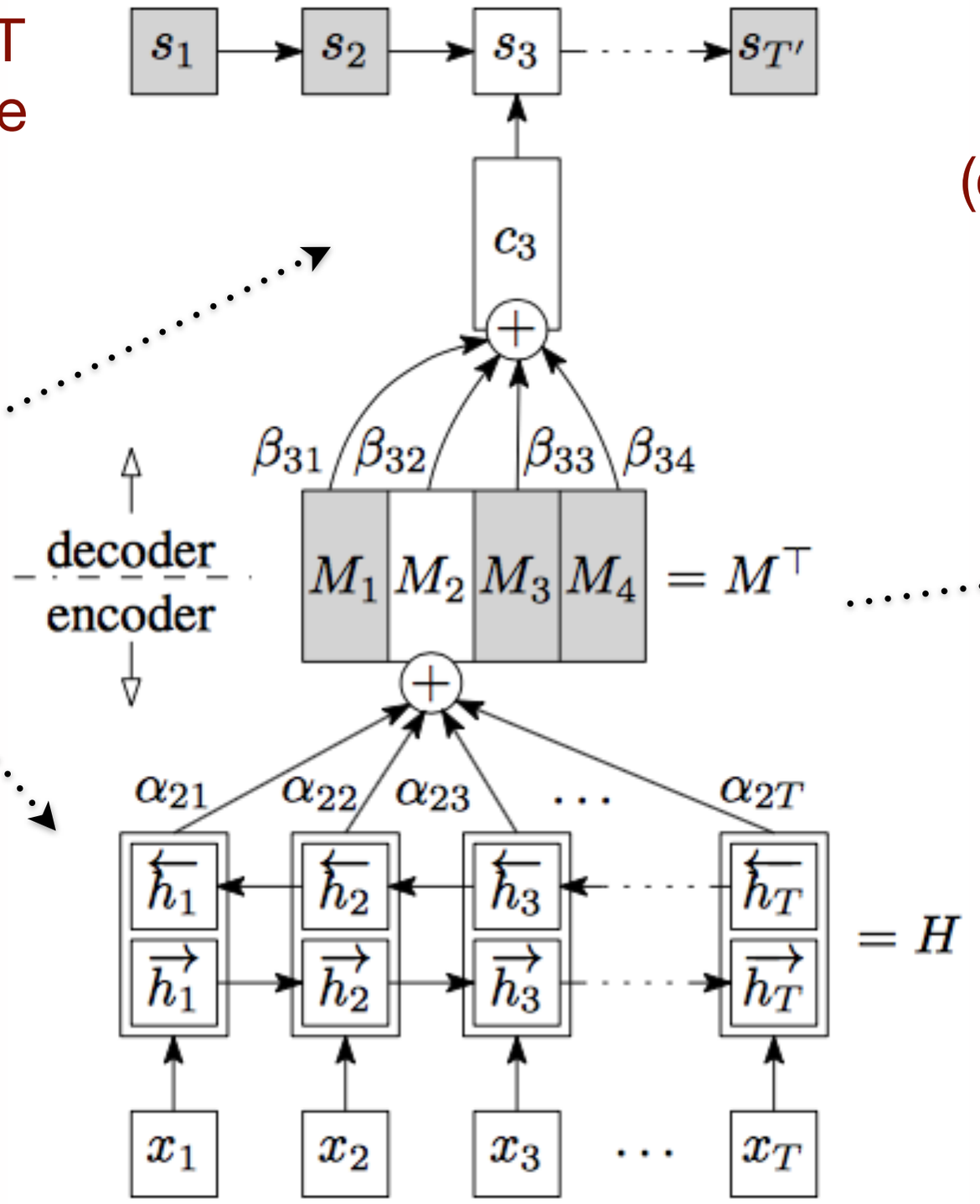
Our implementation in OpenNMT-py (MTM2018)

Implementation of an “Attention Bridge”

Benchmark with MT on unseen language pairs

Benchmark with semantic probing (downstream) tasks

language-specific parameters



shared among all language pairs

Rotate languages in scheduled training!

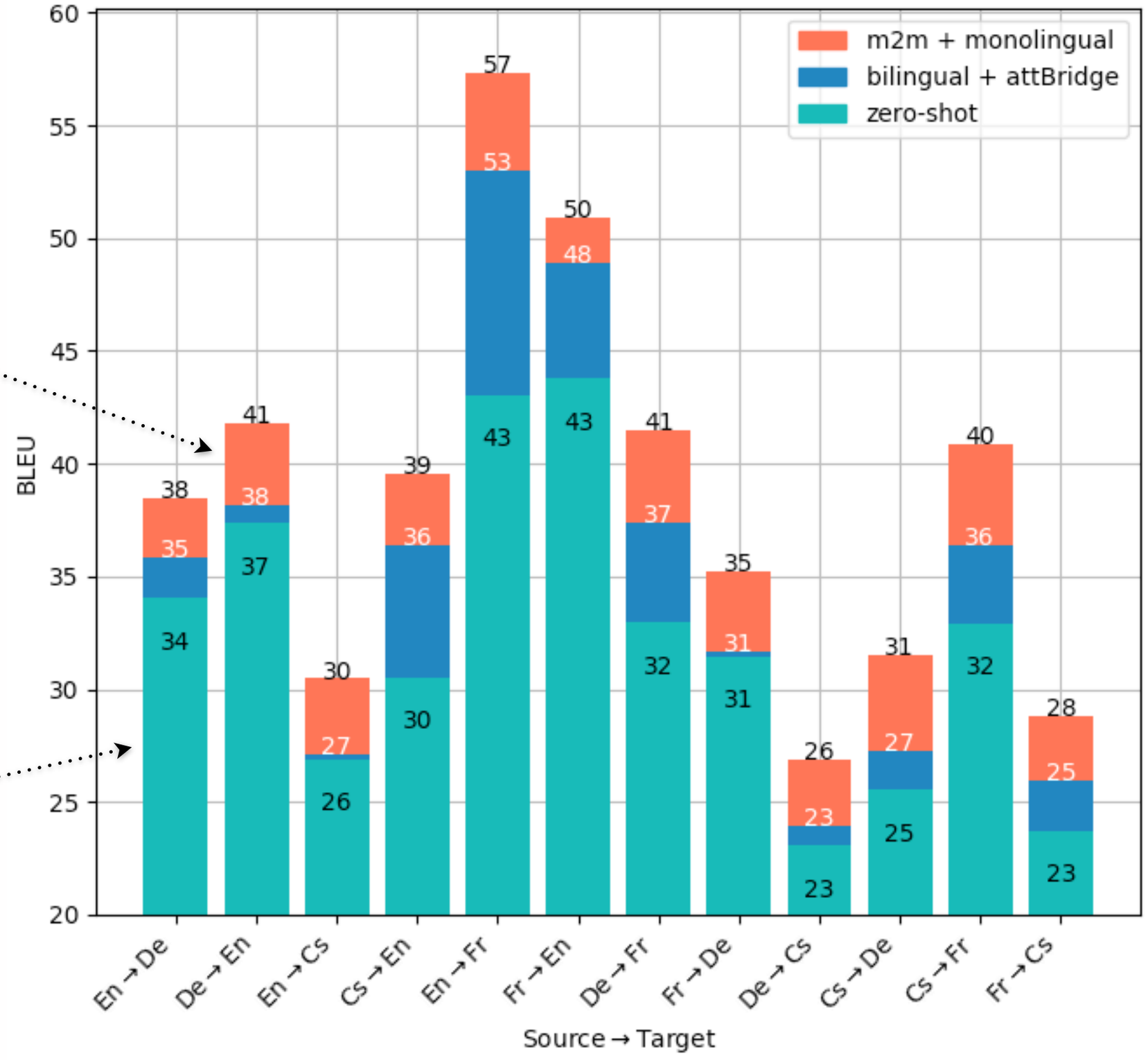
Architecture proposed by Cífka and Bojar (2018).

Our implementation in OpenNMT-py (MTM2018)

Translating image captions (6 languages)

model with all languages

language pair not seen in training data (zero shot)



Natural Language Inference (NLI)

Benchmark for reasoning with language

A black race car starts up in front of a crowd of people.

contradicts

A man is driving down a lonely road.

A soccer game with multiple males playing.

entails

Some men are playing a sport.

Measuring semantic similarity

Examples from the SICK dataset

ENTAILMENT, relatedness score = 4.7

The young boys are playing outdoors and the man is smiling nearby
The kids are playing outdoors near a man with a smile

CONTRADICTION, relatedness score = 3.6

The young boys are playing outdoors and the man is smiling nearby
There is no boy playing outdoors and there is no man smiling

NEUTRAL, relatedness score = 1.7

A lone biker is jumping in the air
A man is jumping into a full pool

Attention bridge in downstream tasks

multilingual models

natural language inference

TASK	EN-DE	EN-CS	EN-FR	M ↔ EN	M-2-M
SNLI	61.45	61.75	60.95	64.52	65.12
SICKE	72.82	73.89	74.85	75.46	76.92
TRAINABLE SEMANTIC SIMILARITY TASKS					
SICKR	0.685	0.720	0.717	0.727	0.740
	0.618	0.652	0.646	0.659	0.677
STS-B	0.578	0.603	0.591	0.629	0.678
	0.564	0.616	0.574	0.618	0.630

Language grounding

visual grounding

En: A *wall* divided the city.
De 1: Eine *Wand* teilte die Stadt. ×
De 2: Eine *Mauer* teilte die Stadt. ✓

translational grounding

Image from multimodal MT at WMT

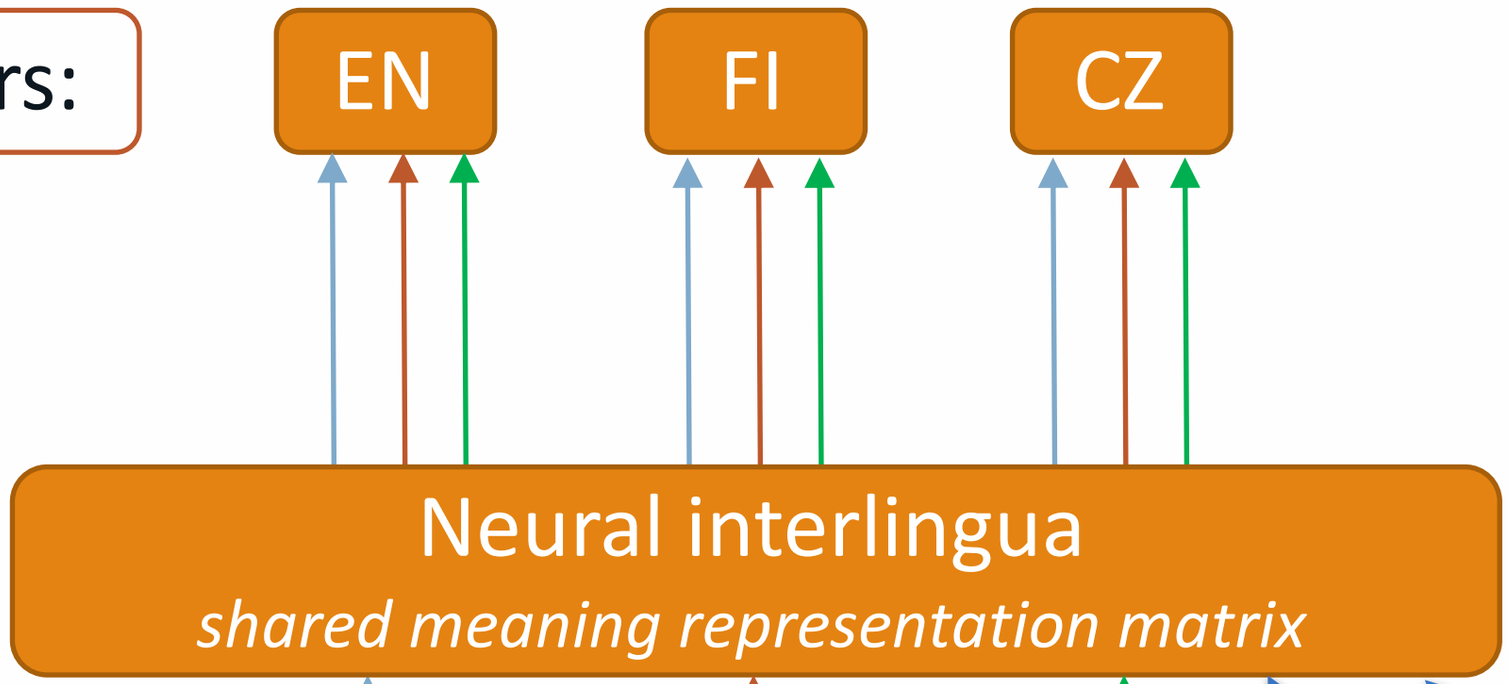
Adding multimodality

MeMAD

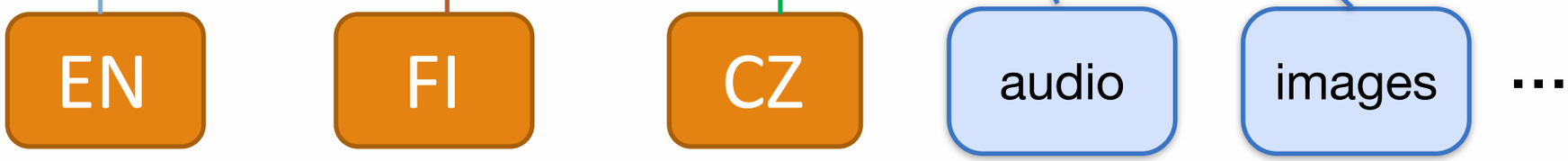
Methods for Managing Audiovisual Data

https://memad.eu

NN Decoders:



NN Encoders:



MeMAD: Access to audio-visual content

automatic transcription of videos accessible in many languages

https://memad.eu

MeMAD
Methods for Managing Audiovisual Data

All of this is powered by CSC!

puhti, cPouta, ObjectStorage, allas, ...

Constant need of billing units ...

OPUS

Billing units remaining:

10.1

103888 / 1030901

NLPL

Billing units remaining:

45.6%

870941 / 1910000

CrossNLP

Billing units remaining:

0 / 3820087

OPUS-MT

Billing units remaining:

42.8%

1073760 / 2510000

MeMAD

Billing units remaining:

28.8%

463936 / 1610000

Multi-MT

Billing units remaining:

0 / 2319155

OPUS-LM

Billing units remaining:

100%

1010000 / 1010000

FoTran

Billing units remaining:

5

614261 / 11500000

LingDA

Billing units remaining:

49.7%

99364 / 200000

> 22 million BUs used in the projects above, mostly in the last 2 years

Storage needs keep on growing ...

```
Project scratch
-----------------------------------------------------
456.5G/1T          14.27k/1000k  CrossNLP
721.1G/1T          13.32k/1000k  NeIC-NLPL
  195G/1T          90.05k/1000k  NLPL-OPUS
3.236T/4.883T     852.59k/1000k  MeMAD
7.664T/10T        308.63k/1000k  FoTran
  208G/1T           7.53k/1000k  LingDA-LT
 4.83T/10T        415.40k/1000k  MultiMT
  168G/1T           6.06k/1000k  SimplifyRussian
 2.55G/1T            .19k/1000k  offenseval
  136G/1T          13.72k/1000k  OPUS-MT
845.3G/1T           7.47k/1000k  OPUS-LM
    4k/1T             0k/1000k  EOSC-NLPL
```

> 20T in total

```
Other data storage
-----------------------------------------------------
       10T                          NLPL (puhti)
       7.5T                         OPUS (cPouta)
    ??? T                           ObjectStorage, IDA
```

Cloud services on cPouta

Instances

Used 12 of 16

VCPUs

Used 30 of 32

RAM

Used 39.1GB of 156.3GB

Floating IPs

Used 5 of 5

Security Groups

Used 4 of 20

Volumes

Used 7 of 10

Volume Storage

Used 5.5TB of 7.3TB

https://translate.ling.helsinki.fi

Cloud services on cPouta

Instances

Used 12 of 16

VCPUs

Used 30 of 32

RAM

Used 39.1GB of 156.3GB

Floating IPs

Used 5 of 5

Security Groups

Used 4 of 20

Volumes

Used 7 of 10

Volume Storage

Used 5.5TB of 7.3TB

OPUS

http://opus.nlpl.eu

NLPL

Nordic Language Processing Laboratory

fiskmö

finsk-svensk korpus & maskinöversättning

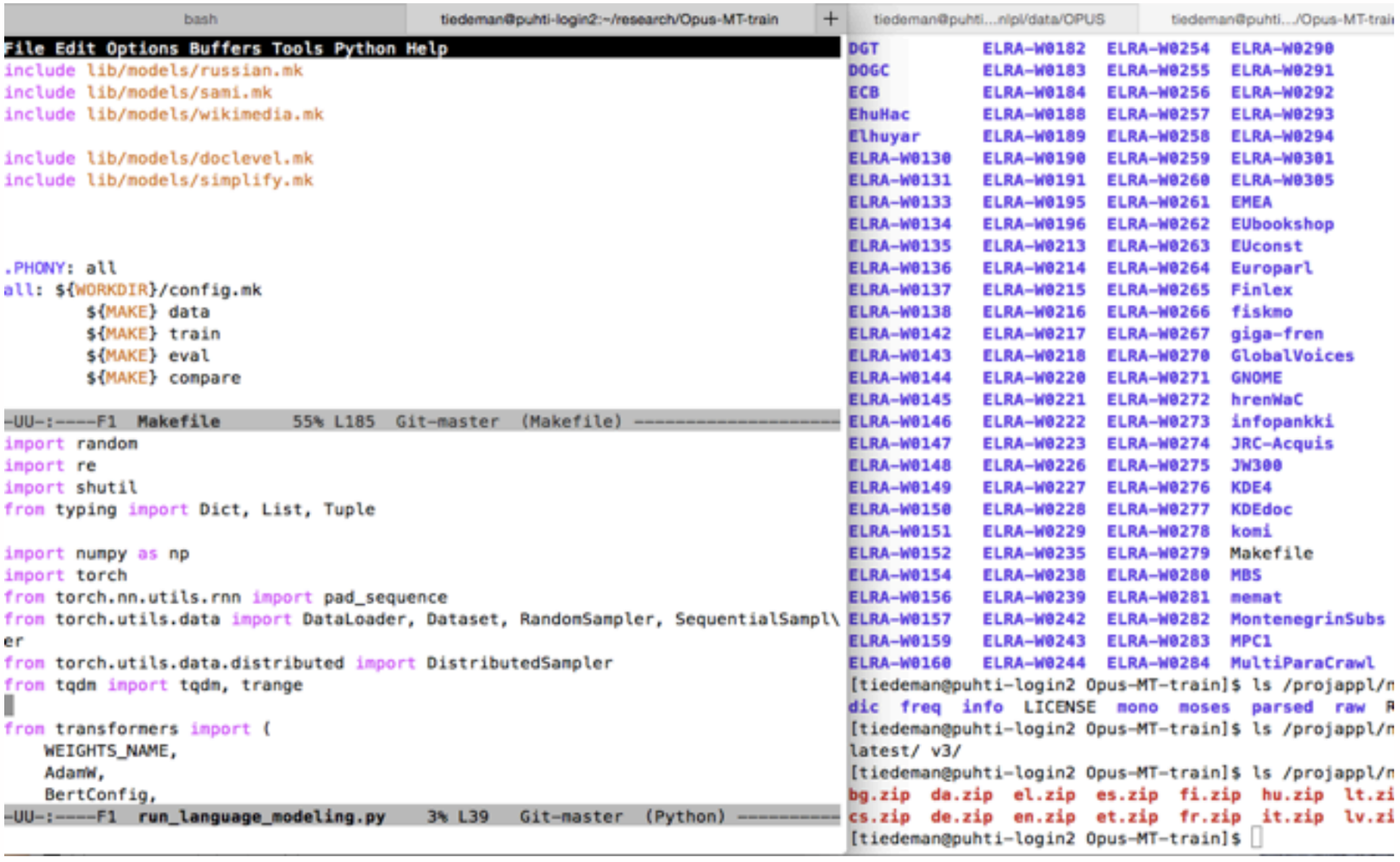
https://translate.ling.helsinki.fi

https://blogs.helsinki.fi/fiskmo-project/

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

How do we work?

- command-line, ssh, bash-scripts, slurm, github, slack
- pyTorch, tensorflow, OpenNMT, MarianNMT, ….
- avoid graphical user interfaces for research & development



To sum up: What do we need?

Computing resources

- more GPU's
- shorter job queues (right now on puhti: 2408 jobs waiting)
- storage for data (general and temporary)
- sometimes fast I/O

On-going work and plans for the future

- cross-border activities (NLPL, EOSC-nordic)
- better and more efficient access to data
- improved replicability (sharing of data, models, code)

CSC - IT CENTER FOR SCIENCE

erc

Thank You!

Questions and discussions?

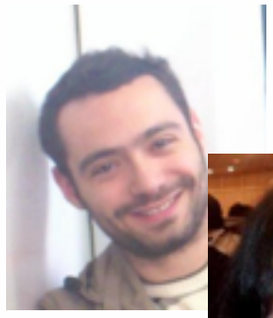
https://blogs.helsinki.fi/language-technology/

ACADEMY OF FINLAND

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

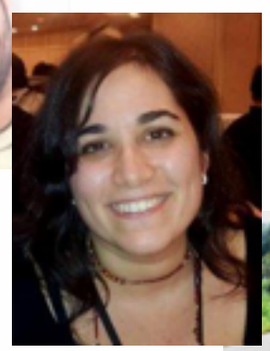
Language Technology in Helsinki

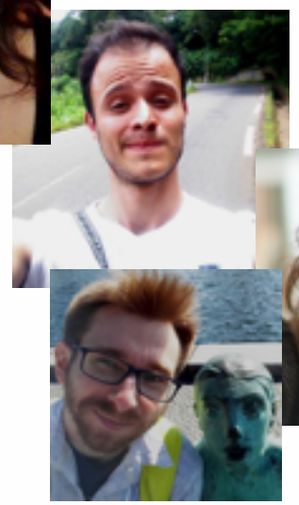
http://blogs.helsinki.fi/language-technology/



fiskmö
finsk-svensk korpus & maskinöversättning

Data collection & MT
2 languages (Finnish/Swedish)
https://blogs.helsinki.fi/fiskmo-project/





FOTRAN
Found in Translation

semantics & MT
> 1,000 languages

https://github.com/Helsinki-NLP

audiovisual data & MT
6 languages

OPUS

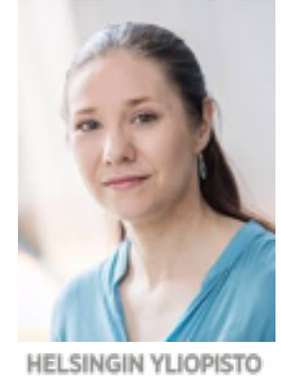
http://opus.nlpl.eu

Data collection for MT
> 200 languages

sentimentator

MeMAD
Methods for Managing Audiovisual Data

https://memad.eu



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI