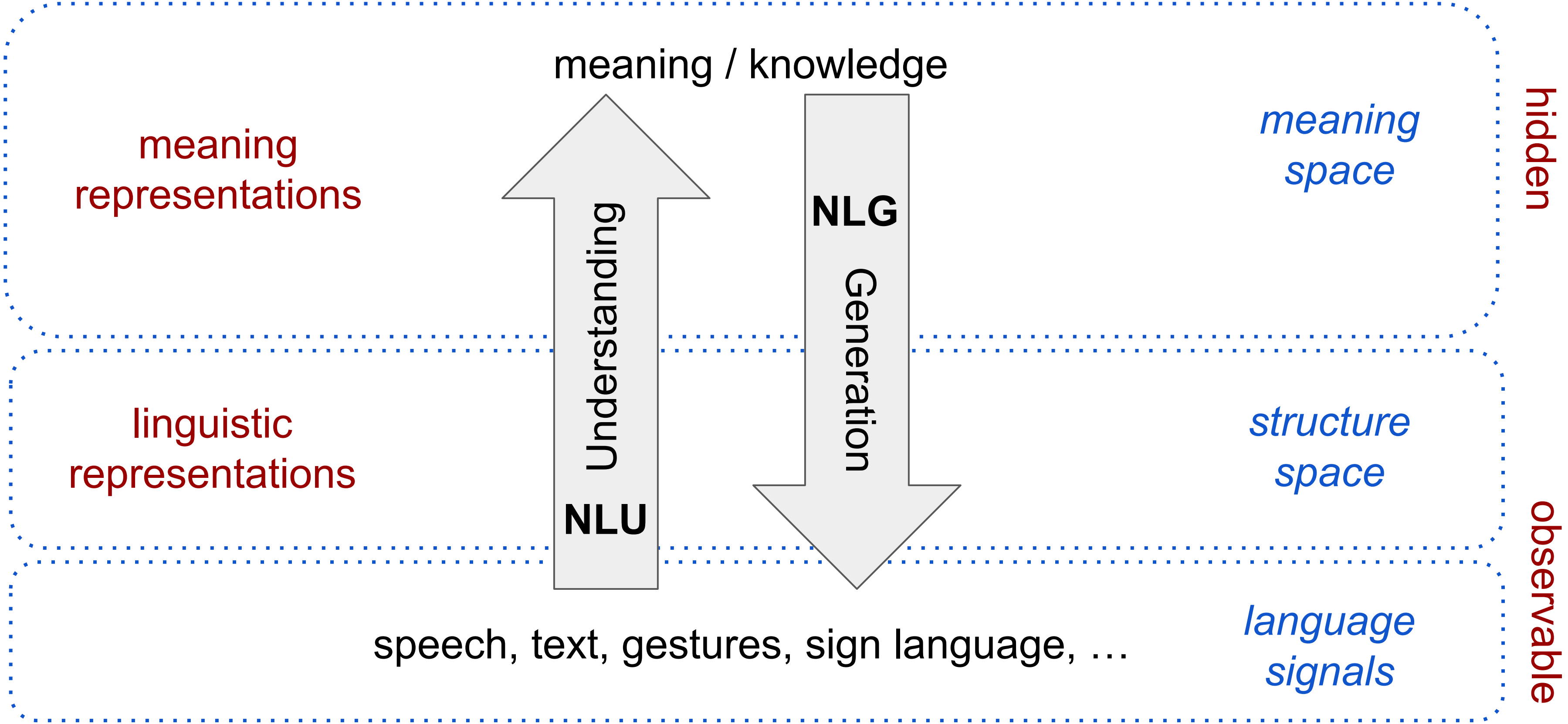
Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

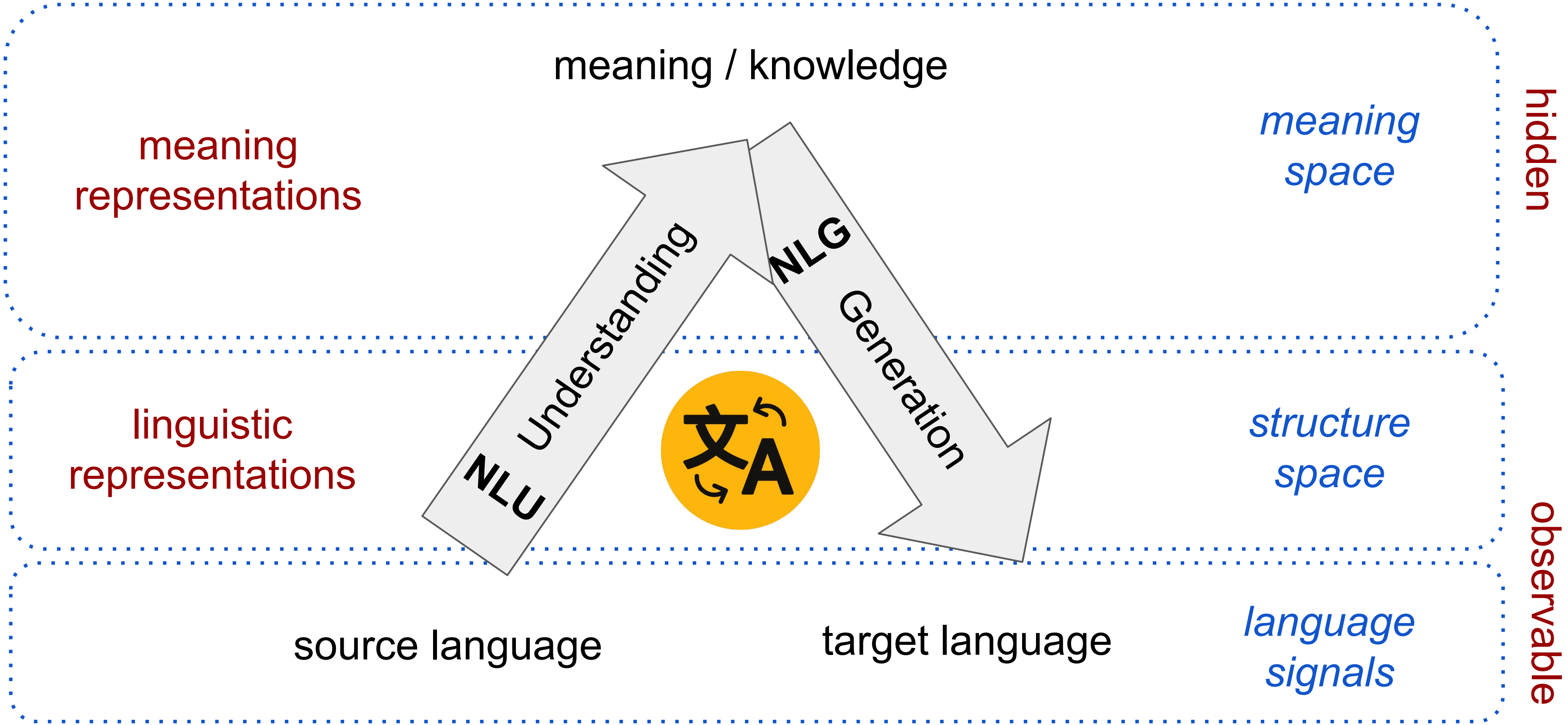
Lost in Meaning - Found in Translation

Natural Language Understanding with Multilingual Data

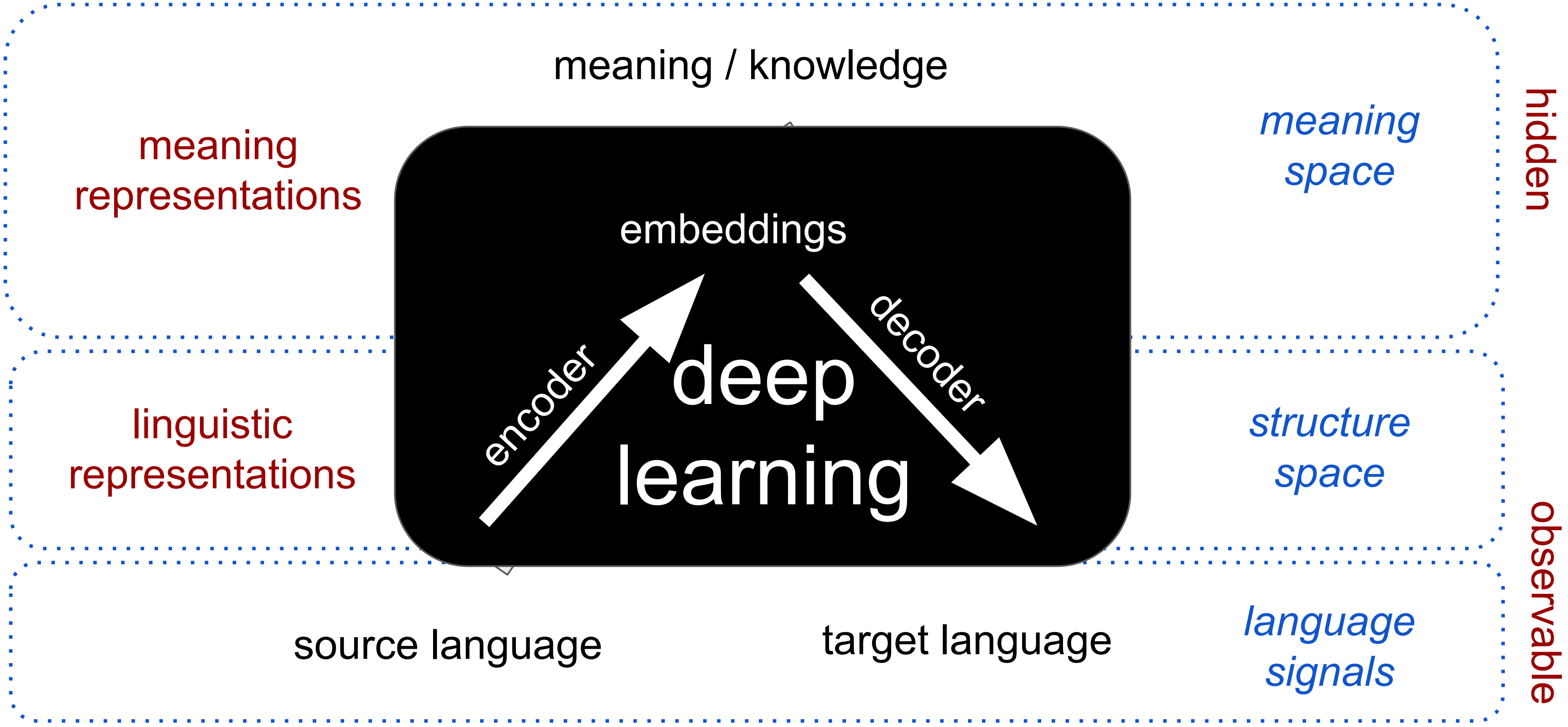
Natural Language Processing



Natural Language Processing: **Machine Translation**

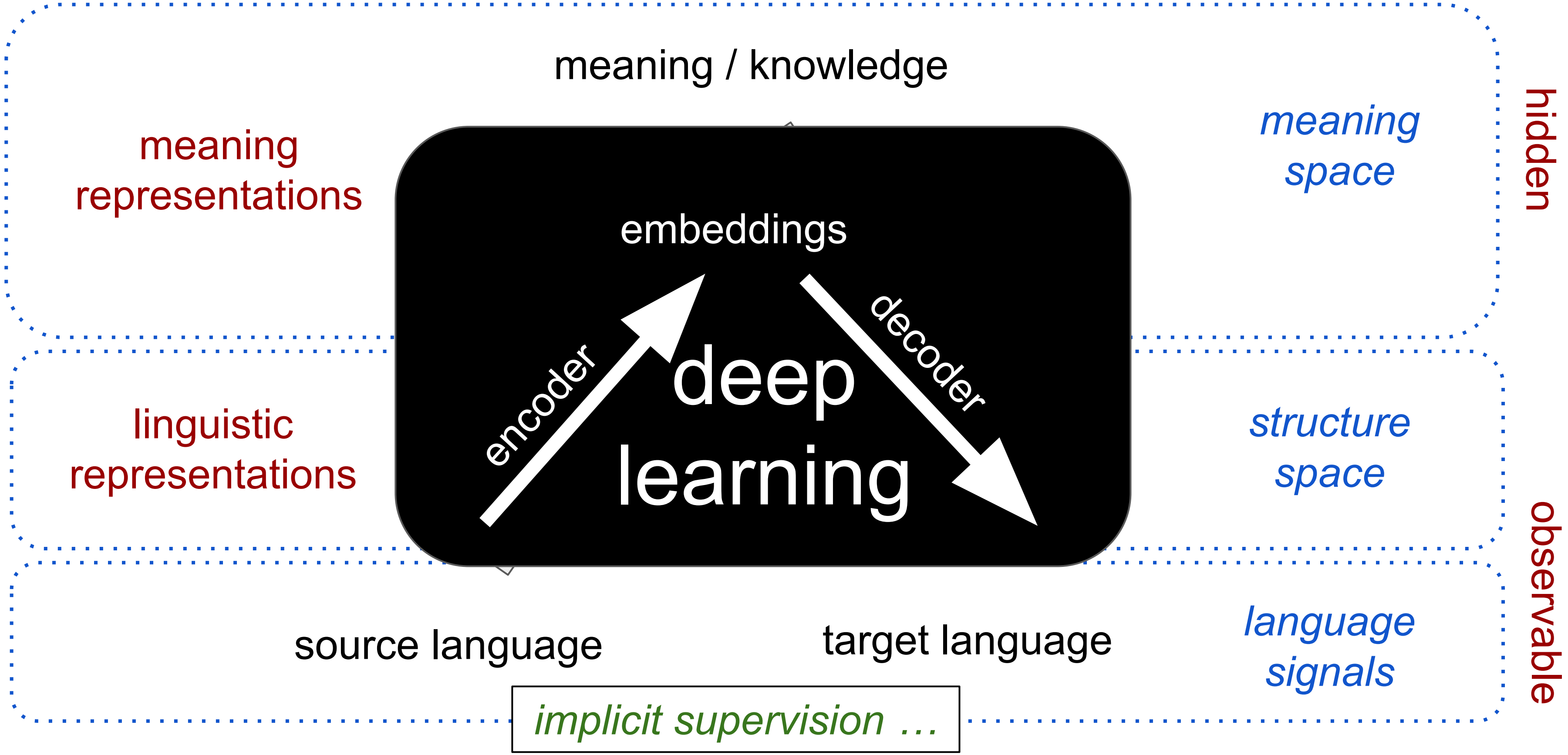


Neural Machine Translation

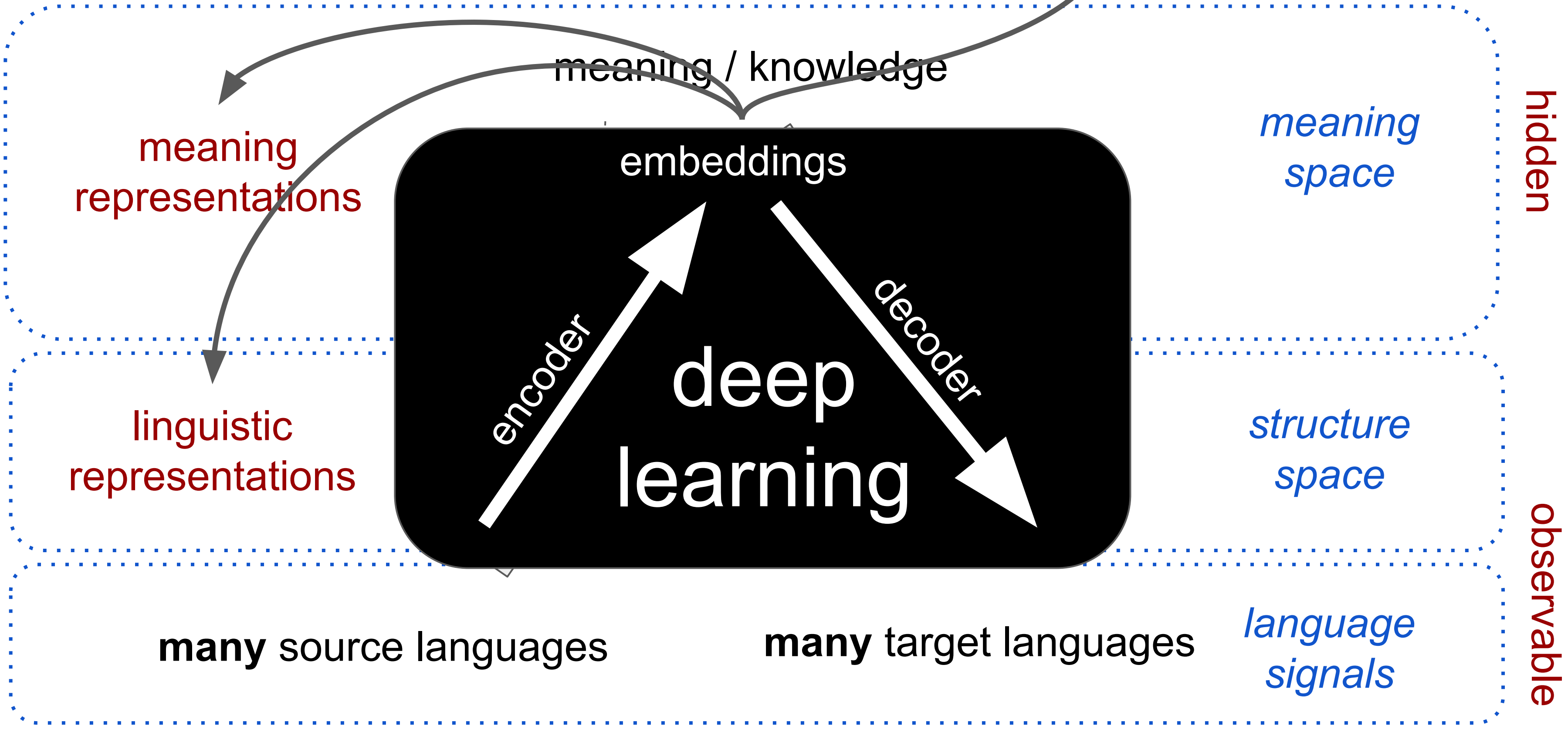


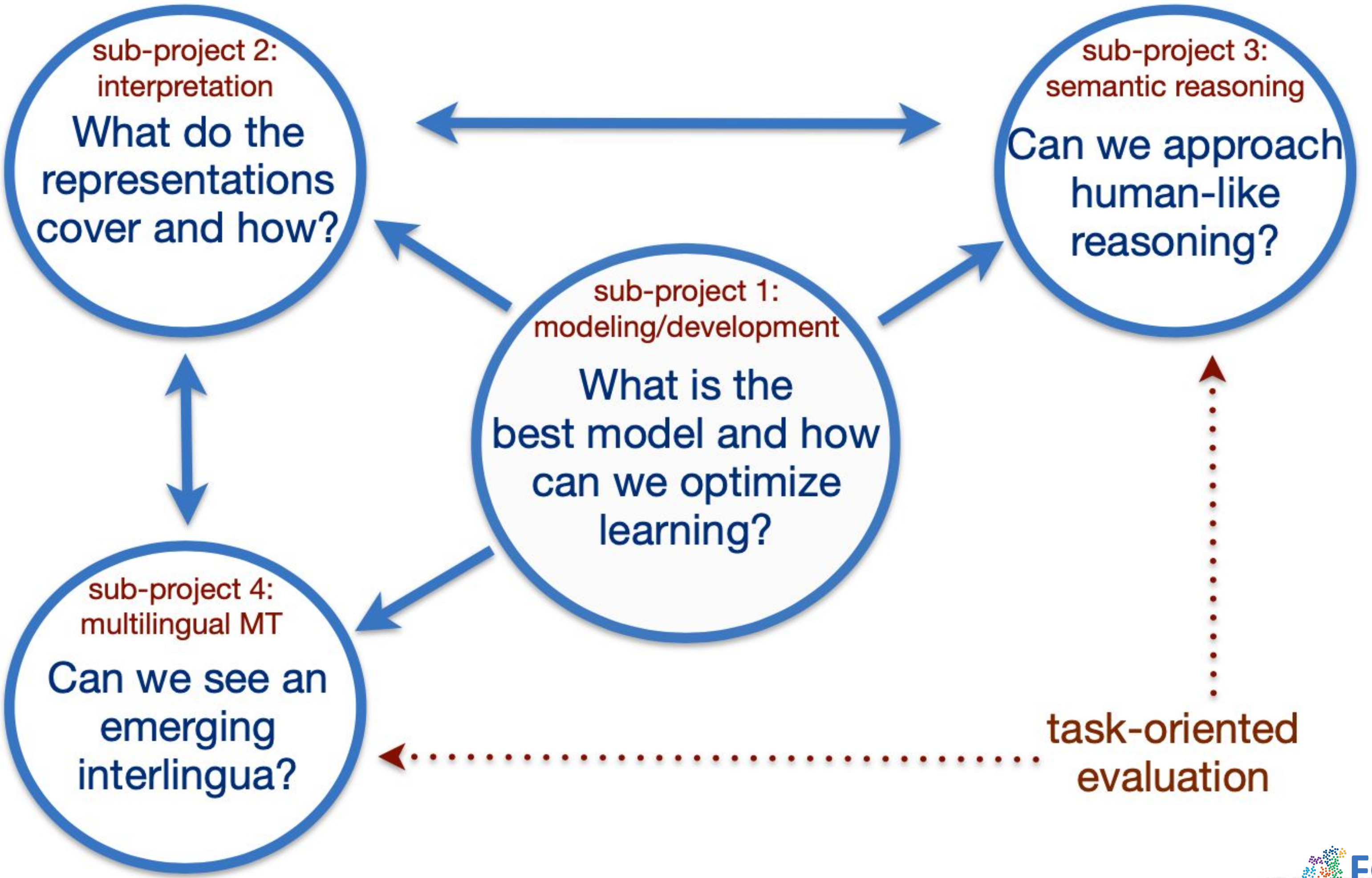
Neural Machine Translation

… for latent representations



Multilingual Neural Machine Translation



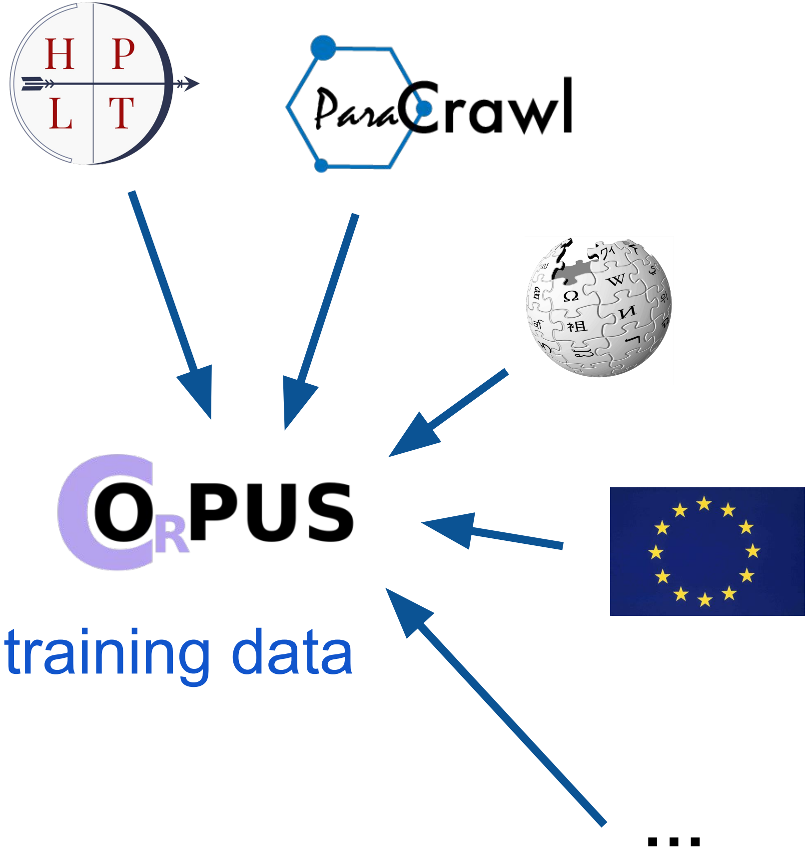


FOTRAN
Found in Translation

(1) Creating the basic environment:
The OPUS ecosystem

https://github.com/Helsinki-NLP/OPUS

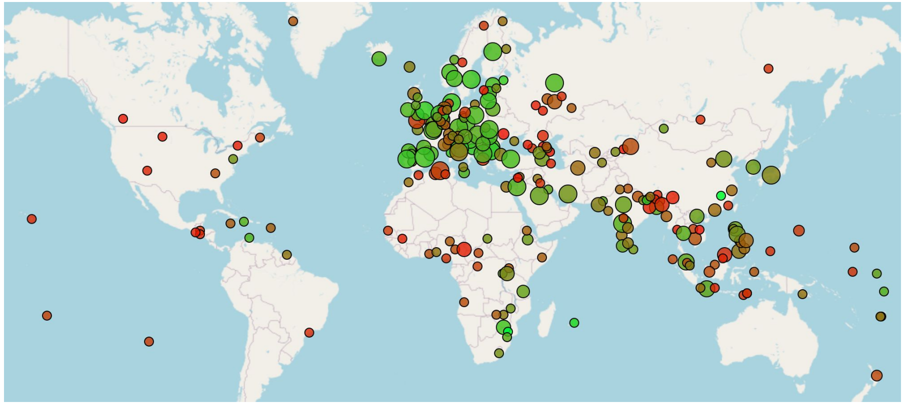
The OPUS corpus




```
pip install opustools
```

Total size of all releases: ca 30 TB

- > 700 languages
- > 40,000 language pairs
- > 45 billion sentences



https://github.com/Helsinki-NLP/OPUS-MT

OPUS-MT: pre-trained translation models

OpusFilter
OpusCleaner
OpusTrainer

Tatoeba Translation Challenge

2,347 released translation models

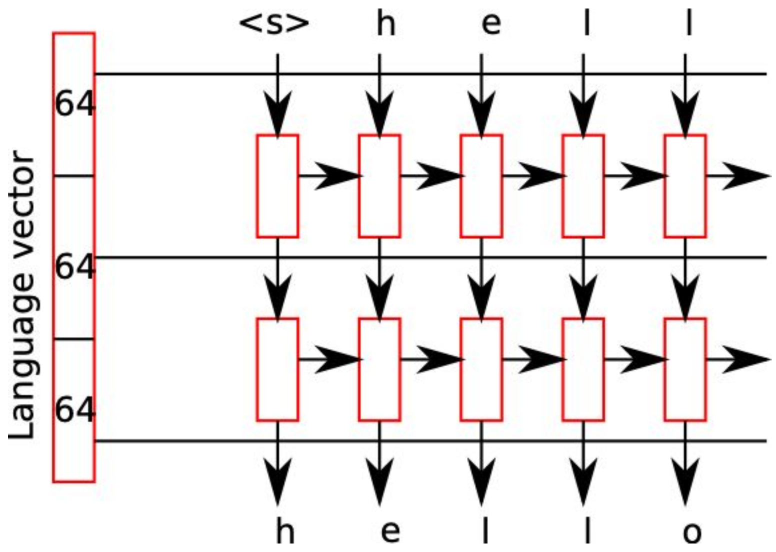
- 758 multilingual models
- base and big transformer models
- compact students models

The Blessings of Multilinguality

The language continuum and language embeddings

Back in 2016:

1303 Bible translations
into 990 languages



Continuous multilinguality with language vectors

Robert Östling
Department of Linguistics*
Stockholm University
robert@ling.su.se

Jörg Tiedemann
Department of Modern Languages
University of Helsinki
jorg.tiedemann@helsinki.fi

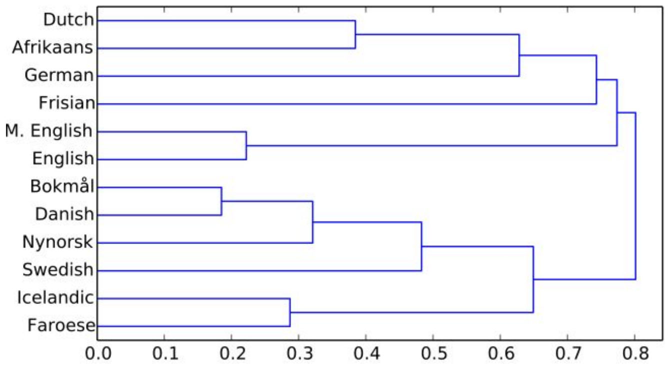
Abstract

Most existing models for multilingual natural language processing (NLP) treat language as a discrete category, and make predictions for either one language or the other. In contrast, we propose using continuous vector representations of language. We show that these can be learned

separate model for each language. This presupposes large quantities of monolingual data in each of the languages that needs to be covered and each model with its parameters is completely independent of any of the other models.

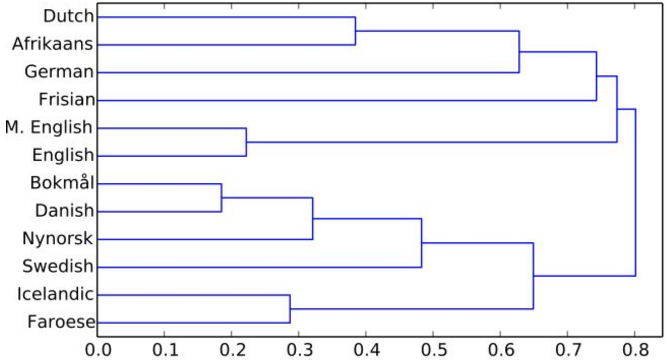
We propose instead to use a single model with real-valued vectors to indicate the language used, and to train this model with a large number of languages. We thus get a language model whose

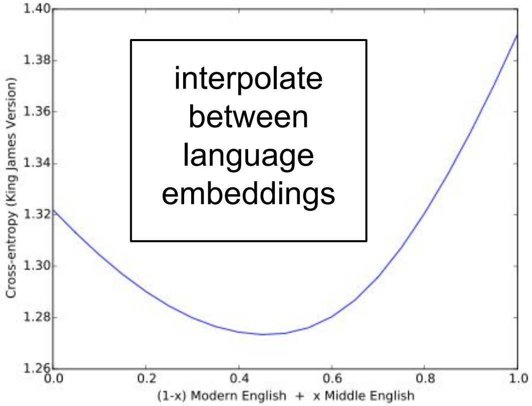
Continuous multilinguality with language embeddings



Language clusters from language embeddings

Continuous multilinguality with language embeddings



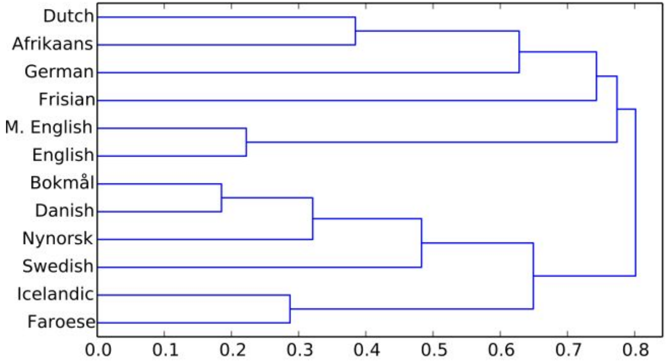


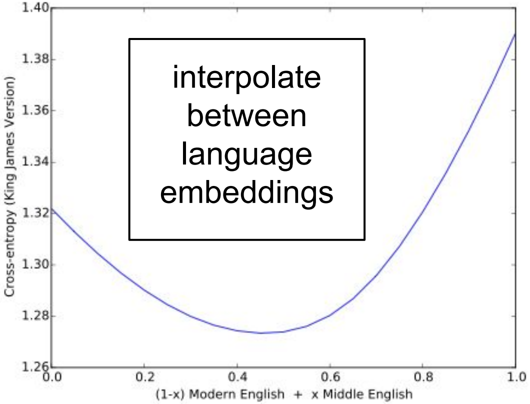
%	Random sample (temperature parameter $\tau = 0.5$)
30	and thei schulen go in to alle these thingis, and schalt endure bothe in the weie
40	and there was a certaine other person who was called in a dreame that he went into a mountaine.
44	and the second sacrifice, and the father, and the prophet, shall be given to it.
48	and god sayd, i am the light of the world, and the powers of the enemies of the most high god may find first for many.
50	but if there be some of the seruants, and to all the people, and the angels of god, and the prophets
52	then he came to the gate of the city, and the bread was to be brought
56	therefore, behold, i will lose the sound of my soul, and i will not fight it into the land of egypt
60	and the man whom the son of man is born of god, so have i therefore already sent to the good news of christ.

middle English

modern English

Continuous multilinguality with language embeddings





Control text generation with language embeddings:

turn on Swedish:

och jehova sade till honom : ” jehova har sagt , och jag skall ...

turn on German:

und er sprach zu ihnen : siehe , ich bin der herr

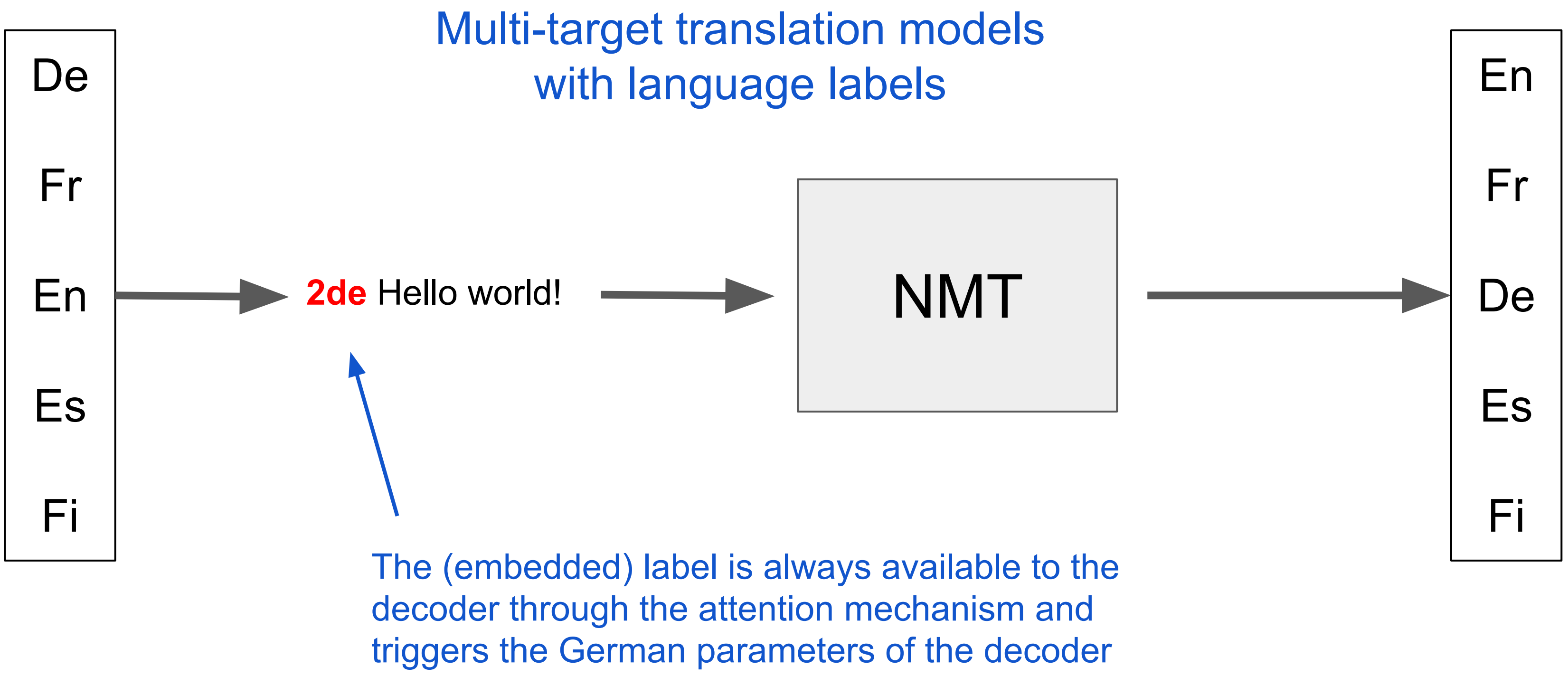
mix Swedish and German:

vocken ånner vocken ånnen söhenöckenföcken ...

average of Scandinavian languages:

og han sa til herrens : " han skal vitnaðus til herrens hjárt

Machine translation with language embeddings



Effective transfer learning

BLEU scores (in %)

Model / test set	Belarusian → English	English → Belarusian
Belarusian - English	10.0	8.2
East Slavic languages - English	38.7	20.8
Slavic languages - English	**42.7**	**22.9**

The Curse of Multilinguality

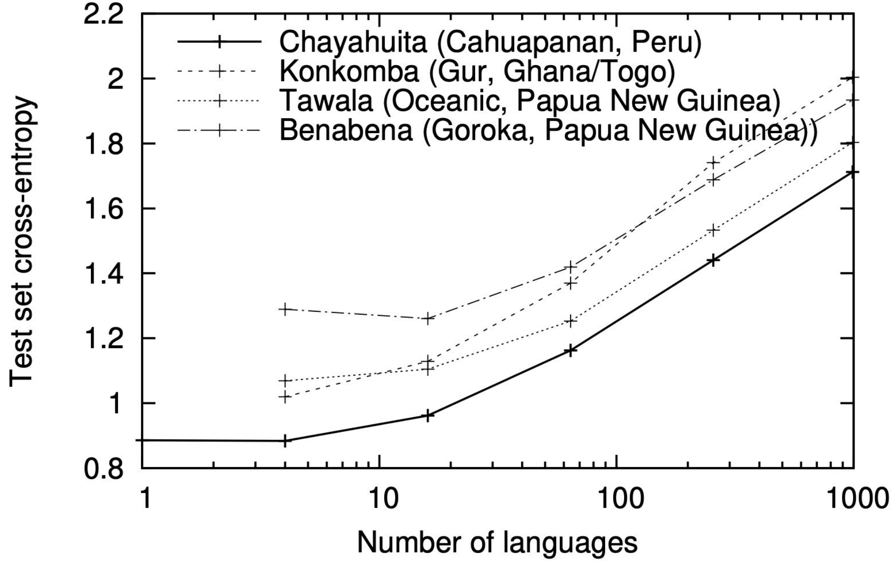
Limits of generalisation & transfer learning

BLEU scores (in %)

Model / test set	Belarusian → English	English → Belarusian
Belarusian - English	10.0	8.2
East Slavic languages - English	38.7	20.8
Slavic languages - English	**42.7**	**22.9**
Indo-European languages - English	41.7	18.1

(increasing language coverage while keeping the model size constant)

(1) Limits of the model capacity

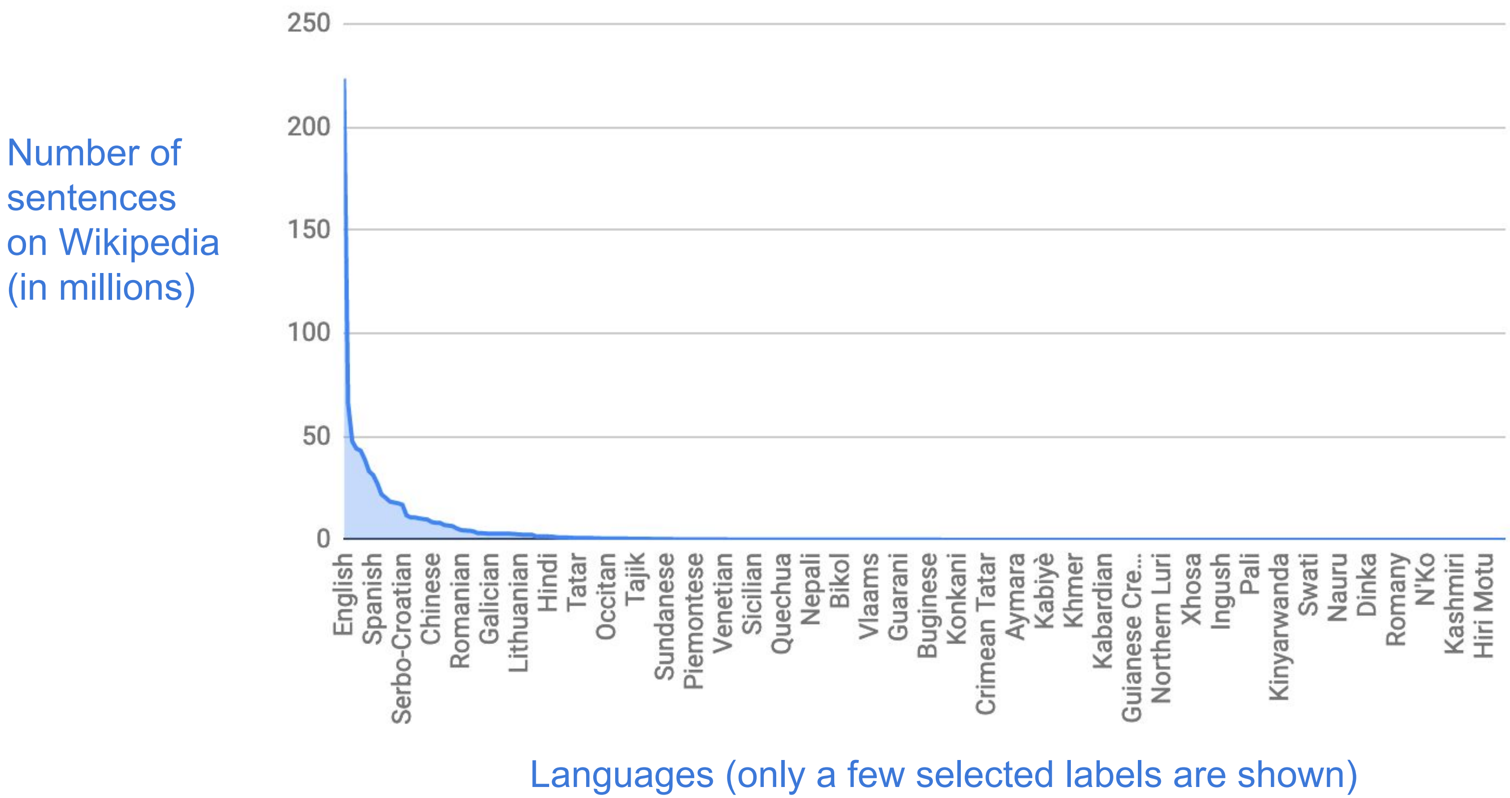


Testing the model capacity when adding more languages

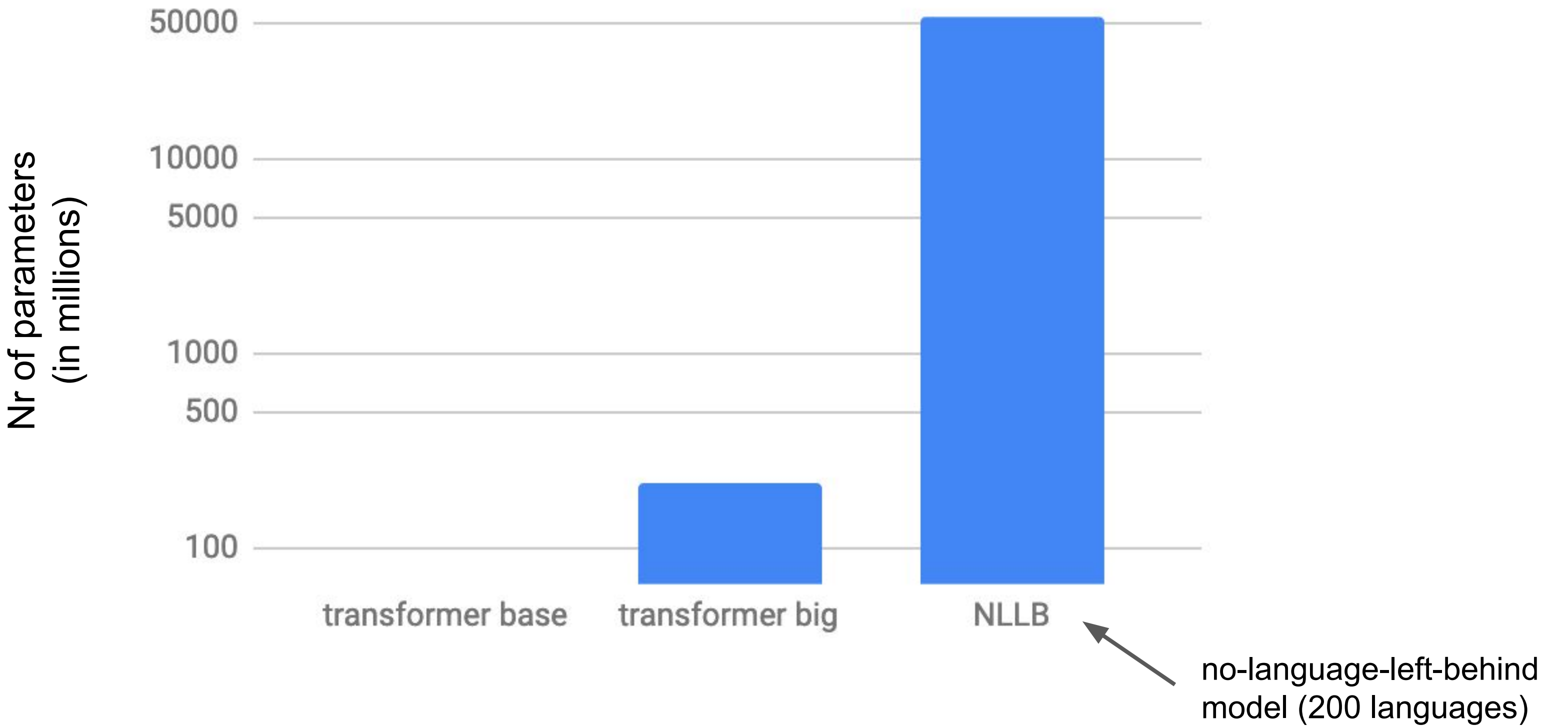
(similar patterns for adding languages in random order or according to typological relationship)

(from *"Continuous multilinguality with language vectors"*)

(2) Limits of training data

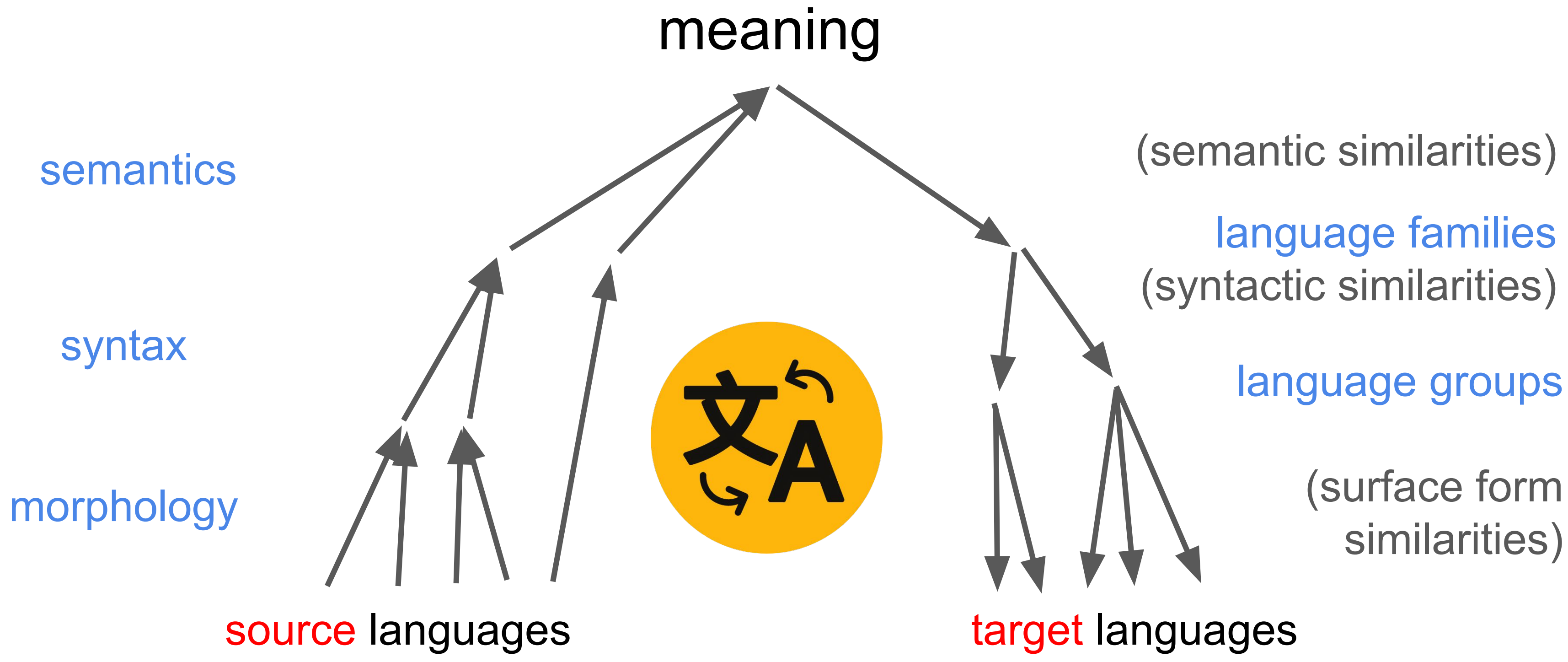


(3) Growing model size also for multilingual MT models



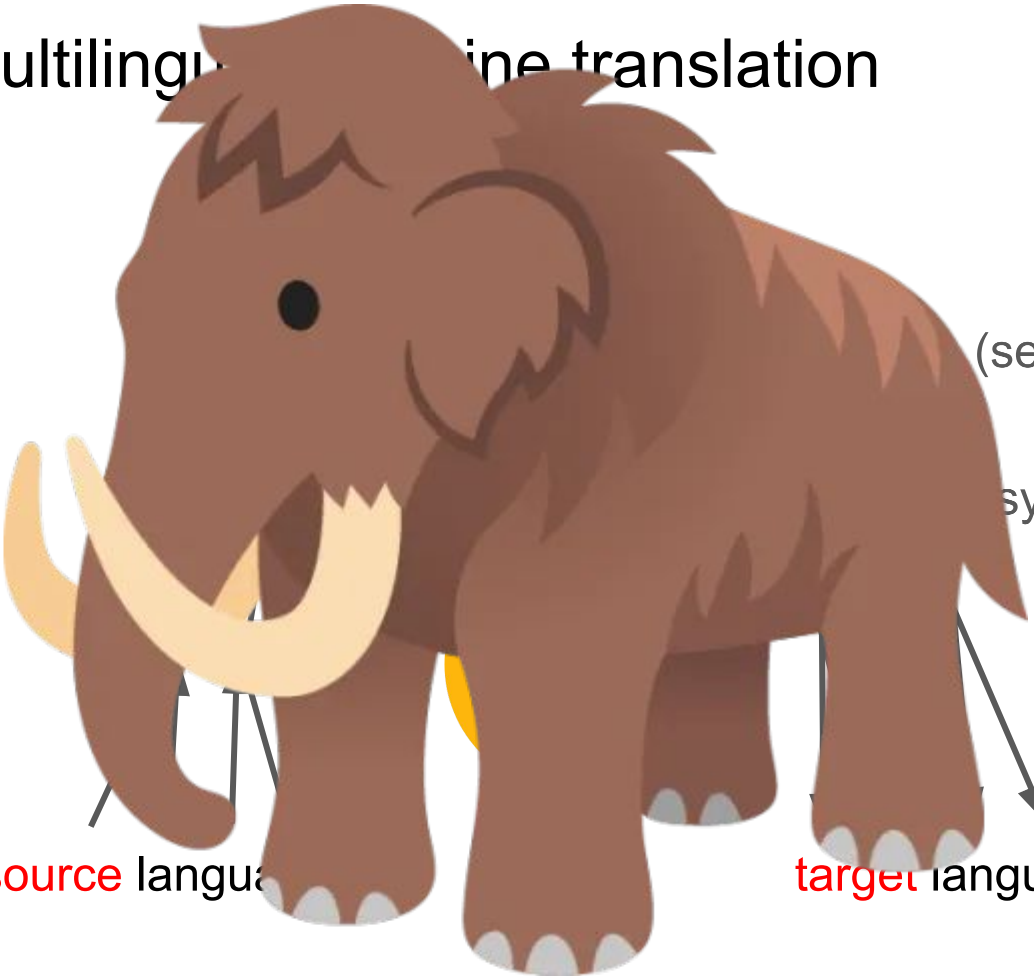
Back to Modularity

Modular multilingual machine translation



https://github.com/Helsinki-NLP/Mammoth

Modular multilingu ine translation



semantics

syntax

morphology

(semantic similarities)

language families

syntactic similarities)

language groups

(surface form similarities)

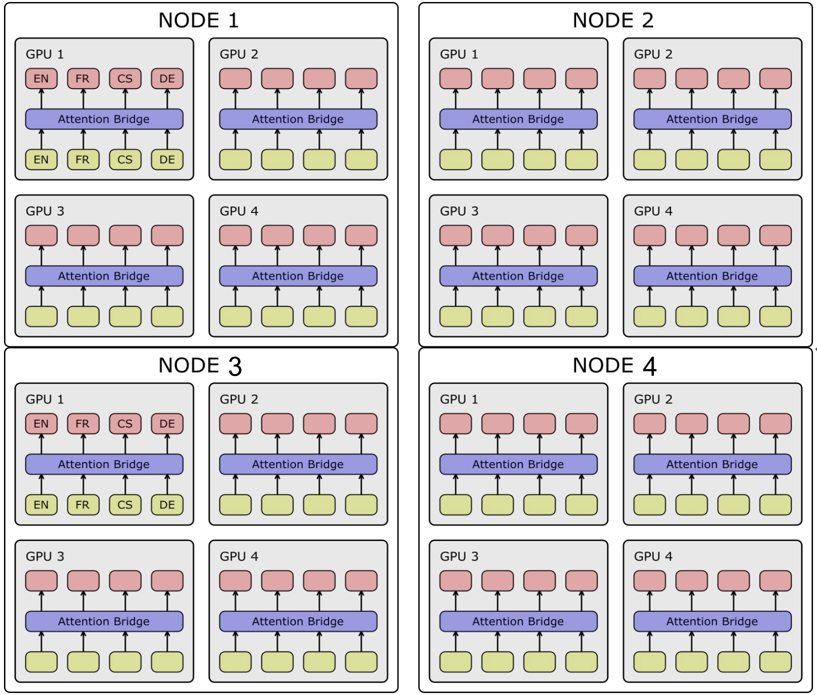
source langua

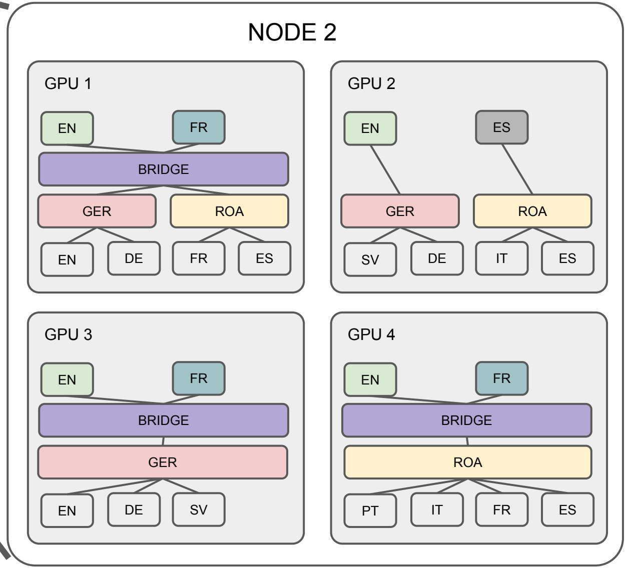
targe languages

https://github.com/Helsinki-NLP/Mammoth

Building scalable modular models

various types of modularity
and parameter sharing



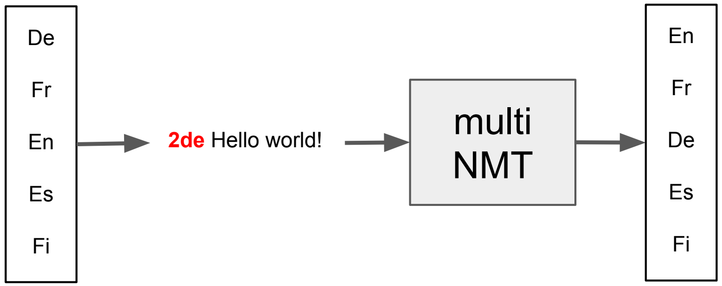


Efficient parallelization and resource allocation

https://github.com/Helsinki-NLP/Mammoth

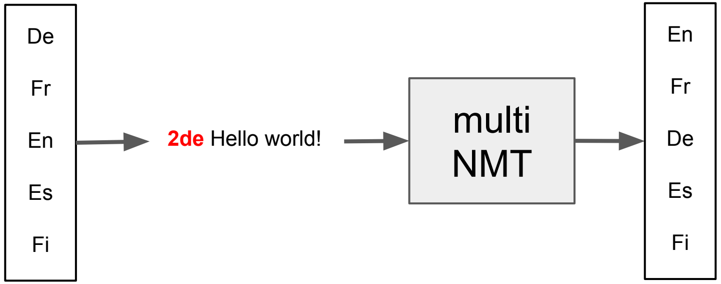
Support for different types of parameter sharing

(1) full sharing with language labels:
(e.g. Johnson et al., 2017)



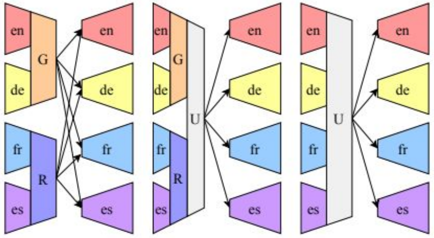
Support for different types of parameter sharing

(1) full sharing with language labels: (e.g. Johnson et al., 2017)



(2) partial sharing schemes

(e.g.,Purason & Tättar, 2022)

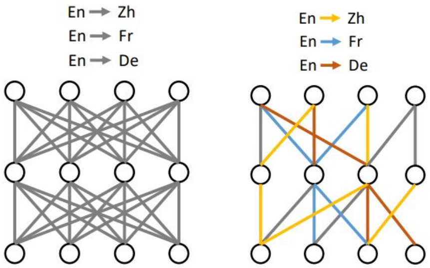


(a) Language group sharing

(b) Language group and universal sharing (tiered)

(c) Universal sharing (between all languages)

(e.g., Lin et al., 2021)



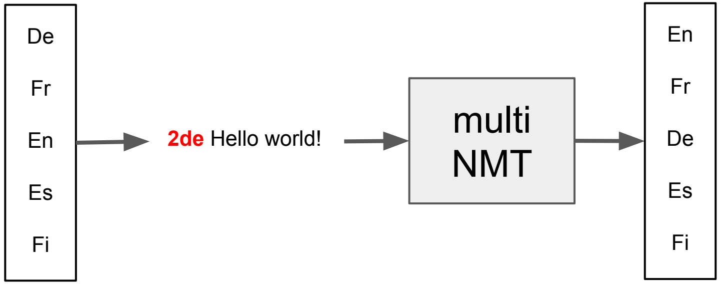
(a) Full network

(b) LaSS

https://github.com/Helsinki-NLP/Mammoth

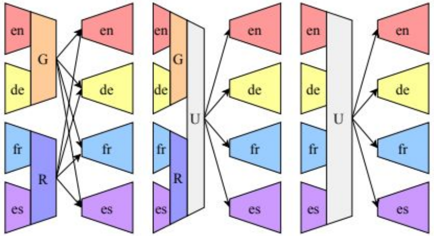
Support for different types of parameter sharing

(1) full sharing with language labels:
(e.g. Johnson et al., 2017)



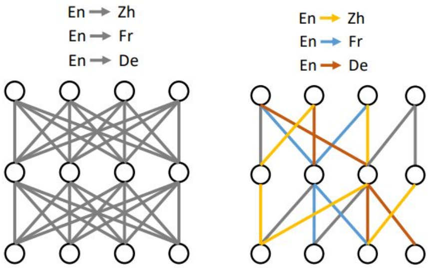
(2) partial sharing schemes

(e.g.,Purason & Tättar, 2022)



(a) Language group sharing **(b)** Language group and universal sharing (tiered) **(c)** Universal sharing (between all languages)

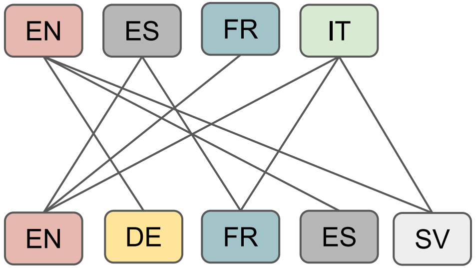
(e.g., Lin et al., 2021)



(a) Full network (b) LaSS

(3) no sharing

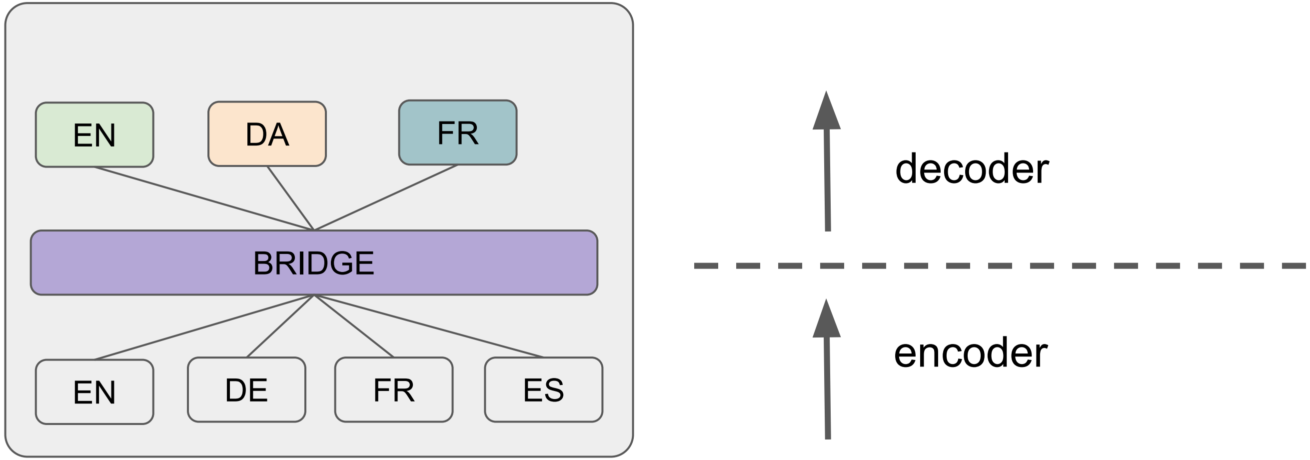
(e.g., Escolano et al., 2021)



https://github.com/Helsinki-NLP/Mammoth

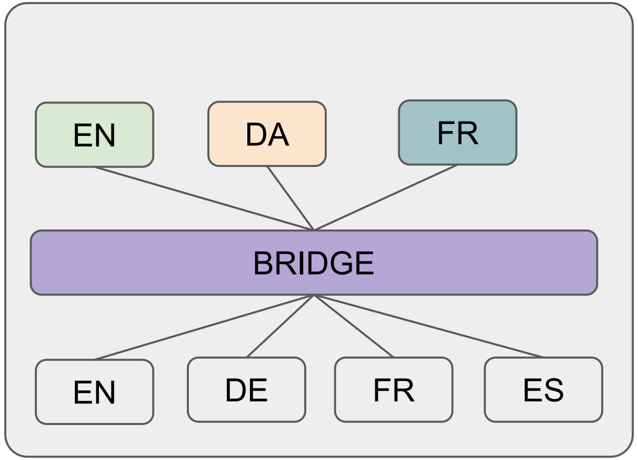
Support for bridges, adapters and hierarchical structures

Using an attention bridge

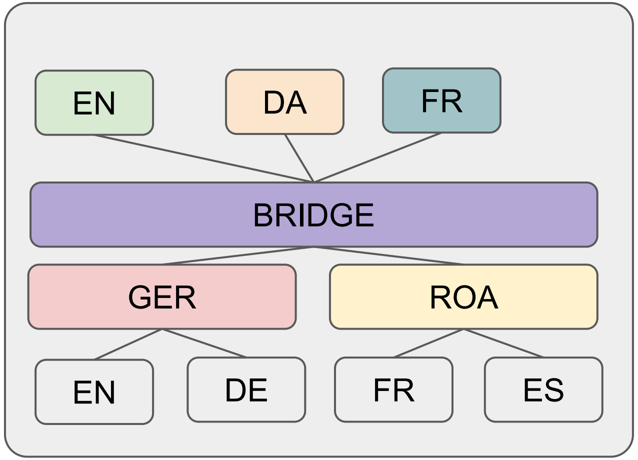


Support for bridges, adapters and hierarchical structures

Using an attention bridge



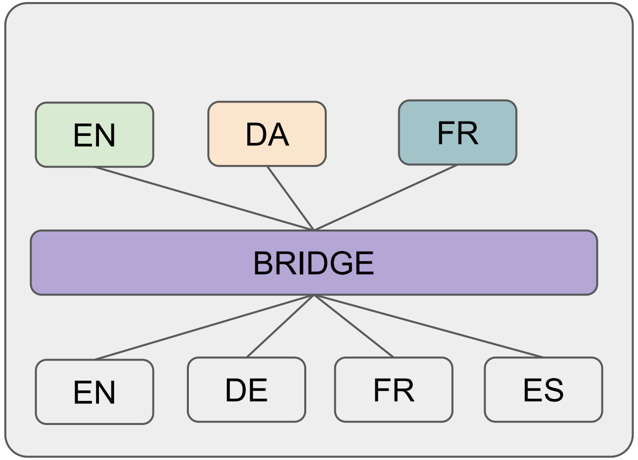
… with language groups



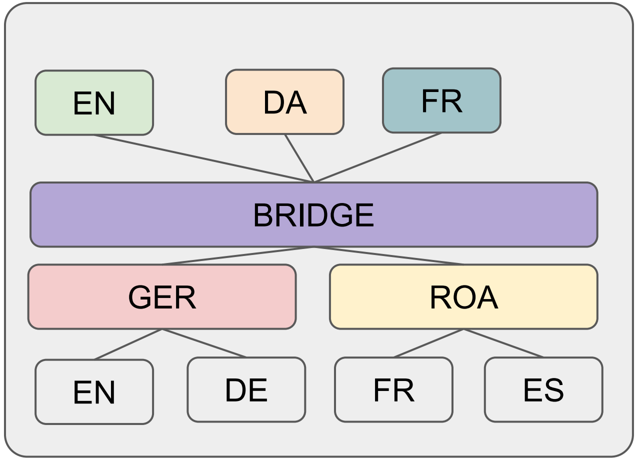
https://github.com/Helsinki-NLP/Mammoth

Support for bridges, adapters and hierarchical structures

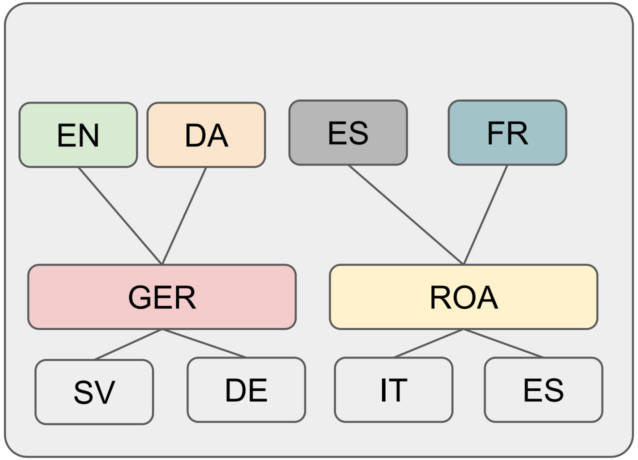
Using an attention bridge



… with language groups



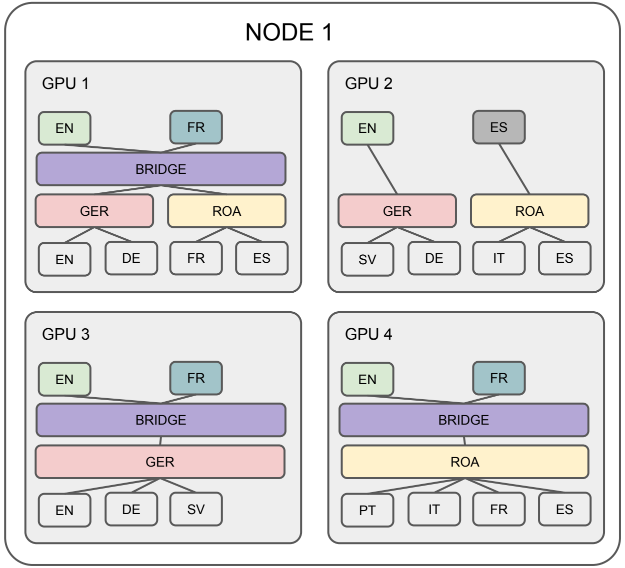
no bridging structure



https://github.com/Helsinki-NLP/Mammoth

Efficient training and resource allocation

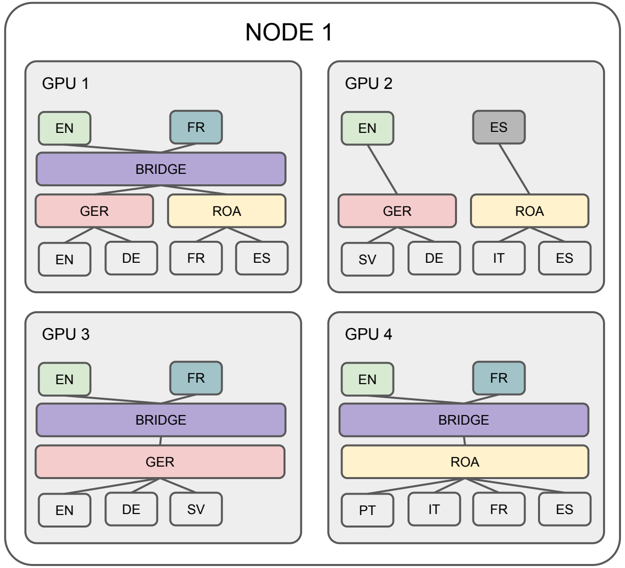
Custom model parallelism increases parameter sharing versatility



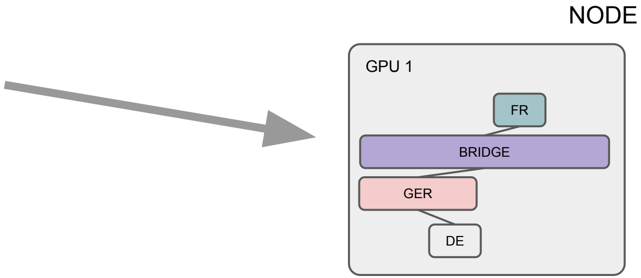
- Modules are synchronized in the GPUs where they are present:
  - AB layer synced in GPUs 1,3 & 4
  - Language-specific components synced as needed (e.g., EN-decoder in all GPUs)
  - Language group-specific components also synced as needed (e.g., GER in GPUs 1,2 & 3)
- Allocation tool: task2gpu

https://github.com/Helsinki-NLP/Mammoth

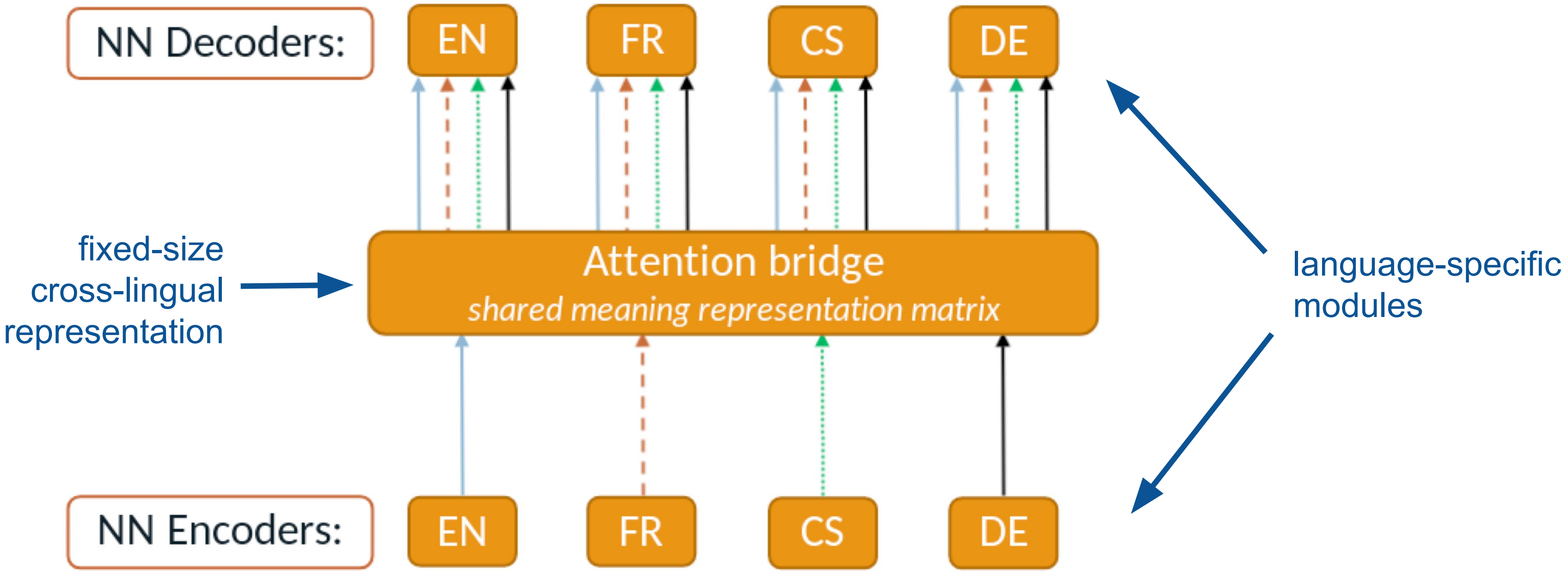
Reusability and inference efficiency through modularity



- All modules are saved independently
- Light inference, e.g., DE → FR only loads



Case study 1: The attention bridge model

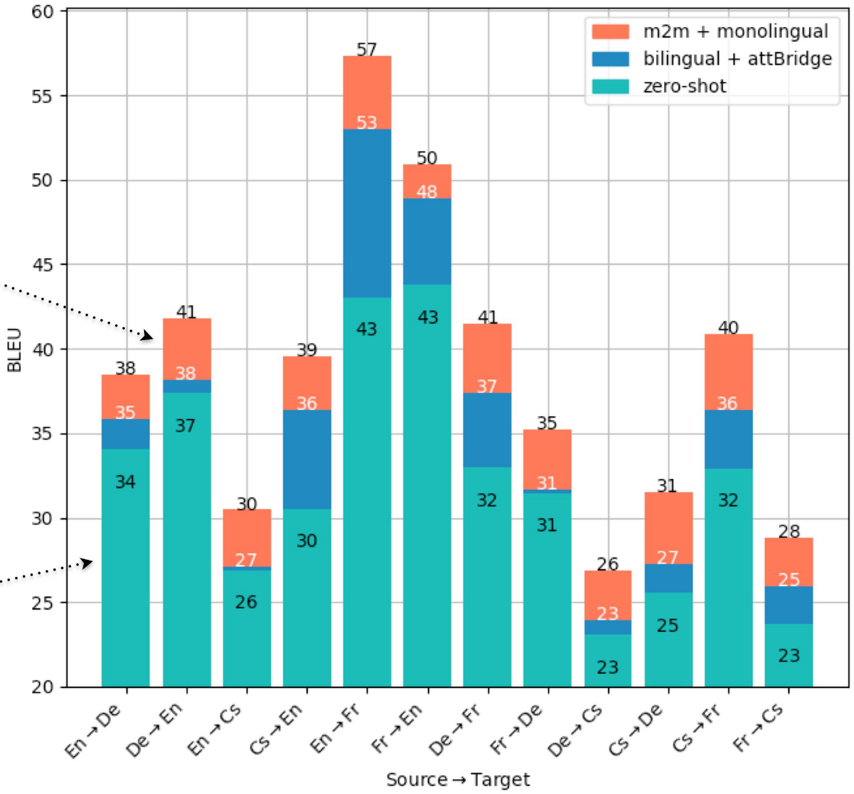


Case study 1: Transfer learning and zero shot

Translating image captions

model with
all languages

language pair
not seen in
training data
(zero shot)



Case study 1: Test with SentEval

Apply intermediate representation to

- 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

Case study 1: Test with SentEval

Apply intermediate representation to

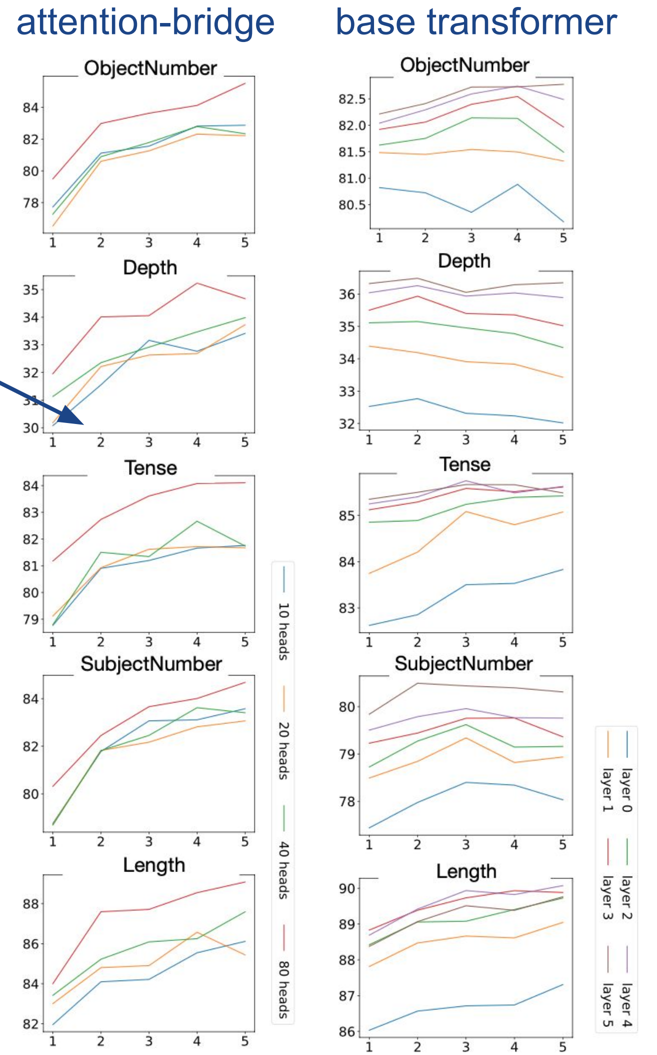
- downstream tasks
- linguistic probing tasks

number of languages

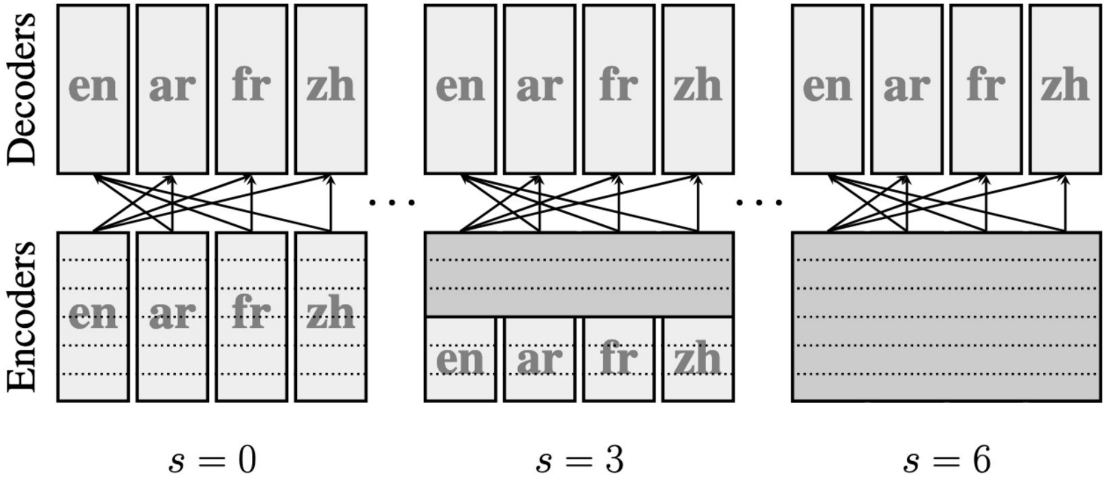
multilingual models

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



Case study 2: Partially shared encoder layers



What happens if we add more languages?

Subset selection:

- maximise the number of datapoints available for training
- the presence of zero-shot translation test sets
- the existence of XNLI data for the languages
- maximize language diversity
- always English-centric

OPUS
100

What happens if we add more languages?

Subset selection:

- maximise the number of datapoints available for training
- the presence of zero-shot translation test sets
- the existence of XNLI data for the languages
- maximize language diversity
- always English-centric

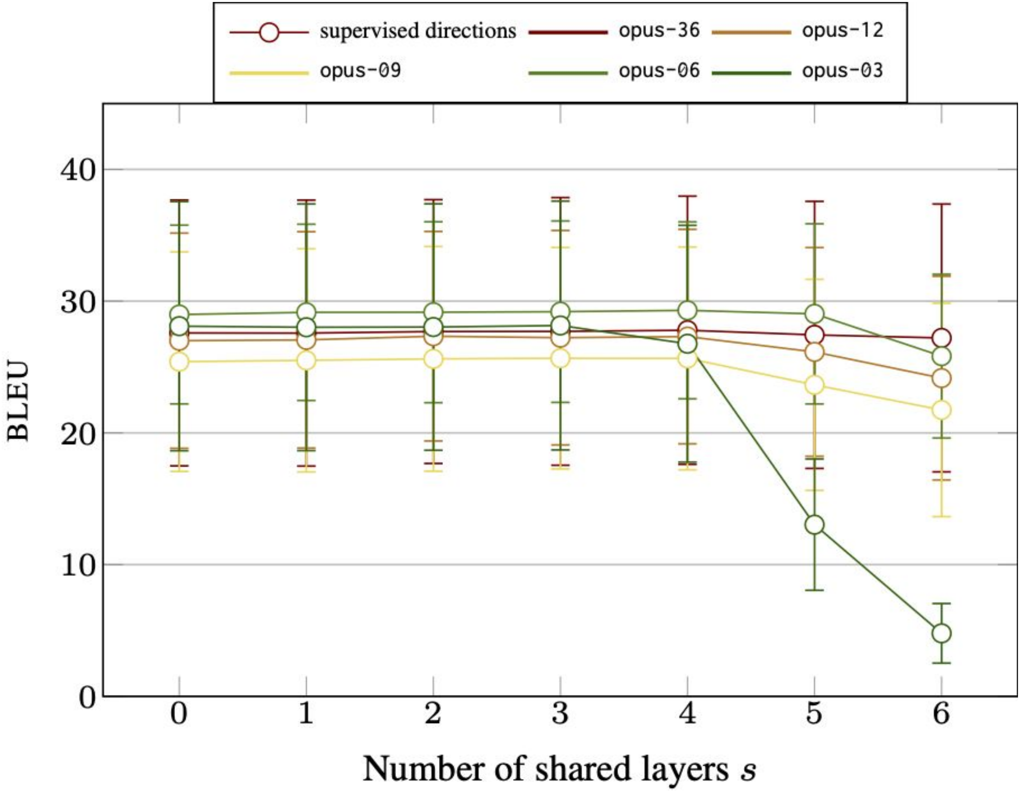
OPUS 100

ISO 2	Dataset	Train size	XNLI
ar	opus-03	1,000,000	✓
fr	opus-03	1,000,000	✓
zh	opus-03	1,000,000	✓
de	opus-06	1,000,000	✓
nl	opus-06	1,000,000	✓
ru	opus-06	1,000,000	✓
th	opus-09	1,000,000	✓
tr	opus-09	1,000,000	✓
vi	opus-09	1,000,000	✓
bg	opus-12	1,000,000	✓
el	opus-12	1,000,000	✓
es	opus-12	1,000,000	✓

bn	bs	opus-36	1,000,000	–
eu	cs	opus-36	1,000,000	–
fa	et	opus-36	1,000,000	–
fi	hu	opus-36	1,000,000	–
he	is	opus-36	1,000,000	–
id	lt	opus-36	1,000,000	–
it	mt	opus-36	1,000,000	–
ja	ro	opus-36	1,000,000	–
ko	sk	opus-36	1,000,000	–
lv	sq	opus-36	1,000,000	–
mk	sr	opus-36	1,000,000	–
sv	uk	opus-36	1,000,000	–

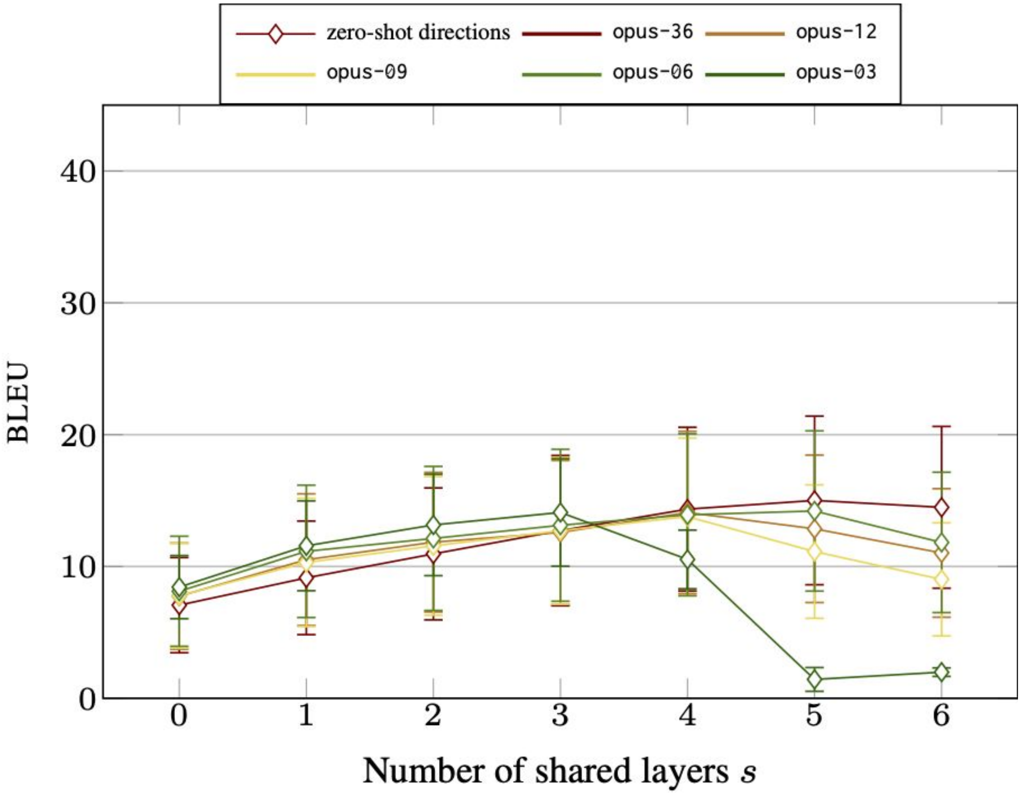
Task Fitness: Translation

The effect of parameter sharing (supervised)



The effect of parameter sharing (zero-shot)

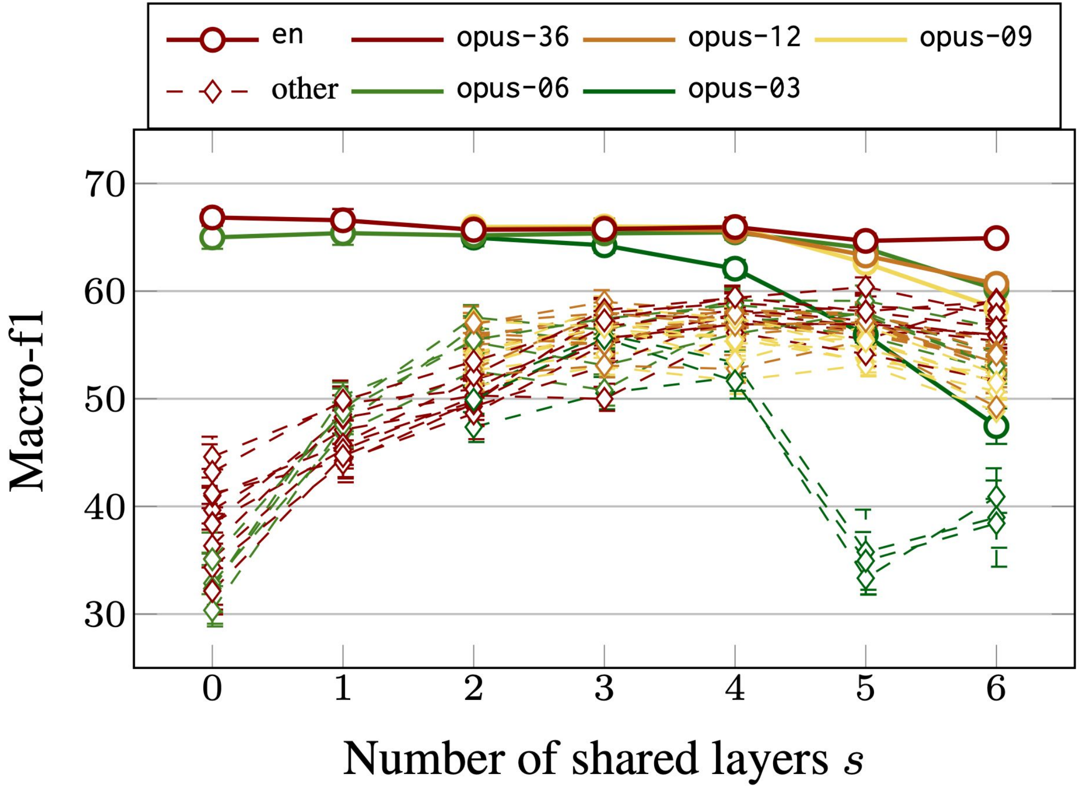
testing language pairs not seen during training



Language Independence:

Testing Cross-Lingual NLI (XNLI)

The effect of parameter sharing (average XNLI scores)



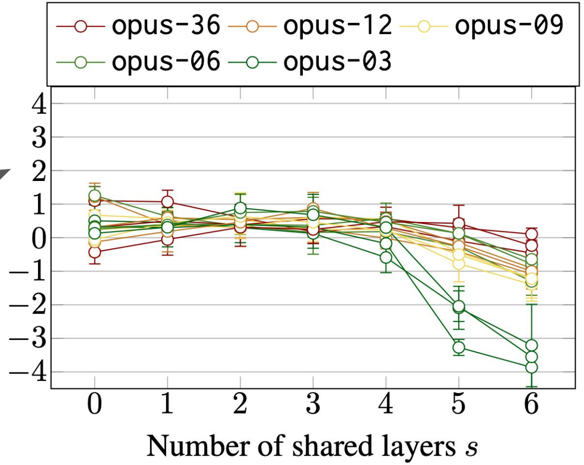
Semantic Content: NLU Benchmarks

		Dataset	Task	Size
ALUE	ar	NSURL-2019 Task 8	question similarity	10,797
		OSACT4 Task-A	offensive speech detection	6,839
		OSACT4 Task-B	hate speech detection	6,839
GLUE	en	COLA	linguistic acceptability	8,551
		MRPC	sentence similarity	3,668
		QNLI	NLI	104,743
		QQP	question similarity	363,846
FLUE	fr	PAWSX	paraphrase detection	49,399
		STSB	paraphrase detection	5,749
		XNLI	NLI	392,702
CLUE	zh	AFQMC	question similarity	34,334
		CMNLI	NLI	391,783
		TNEWS	news topic classification	53,360

The effect of parameter sharing

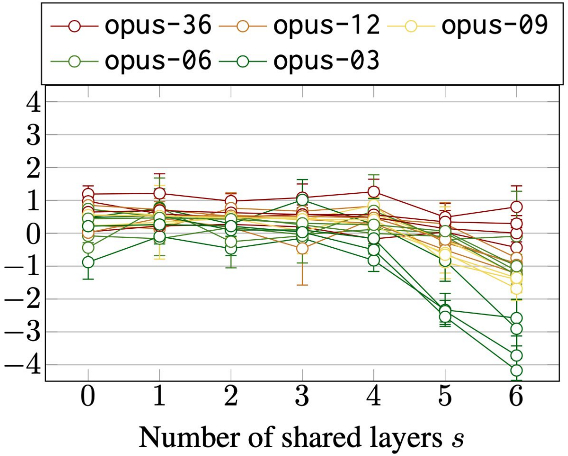
(z-scaled averaged macro-F1 scores)

French

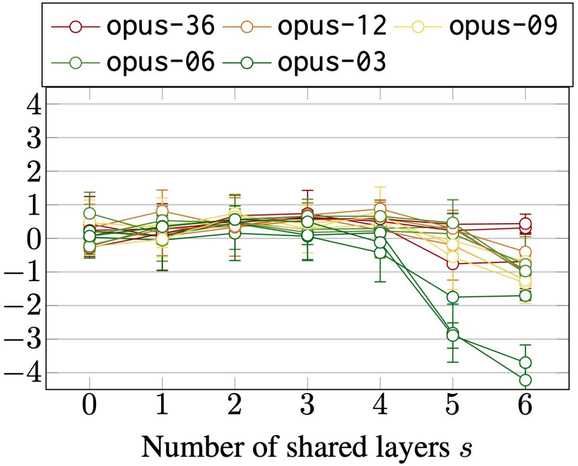


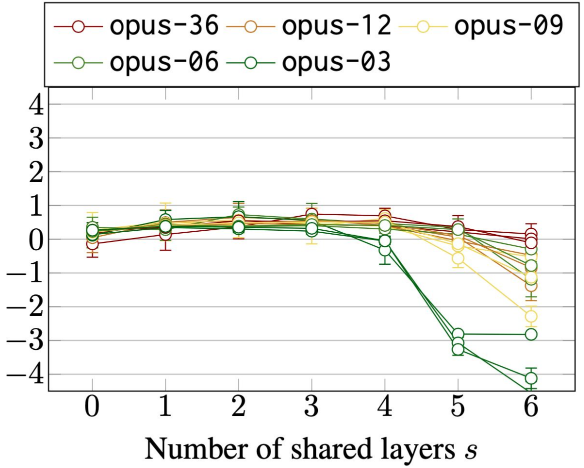
Chinese

English



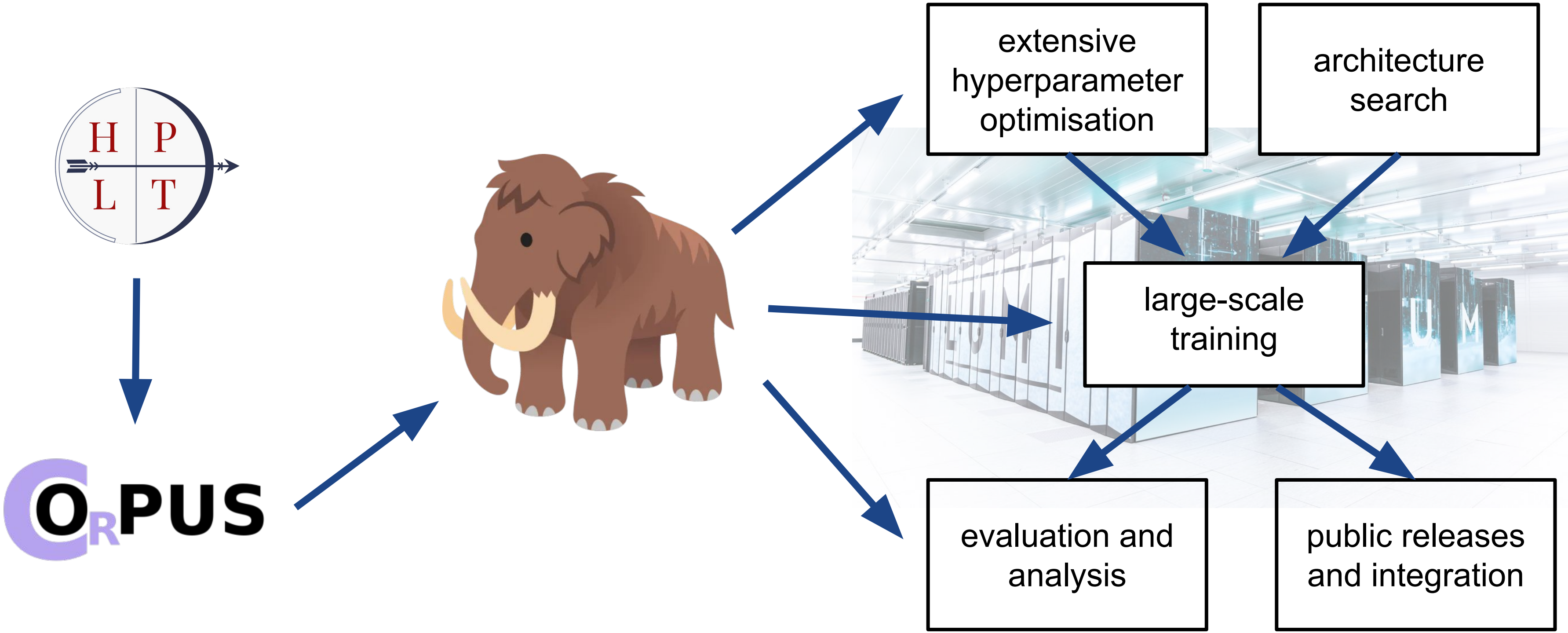
Arabic





What is next?

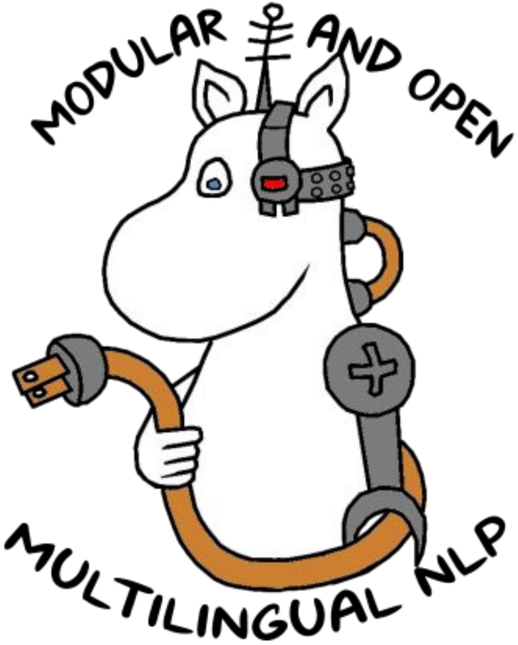
Building a MAMMOTH flagship model



Upcoming EACL workshop: **MOOMIN**

FOTRAN Found in Translation

GreenNLP



Focus on

- Scalability and Language Coverage
- Efficiency and Re-usability

Submit papers on topics like

- mixture of expert models and gated routing
- modular pre-training of multilingual language and translation models
- effective transfer with modular architectures such as adapters and hypernetworks
- efficient parallelization and distribution of modular model training
- modular frameworks and architecture implementations
- massively multilingual models with large language coverage
- subnet selection and pruning
- modular distillation
- modular extensions of existing NLP models systems, especially in low-resource settings and for low-resource languages
- evaluation of modular systems in terms of performance, efficiency, and computational costs
- platforms for distributing, sharing, and integrating NLP components

https://moomin-workshop.github.io/

Summing up: What to remember from this talk

OPUS

- Is a huge and very interesting resource not only for MT research
- The OPUS ecosystem is much more than just data
- Please contribute to make it even more useful

https://opus.nlpl.eu/

Summing up: What to remember from this talk

OPUS

- Is a huge and very interesting resource not only for MT research
- The OPUS ecosystem is much more than just data
- Please contribute to make it even more useful

MAMMOTH

- Flexible framework for building modular multilingual NLP
- Scalable training and efficient light-weight inference
- Reuse and contributions welcome

https://github.com/Helsinki-NLP/Mammoth

Thank you! Any question?

HELSINKI NLP

H P L T

FOTRAN
Found in Translation

erc

ACADEMY OF FINLAND

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

NVIDIA

CSC
ICT Solutions for Brilliant Minds

EUROPEAN LANGUAGE GRID