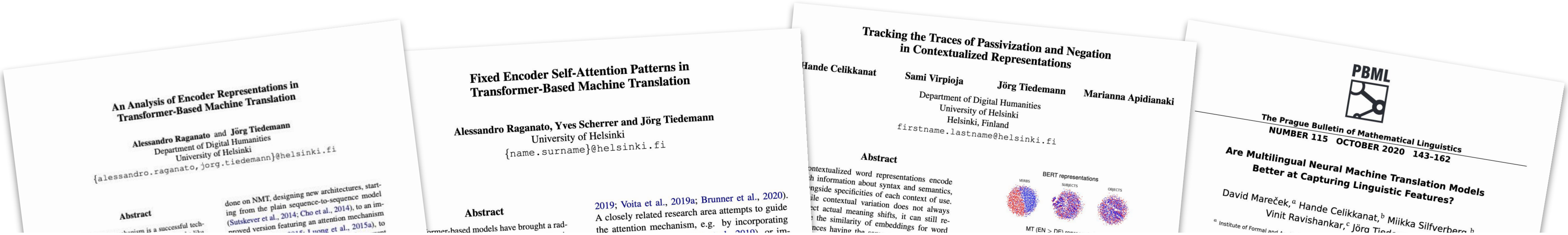
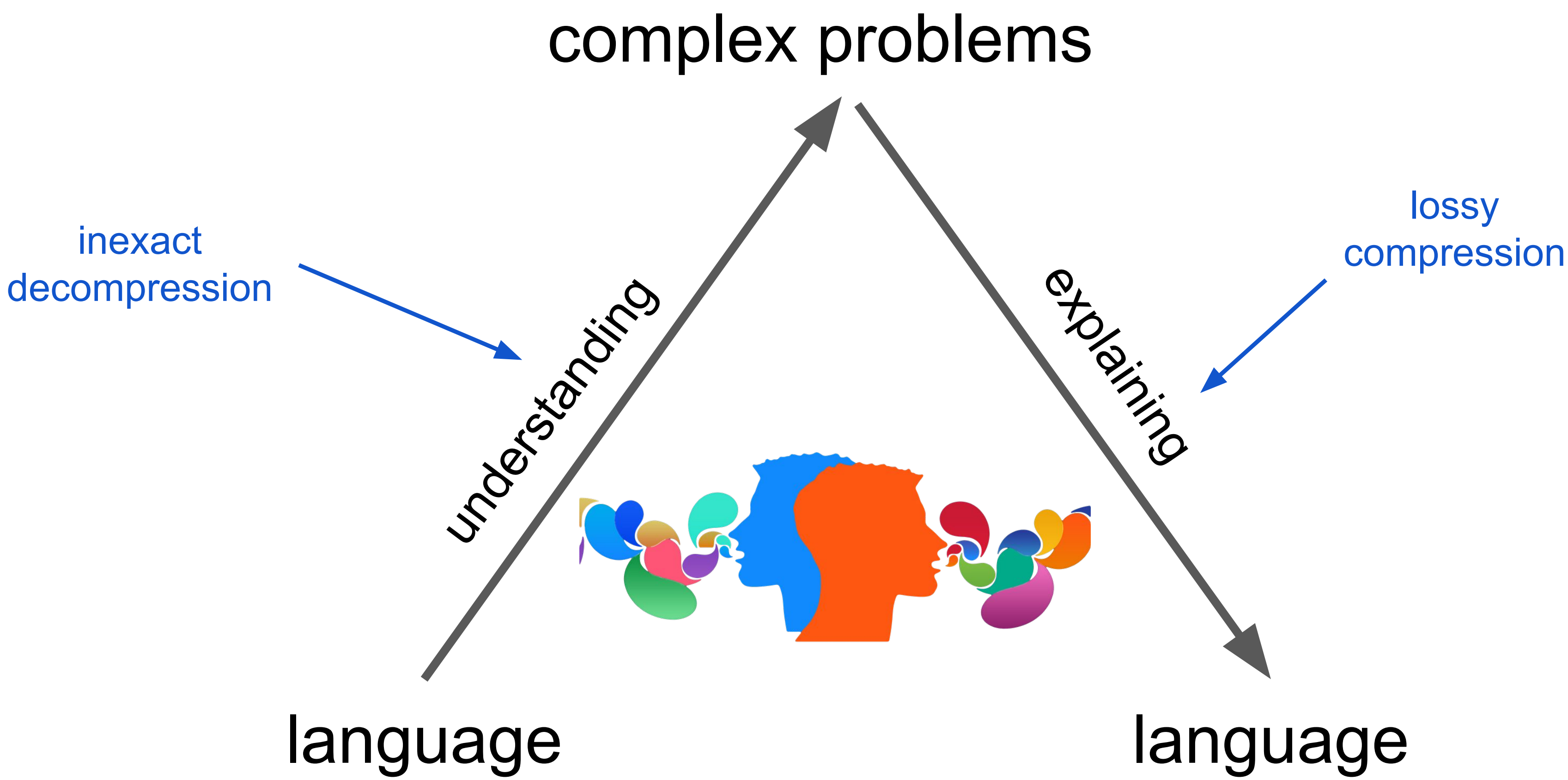
Jörg Tiedemann
Department of Digital Humanities
University of Helsinki

What's in a translation model?

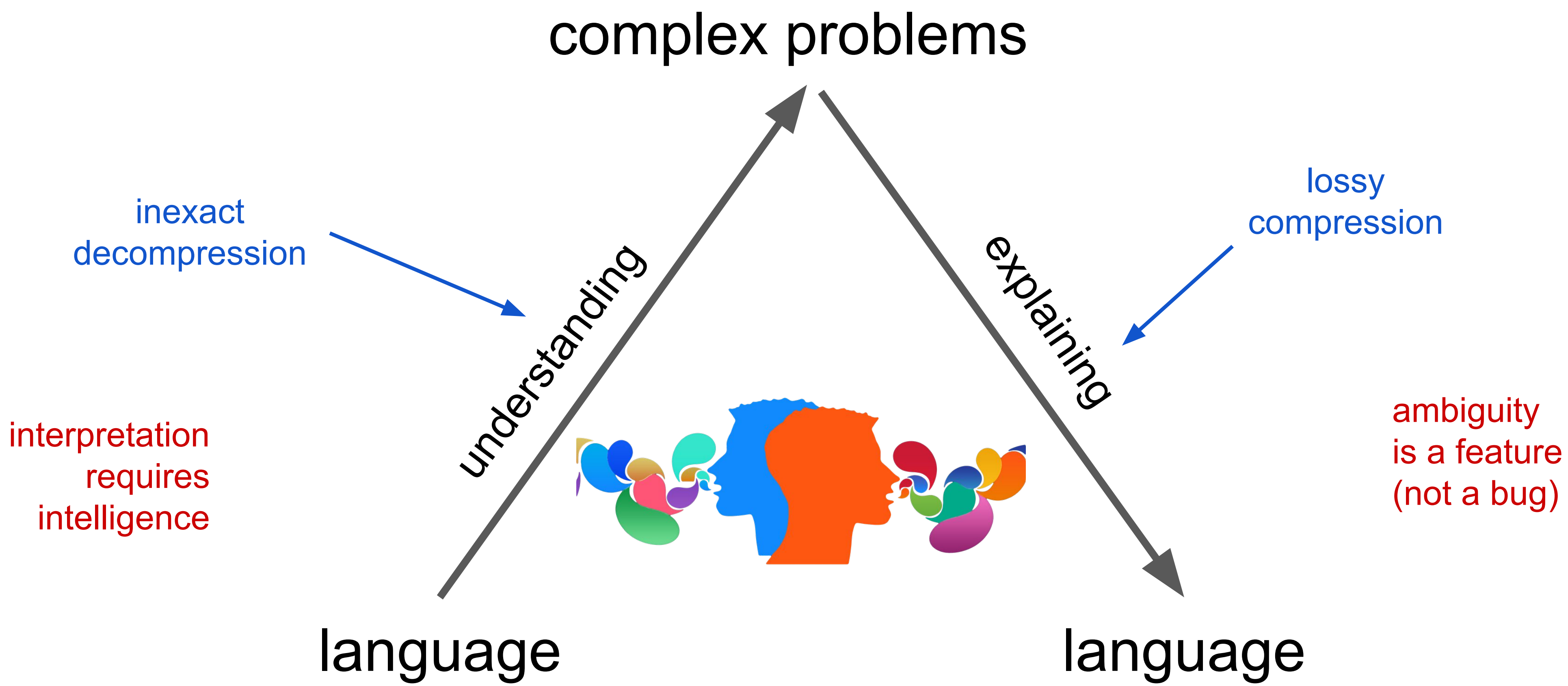
Analyzing neural seq2seq models and the representations they learn



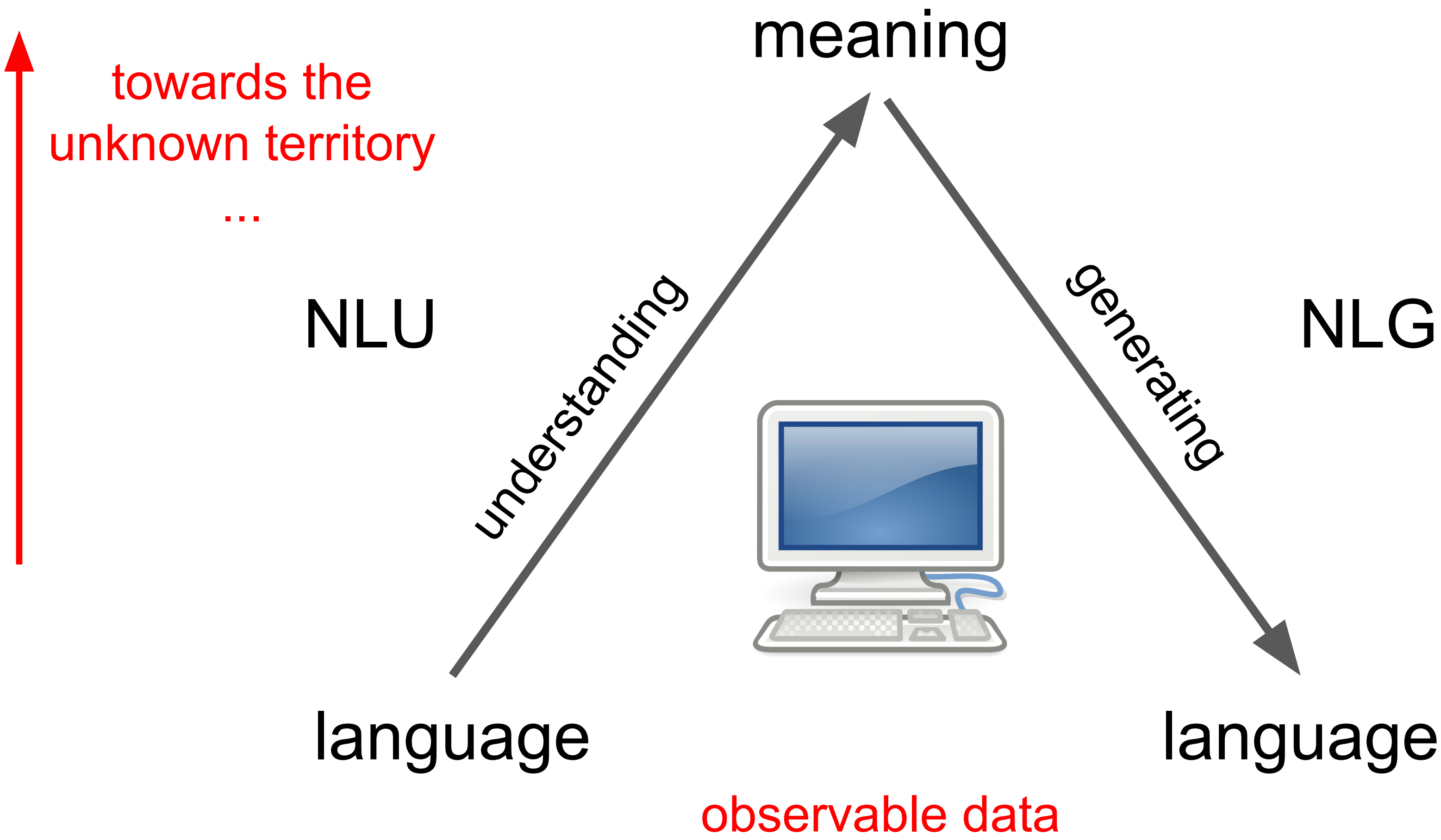
Language, communication and intelligence



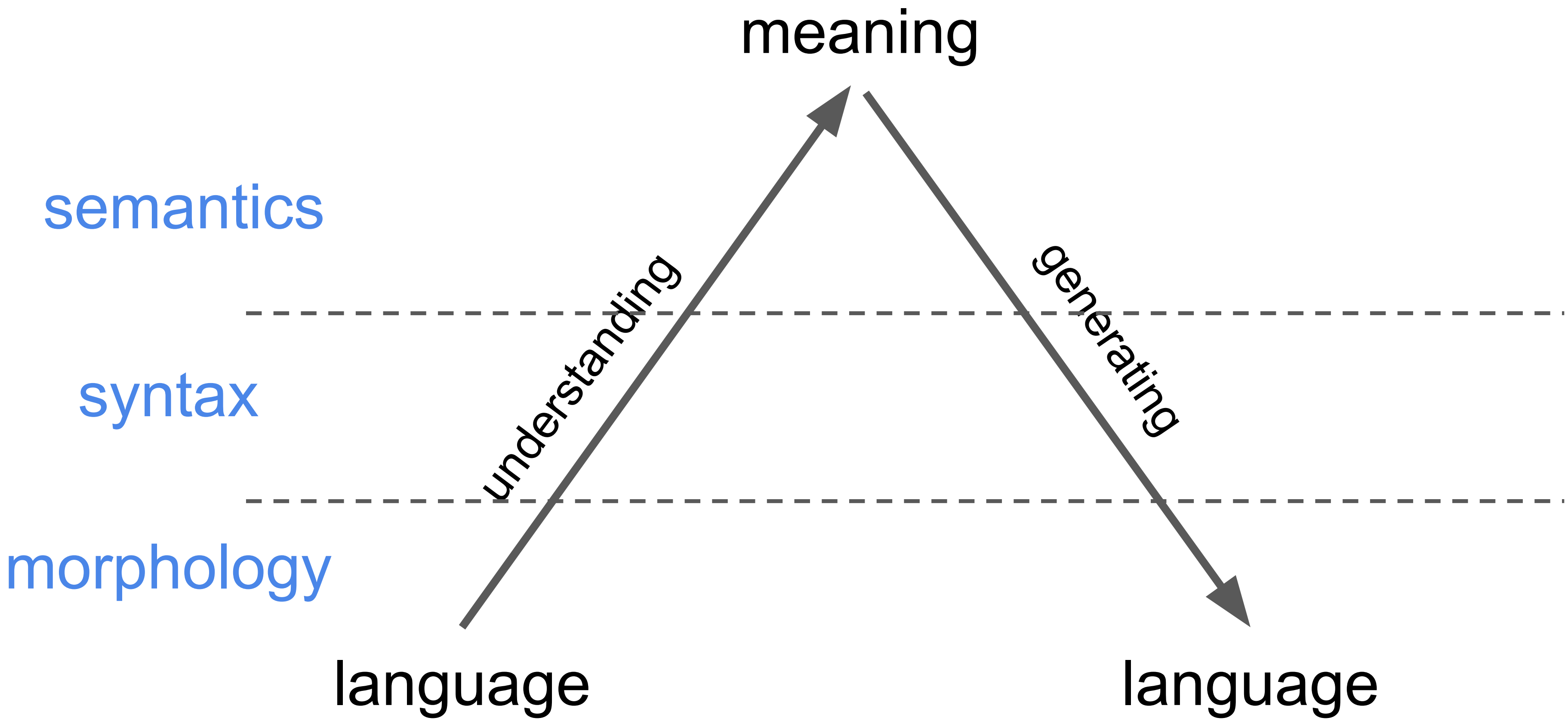
Language, communication and intelligence



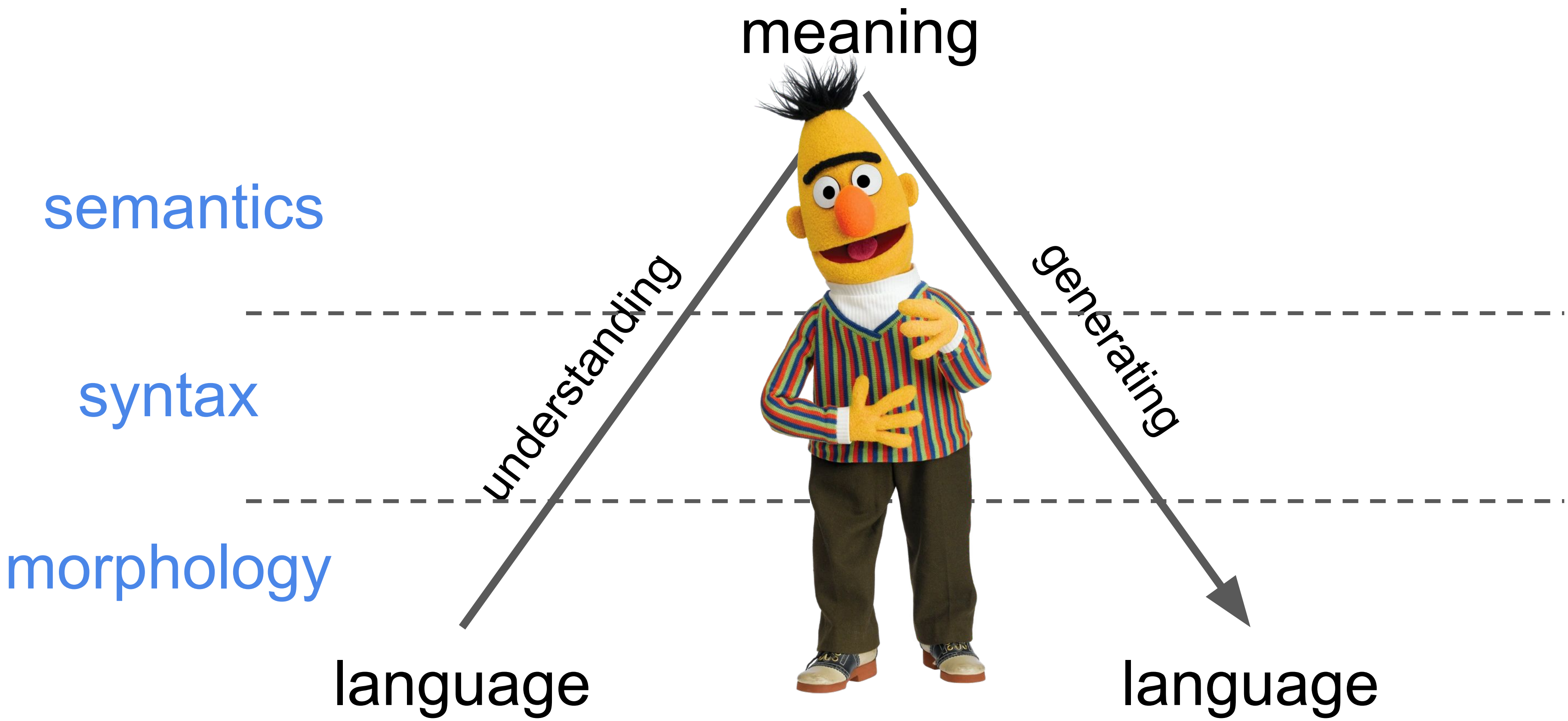
What is language technology?



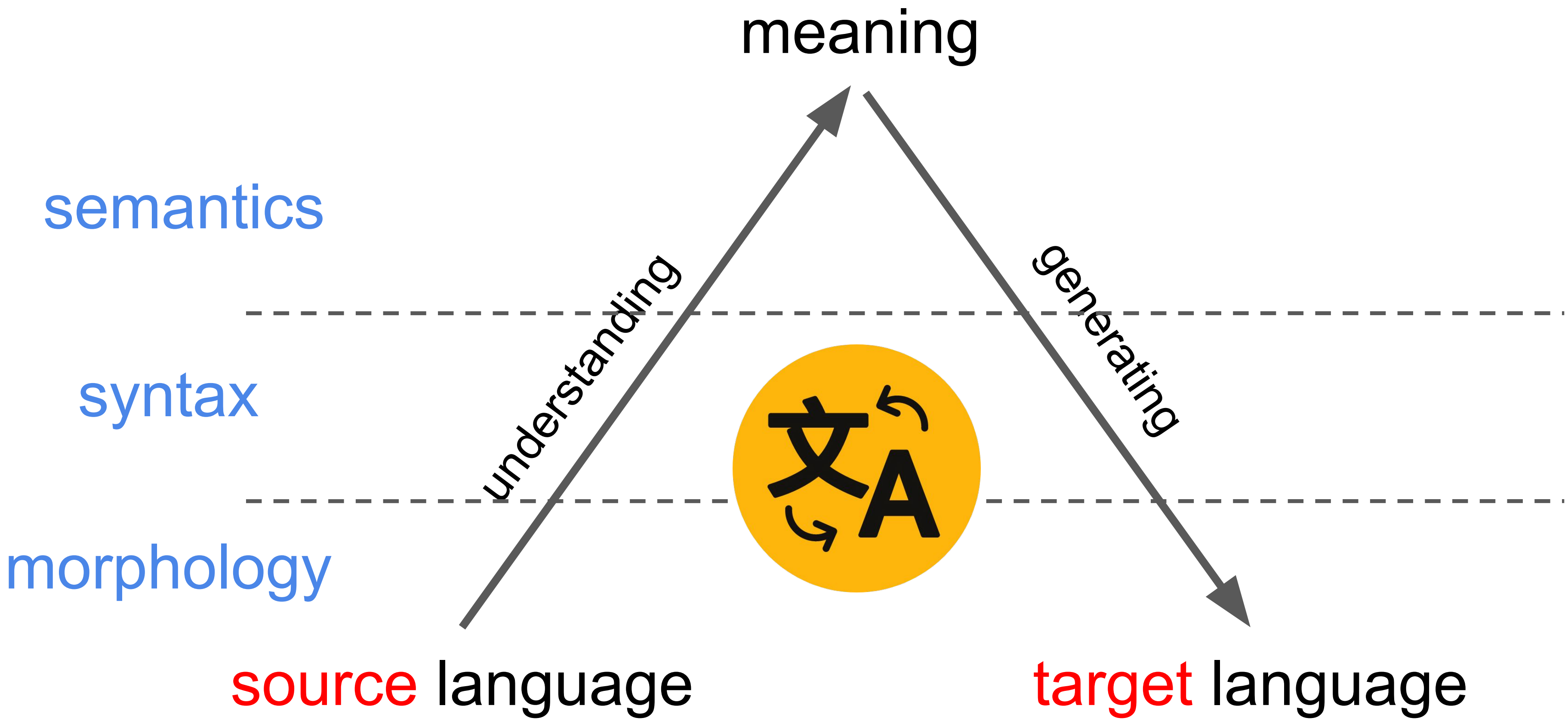
What is language technology?



Bertology: What does my language model learn?

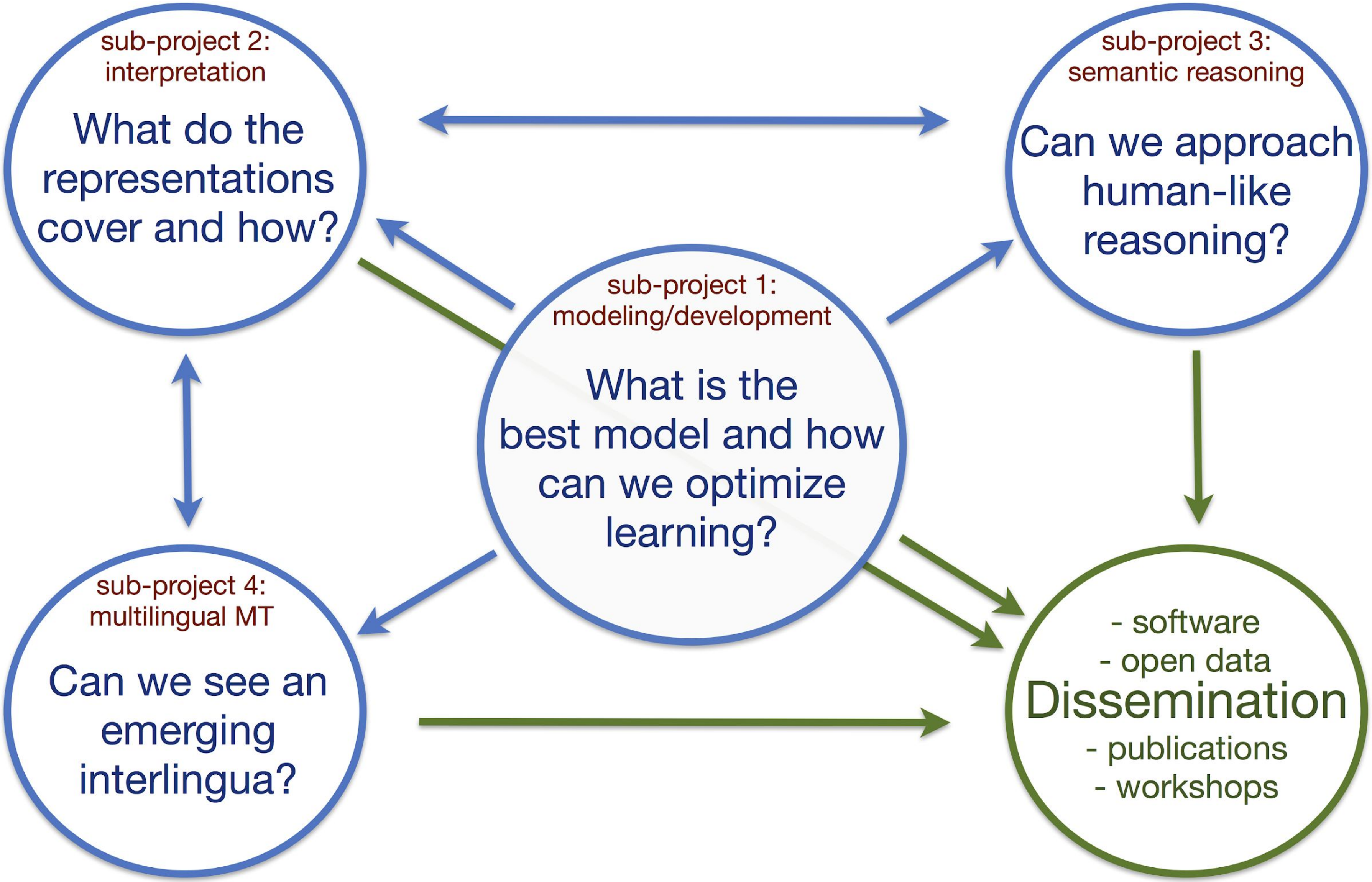


Machine translation: Naturally combine NLU and NLG



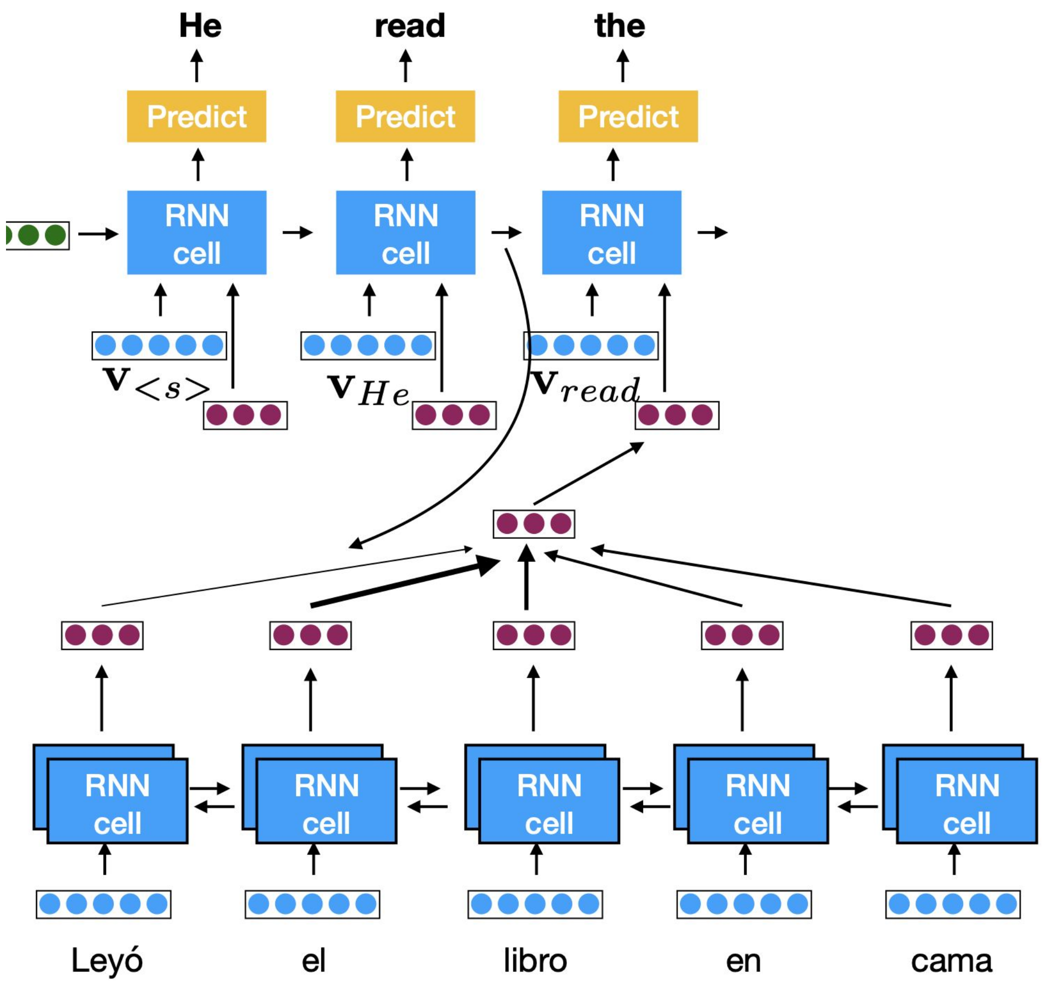
FOTRAN

Found in Translation

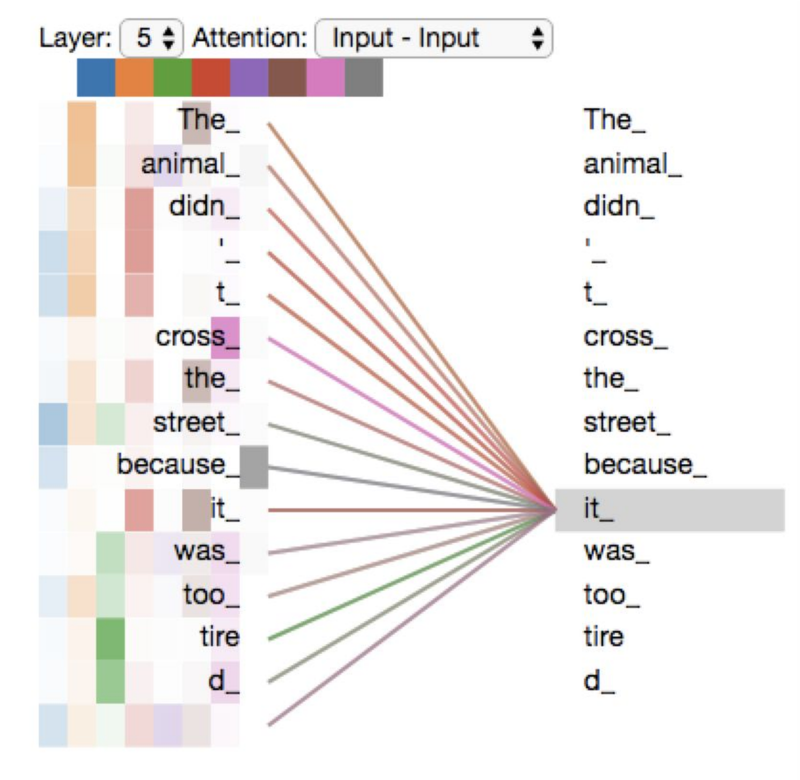

sub-project 2: interpretation
What do the representations cover and how?
sub-project 3: semantic reasoning
Can we approach human-like reasoning?
sub-project 1: modeling/development
What is the best model and how can we optimize learning?
sub-project 4: multilingual MT
Can we see an emerging interlingua?
- software
- open data
Dissemination
- publications
- workshops

Translation Models

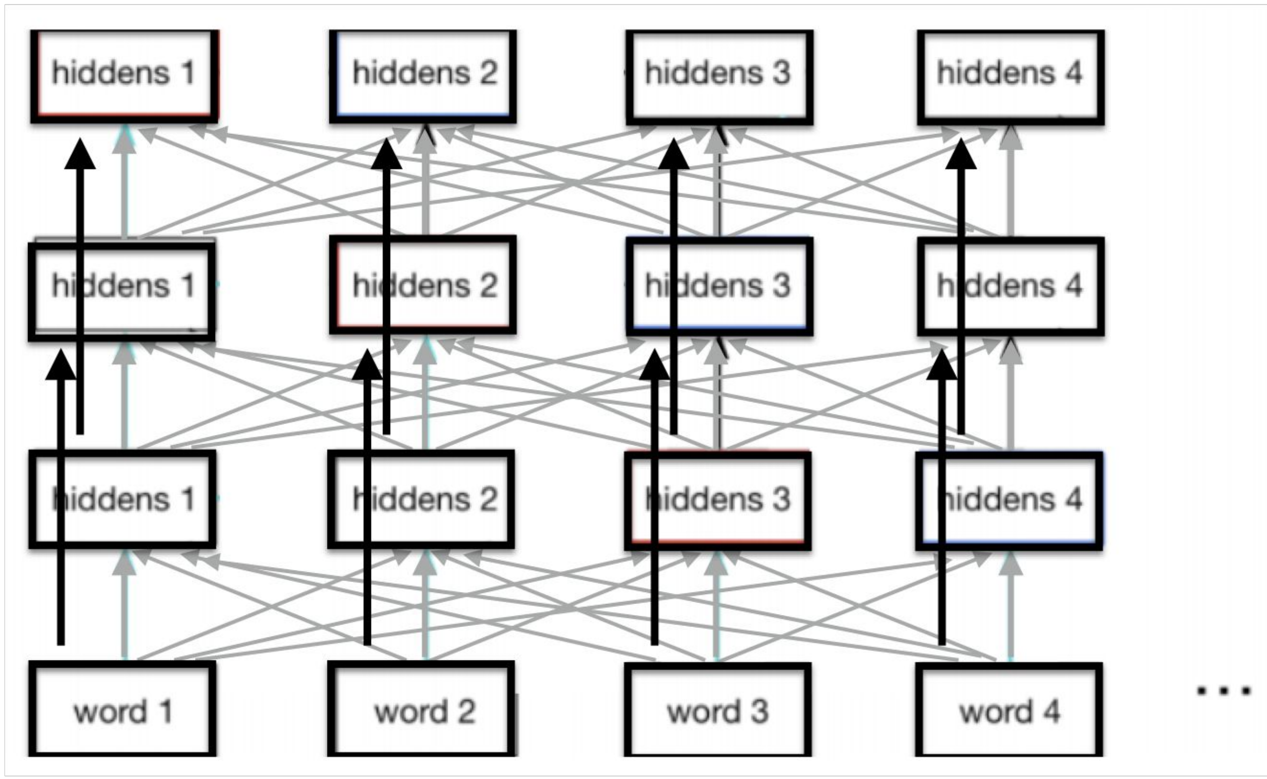
Recurrent sequence-to-sequence models with attention



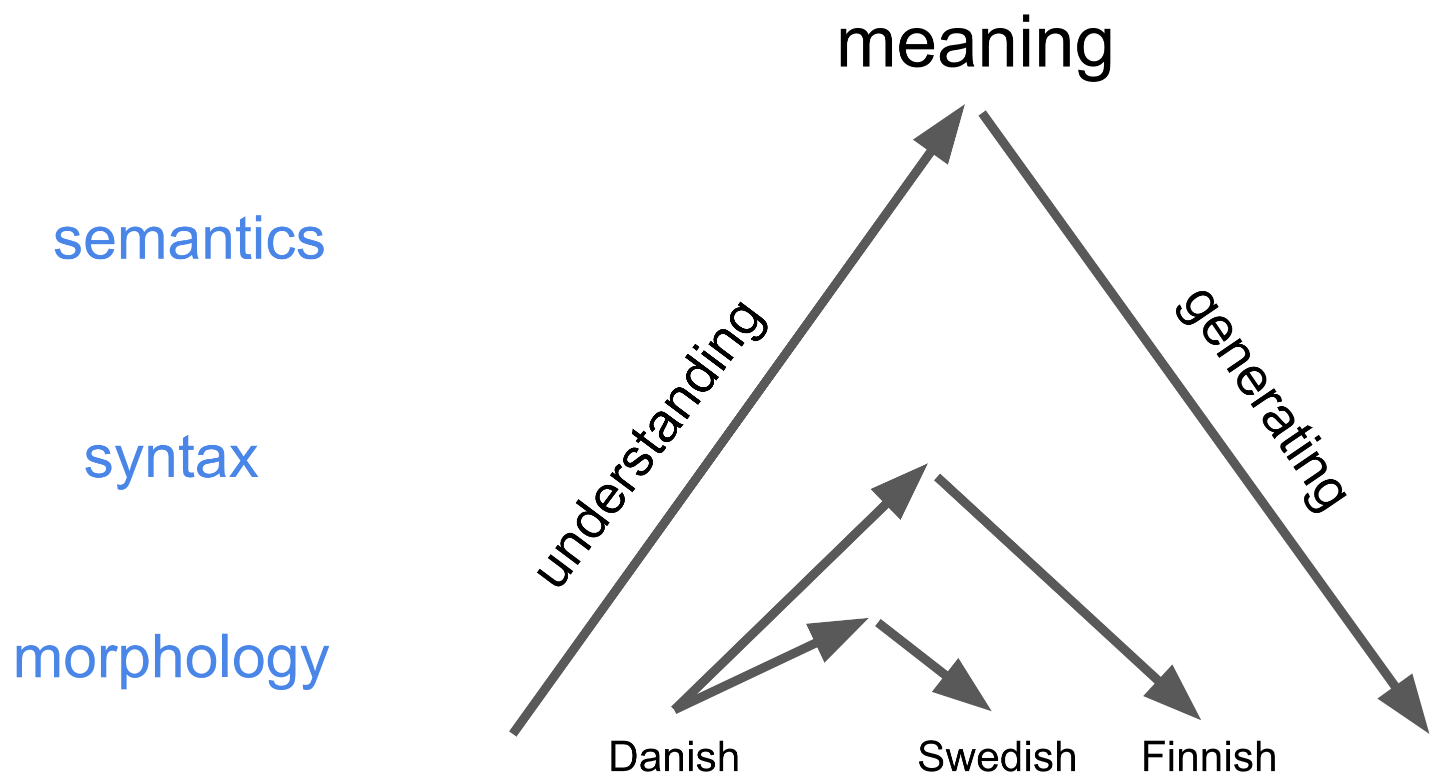
Transformer-based encoders and decoders



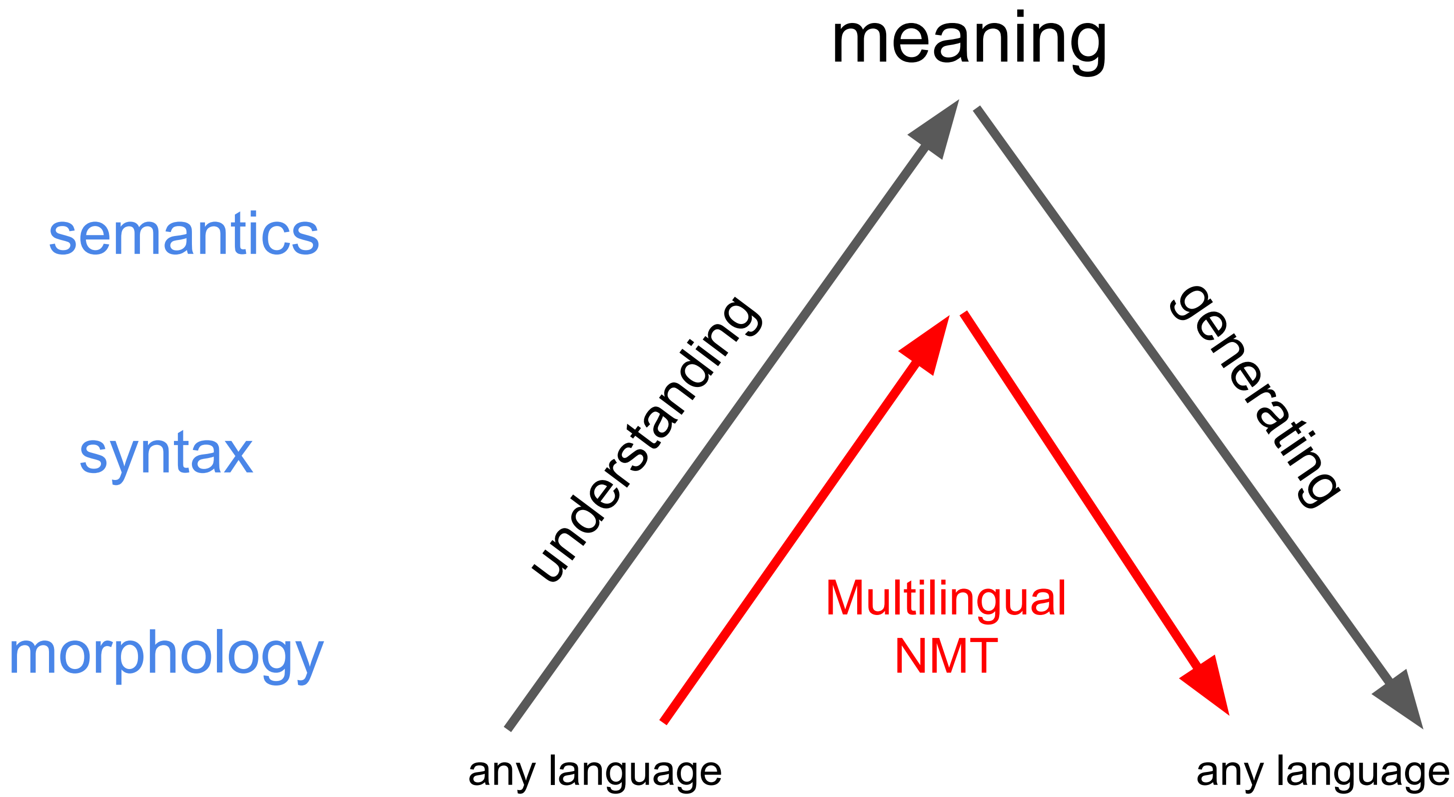
http://jalammar.github.io/illustrated-transformer/



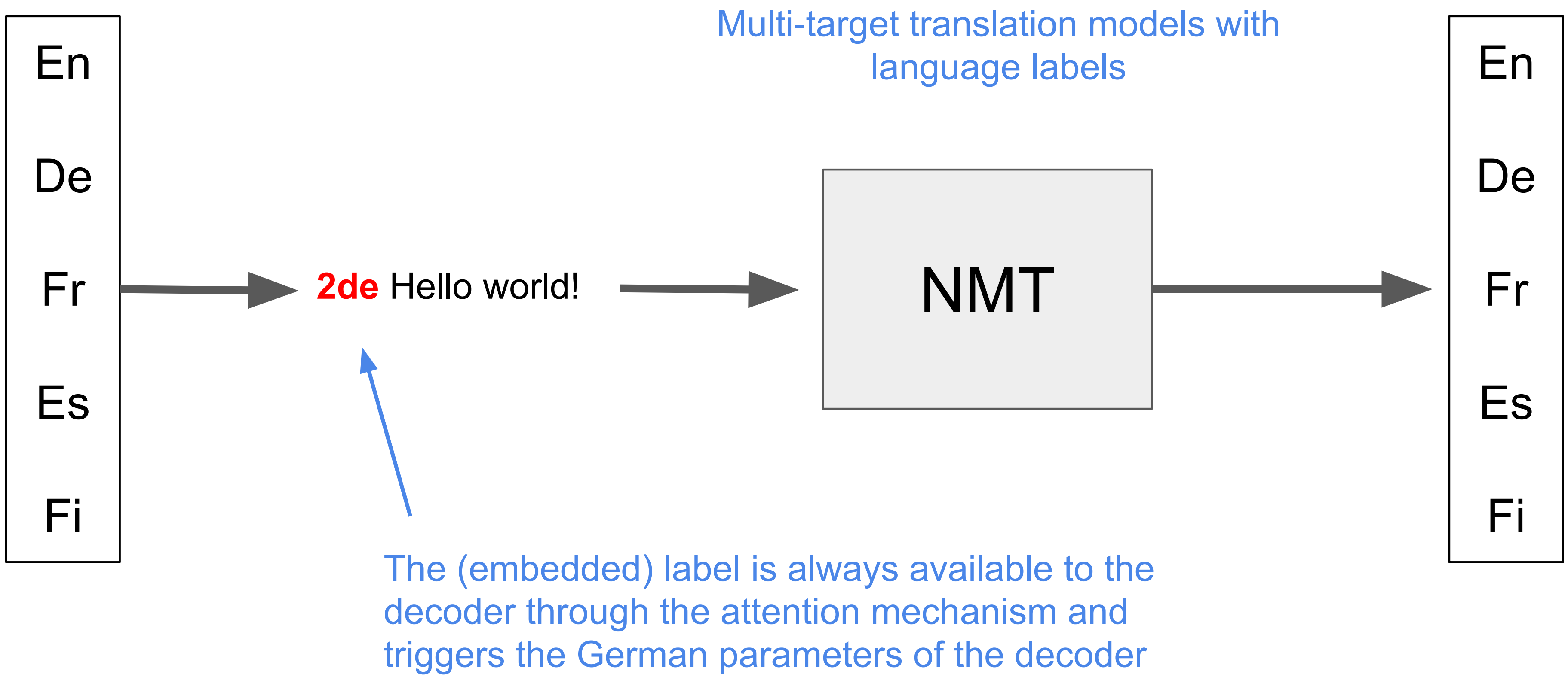
How can we force MT to really learn the semantics?



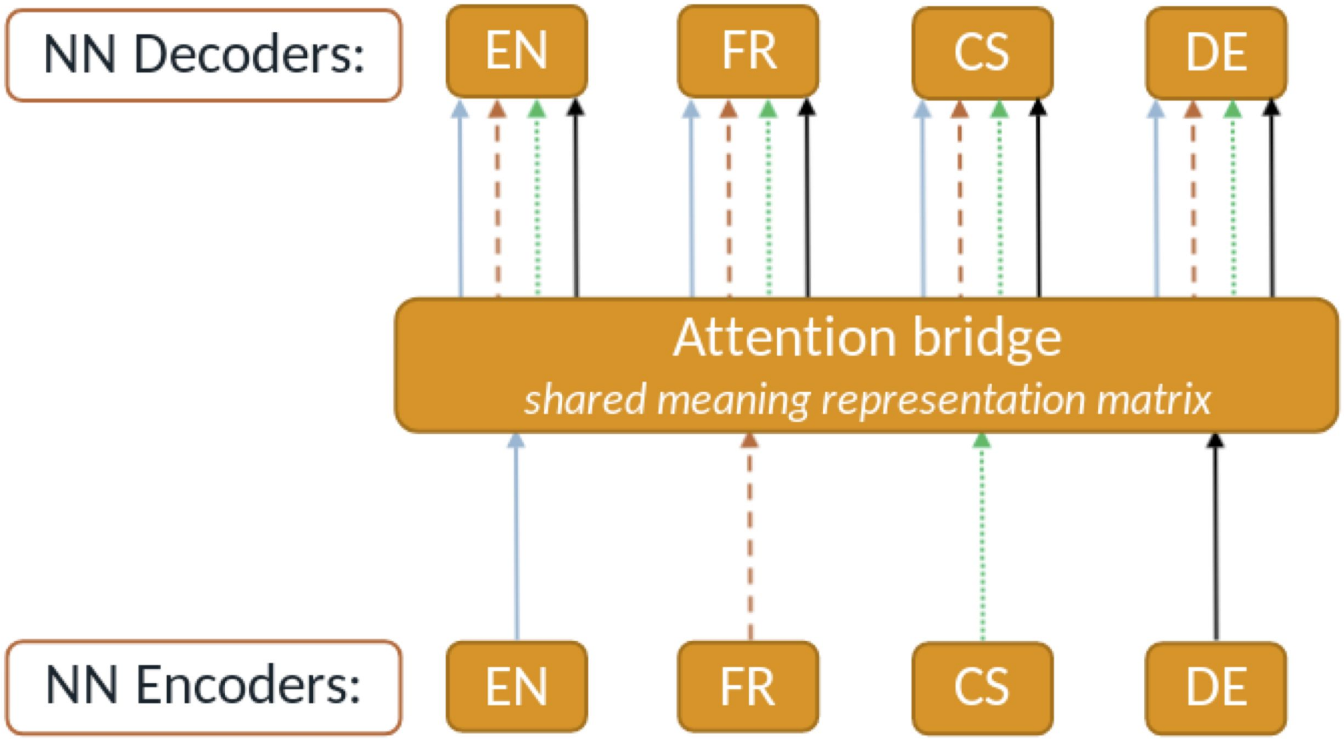
How can we force MT to really learn the semantics?



(1) Language labels and completely shared parameters



(2) Language-specific components



Multilingual NMT and language embeddings

Emerging Language Spaces Learned From Massively Multilingual Corpora

Jörg Tiedemann

University of Helsinki
jorg.tiedemann AT helsinki.fi

Abstract. Translations capture important information about languages that can be used as implicit supervision in learning linguistic properties and semantic representations. In an information-centric view, translated texts may be considered as semantic mirrors of the original text and the significant variations that we can observe across various languages can be used to disambiguate a given expression using the linguistic signal that is grounded in translation. Parallel corpora consisting of massive amounts of human translations with a large linguistic variation can be applied to increase abstractions and we propose the use of highly multilingual machine translation models to find language-independent meaning representations. Our initial experiments show that neural machine translation models can indeed learn in such a setup and we can show that the learning algorithm picks up information about the relation between languages in order to optimize transfer leaning with shared parameters. The model creates a continuous language space that represents relationships in terms of geometric distances, which we can visualize to illustrate how languages cluster according to language families and groups. Does this open the door for new ideas of data-driven language typology with ...odels and techniques in empirical cross-linguistic research?

Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks

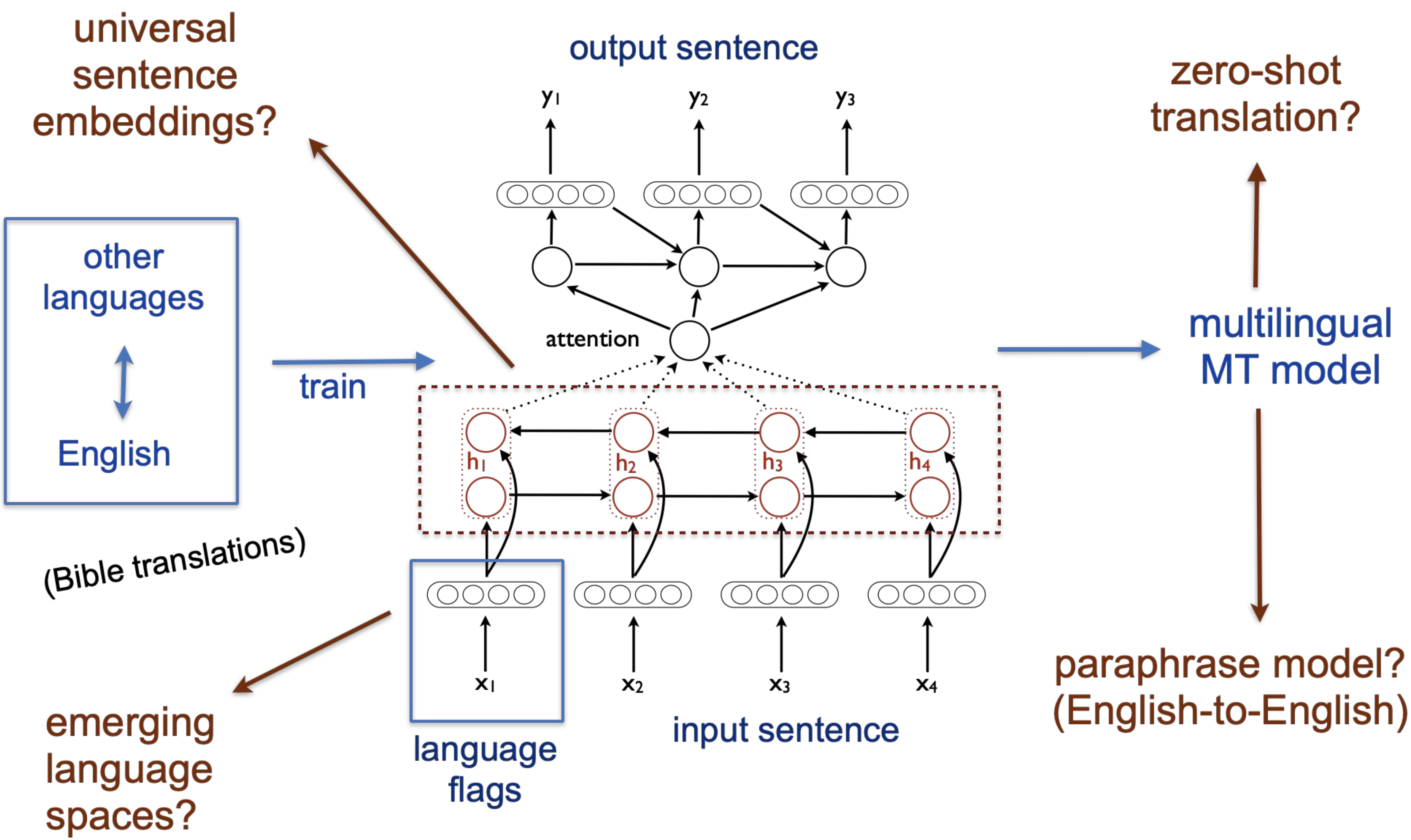
Jörg Tiedemann and **Yves Scherrer**
Department of Digital Humanities / HELDIG
University of Helsinki

Abstract

In this paper, we investigate whether multilingual neural translation models learn stronger semantic abstractions of sentences than bilingual ones. We test this hypotheses by measuring the perplexity of such models when applied to paraphrases of the source language. The intuition is that an encoder produces better representations if a decoder is capable of recognizing synonymous sentences in the same language even though the model is never trained for that task. In our setup, we add 16 different auxiliary languages to a bidirectional bilingual baseline model (English-French) and test it with in-domain and out-of-domain paraphrases in English. The results show that the perplexity is significantly reduced in each of the cases, indicating that meaning can be grounded in translation. This is further sup-

representations learned from multilingual data sets covering a larger linguistic diversity better reflect semantics than representations learned from less diverse material. This hypothesis is supported by the findings of related work focusing on universal sentence representation learning from multilingual data (Artetxe and Schwenk, 2018; Artetxe and Schwenk, 2018; Schwenk and Douze, 2017) to be used in natural language inference or other downstream tasks. In contrast to related work, we are not interested in fixed-size sentence representations that can be fed into external classifiers or regression models. Instead, we would like to fully explore the use of the encoded information in the attentive recurrent layers as they are produced by the seq2seq model.

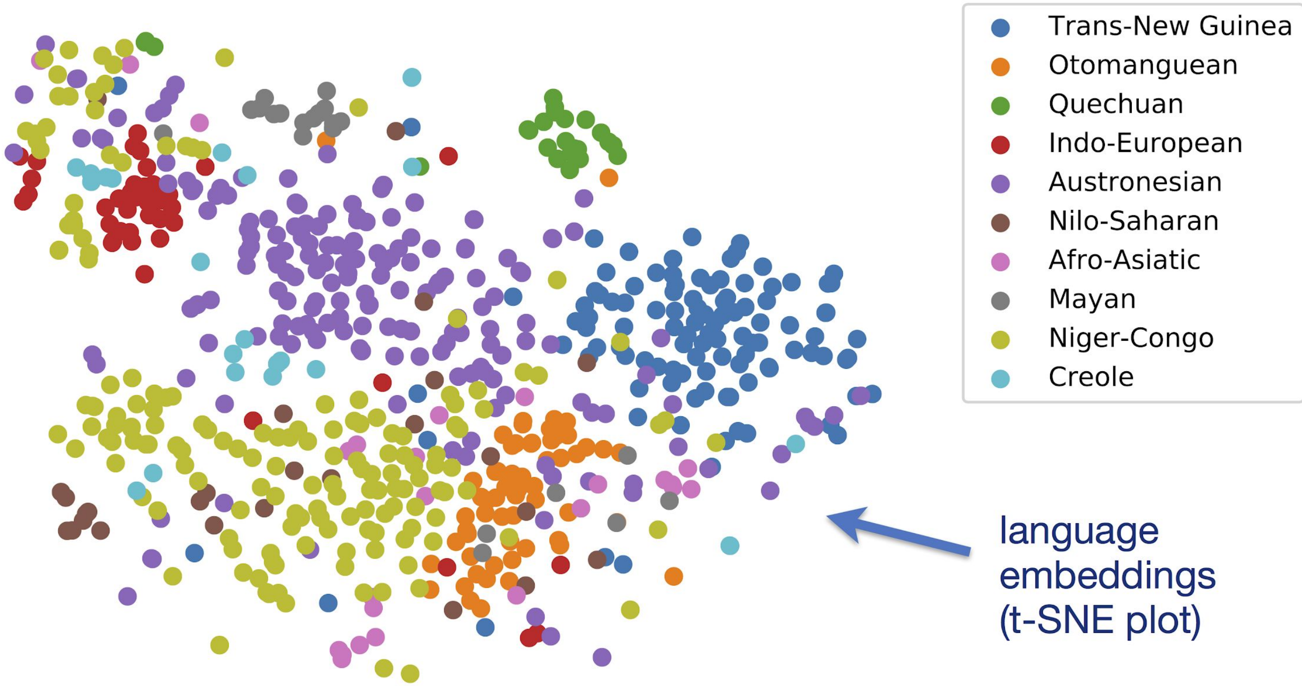
Our basic framework consist... attentional seq...



universal sentence embeddings?
output sentence
y1
y2
y3
zero-shot translation?
other languages
English
train
attention
multilingual MT model
h1
h2
h3
h4
(Bible translations)
x1
x2
x3
x4
paraphrase model? (English-to-English)
emerging language spaces?
language flags
input sentence

Emerging language spaces

Rough clusters of language families

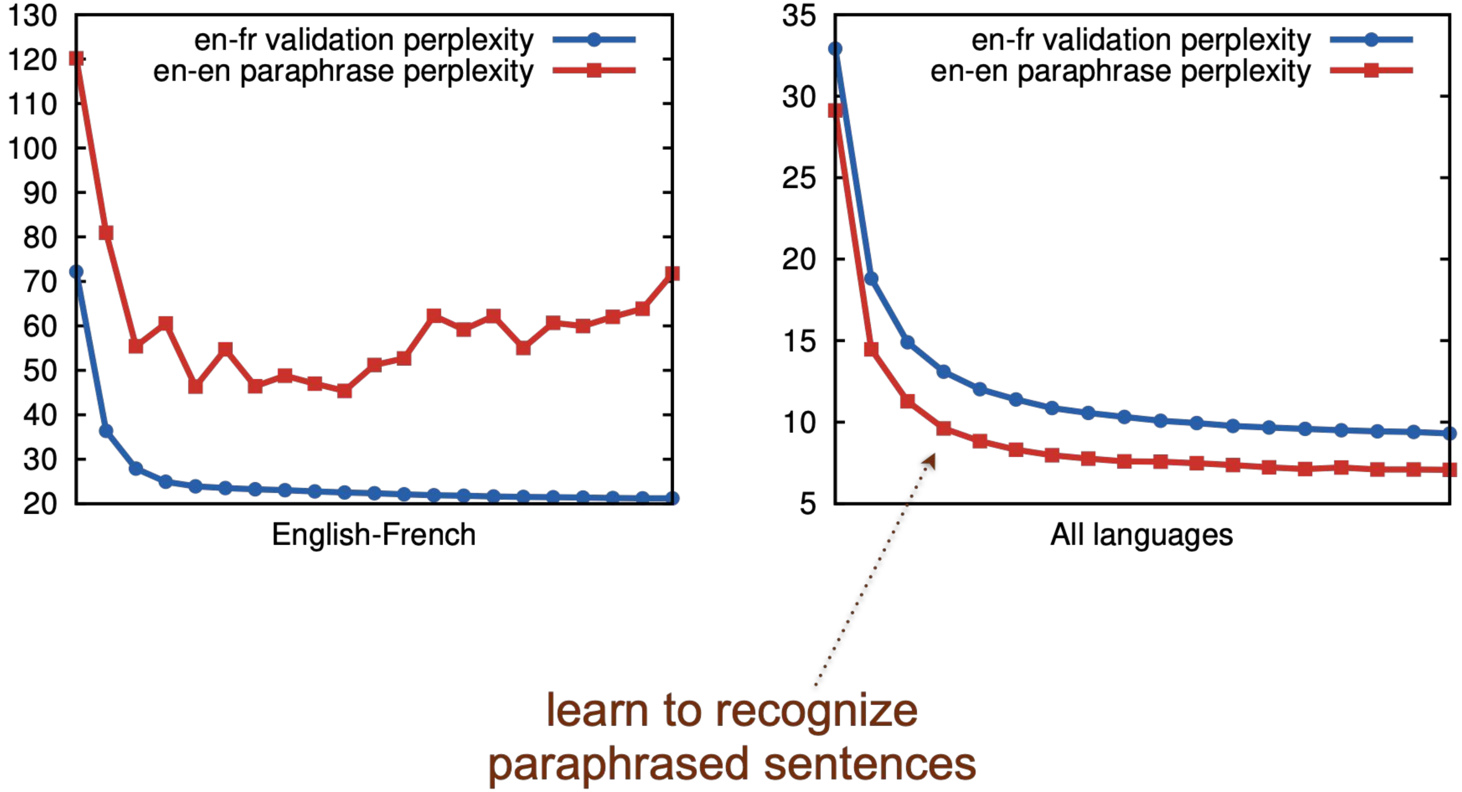


language embeddings (t-SNE plot)

Emerging Language Spaces Learned From Massively Multilingual Corpora (https://arxiv.org/abs/1802.00273)

Multilingual NMT as zero-shot paraphrase model

Learning curves during training:



Generating paraphrases with multilingual NMT

in-domain (Bible)

Source	But even as he was on the road going down, his servants met him and reported, saying, Your son lives!
+NLD	And as he was on the road, his servants went down with him, and reported, saying, Thy son lives!
+SPA	But as it was on the road, his servants came to him and told him, “Your own Son lives!”
+ALL	And while he was on the way, his servants came to him, saying, “Your son lives!”

Generating paraphrases with multilingual NMT

out-of-domain (Tatoeba)

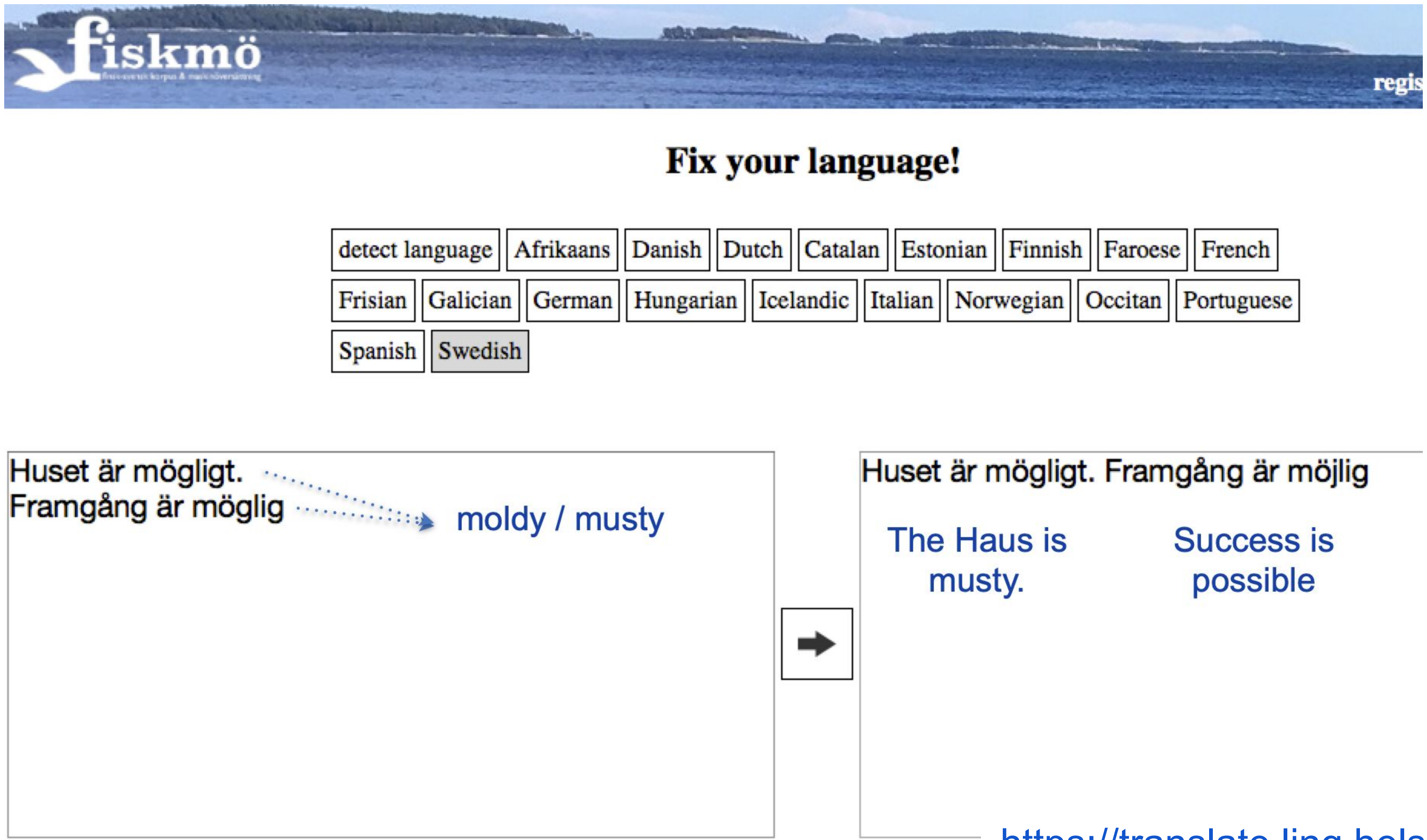
Source	He slept soundly.
Eng-Fra	Et il se prosterna devant soi.
+BRE	And, behold, he rose up quickly.
+DEU	And he began to sleep.
+ELL	He was sleeping.
+ALL	And when he had died, he was asleep.
Source	She has no brothers.
Eng-Fra	Elle n'a point de frères.
+BRE	Or, elle n'a pas de frères.
+DEU	For she has no brothers.
+OSS	No, brothers.
+ALL	You have no brothers.

Source	Have you never eaten a kiwi?
+AFR	Have you not eaten sour grapes?
Source	Do you have a cellphone?
+HIN	Do you have a scorpion?
Source	Do your children speak French?
+SPA	Do your children speak Greek?
Source	Could I park my car here?
+ITA	Do I get up here with my cavalry?
Source	Birds fly.
+DEU	The flying creatures shall fly away .

Multilingual NMT for text normalisation

https://translate.ling.helsinki.fi/fix_language

Multilingual NMT for contextualized spell checking



https://translate.ling.helsinki.fi/fix_language

Completely shared or language-specific components?

A Systematic Study of Inner-Attention-Based Sentence Representations in Multilingual Neural Machine Translation

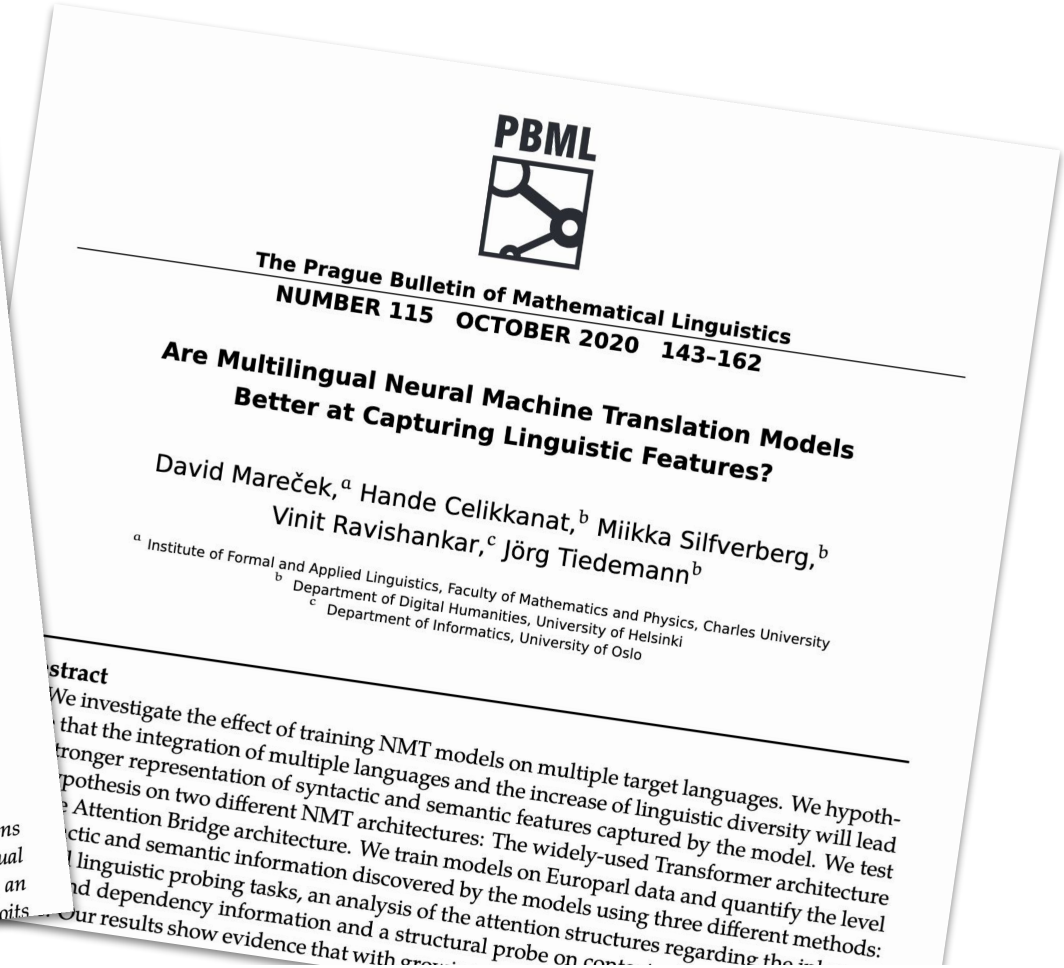
Raúl Vázquez
University of Helsinki
Department of Digital Humanities
raul.vazquez@helsinki.fi

Alessandro Raganato
University of Helsinki
Department of Digital Humanities
alessandro.raganato@helsinki.fi

Mathias Creutz
University of Helsinki
Department of Digital Humanities
mathias.creutz@helsinki.fi

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

Neural machine translation has considerably improved the quality of automatic translations by learning good representations of input sentences. In this article, we explore a multilingual translation model capable of producing fixed-size sentence representations by incorporating an ... layer, which we refer to as attention bridge. This layer exploits



PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 143–162

Are Multilingual Neural Machine Translation Models Better at Capturing Linguistic Features?

David Mareček,[a] Hande Celikkanat,[b] Miikka Silfverberg,[b] Vinit Ravishankar,[c] Jörg Tiedemann[b]

[a] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
[b] Department of Digital Humanities, University of Helsinki
[c] Department of Informatics, University of Oslo

...stract

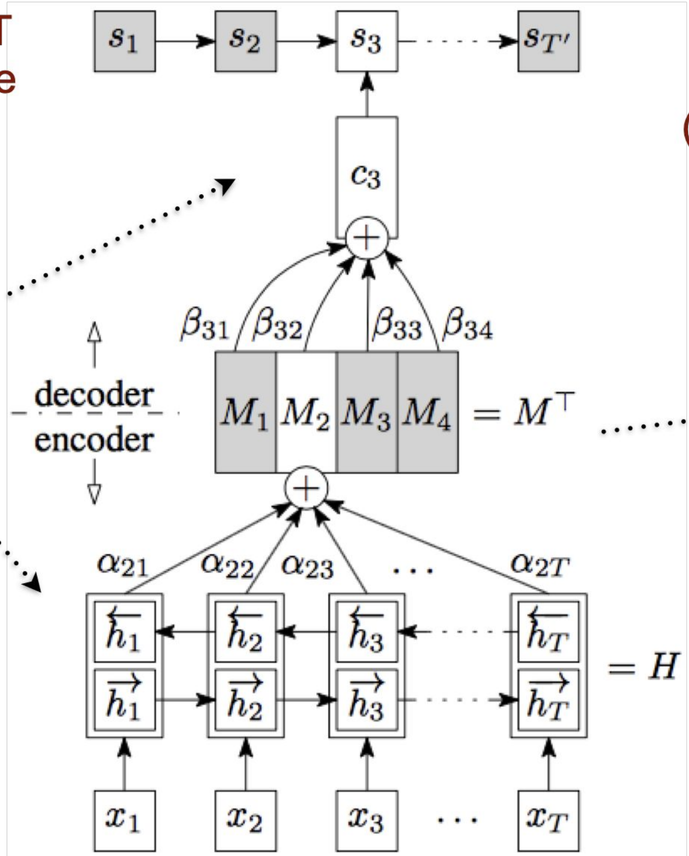
We investigate the effect of training NMT models on multiple target languages. We hypoth-
... that the integration of multiple languages and the increase of linguistic diversity will lead
...tronger representation of syntactic and semantic features captured by the model. We test
...pothesis on two different NMT architectures: The widely-used Transformer architecture
... Attention Bridge architecture. We train models on Europarl data and quantify the level
...ctic and semantic information discovered by the models using three different methods:
... linguistic probing tasks, an analysis of the attention structures regarding the in...
...nd dependency information and a structural probe on conte...
... Our results show evidence that with grow...

The attention bridge model

Benchmark with MT on unseen language pairs (zero-shot)

language-specific parameters

Rotate languages in scheduled training!



Benchmark with semantic probing (downstream) tasks

shared among all language pairs

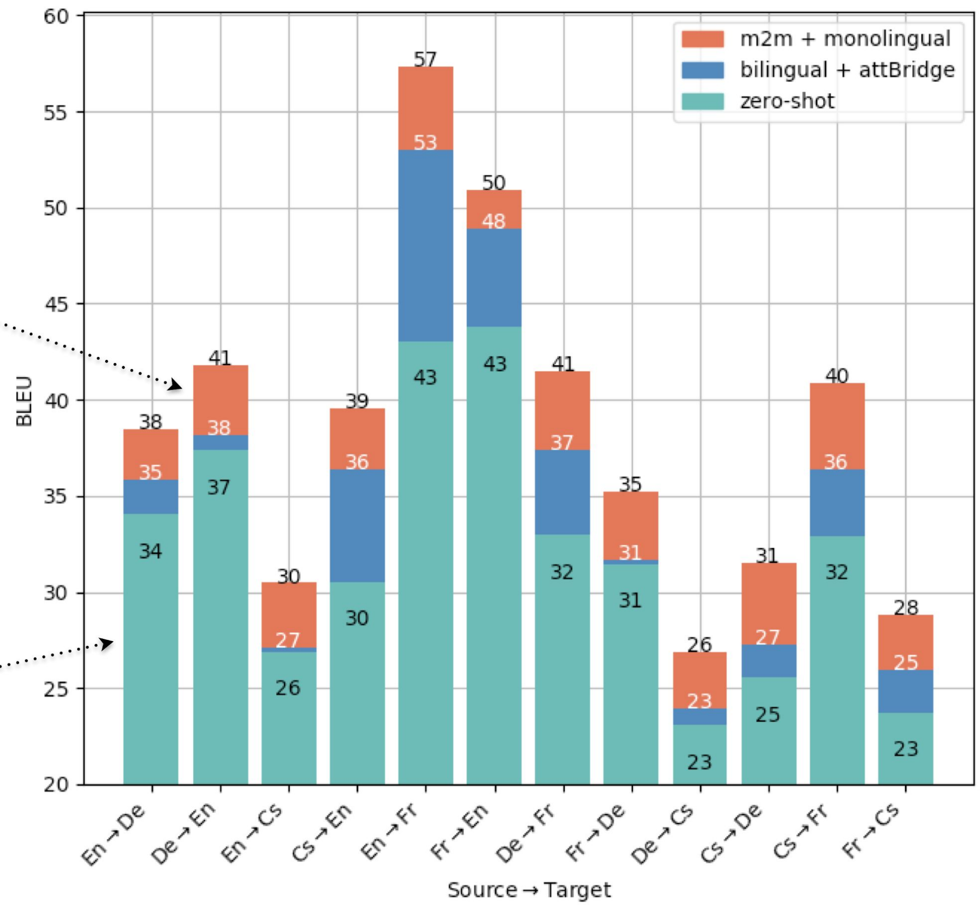
Architecture proposed by Cífka and Bojar (2018).

Our implementation in OpenNMT-py (MTM2018)

Multilingual image caption translation

model with
all languages

language pair
not seen in
training data
(zero shot)



Shared representation layer in other downstream tasks

natural language inference

multilingual models

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

Note: trained on very small data only (mult30k)

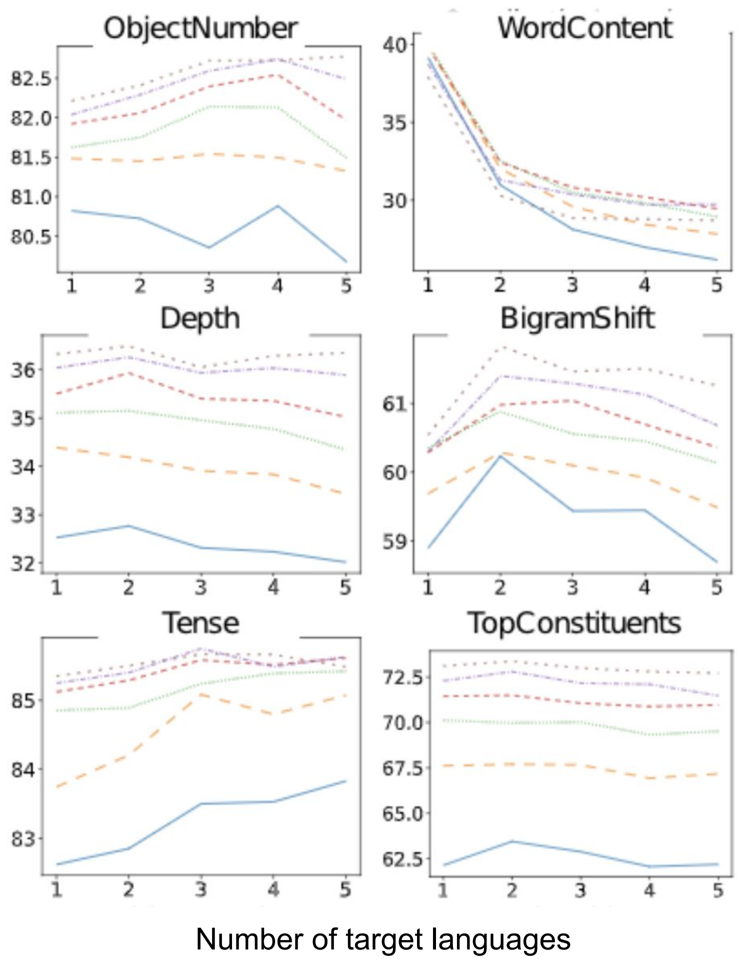
Linguistic properties in multilingual MT

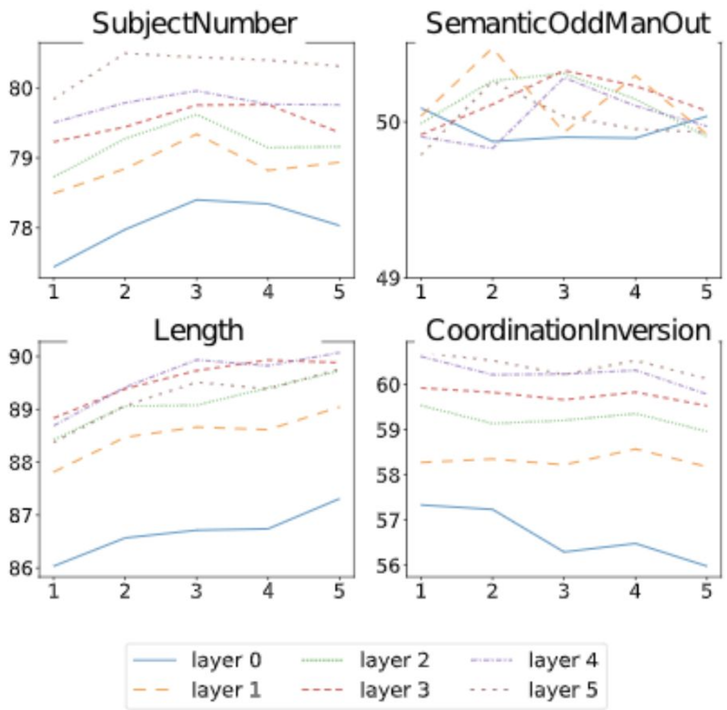
Multi-parallel subset from Europarl corpus (Koehn, 2005)

Spanning 391,306 sentences in EN, CS, FI, DE, EL, IT (100k joint vocabulary)

Source	**Target**	
{En}	**1 tgt**	{Cs}, {De}, {El}, {Fi}, {It}
	2 tgts	{Cs, De}, {De, El}, {El, Fi}, {Fi, It}, {It, Cs}
	3 tgts	{Cs, De, El}, {De, El, Fi}, {El, Fi, It}, {Fi, It, Cs}, {It, Cs, De}
	4 tgts	{Cs, De, El, Fi}, {De, El, Fi, It}, {El, Fi, It, Cs}, {Fi, It, Cs, De}, {It, Cs, De, El}
	5 tgts	{Cs, De, El, Fi, It}

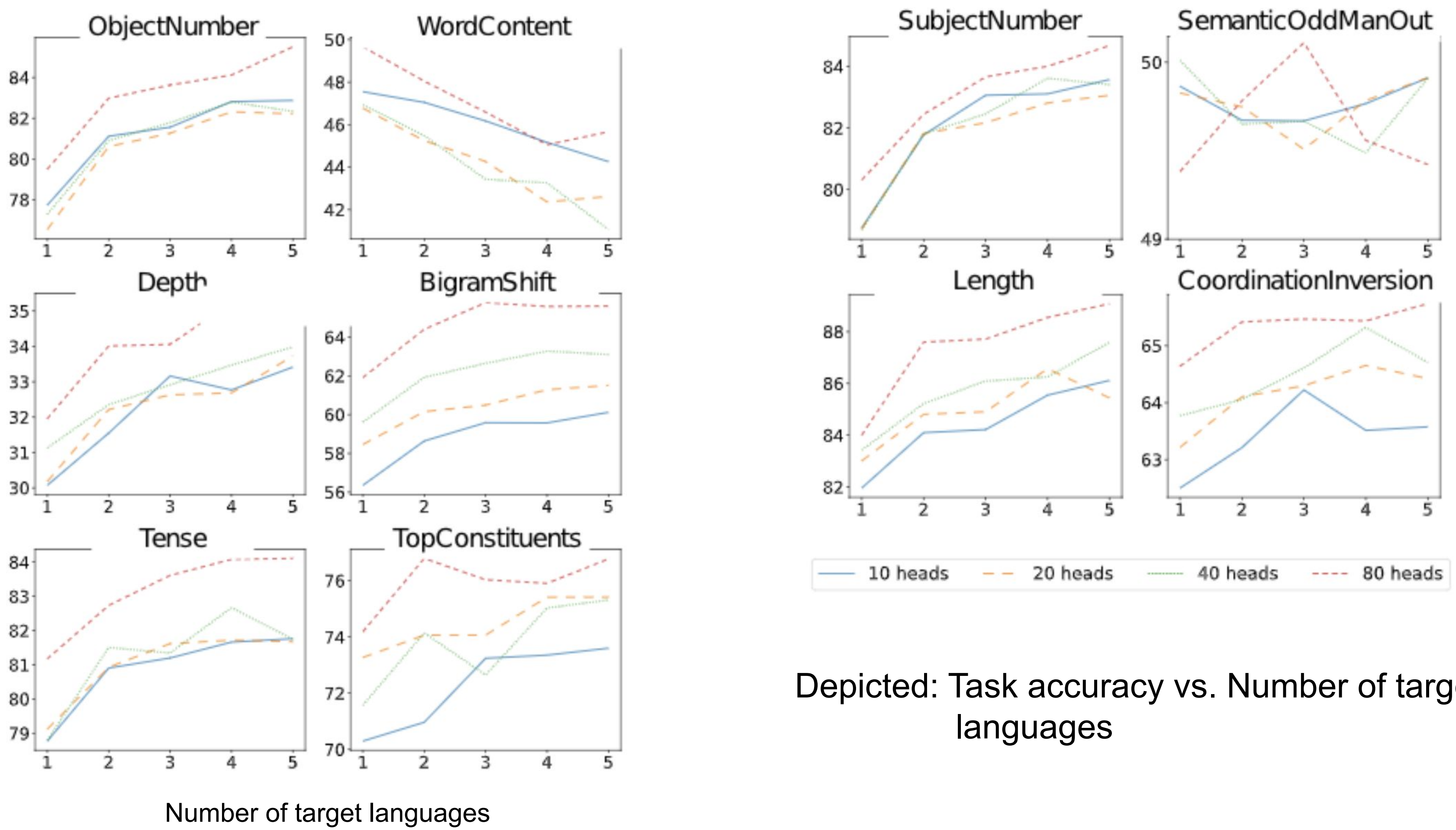
SentEval: Linguistic probing tasks (transformer)





Depicted: Task accuracy vs. Number of target languages

SentEval: Linguistic probing tasks (attention bridge)



Depicted: Task accuracy vs. Number of target languages

Intermediate takeaways

Multilingual transformers and shared parameters

- Simple and effective with emerging language spaces
- No significant difference in linguistic abstractions according to probing tasks
- Higher layers provide more abstract linguistic information

Multilingual bridge models

- Modularity and fixed-size “language agnostic” semantic representation
- Improved linguistic encoding with additional languages
- Bigger attention bridge leads to better performance

How do neural translation models encode information?

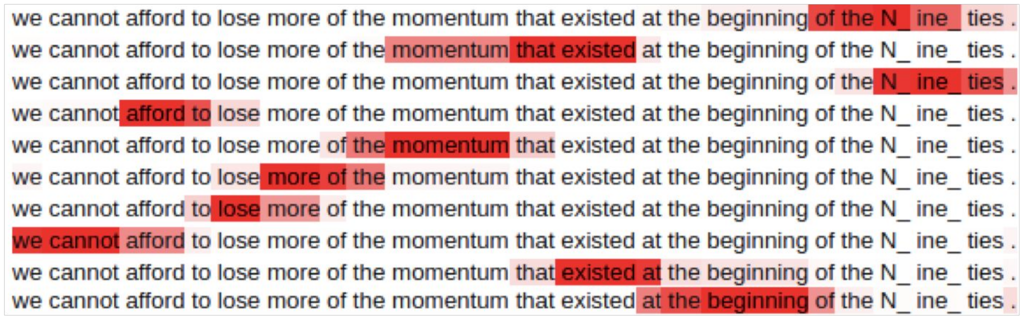
Where does the attention-bridge look at?

Attention weight of individual heads:

size of attention-bridge

(a) k = 1

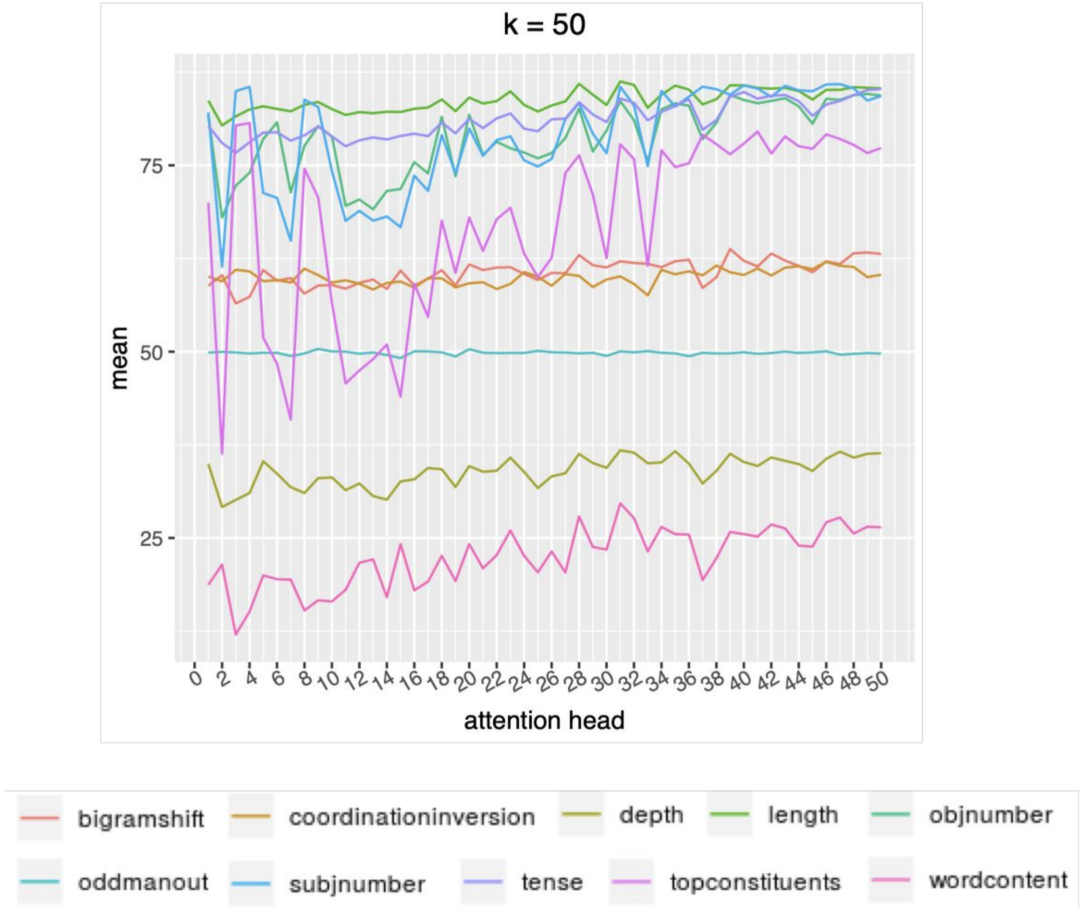
Very focused attention!



(b) k = 10

(d) k = 50

Probing individual attention-bridge heads

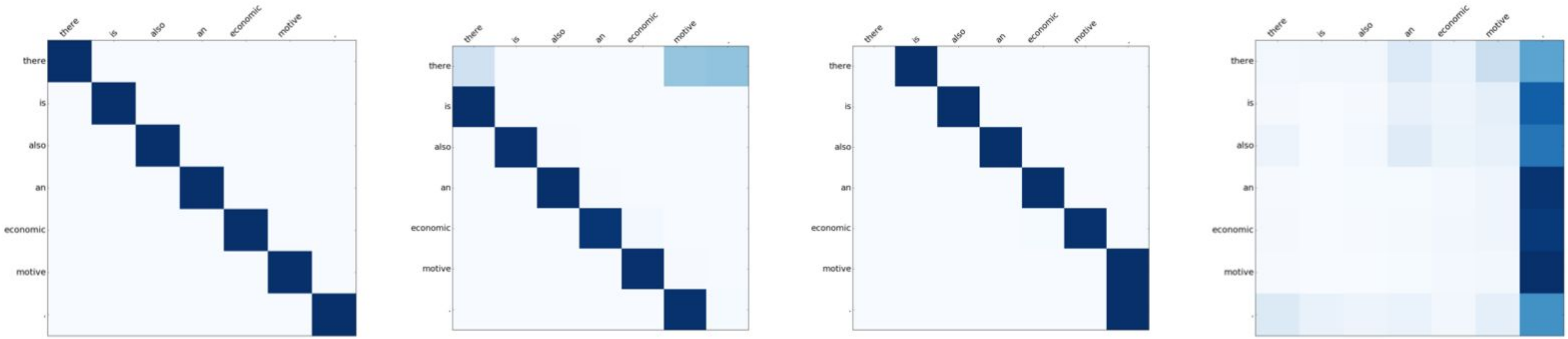


Does self-attention encode syntactic information?

		en → cs	en → de	en → et	en → fi	en → ru	en → tr	en → zh
Layer 0	attention head 0	15.06	10.67	8.79	**31.63**	**17.13**	10.99	13.00
	attention head 1	9.94	**32.90**	8.68	12.58	12.02	10.74	15.76
	attention head 2	15.84	10.62	9.60	10.12	12.08	13.69	15.50
	attention head 3	10.62	15.39	**31.38**	8.31	11.08	9.78	22.79
	attention head 4	17.25	18.12	7.76	25.10	11.75	13.20	10.28
	attention head 5	16.71	14.47	24.24	13.63	12.39	27.55	17.19
	attention head 6	**30.26**	26.28	11.76	10.43	11.55	9.90	**33.26**
	attention head 7	15.17	15.31	9.61	9.51	12.13	**31.81**	9.69
Layer 5	attention head 0	**36.02**	29.80	17.37	17.49	**35.56**	16.91	16.75
	attention head 1	28.02	27.23	16.68	28.25	13.04	**28.23**	17.71
	attention head 2	20.20	11.14	19.02	33.38	18.49	7.98	13.45
	attention head 3	11.86	8.30	22.45	14.71	19.17	15.76	19.16
	attention head 4	31.71	19.62	**33.68**	**31.87**	26.42	13.61	27.50
	attention head 5	13.55	15.20	30.73	17.35	11.98	23.13	26.70
	attention head 6	26.02	**35.32**	14.83	24.99	9.77	16.99	**29.73**
	attention head 7	18.63	10.33	15.71	11.01	12.59	25.67	14.79

Unlabeled attachment scores compared with verified syntactic treebank trees (CoNLL2017)

Typical self-attention patterns in transformer-based NMT



Often pretty sharp attention patterns related to positional information!

Replace self-attention with fixed attention patterns

High resource scenario:

- German -- English
- 11.5M training sentences

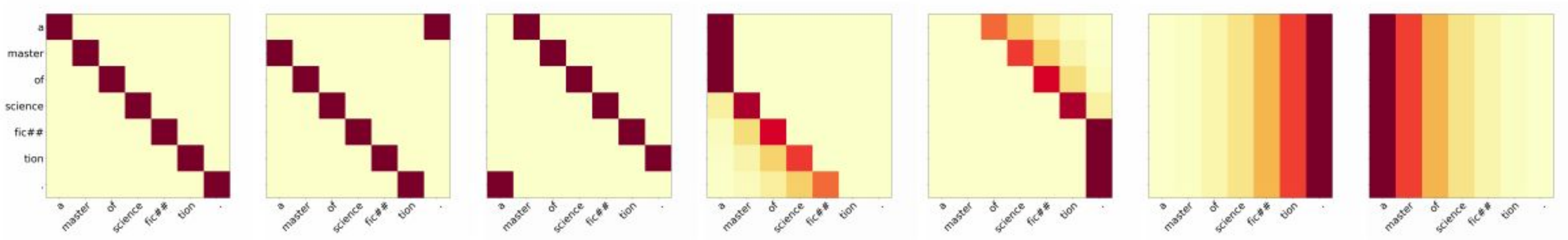
Encoder heads	EN–DE	DE–EN
8L	26.75	**34.10**
$7F_{token}$+1L	26.52	33.50
$7F_{word}$+1L	**26.92**	33.17
1L	26.26	32.91

xL = x learnable attention heads
xF = x fixed attention heads

Low resource scenario:

- German -- English, 159K
- Korean -- English, 90K
- Vietnamese -- English, 133K

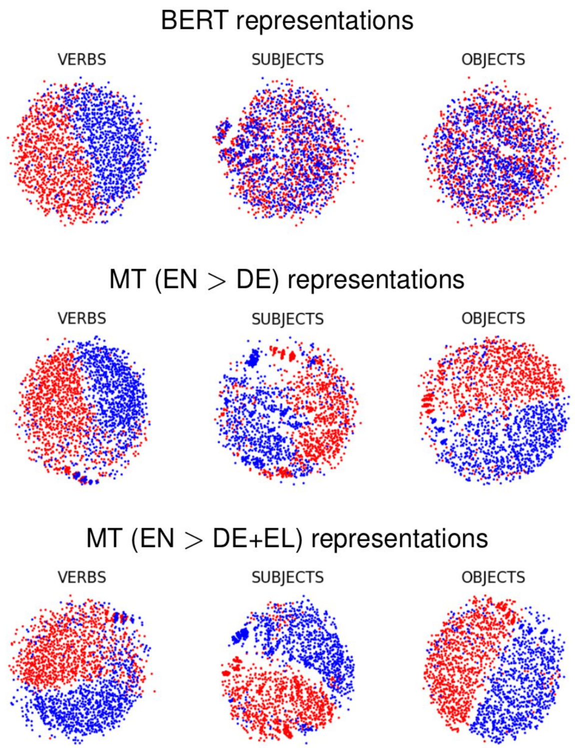
Enc. heads	DE–EN	KO–EN	EN–VI	VI–EN
8L	30.86	6.67	29.85	26.15
$7F_{token}$+1L	**32.95**	8.43	31.05	**29.16**
$7F_{word}$+1L	32.56	**8.70**	**31.15**	28.90
1L	30.22	6.14	28.67	25.03
Prior work	† 33.60	† 10.37	⊎ 27.71	⊎ 26.15

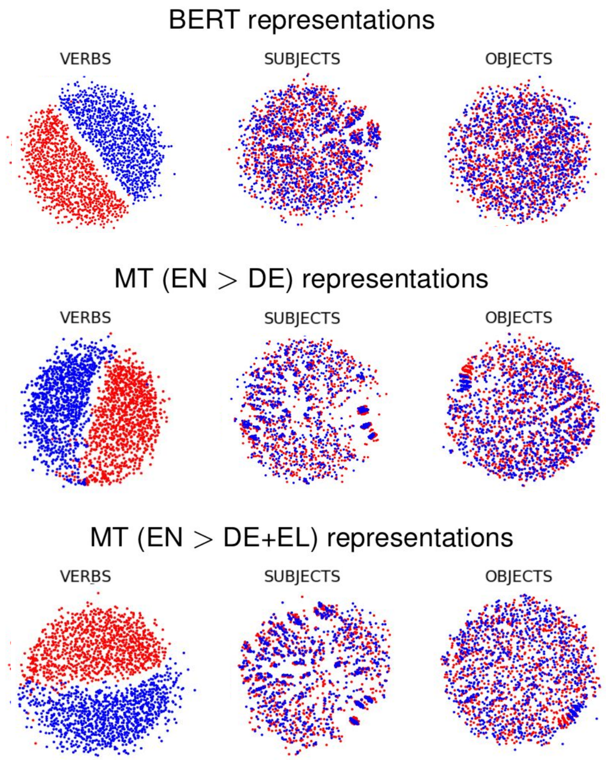


Imprint of **Passivization** and **Negation** on Contextualized Representations

(1) The **mafia kidnapped** the **millionaire**.
(2) The **millionaire** was **kidnapped** by the **mafia**.

(1) The **boy** is **playing** the **piano**.
(2) The **boy** is *not* **playing** the **piano**.





Data: contrastive pairs from SICK and template based synthetic examples

The “De-biasing” Procedure

Ravfogel at el., 2020, Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. ACL.

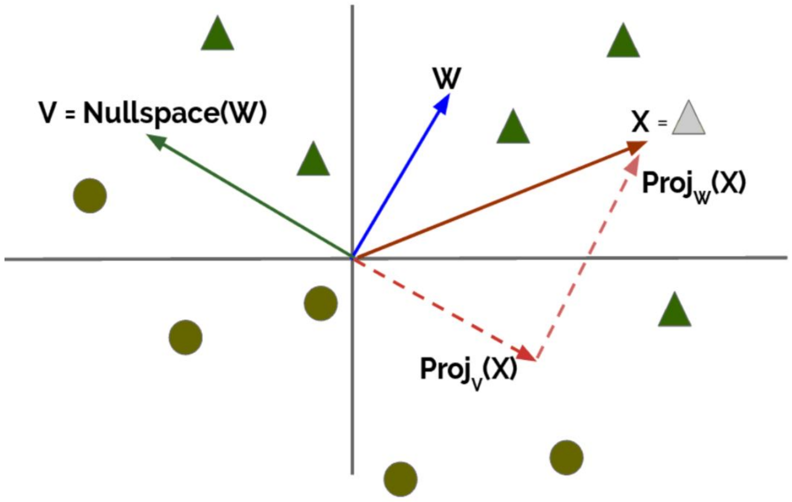


Fig. from Ravfogel at el., 2020,
Null It Out: Guarding Protected Attributes
by Iterative Nullspace Projection. ACL.

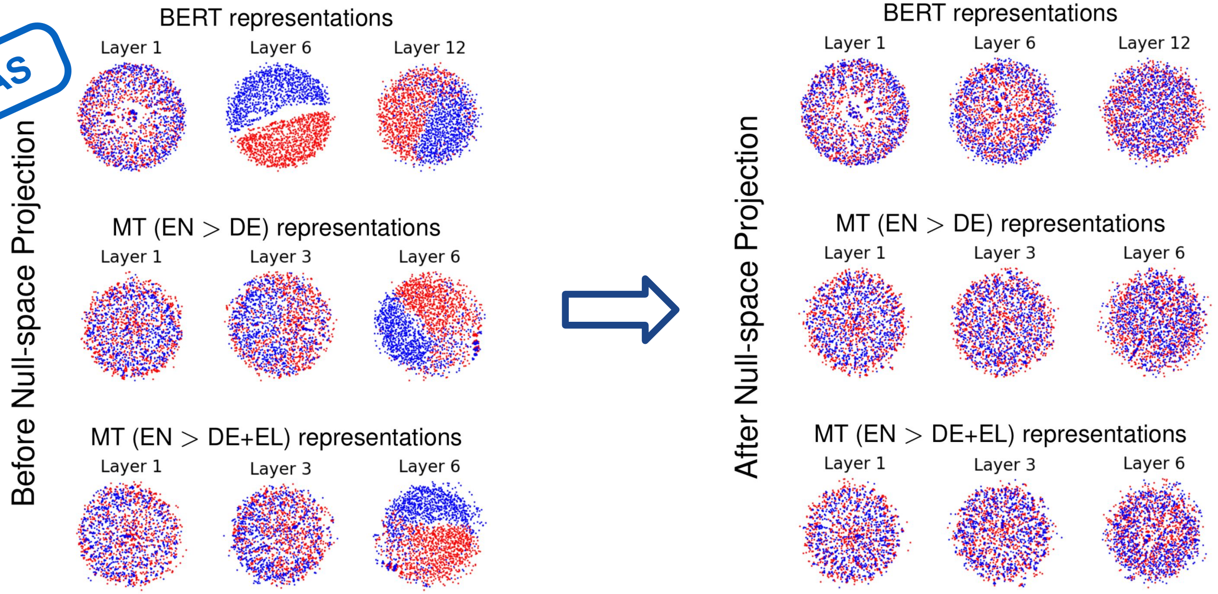
Iterative Null-Space Projection (INLP):

1. Train **linear classifier** with weight matrix W
2. **Find nullspace** of the classifier N(W) and projection matrix $P_{N(W)}$ st. $W(P_{N(W)}x) = 0 \ \forall x$
3. **Project** data **on nullspace** using $P_{N(W)}$
4. Repeat 1-3 until classifier training fails

Before vs. After

		Active-Passive					
		VERB		A-SUBJ/P-AG		A-OBJ/P-SUBJ	
		It-0	It-2	It-0	It-2	It-0	It-2
BERT	L-1	0.99	0.50	1.00	0.50	0.99	0.50
	L-6	1.00	0.49	1.00	0.50	1.00	0.50
	L-12	0.99	0.50	0.99	0.50	0.95	0.50
MT (EN > DE)	L-1	0.86	0.49	0.98	0.47	0.91	0.50
	L-3	0.87	0.49	1.00	0.49	0.96	0.50
	L-6	0.90	0.49	1.00	0.53	0.97	0.50
MT (EN > DE+EL)	L-1	0.86	0.48	0.98	0.48	0.92	0.50
	L-3	0.86	0.49	0.98	0.49	0.96	0.50
	L-6	0.91	0.49	0.99	0.49	0.98	0.51

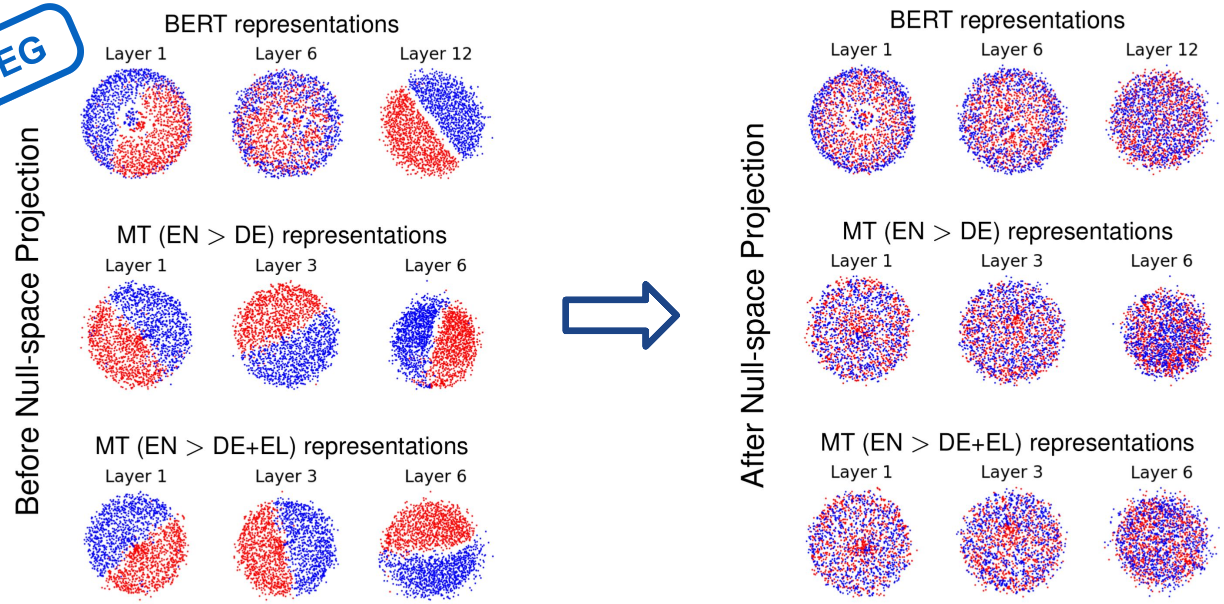
TEMPL-PAS



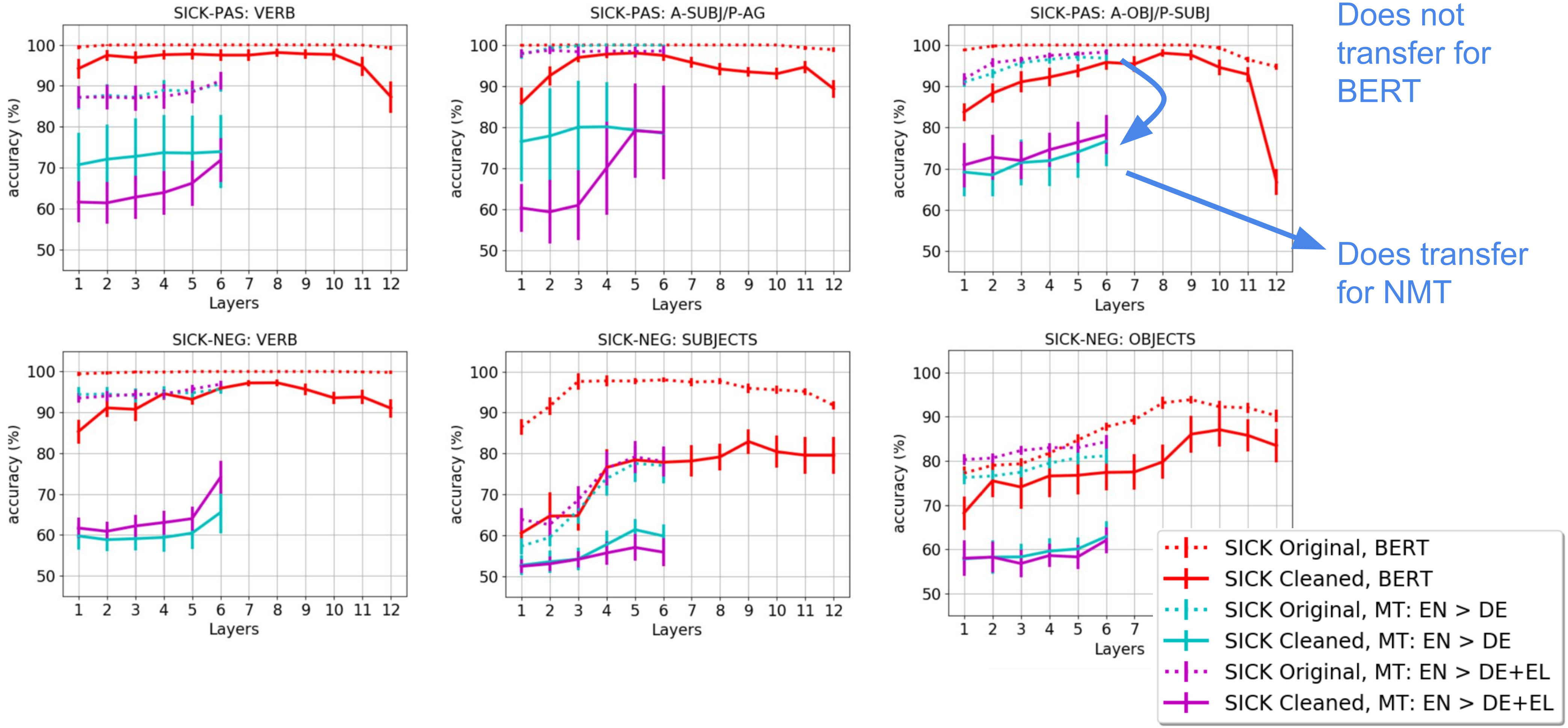
*classification accurracies before and after 2 iterations

TEMPL-NEG

		Positive-Negative					
		VERB		SUBJECT		OBJECT	
		It-0	It-2	It-0	It-2	It-0	It-2
BERT	L-1	0.99	0.49	0.86	0.50	0.77	0.50
	L-6	1.00	0.50	0.98	0.50	0.88	0.50
	L-12	1.00	0.50	0.92	0.50	0.90	0.50
MT (EN > DE)	L-1	0.94	0.49	0.57	0.50	0.76	0.51
	L-3	0.94	0.51	0.66	0.50	0.77	0.50
	L-6	0.96	0.47	0.77	0.50	0.81	0.49
MT (EN > DE+EL)	L-1	0.93	0.52	0.64	0.50	0.80	0.50
	L-3	0.94	0.49	0.69	0.50	0.83	0.50
	L-6	0.97	0.47	0.78	0.50	0.85	0.50



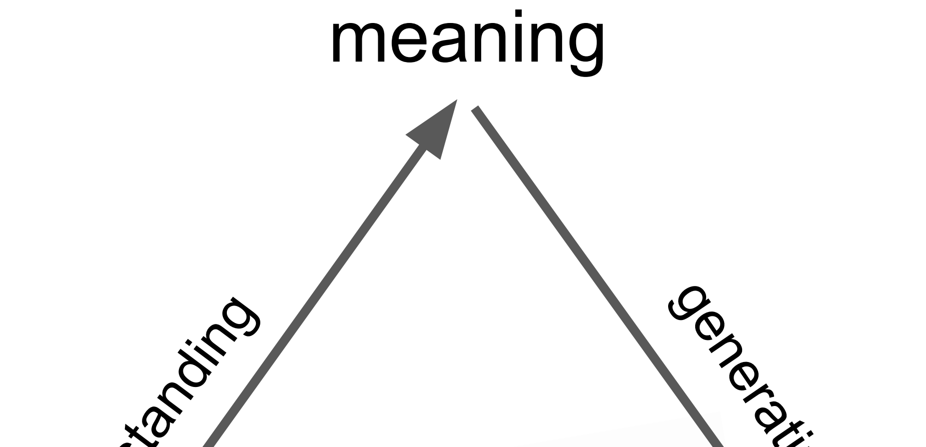
Transferring the projection between datasets (TEMPL → SICK)

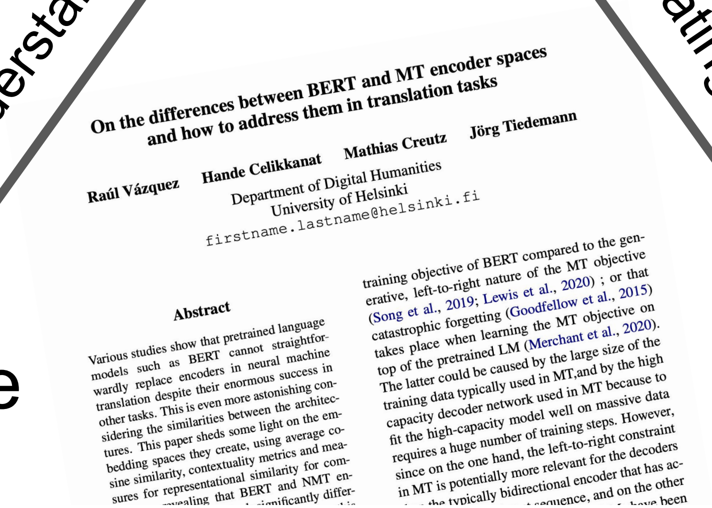


Does not transfer for BERT

Does transfer for NMT

What is the difference between LM and MT encoders?





On the differences between BERT and MT encoder spaces and how to address them in translation tasks

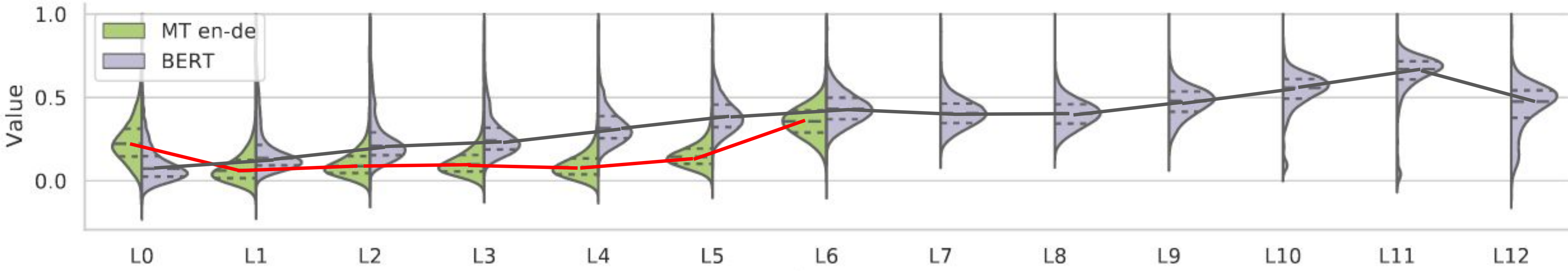
Raúl Vázquez **Hande Celikkanat** **Mathias Creutz** **Jörg Tiedemann**
Department of Digital Humanities
University of Helsinki
firstname.lastname@helsinki.fi

Abstract

Various studies show that pretrained language models such as BERT cannot straightforwardly replace encoders in neural machine translation despite their enormous success in other tasks. This is even more astonishing considering the similarities between the architectures. This paper sheds some light on the embedding spaces they create, using average cosine similarity, contextuality metrics and measures for representational similarity for com-

training objective of BERT compared to the generative, left-to-right nature of the MT objective (Song et al., 2019; Lewis et al., 2020) ; or that catastrophic forgetting (Goodfellow et al., 2015) takes place when learning the MT objective on top of the pretrained LM (Merchant et al., 2020). The latter could be caused by the large size of the training data typically used in MT,and by the high capacity decoder network used in MT because to fit the high-capacity model well on massive data requires a huge number of training steps. However, since on the one hand, the left-to-right constraint in MT is potentially more relevant for the decoders

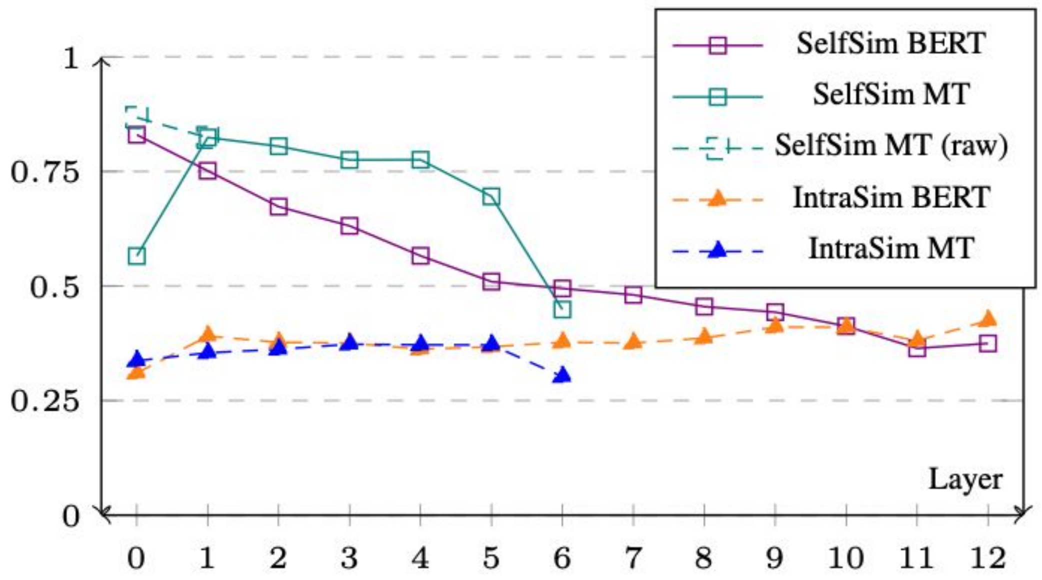
Comparing the shape of the embedding spaces



Measure of anisoptropy of the representation space:
Average cosine similarity between randomly sampled words

Method of (Ethayarajh, 2019)

Comparing the contextualisation of embeddings



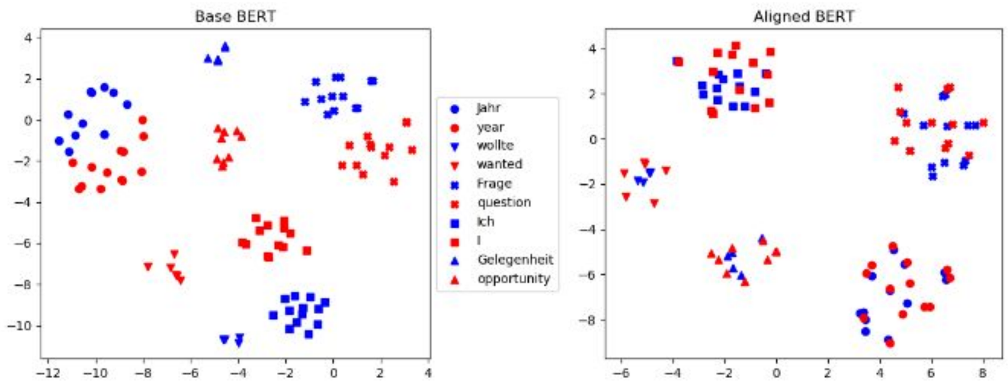
SelfSim: average cosine similarity of words in different contexts

IntraSim: average cosine similarity of words to the mean sentence vector

How to turn BERT into an MT encoder

	Encoder	Explicit alignment	Fine-tuning
MTbaseline	6-layers	✗	✗
huggingface en-de	Trf	✗	✗
M1:align	BERT (12-layers)	✓	✗
M2:fine-tune		✗	✓
M3:align+fine-tune		✓	✓

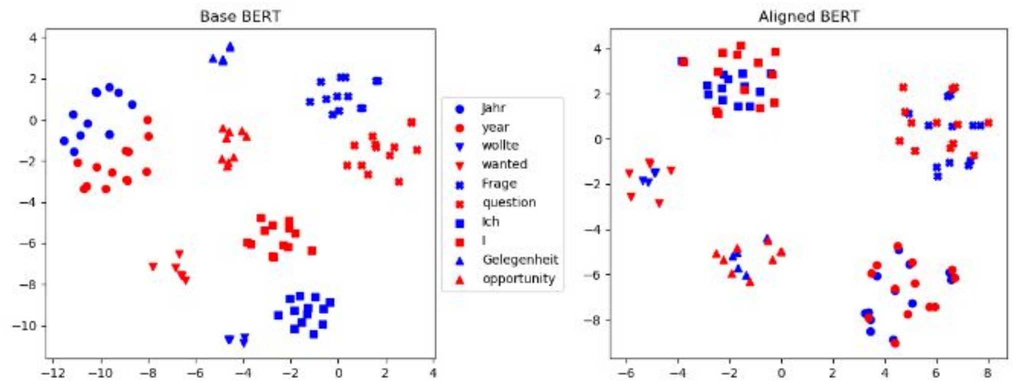
t-SNE view of the embedding space o multilingual BERT for english-german before(left) and after (right) alignment (Cao et al., 2020).



How to turn BERT into an MT encoder

	Encoder	Explicit alignment	Fine-tuning
MTbaseline	6-layers	✗	✗
huggingface en-de	Trf	✗	✗
M1:align	BERT (12-layers)	✓	✗
M2:fine-tune		✗	✓
M3:align+fine-tune		✓	✓

	Train		Val.
	Explicit Alignment	Fine-Tuning	
Europarl	45K	150K	1.5K
MuST-C	45K	150K	1.5K
newstest	13K	13K	500
Total	102K	313K	3.5K



t-SNE view of the embedding space o multilingual BERT for english-german before(left) and after (right) alignment (Cao et al., 2020).

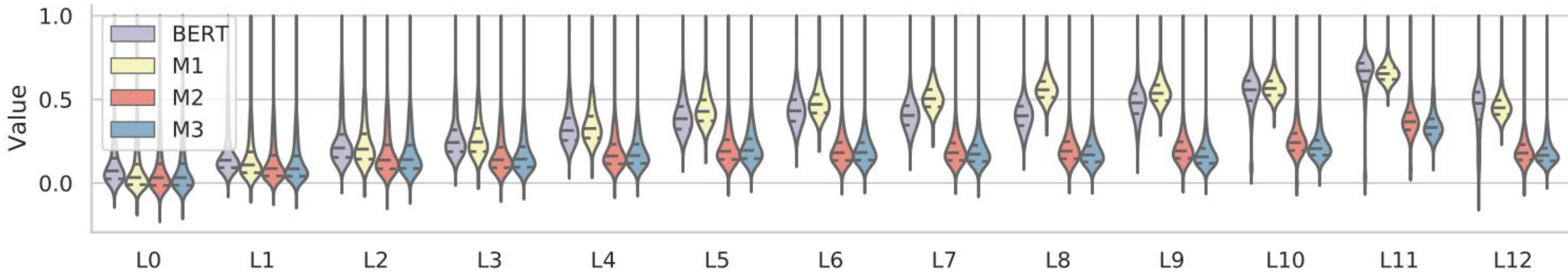
How to turn BERT into an MT encoder

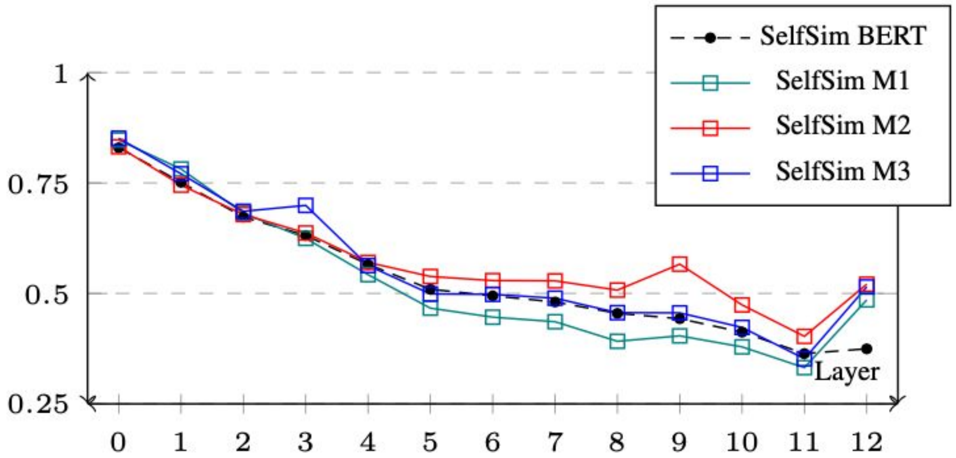
	Encoder	Explicit alignment	Fine-tuning
MTbaseline	6-layers	✗	✗
huggingface en-de	Trf	✗	✗
M1:align	BERT (12-layers)	✓	✗
M2:fine-tune		✗	✓
M3:align+fine-tune		✓	✓

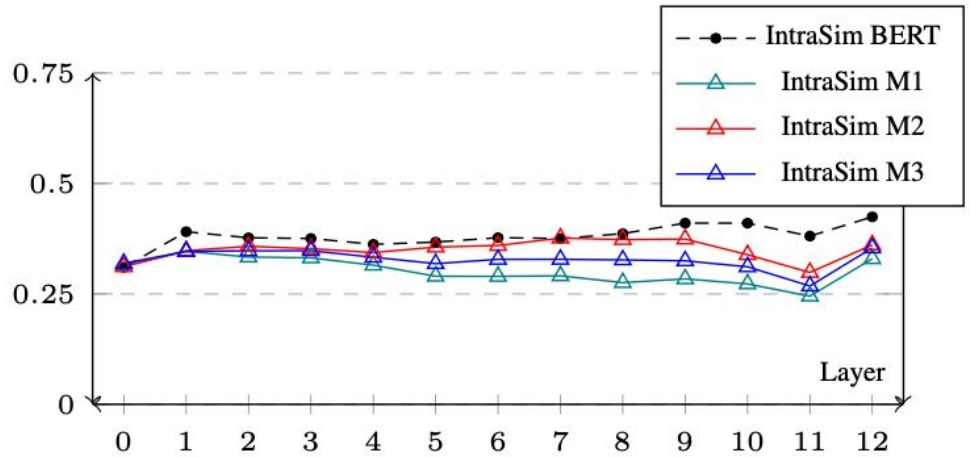
	Train		Val.
	Explicit Alignment	Fine-Tuning	
Europarl	45K	150K	1.5K
MuST-C	45K	150K	1.5K
newstest	13K	13K	500
Total	102K	313K	3.5K

	MuST-C	newstest2014
MTbaseline	29.9	14.5
huggingface en-de	33.7	28.3
M1:align	21.4	18.1
M2:fine-tune	33.8	23.9
M3:align+fine-tune	34.1	25.0

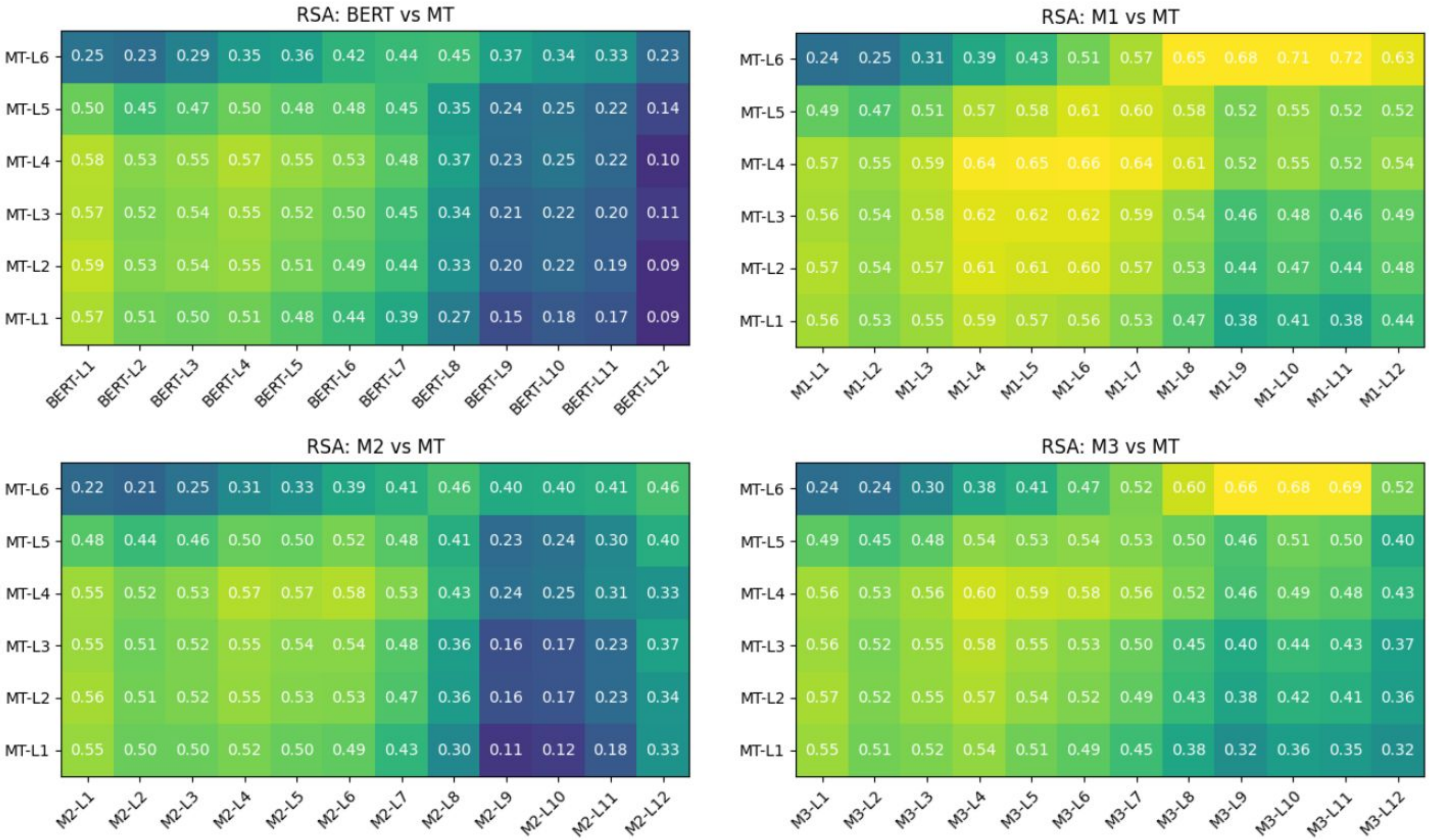
What happens to the embedding spaces?



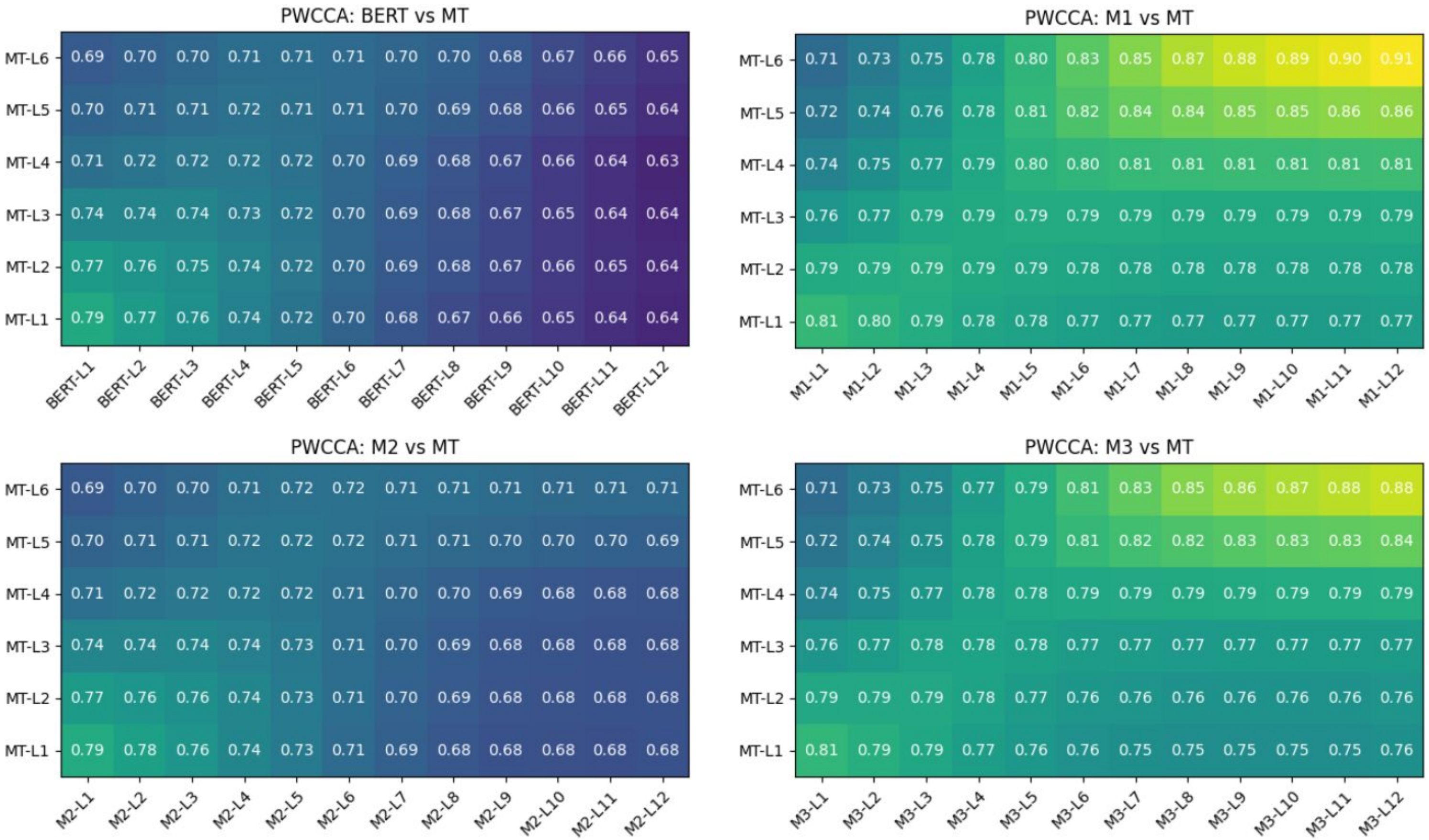




Representation similarity analysis (RSA)



Projection-Weighted Canonical Correlation Analysis



Takeaways

FOTRAN
Found in Translation
http://helsinki.fi/fotran

It's easy

- … to train a translation model
- … to include additional languages
- … to use multilingual models for various tasks

It's difficult

- … to do something smarter than adding more data and training from scratch
- … to understand what is going on in the model
- … to design probing tasks and benchmarks that lead to reliable conclusions

Possible conclusions

FOTRAN
Found in Translation
http://helsinki.fi/fotran

Multilinguality is useful

- Knowledge transfer works to some extent
- Zero-shot learning is possible (but weak)
- May lead to more abstraction

Linguistic information

- Is spread all over the place without very clear patterns
- Local dependencies dominate a lot (which is no big surprise)
- Certain phenomena can be extracted from distributed representations

Many things left to do ...

Next steps

FOTRAN
Found in Translation
http://helsinki.fi/fotran

Scale up and extend

- Massively multilingual models (with modular architectures?)
- Add multimodality (we already have an audio encoder for the attention bridge)
- Hierarchically-shared bridge models (typological hierarchies?)
- Properly model uncertainty

Continue the analyses of NMT representations (and benchmarks)

- Difference between LMs and translation models
- Monitor representations during training with different objectives
- Understand what benchmarks really test and reveal

meaning
understanding
generating
HELSINKI NLP
the World's languages

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

ACADEMY OF FINLAND
CSC
erc European Research Council


Thank you!

Special thanks to
Hande Celikkanat, Raul Vazquez
Yves Scherrer, Alessandro Raganato
and the entire Helsinki-NLP team