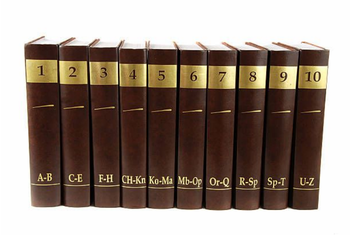
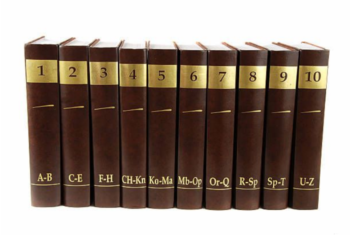
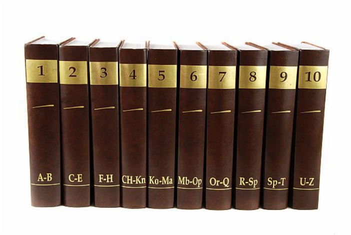
ATNLP: Multilinguality

Anders Søgaard

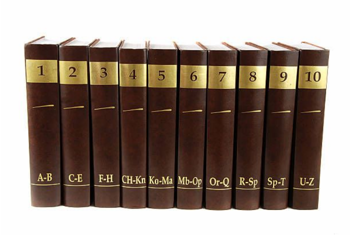
What are language models for?


1
A-B
2
C-E
3
F-H
4
CH-Kn
5
Ko-Ma
6
Mb-Op
7
Or-Q
8
R-Sp
9
Sp-T
10
U-Z


1
2
3
4
5
6
7
8
9
10
A-B
C-E
F-H
CH-Kn
Ko-Ma
Mb-Op
Or-Q
R-Sp
Sp-T
U-Z


1 2 3 4 5 6 7 8 9 10
A-B C-E F-H CH-Kn Ko-Ma Mb-Op Or-Q R-Sp Sp-T U-Z

+


1 2 3 4 5 6 7 8 9 10
A-B C-E F-H CH-Kn Ko-Ma Mb-Op Or-Q R-Sp Sp-T U-Z

+ +

	4	4-6	4-6-6	4-6-6-3
good	0.2	0.2	0.2	0.1
image	0.1	0.1		
home	0.1	0.1	0.1	0.1
hope	0.2	0.2	0.2	

He _____	walks, talks, …
{*Legal*} He _____	appeals, adjudicates, …
{*9th century*}, He _____	sayeth, hath, …
{*Translation, English-German*}, He _____	Er
Translate ‘He’ into German…	Er

2017

Transformers

Good scaling properties, exploiting GPUs, ability to model long-range dependencies.

2018

Conditional language models

Language prefixes in machine translation, transfer learning, style transfer.

2019

Seeing everything as question answering

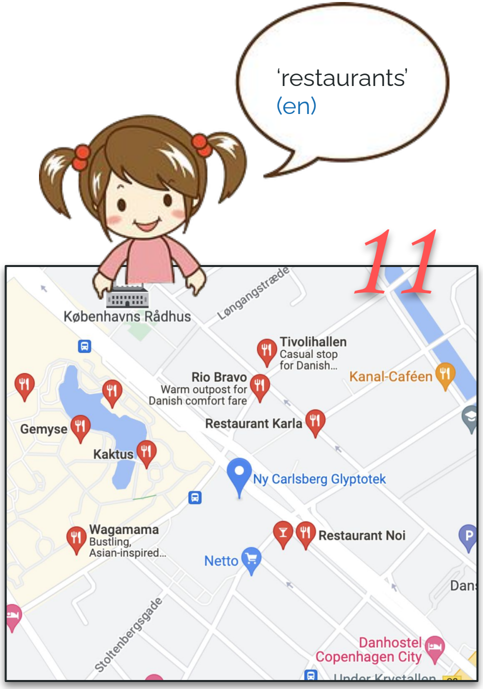
What is the German translation (or the grammatical analysis) of ‘Mary bought a house’?

2021

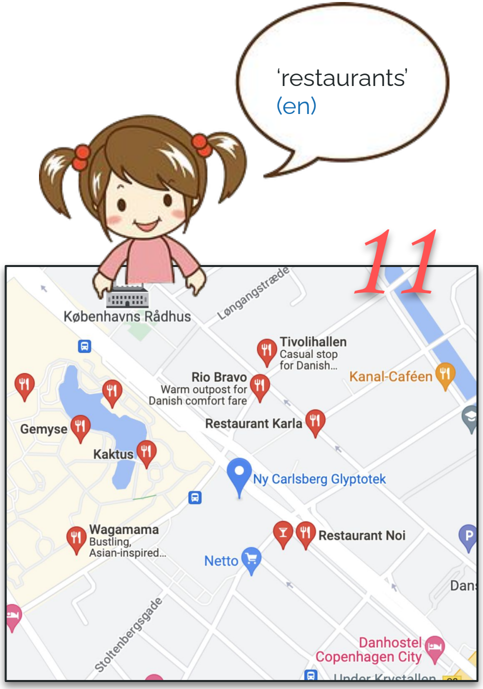
In-context learning

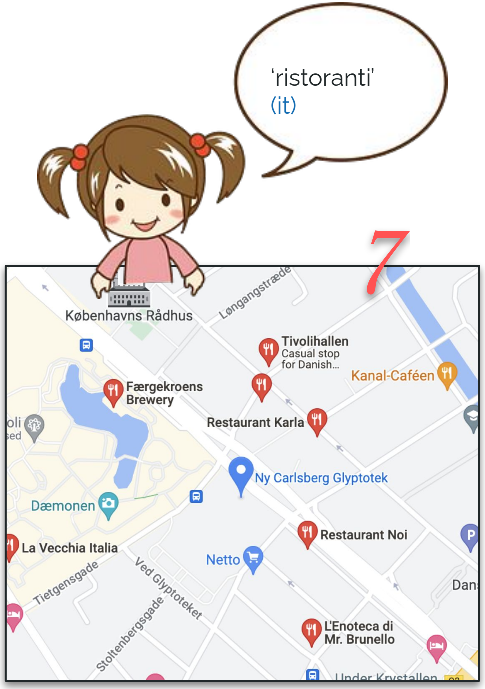
Moving away from training one model for each task, toward training models that can learn on the fly.

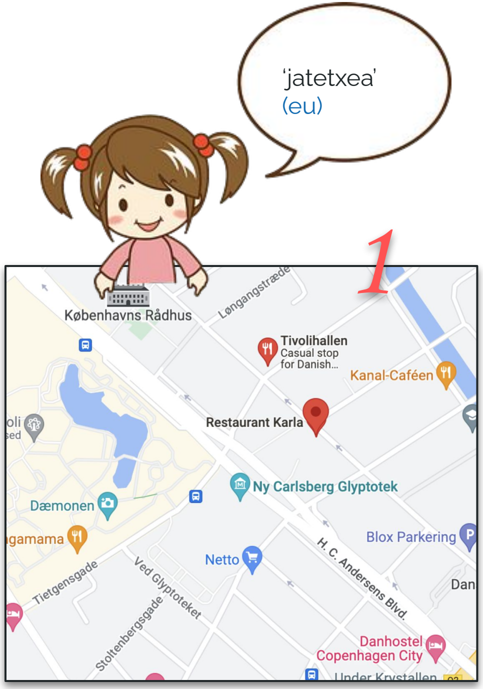
The Digital Language Divide

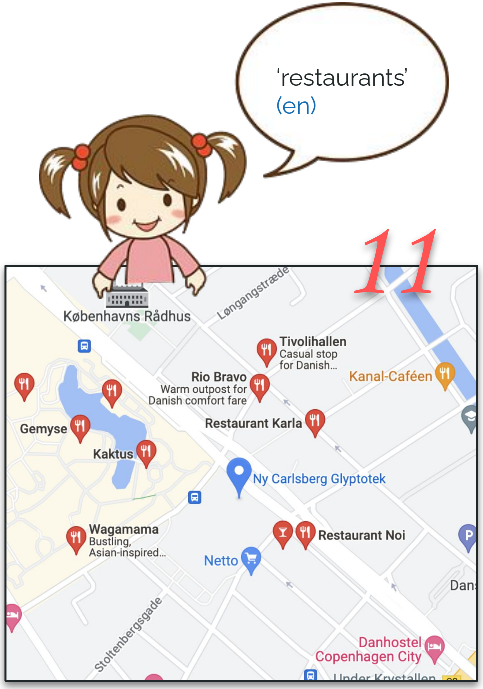


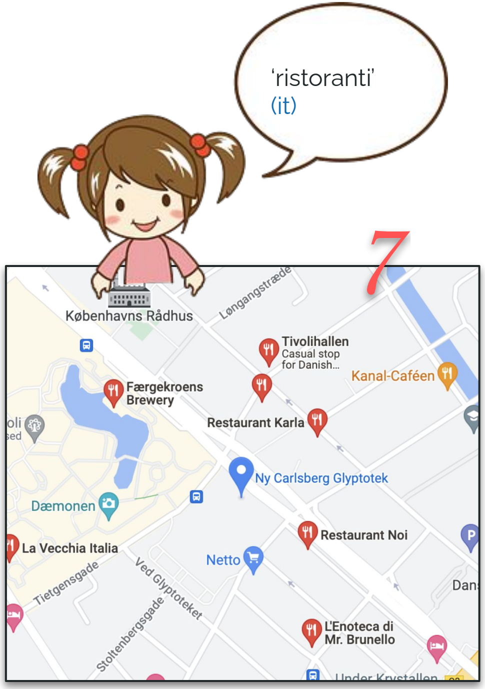
'restaurants'
(en)
11
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Rio Bravo
Warm outpost for
Danish comfort fare
Kanal-Caféen
Restaurant Karla
Gemyse
Kaktus
Ny Carlsberg Glyptotek
Wagamama
Bustling,
Asian-inspired...
Restaurant Noi
Netto
Stoltenbergsgade
Danhostel
Copenhagen City


'restaurants'
(en)
11
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Rio Bravo
Warm outpost for
Danish comfort fare
Kanal-Caféen
Restaurant Karla
Gemyse
Kaktus
Ny Carlsberg Glyptotek
Wagamama
Bustling,
Asian-inspired...
Restaurant Noi
Netto
Stoltenbergsgade
Danhostel
Copenhagen City

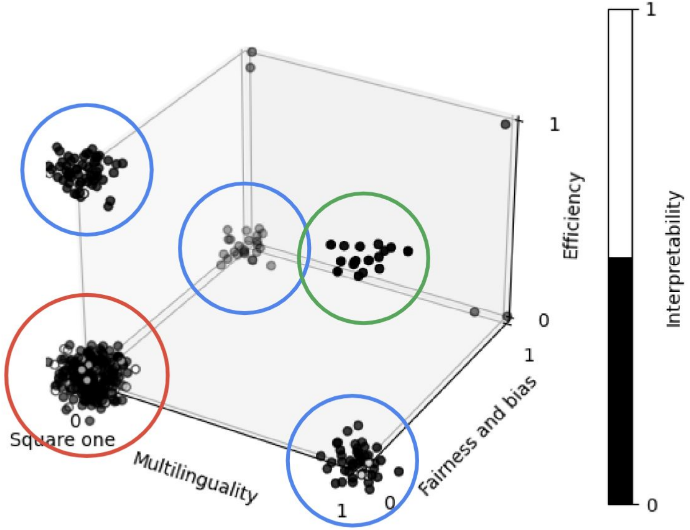

'ristoranti'
(it)
7
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Færgekroens
Brewery
Kanal-Caféen
Restaurant Karla
Ny Carlsberg Glyptotek
Dæmonen
La Vecchia Italia
Restaurant Noi
Netto
Tietgensgade
Ved Glyptoteket
Stoltenbergsgade
L'Enoteca di
Mr. Brunello


'jatetxea'
(eu)
1
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Kanal-Caféen
Restaurant Karla
Ny Carlsberg Glyptotek
Dæmonen
Blox Parkering
Netto
Tietgensgade
Ved Glyptoteket
H. C. Andersens Blvd.
Stoltenbergsgade
Danhostel
Copenhagen City
Under Krystallen


'restaurants'
(en)
11
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Rio Bravo
Warm outpost for
Danish comfort fare
Kanal-Caféen
Gemyse
Restaurant Karla
Kaktus
Ny Carlsberg Glyptotek
Wagamama
Bustling,
Asian-inspired...
Restaurant Noi
Netto
Stoltenbergsgade
Danhostel
Copenhagen City
Under Krystallen

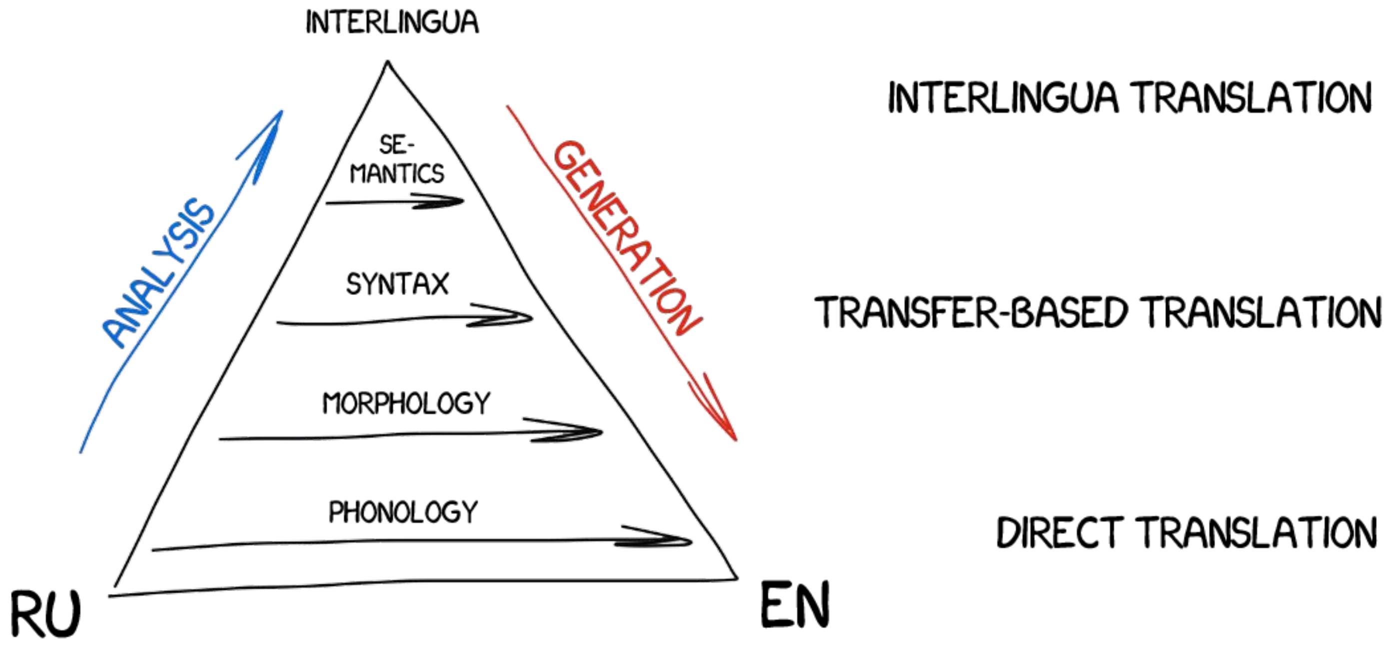

'ristoranti'
(it)
7
Københavns Rådhus
Løngangstræde
Tivolihallen
Casual stop
for Danish...
Færgekroens
Brewery
Kanal-Caféen
Restaurant Karla
Ny Carlsberg Glyptotek
Dæmonen
Restaurant Noi
La Vecchia Italia
Netto
Tietgensgade
Ved Glyptoteket
Stoltenbergsgade
L'Enoteca di
Mr. Brunello
Under Krystallen

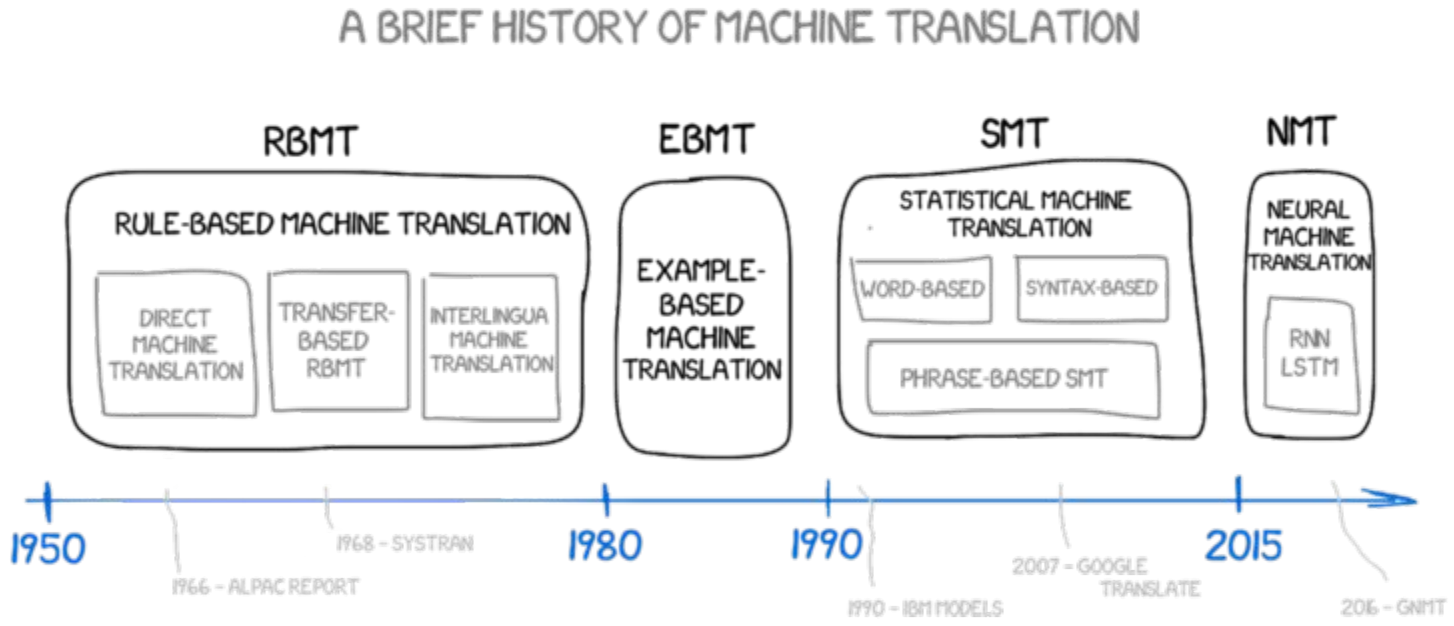
Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%
Sentiment Analysis	18	100.0%	72.2%	0.0%	0.0%	11.1%	11.1%	0.0%
Summarization	12	91.7%	0.0%	0.0%	8.3%	0.0%	8.3%	0.0%



A Brief History of Machine Translation

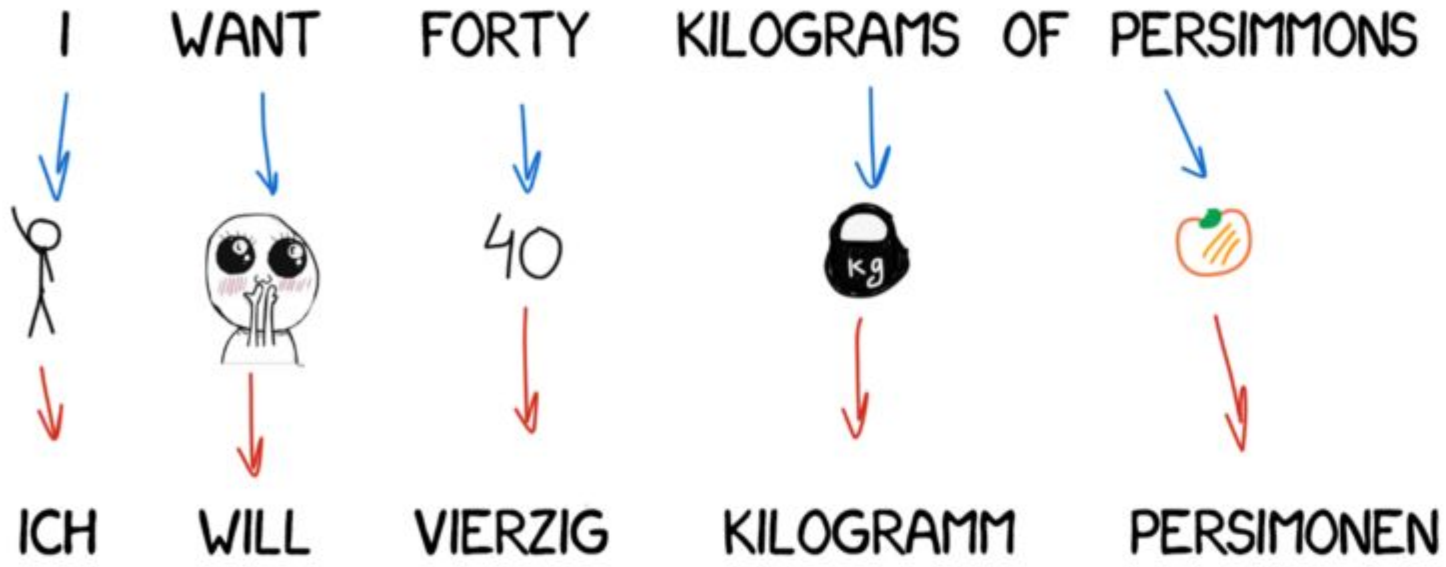
VAUQUOIS TRIANGLE

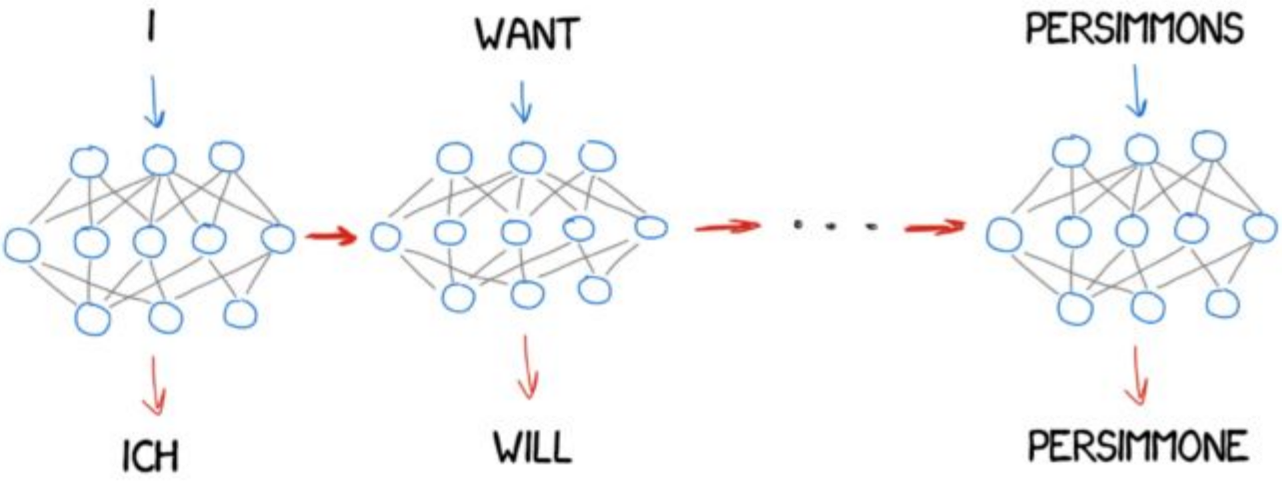


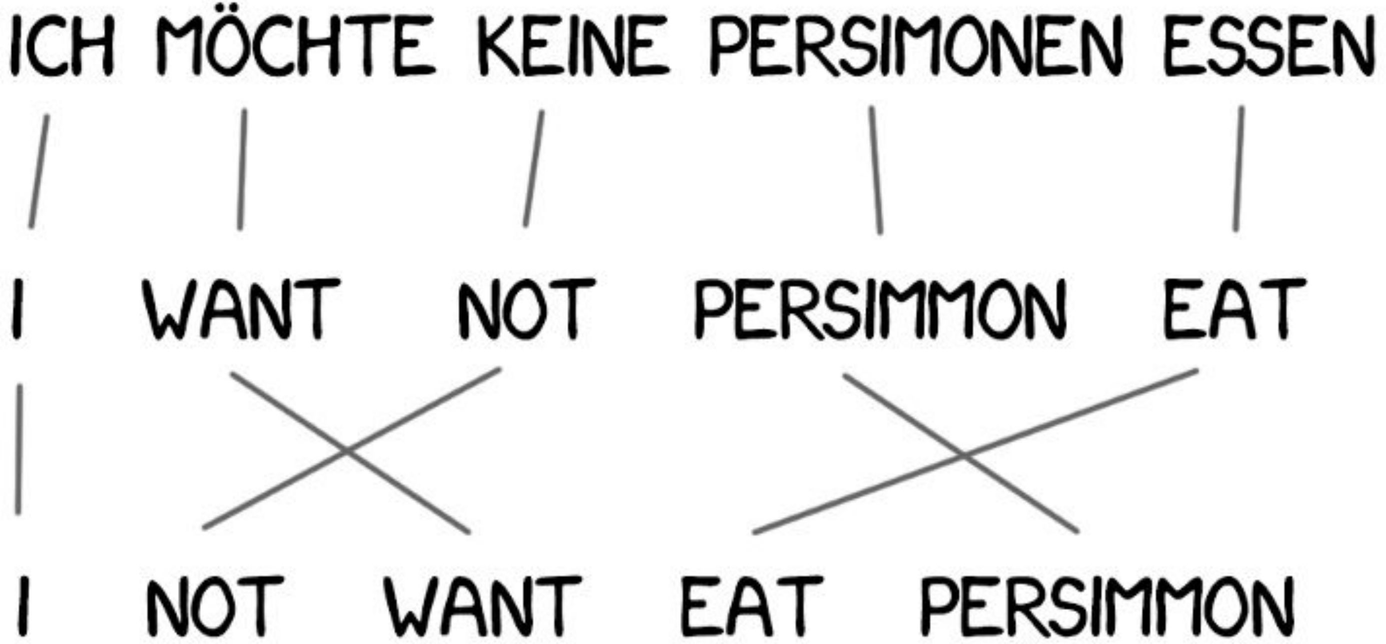

A BRIEF HISTORY OF MACHINE TRANSLATION
RBMT
RULE-BASED MACHINE TRANSLATION
DIRECT MACHINE TRANSLATION
TRANSFER-BASED RBMT
INTERLINGUA MACHINE TRANSLATION
EBMT
EXAMPLE-BASED MACHINE TRANSLATION
SMT
STATISTICAL MACHINE TRANSLATION
WORD-BASED
SYNTAX-BASED
PHRASE-BASED SMT
NMT
NEURAL MACHINE TRANSLATION
RNN
LSTM
1950
1966 - ALPAC REPORT
1968 - SYSTRAN
1980
1990
1990 - IBM MODELS
2007 - GOOGLE TRANSLATE
2015
2016 - GNMT

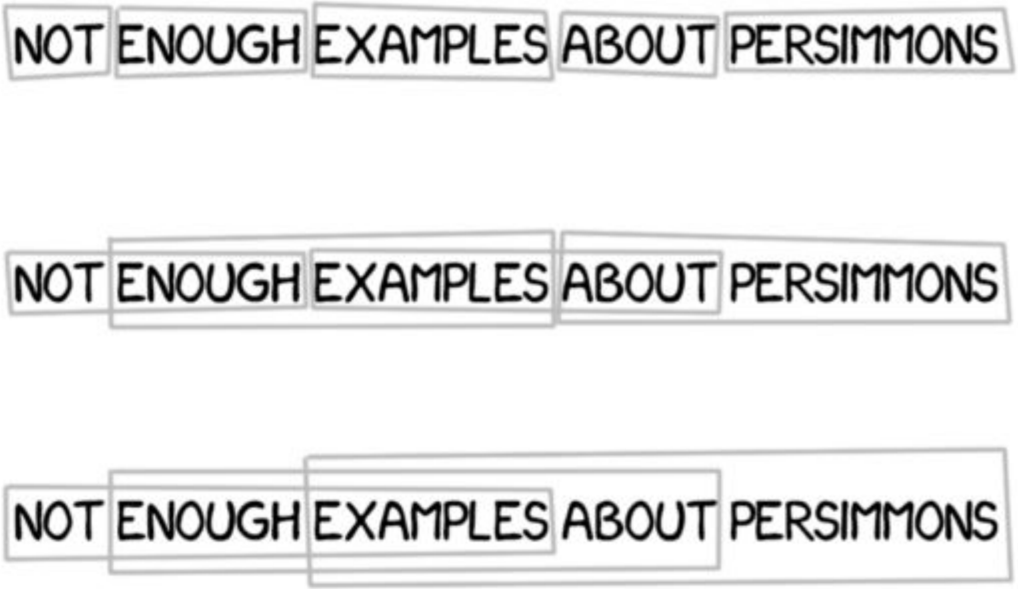
I
WANT
FORTY
KILOGRAMS OF
PERSIMMONS
ICH
WOLLEN
VIERZIG
KILOGRAMM
PERSIMONEN

I | CRAVE | 40 | KG | KAKI


I WANT FORTY KILOGRAMS OF PERSIMMONS
40
KG
ICH WILL VIERZIG KILOGRAMM PERSIMONEN


I
WANT
PERSIMMONS
ICH
WILL
PERSIMMONE


ICH MÖCHTE KEINE PERSIMONEN ESSEN
I WANT NOT PERSIMMON EAT
I NOT WANT EAT PERSIMMON



UNIGRAMS:

1. NOT
2. ENOUGH
3. EXAMPLES
4. ABOUT
5. PERSIMMONS

BIGRAMS:

1. NOT ENOUGH
2. ENOUGH EXAMPLES
3. EXAMPLES ABOUS
4. ABOUT PERSIMMONS

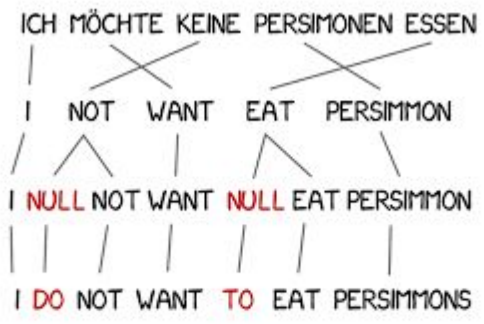
TRIGRAMS:

1. NOT ENOUGH EXAMPLES
2. ENOUGH EXAMPLES ABOUT
3. EXAMPLES ABOUS PERSIMMONS

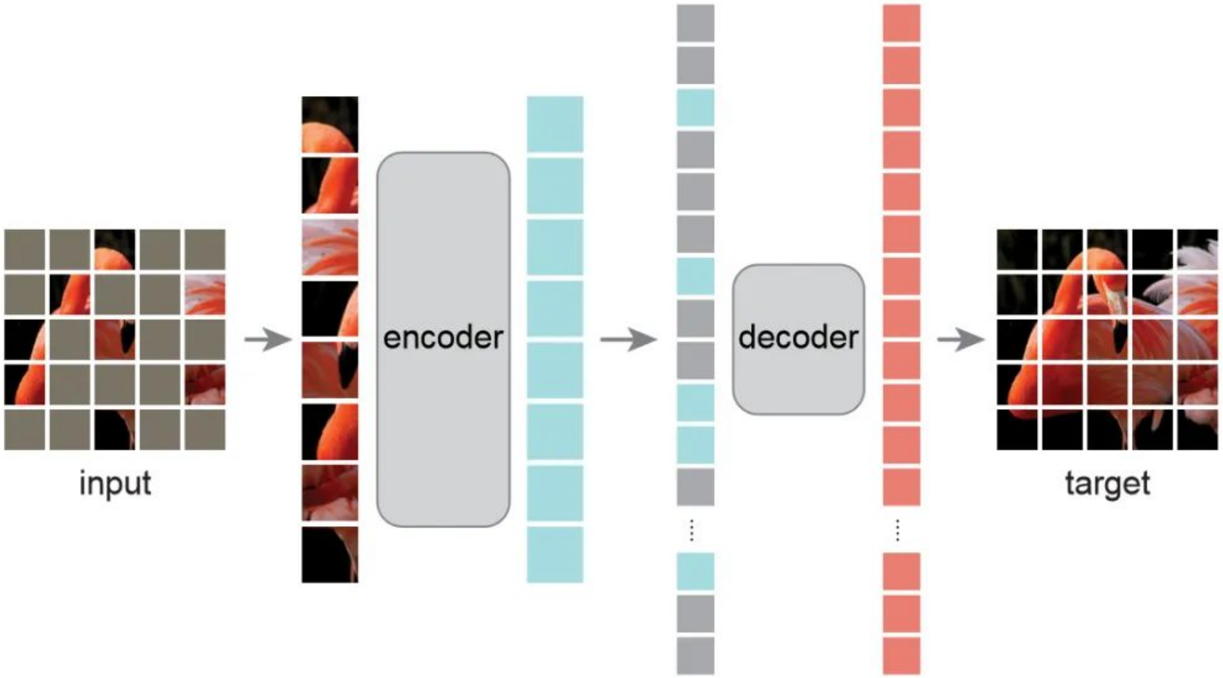
Why translation ends up ugly

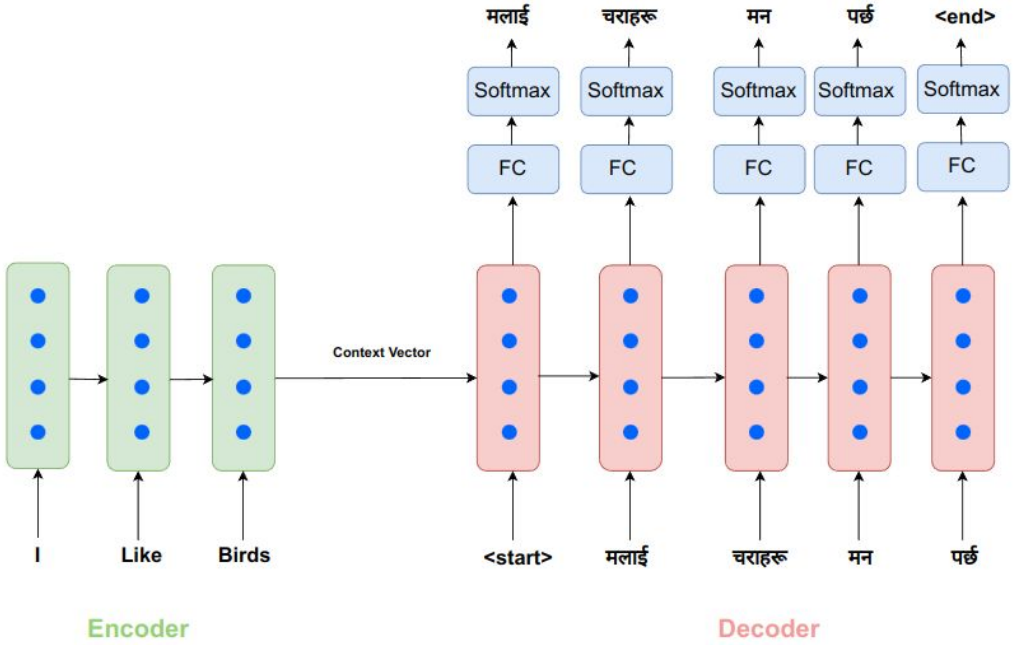
Standard Challenges in Machine Translation

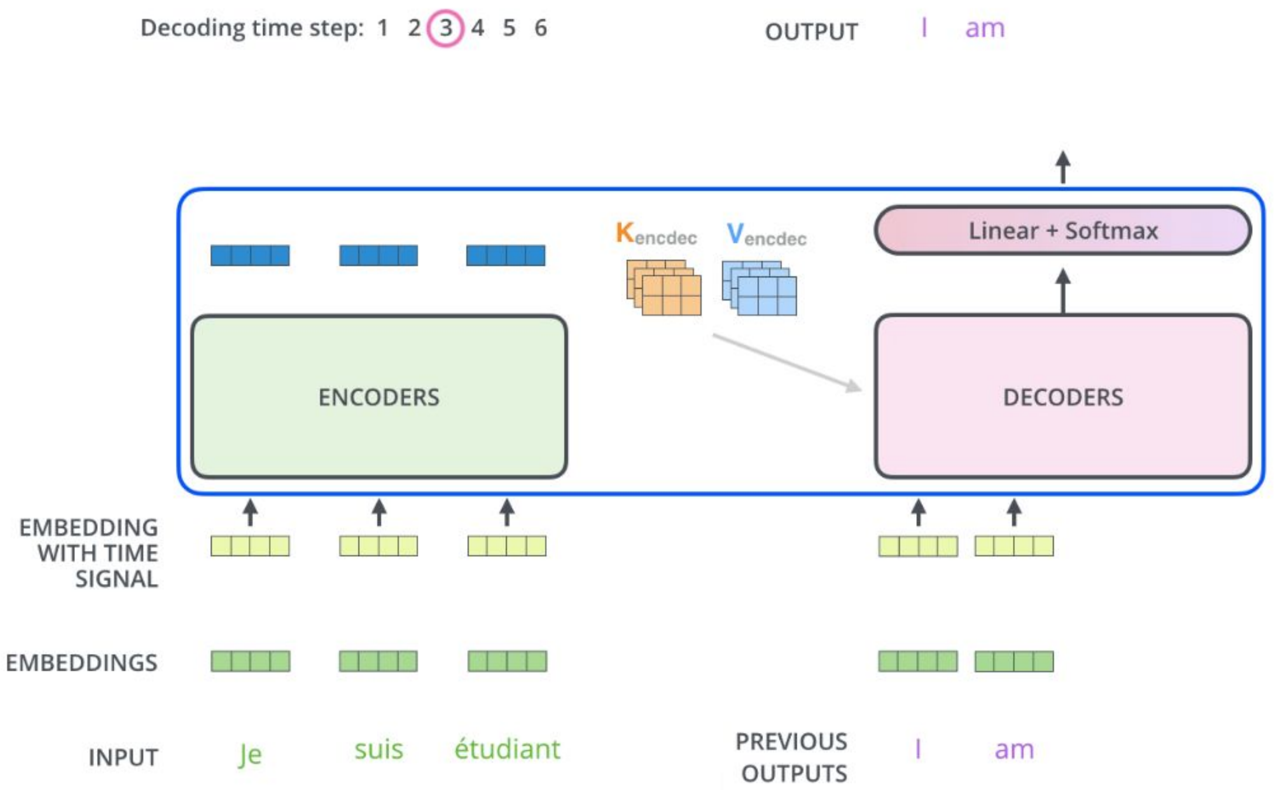
1. Correspondences are not 1:1.
2. The units of translation are not words, but concepts.
3. Words-to-concepts is not 1:1, because
   a. Synonymy
   b. Ambiguity
   c. Rich morphology and word order



Supervised Encoder-Decoder Machine Translation


input
encoder
decoder
target


मलाई
चराहरू
मन
पर्छ
<end>
Softmax
Softmax
Softmax
Softmax
Softmax
FC
FC
FC
FC
FC
Context Vector
I
Like
Birds
<start>
मलाई
चराहरू
मन
पर्छ
Encoder
Decoder

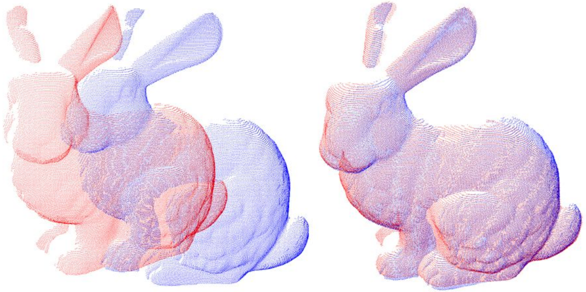


Decoding time step: 1 2 3 4 5 6
OUTPUT
I am
Linear + Softmax
Kencdec
Vencdec
ENCODERS
DECODERS
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je suis étudiant
PREVIOUS OUTPUTS
I am

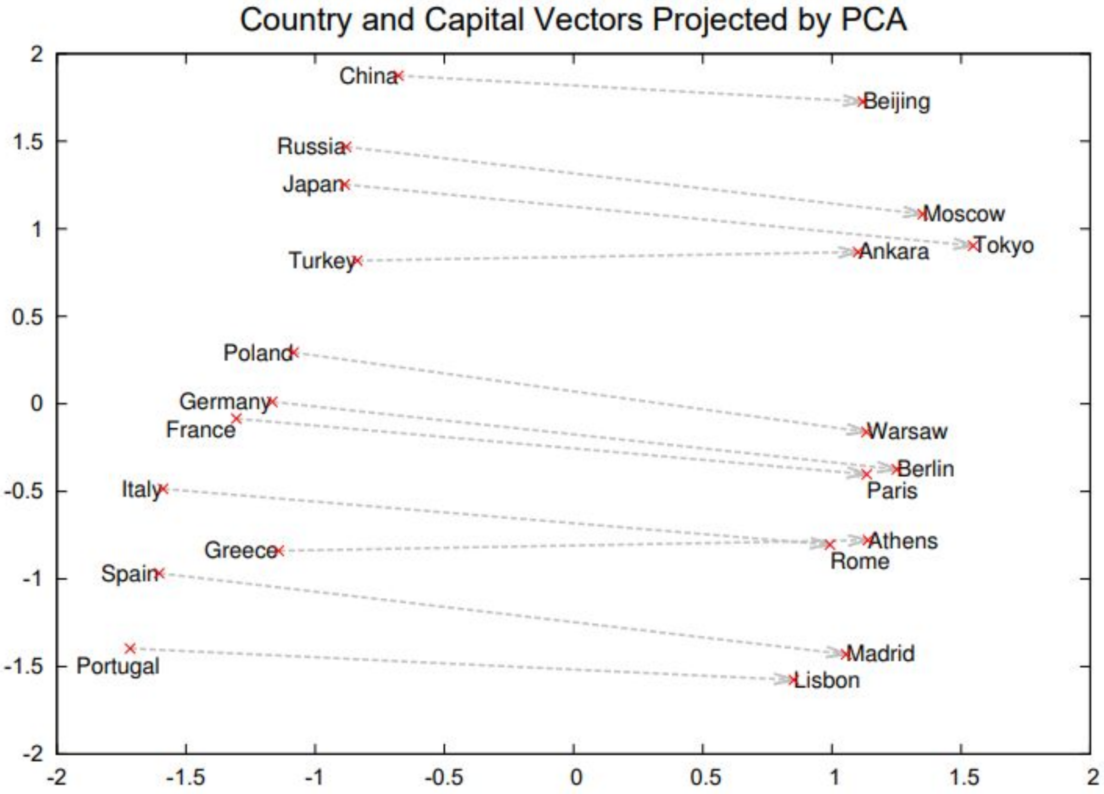
Unsupervised Machine Translation?

Multilingual language models

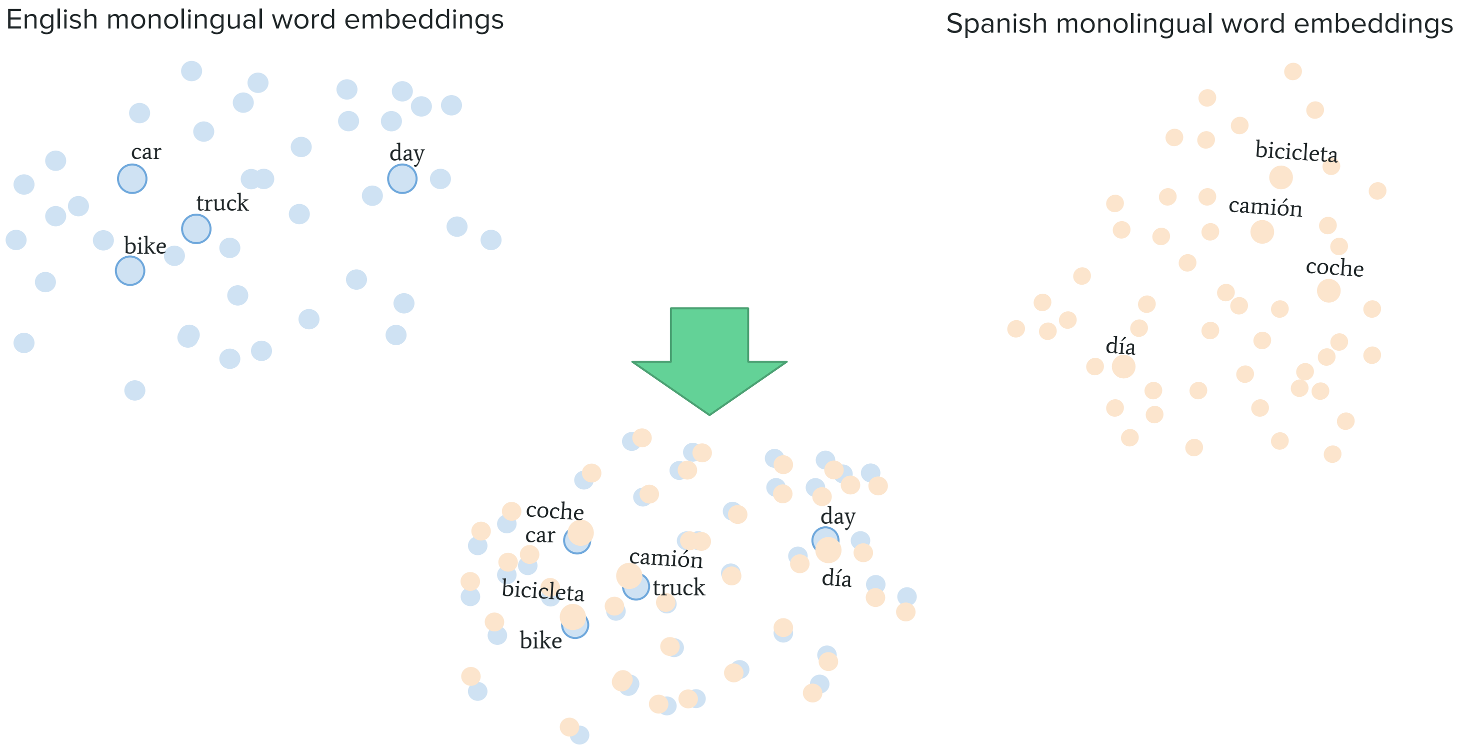
Observation: Because relations are encoded systematically, in the limit language-specific embedding spaces will be isomorphic. This means we can learn linear mappings to align them.



(a) (b)



Cross-Lingual Word-Embedding **Alignment**



Alignment

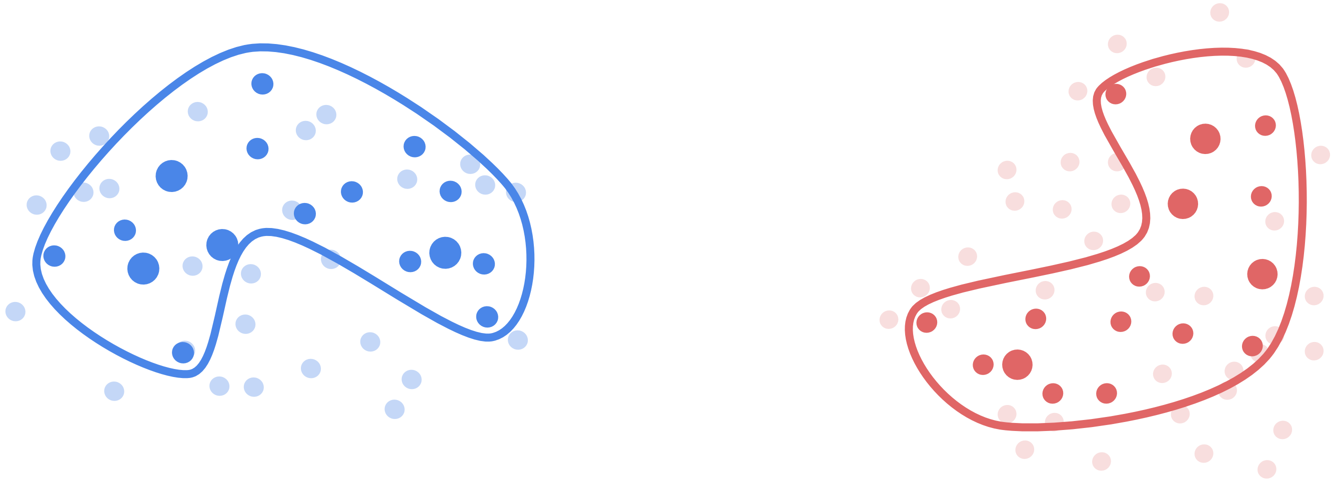
Train dictionary:

car	coche
truck	camión
bike	bicicleta
day	día
	…

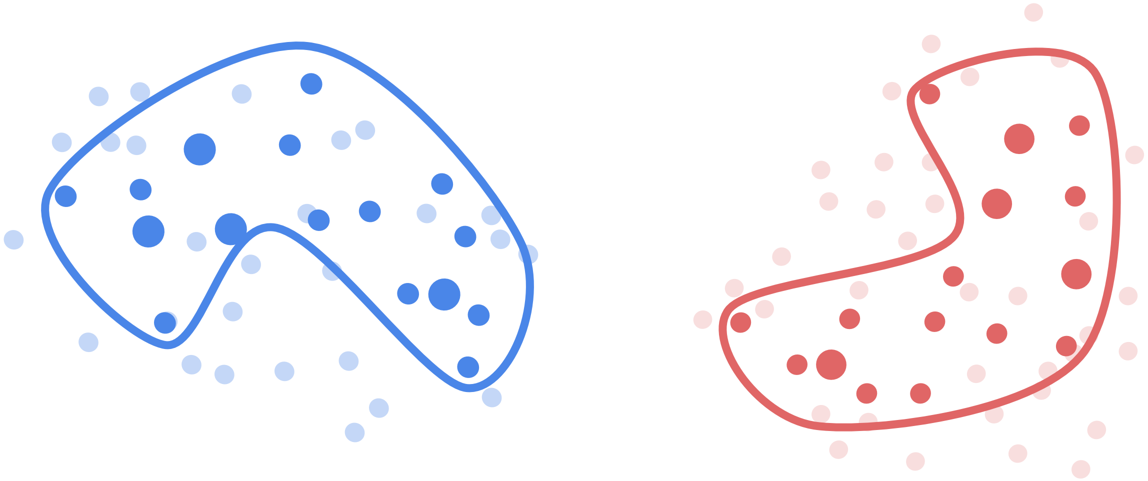
…

Alignment

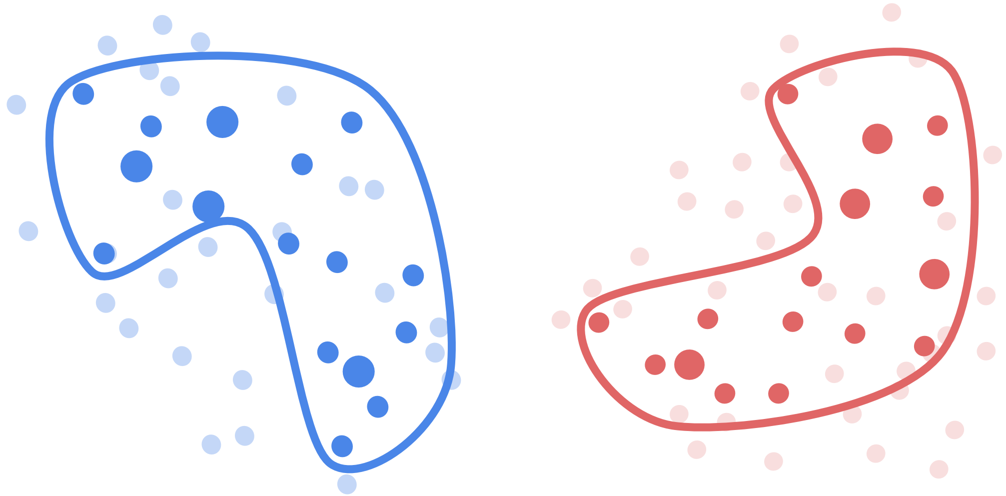
Alignment



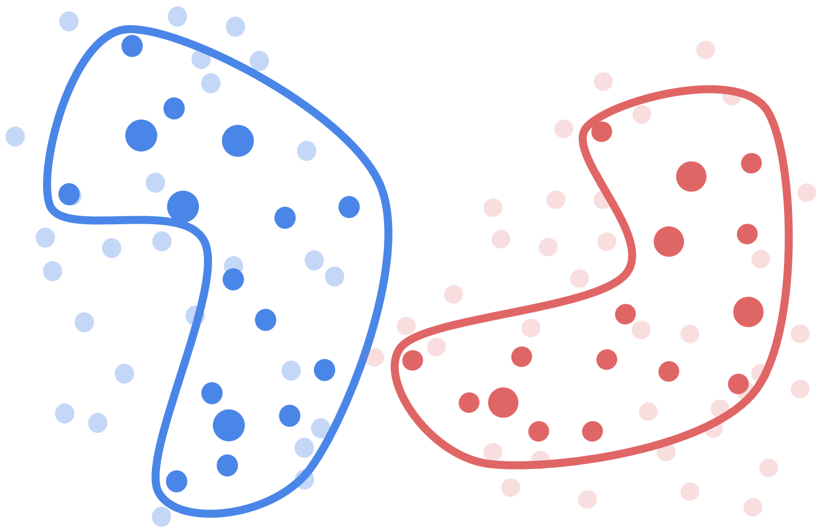
Alignment



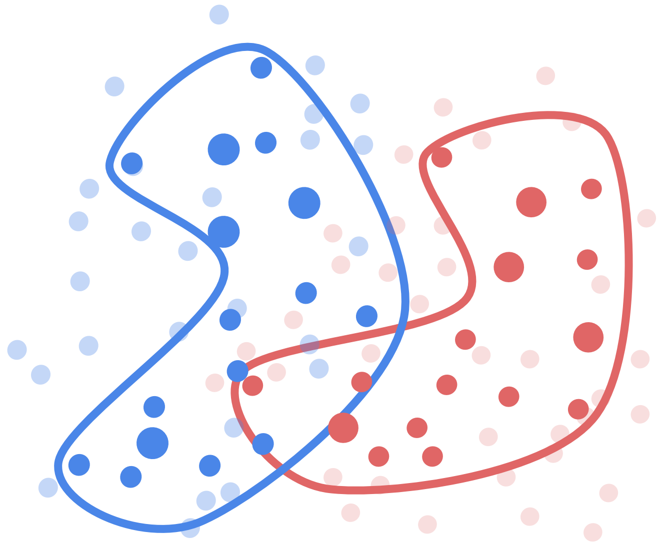
Alignment



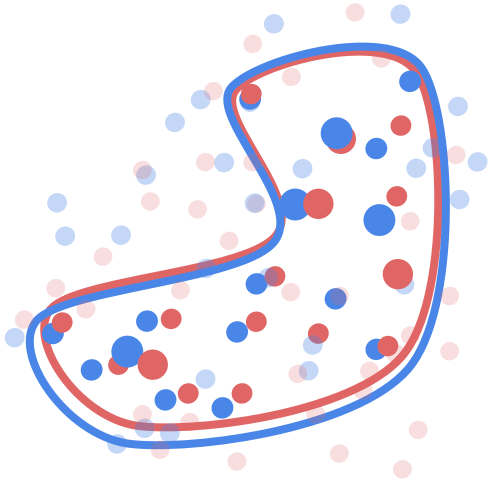
Alignment



Alignment



Alignment



Alignment

X'

car	-3.2	1.2
truck	-3.4	-1.1
bike	-2.9	0.5
day	2.3	1.3
...	...	...

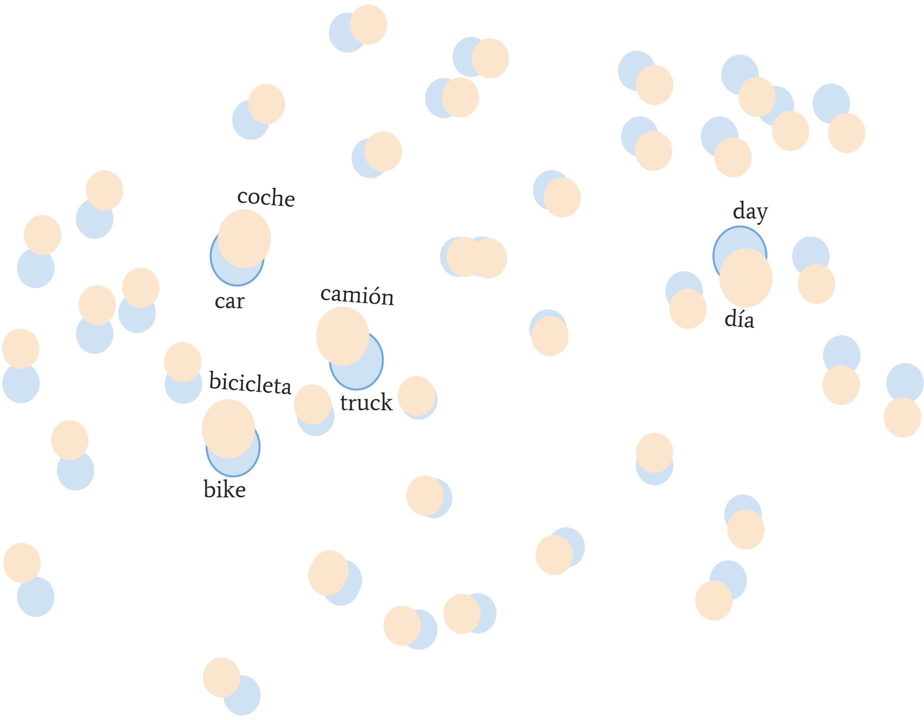
Z'

coche	1.2	-2.3
camión	1.7	-1.1
bicicleta	1.4	3.1
día	4.2	2.1
...	...	...

Alignment

$$
\begin{array}{c}
X' \\
\begin{array}{l|cc}
\text{car} & -3.2 & 1.2 \\
\text{truck} & -3.4 & -1.1 \\
\text{bike} & -2.9 & 0.5 \\
\text{day} & 2.3 & 1.3 \\
\ldots & \ldots & \ldots
\end{array}
\end{array}
\quad
\begin{array}{c}
W \\
\begin{array}{|c|c|}
\hline
? & ? \\
\hline
? & ? \\
\hline
\end{array}
\end{array}
\quad = \quad
\begin{array}{c}
Z' \\
\begin{array}{l|cc}
\text{coche} & 1.2 & -2.3 \\
\text{camión} & 1.7 & -1.1 \\
\text{bicicleta} & 1.4 & 3.1 \\
\text{día} & 4.2 & 2.1 \\
\ldots & \ldots & \ldots
\end{array}
\end{array}
$$

Bilingual Dictionary Induction (BDI)



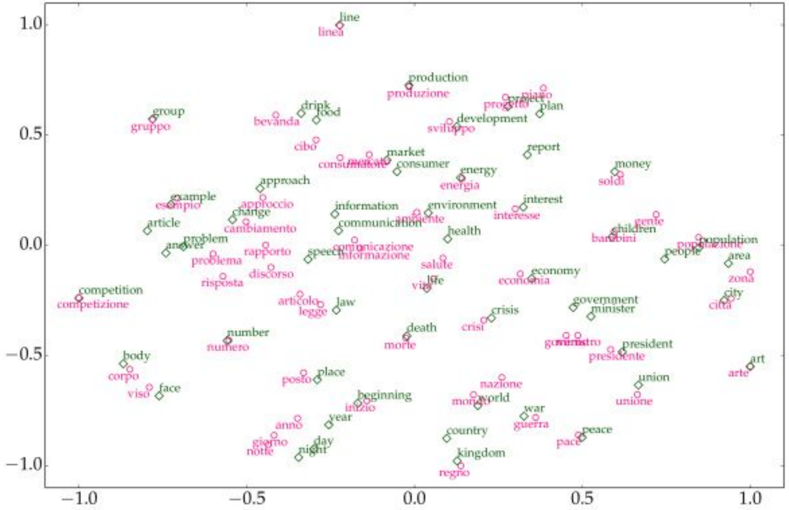
car coche
bike bicicleta
truck camión
day día

table mesa
toy juguete
hour hora

Unsupervised machine translation

Unsupervised machine translation begins with vocabulary alignment, using point set registration algorithms. Once you know that ‘line’ and ‘linea’ are co-referential, we can begin to translate.

Key idea: If LM and CV models were aligned in the same way, we could translate and do VQA.



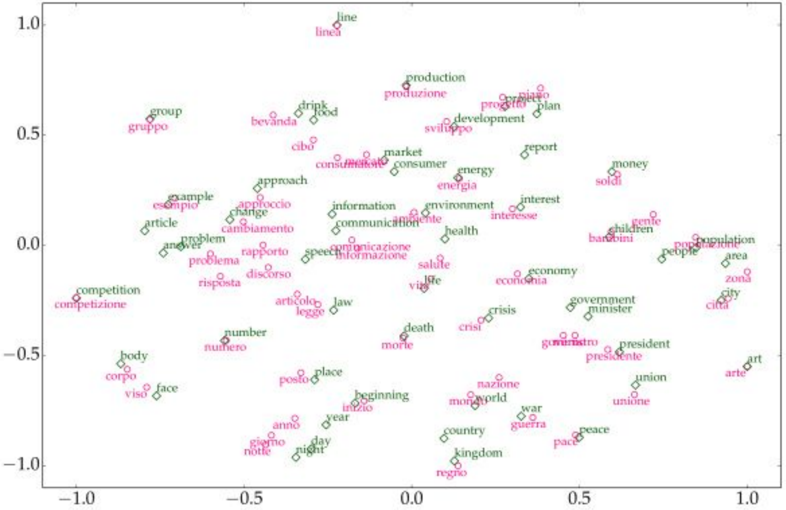
Unsupervised machine translation

Unsupervised machine translation begins with vocabulary alignment, using point set registration algorithms. Once you know that ‘line’ and ‘linea’ are co-referential, we can begin to translate.

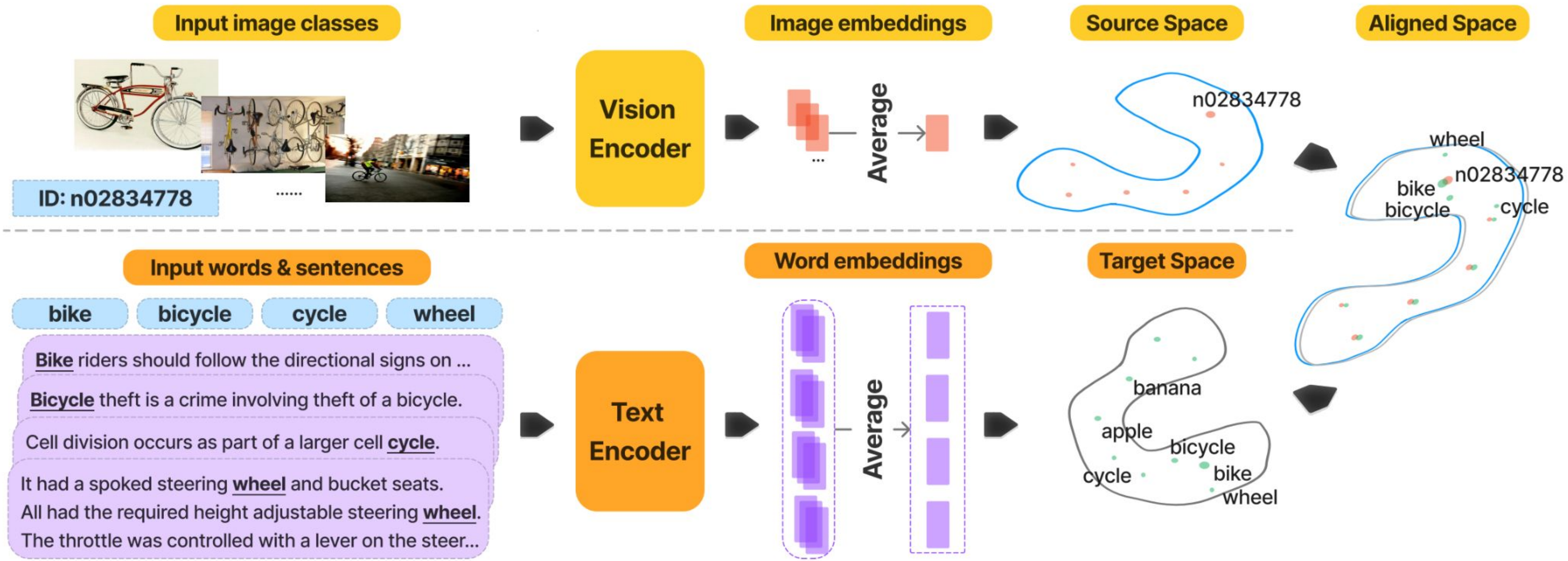
Key idea: If LM and CV models were aligned in the same way, we could translate and do VQA.

Learned lesson: Unsupervised alignment (e.g., using GANs) only work when spaces are *very* similar.

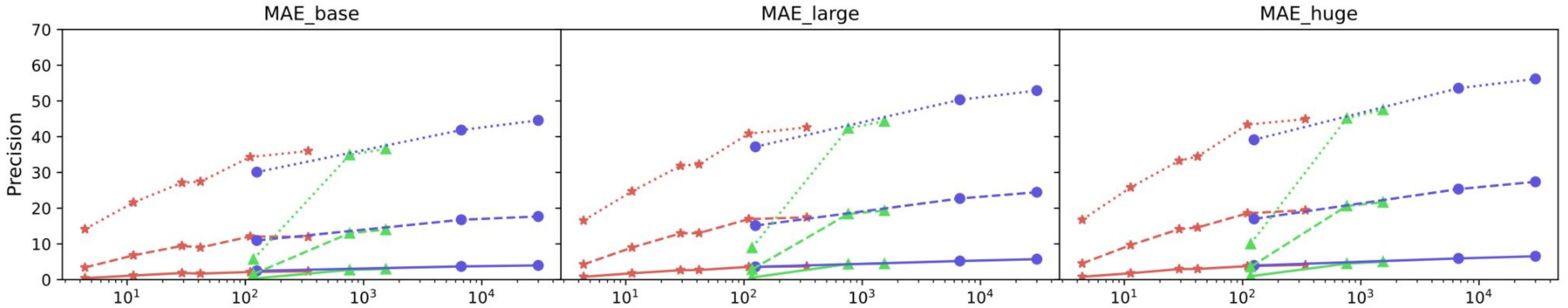
	Unsupervised (Adversarial)	**Supervised** (Identical)
EN-ES	81.89	**82.62**
EN-ET	00.00	**31.45**
EN-FI	00.09	**28.01**
EN-EL	00.07	**42.96**
EN-HU	45.06	**46.56**
EN-PL	46.83	**52.63**
EN-TR	32.71	**39.22**
ET-FI	**29.62**	24.35



Søgaard et al. (2019): ‘**On the limits of unsupervised bilingual dictionary induction**. ACL.



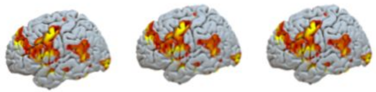
Input image classes
ID: n02834778
Vision Encoder
Image embeddings
Average
Source Space
n02834778
Aligned Space
wheel
bike
n02834778
bicycle
cycle
Input words & sentences
bike
bicycle
cycle
wheel
Bike riders should follow the directional signs on ...
Bicycle theft is a crime involving theft of a bicycle.
Cell division occurs as part of a larger cell cycle.
It had a spoked steering wheel and bucket seats.
All had the required height adjustable steering wheel.
The throttle was controlled with a lever on the steer...
Text Encoder
Word embeddings
Average
Target Space
banana
apple
bicycle
cycle
bike
wheel



⋆ BERT models, ▲ GPT2 models, ● OPT models; Dotted line: P@100, dashed line: P@10, solid line: P@1.

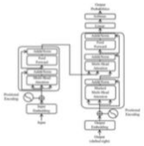
fMRI obtained while participants read or listen to language.

fMRI vectorized and aligned at word level, through Gaussian smoothing.



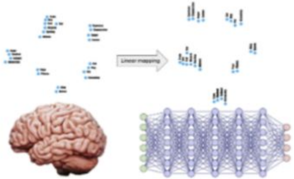
$f_{letter} = \langle 0.88373, 0.75805, 0.00328, \ldots \rangle$

Decontextualized word embeddings obtained from LLMs.

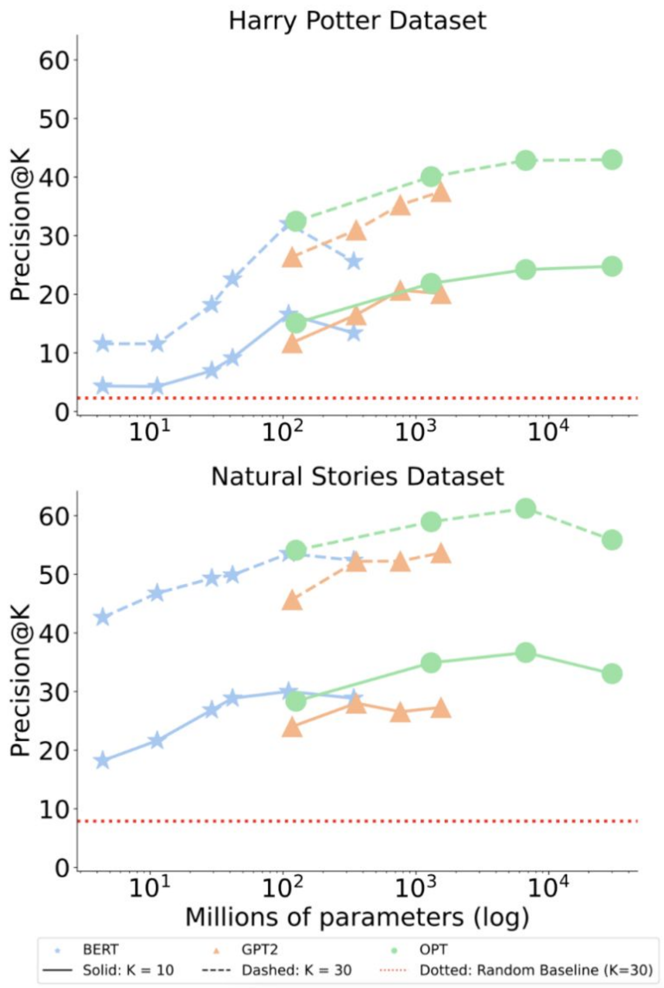


$\mathbf{v}_{letter} = \langle 0.14723, 0.16827, 0.00328, \ldots \rangle$

Alignment with Procrustes Analysis or ridge regression.



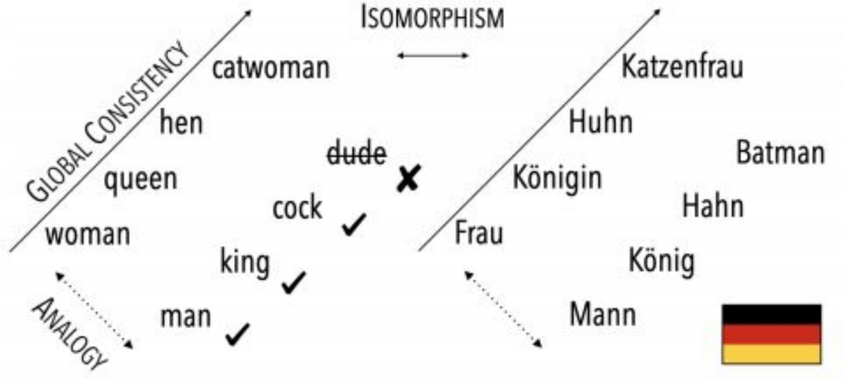
Results, retrieval precision at 50% (see plots ⟶).



Analogy training multilingual language models

Idea: Use knowledge bases to ground language models, thereby debiasing them.

Collaboration with Deepmind and Cambridge University.



Language	Baselines				Analogy Training					
	Fasttext		mBERT		Fasttext		mBERT-WiQueen		mBERT-WiQueen$^+$	
	P@1	ρ	P@1	ρ	P@1	ρ	P@1	ρ	P@1	ρ
Danish	0.1511	0.3001	0.2835	0.3221	0.1688	0.2909	0.3863	0.3010	**0.3935**	**0.2461**
German	0.0997	0.3604	0.2658	0.3548	0.1104	0.3702	0.3894	0.3257	**0.4538**	**0.2868**
English	0.1255	0.2854	0.2897	0.3107	0.1513	0.2550	0.4091	0.2960	**0.4787**	0.2821
Spanish	0.0899	0.3383	0.2596	0.3441	0.1194	0.3573	0.3832	0.3198	**0.3936**	**0.3012**
Finnish	0.1258	0.3908	0.2679	0.3535	0.1682	0.3731	0.3728	0.3192	**0.4019**	**0.2703**
French	0.0943	0.3659	0.2617	0.3545	0.1146	0.3459	0.3707	0.3375	**0.4195**	**0.2991**
Italian	0.0731	0.3979	0.2773	0.3711	0.0949	0.3883	0.3821	**0.3338**	**0.4372**	0.3722
Dutch	0.1291	0.3520	0.2669	0.3443	0.1497	0.3384	0.3811	**0.3202**	**0.4424**	0.3609
Polish	0.1165	0.3397	0.2648	0.3656	0.1456	0.3287	0.3853	0.3468	**0.3894**	**0.2718**
Portuguese	0.0898	0.3614	0.2523	0.3536	0.1072	0.3640	0.3697	0.3409	**0.3718**	**0.2653**
Swedish	0.1071	0.3449	0.2856	0.3378	0.1415	0.3270	0.3832	0.3128	**0.4071**	**0.2672**
Averages	0.1093	0.3488	0.2704	0.3435	0.1338	0.3399	0.3830	0.3231	**0.4171**	**0.2930**

Garneau et al. (2019): '**Analogy Training Multilingual Encoders**'. AAAI.

?