Juan Antonio Pérez-Ortiz

Antonio Valerio Miceli Barone

LESSONS LEARNED

HOW THE GOURMET PROJECT DEVELOPED A MACHINE TRANSLATION MODEL FOR A LOW-RESOURCE LANGUAGE IN 8 WEEKS

THE GOURMET.EU PROJECT

- Models and resources for neural machine translation (NMT) between English and low-resource languages.
- Integration into tools for media analysts and journalists.
- Systems already developed for Gujarati, Bulgarian, Turkish, Swahili, Amharic, Kyrgyz, Serbian, Tamil, Hausa, Macedonian, Igbo, Tigrinya, Pashto. More to come!

gourmet

THE UNIVERSITY of EDINBURGH

Universitat d'Alacant
Universidad de Alicante

UNIVERSITEIT VAN AMSTERDAM

BBC

DW

SURPRISE LANGUAGE CHALLENGE

- Inspiration: US DARPA events.
- Simulate sudden need of assimilation or dissemination of information in regions of the world with languages not included in the digital workflows.
- Pashto was chosen by BBC and DW as a language of their interest that complements the goals of the project.
- NMT development period: February-March 2021.

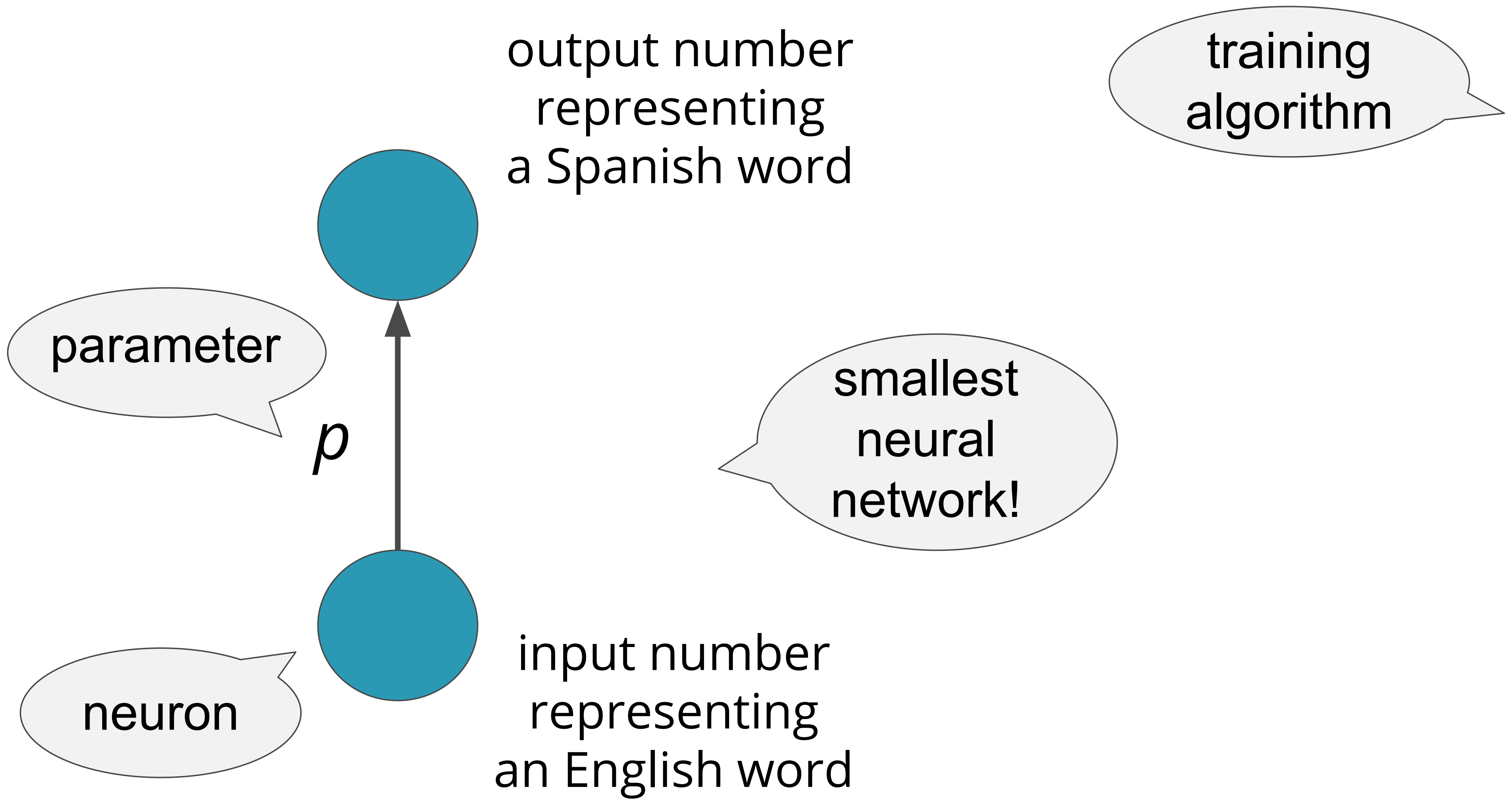
PASHTO

Rich morphology

Indo-Iranian language

45-50 million speakers worldwide

HOW NMT WORKS 101



training set

English	input number	Spanish	output number
tomato	1	tomate	2
red	4	rojo	8

Initialize p to a random number, for example, p = 1.2
Training starts!

Epoch 1
if the input is *tomato*, the neural network produces…
1 x p = 1 x 1.2 = 1.2
It should have been 2
Error is 2 - 1.2 = 0.8
If the input is *red*, the neural network produces…
4 x p = 4 x 1.2 = 4.8
It should have been 8
Error is 8 - 4.8 = 3.2
Total error is 0.8 + 3.2 = 4
A mathematical optimizer uses the error to find a better p, say, p = 1.5

Epoch 2
Evaluate the error again but with the updated p
1 x p = 1 x 1.5 = 1.5 (error: 2 - 1.5 = 0.5)
4 x p = 4 x 1.5 = 6 (error: 8 - 6 = 2)
Total error is 0.5 + 2 = 2.5 (smaller!)
A mathematical optimizer uses the error to find a better p, say, p = 1.72

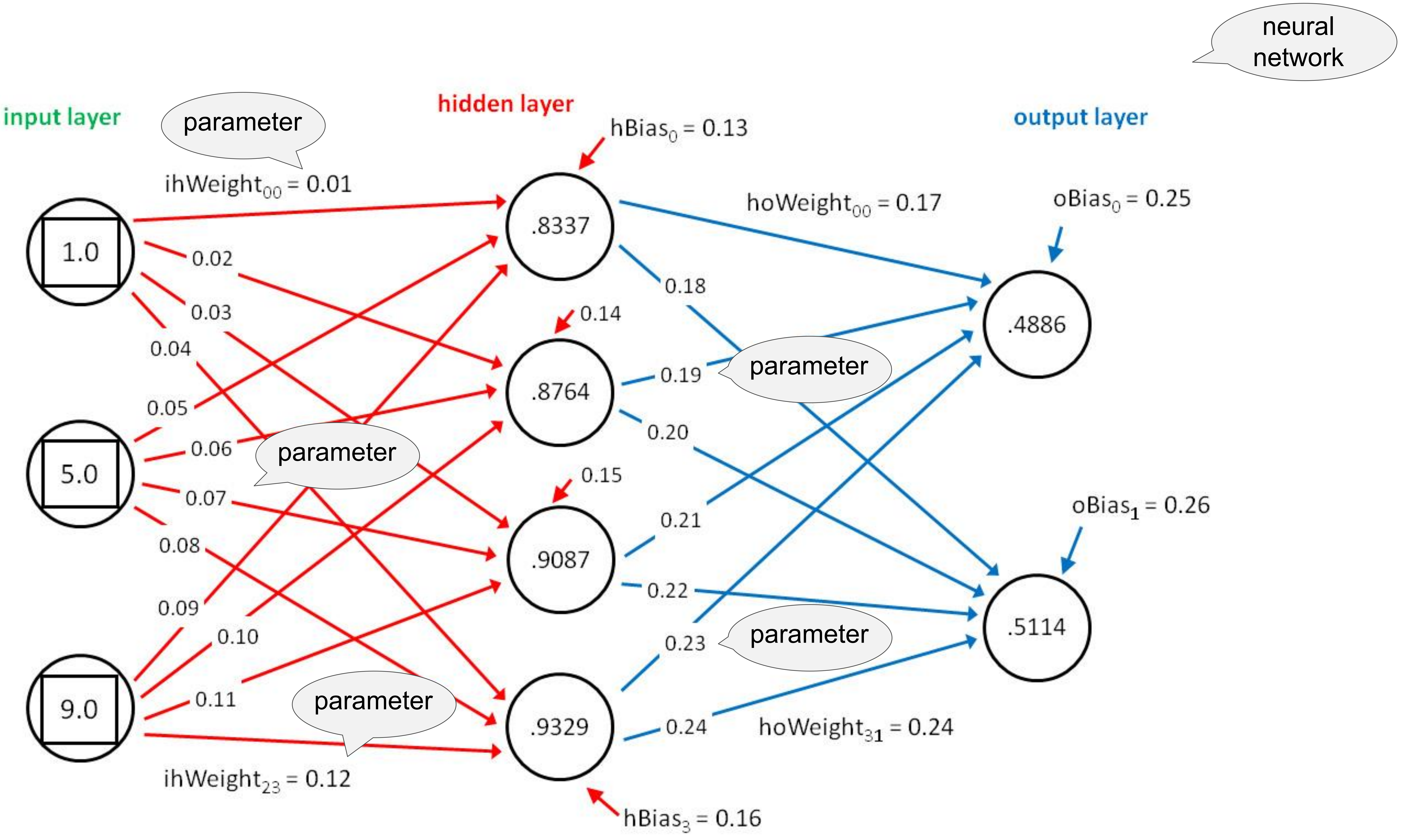
Epoch 3
…
Epoch 4
…

LITTLE WHITE LIES

- An error of 0 is usually not attained, not even desirable.
- We want sentence translation, not word-for-word.
- Real neural networks may have billions of parameters.
- Training may take centuries on a desktop computer.
- Words are not represented with a single number.
- Outputs are not words, but probabilities over all words.

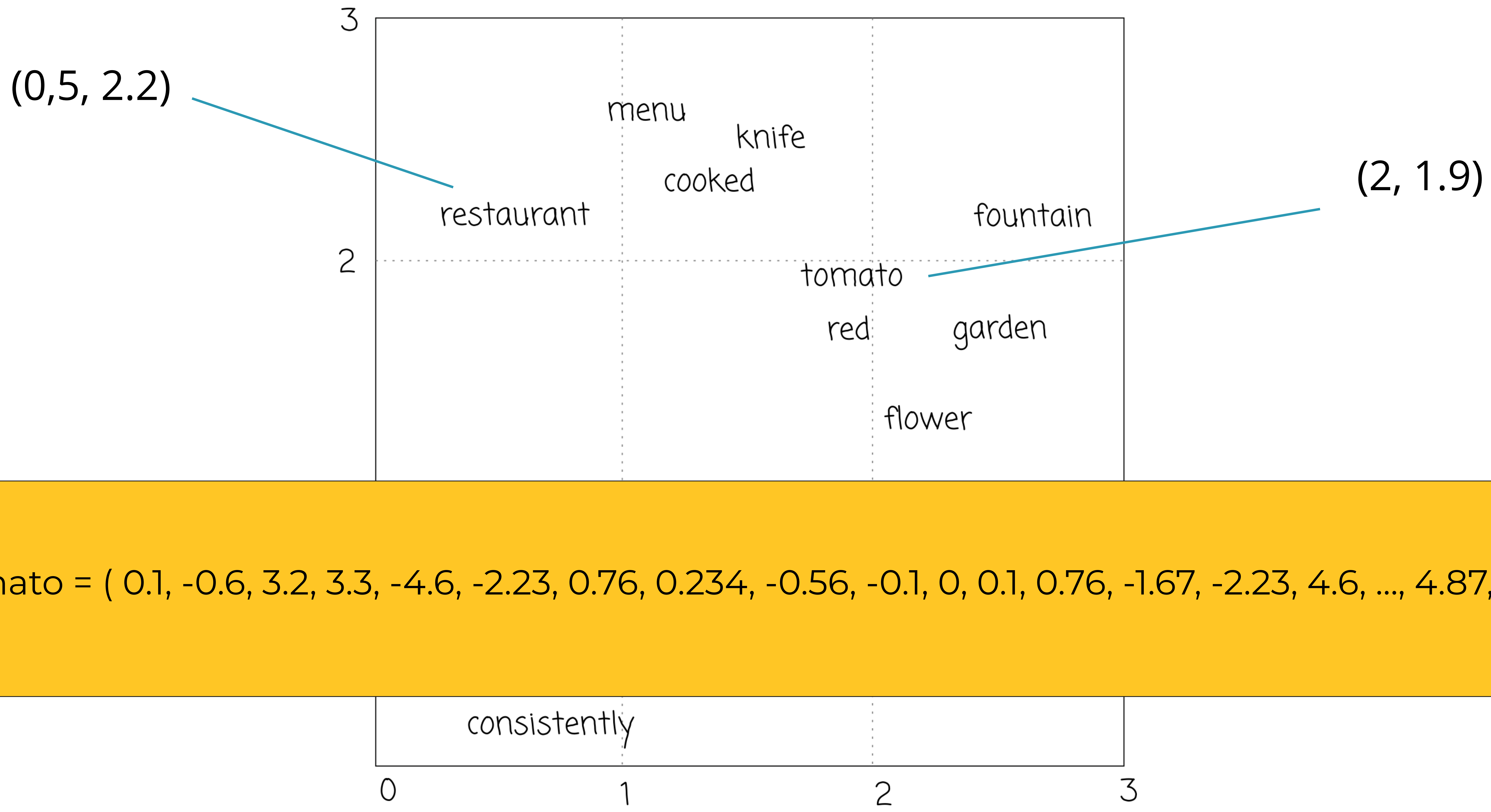
sentence level training set

English	input numbers	Spanish	output numbers
Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much.	???	El señor y la señora Dursley, que vivían en el número 4 de Privet Drive, estaban orgullosos de decir que eran muy normales, afortunadamente.	???
All human beings are born free and equal in dignity and rights.	???	Todos los seres humanos nacen libres e iguales en dignidad y derechos.	???
...	...	...	...



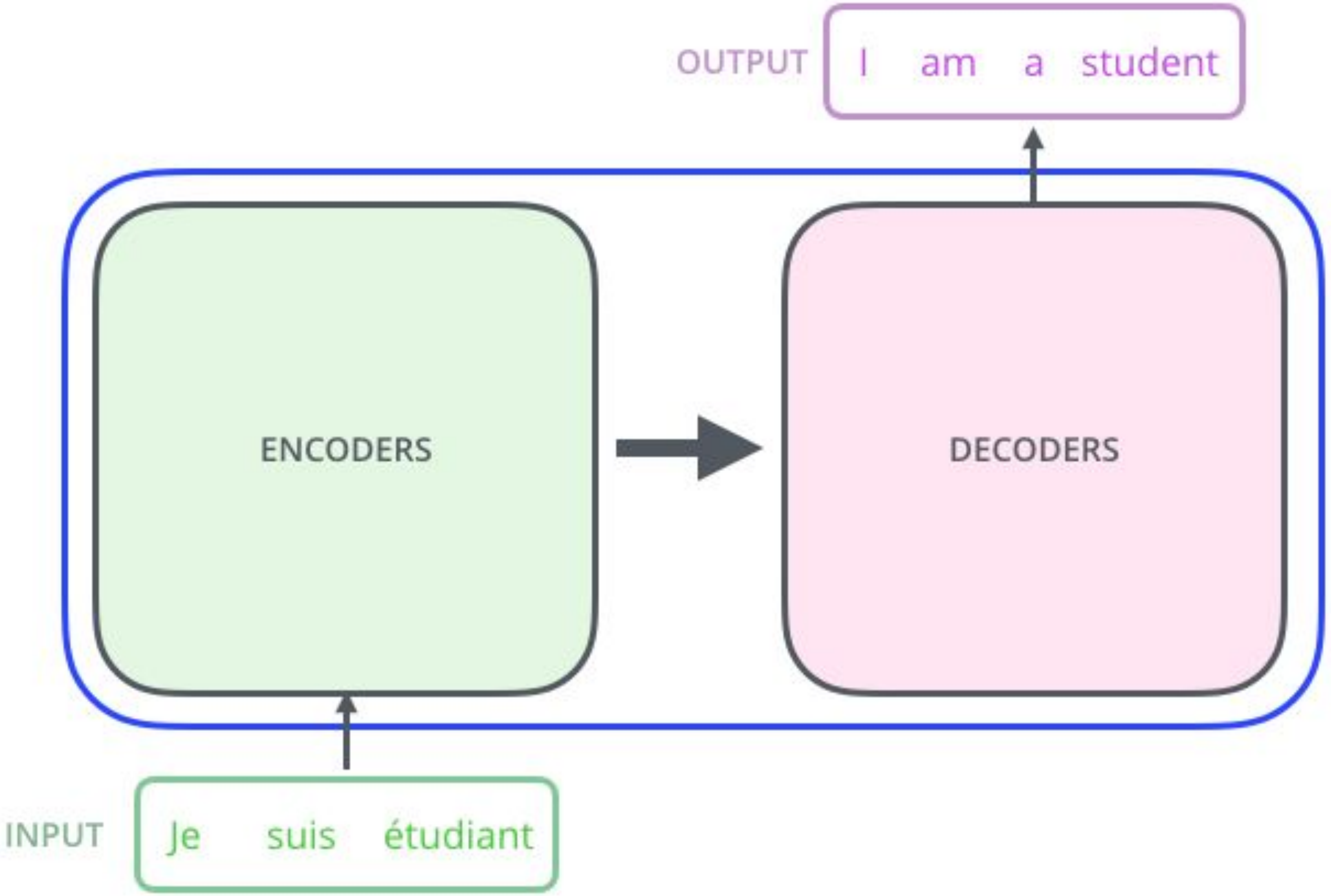
Source: “Use Python with Your Neural Networks”, James McCaffrey

EMBEDDINGS



tomato = (0.1, -0.6, 3.2, 3.3, -4.6, -2.23, 0.76, 0.234, -0.56, -0.1, 0, 0.1, 0.76, -1.67, -2.23, 4.6, ..., 4.87, 5.34)

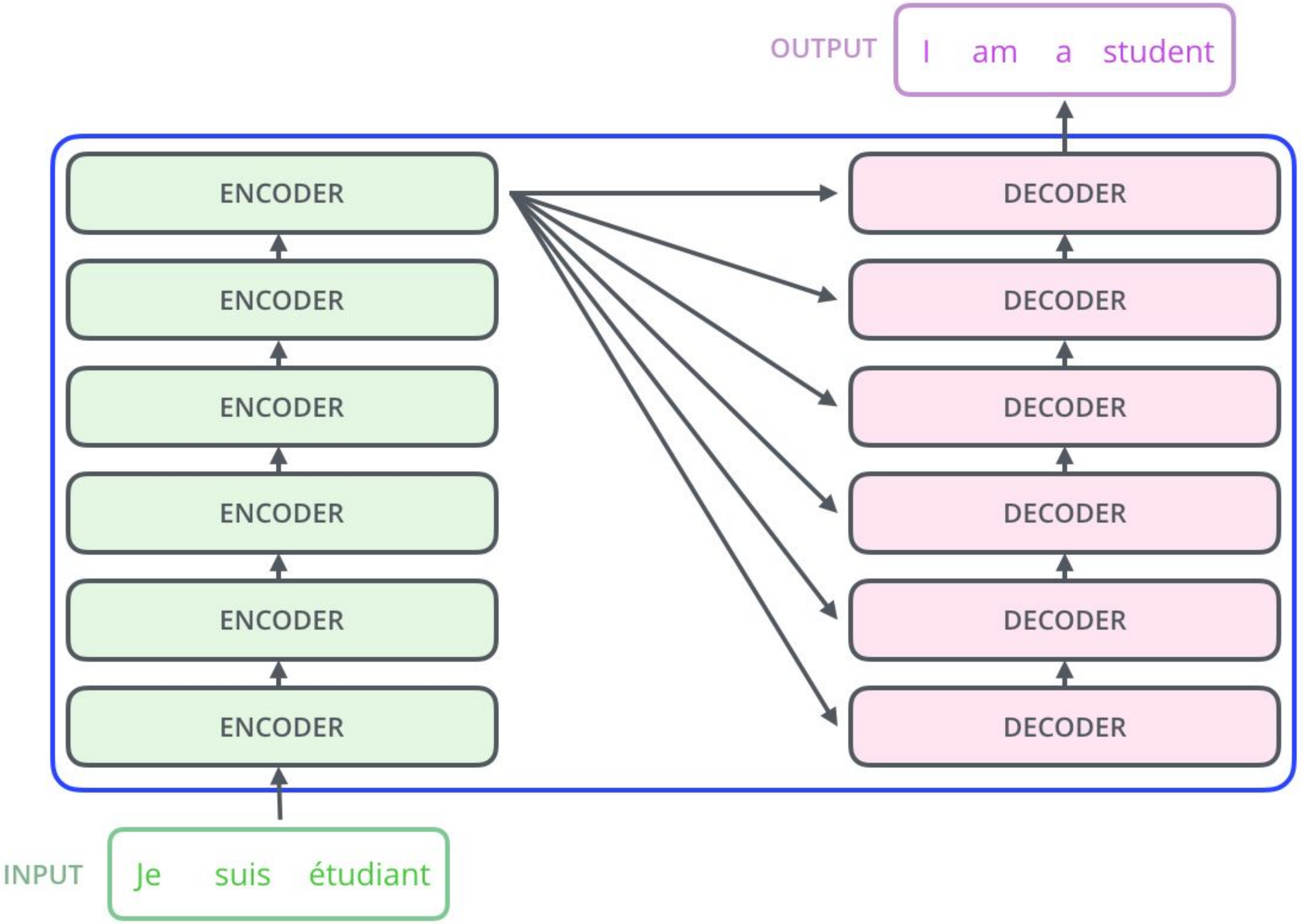
TRANSFORMER



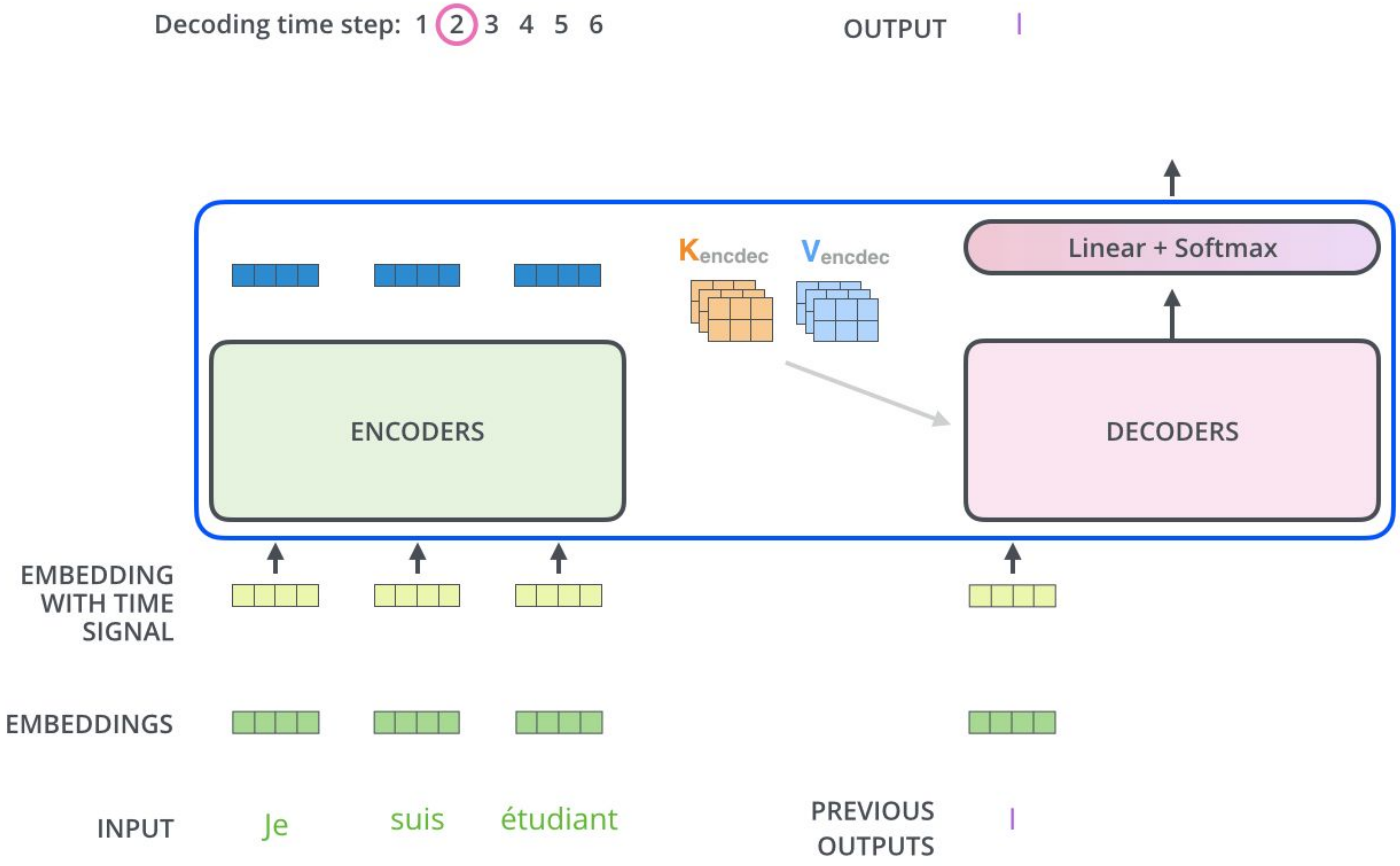
neural network

Source: "The Illustrated Transformer", Jay Alammar

TRANSFORMER

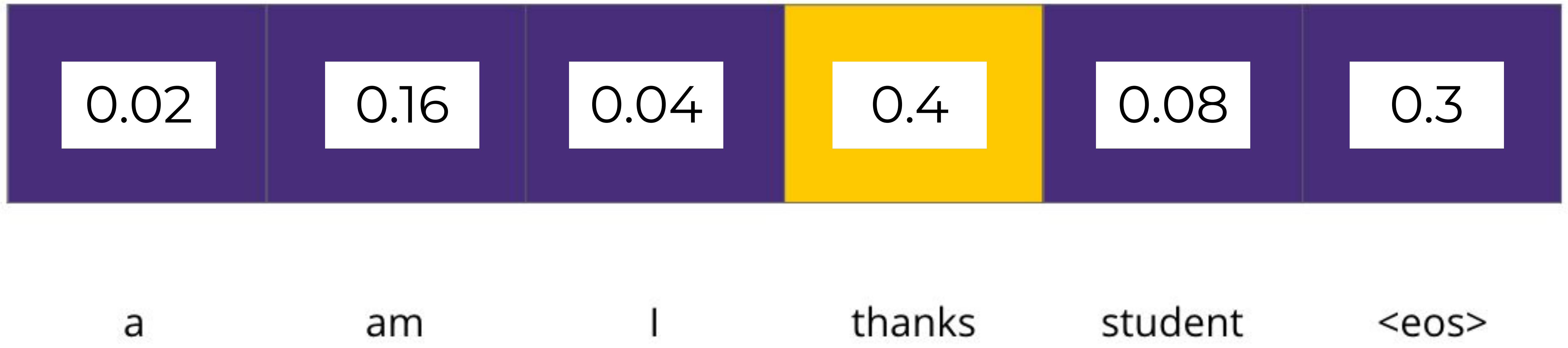


Source: “The Illustrated Transformer”, Jay Alammar



Source: "The Illustrated Transformer", Jay Alammar

OUTPUT PROBABILITIES



a	am	I	thanks	student	<eos>
0.02	0.16	0.04	0.4	0.08	0.3

Source: "The Illustrated Transformer", Jay Alammar

THE HOLY GRAIL: PARALLEL TEXT

DATA DOWNLOADING & CRAWLING

- Existing collections (OPUS, ParaCrawl...) vs ad-hoc crawling.
- Monolingual data: crawling ↦ language identification ↦ sentence splitting.
- Bilingual data: crawling ↦ language identification ↦ document alignment ↦ sentence splitting ↦ sentence-level alignment.
- Tools: LinguaCrawl, Bitextor.

DATA DOWNLOADING & CRAWLING

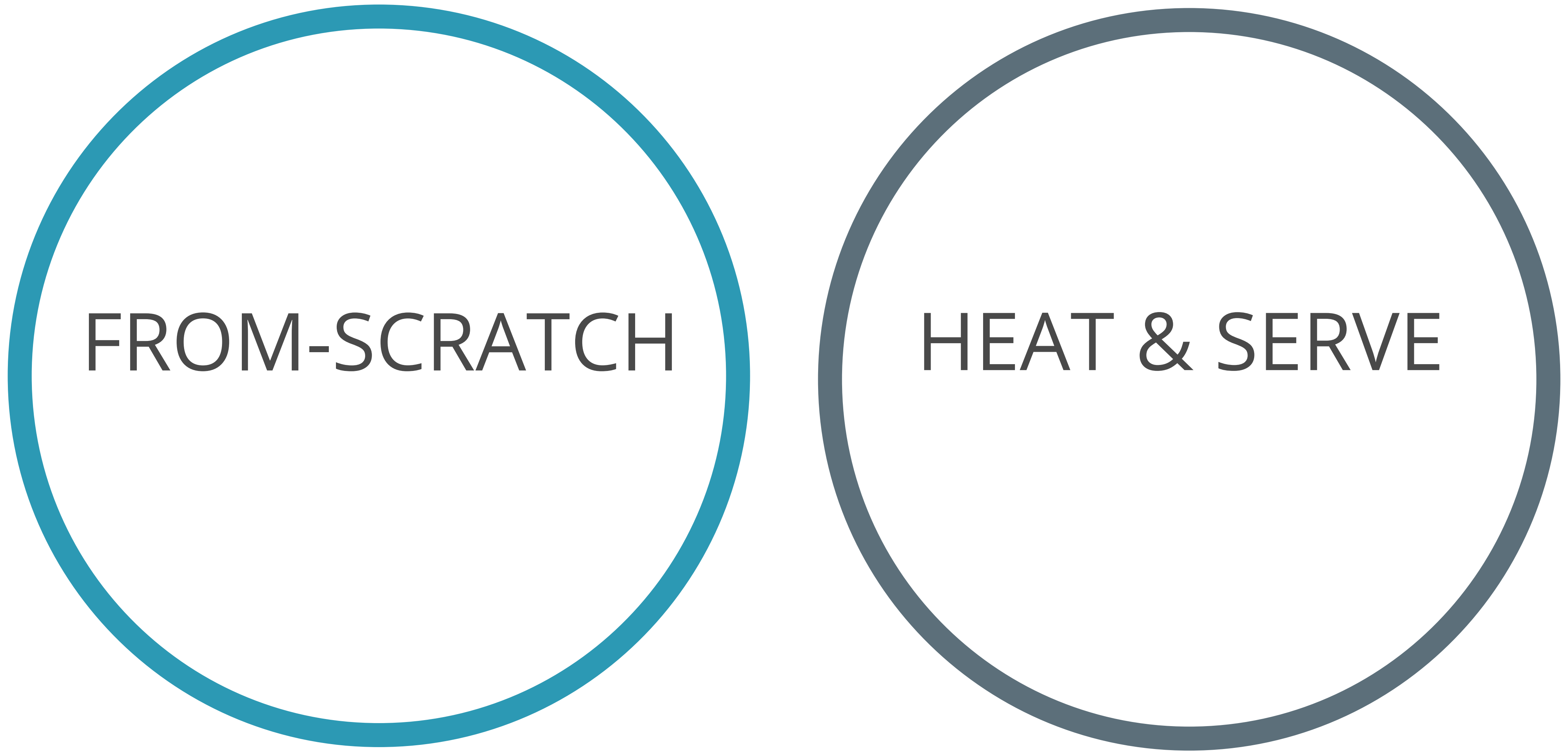
- Crawling origins: manual list of 50 websites + 138 websites of the internet top-level domain `.af` ↦ **60,000** English-Pashto bilingual sentences.

- Already existing parallel corpora: **340,000** sentences.

- **3,000** parallel sentences (news domain) manually checked by our media partners and used as development and test sets.

NMT MODELS

DEVELOPED NMT MODELS



FROM-SCRATCH

HEAT & SERVE

FROM-SCRATCH MODEL

MAKE USE OF ALL AVAILABLE DATA

- Good machine translation requires millions of parallel sentence pairs.

- We only have tens of thousands English-Pashto sentence pairs.

- But we have other types of data:

  - ❑ Parallel data between English and other languages.
  - ❑ “Monolingual” data (raw text).

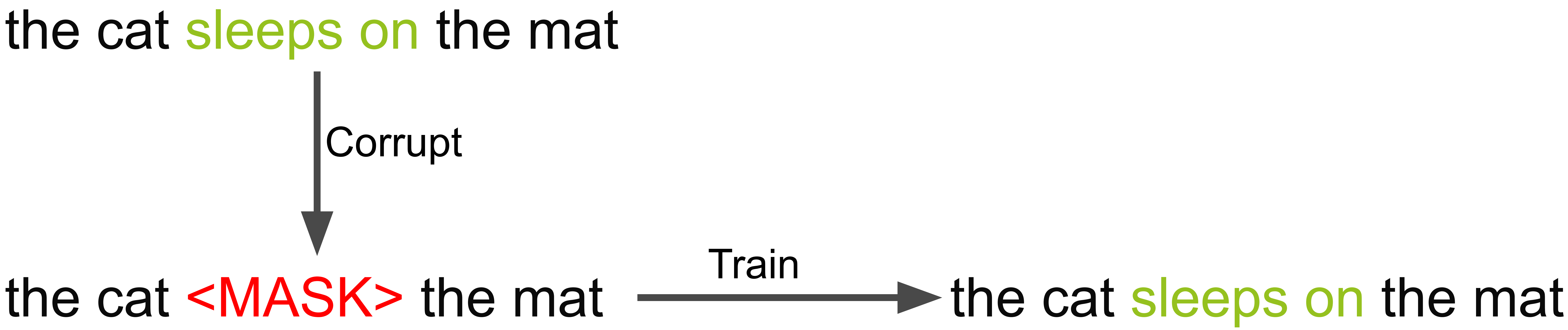
MAKE USE OF ALL AVAILABLE DATA

- Neural models can be trained on a “curriculum” of related tasks:
  - ❏ Start training on tasks that are easier or have more training data.
  - ❏ Fine-tune on the final task of interest (English-Pashto translation).
- Pre-training tasks:
  - ❏ Monolingual gap-filling.
  - ❏ English-German machine translation.

GAP-FILLING PRETRAINING

- Corrupt monolingual sentences by masking spans of words.

- Train the model to reconstruct the original sentences from the corrupted input.



the cat sleeps on the mat

Corrupt

the cat <MASK> the mat

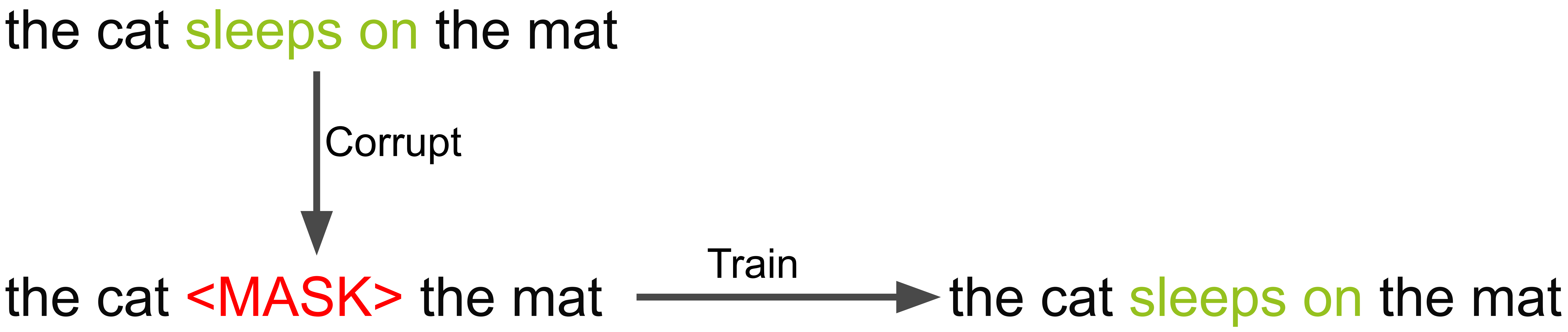
Train

the cat sleeps on the mat

GAP-FILLING PRETRAINING

- We train on gap-filling on both English and Pashto “monolingual” data.

- The model learns how these two languages work, but not how to translate between them ↦ “mBART” approach.



the cat sleeps on the mat

Corrupt

the cat <MASK> the mat

Train

the cat sleeps on the mat

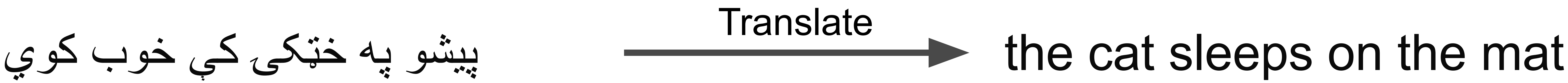
ENGLISH-GERMAN PRETRAINING

- Train to translate between English and some other language(s).
- Ideally the other language(s) should be related to Pashto and high-resource:
  - ❑ Not really available.
- We pretrain on English-German:
  - ❑ English-German is high resource and very well studied.

TRAIN ON SYNTHETIC DATA

- Once we have pre-trained and fine-tuned we can further exploit monolingual data by “back-translation”.

- Translate Pashto sentences to English:

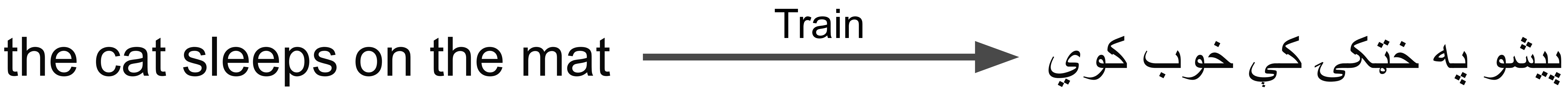


پیشو په خټکۍ کې خوب کوي

Translate

the cat sleeps on the mat

- Flip it around and use it as English ↦ Pashto parallel data:



the cat sleeps on the mat

Train

پیشو په خټکۍ کې خوب کوي

- Repeat the process in the other direction.

TRAIN ON SYNTHETIC DATA

- With “back-translation” training, the model is always trained to generate natural output sentences, although the inputs are synthetic and can contain errors.
- Not as good as training on the same amount of true parallel data.
- But there is much more monolingual data than parallel data, especially for English.

COMBINED APPROACH

- We run multiple iterations of generating back-translation data and training on this data + true parallel data.
- We start each run from a model pre-trained on either “mBART” gap-filling or English-German translation.
- We do a total of 4 rounds from mBART pre-training followed by 2 rounds from English-German pretraining.

HEAT & SERVE MODEL

PRE-TRAINED MODELS

- Large already-trained neural networks available for download.
- Different models available: BERT, BART, GPT-3, T5...
- Multilingual versions: mBERT, mBART, **mBART50**.
- Universal representations arise.
- Fine-tuning (heat-and-serve): fast training starting from the pre-trained model.

mBART50

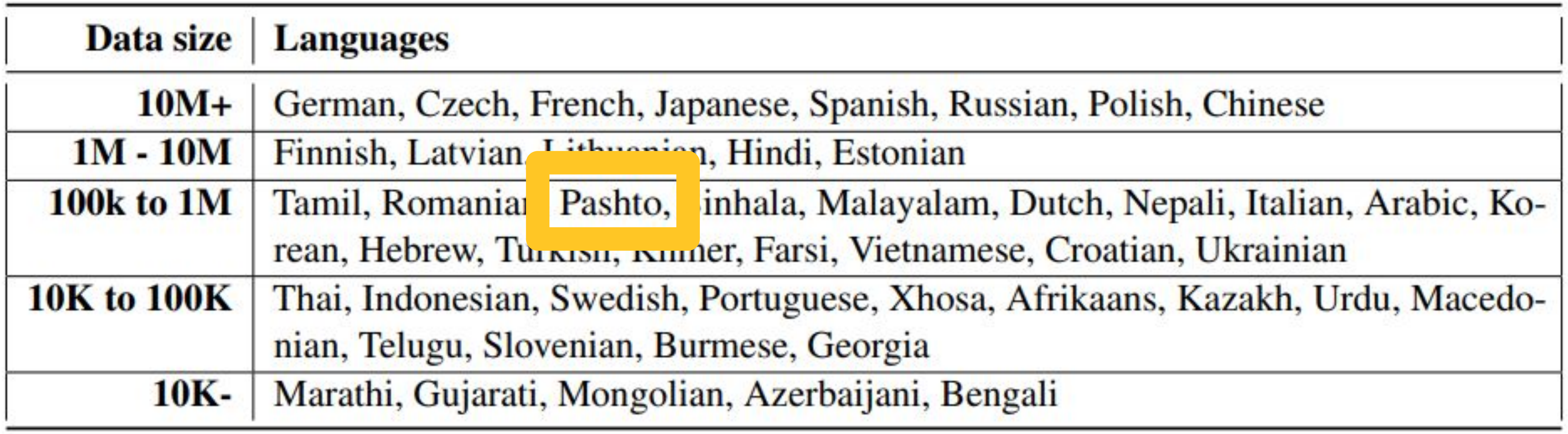
- Pre-trained model released by Facebook on January 2021.
- Transformer first trained mBART-style with monolingual data and then trained to translate between English and 49 languages (both directions).
- Fine-tuning (heat-and-serve) on Pashto-English data in a few hours leads the system towards better parameter values for our language pair at the expense of some quality loss for the others.

mBART50 PRE-TRAINING SET

training set

Input	Output
Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much.	El señor y la señora Dursley, que vivían en el número 4 de Privet Drive, estaban orgullosos de decir que eran muy normales, afortunadamente.
Longtemps, je me suis couché de bonne heure.	For a long time I would go to bed early.
Все люди рождаются свободными и равными в своем достоинстве и правах.	All human beings are born free and equal in dignity and rights.
...	...

mBART50 LANGUAGES



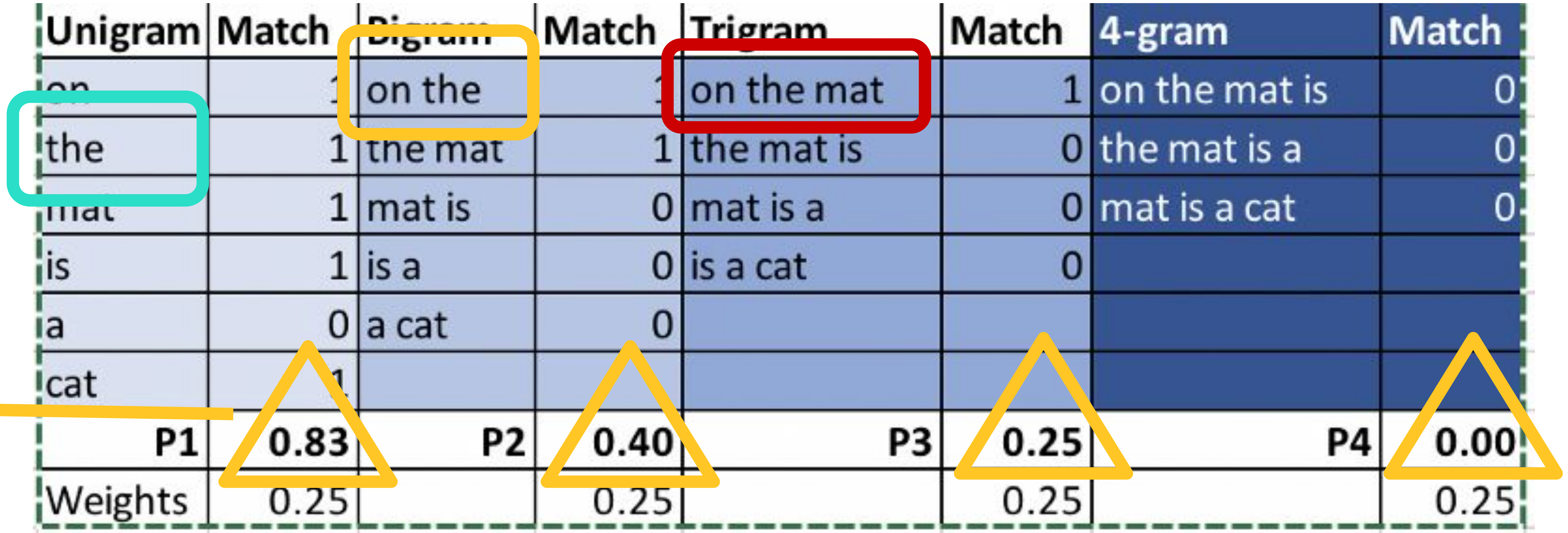
Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

RESULTS

AUTOMATIC EVALUATION WITH THE BLEU SCORE

Machine translation: On the mat is a cat

Reference: The cat is sitting on the mat



Unigram	Match	Bigram	Match	Trigram	Match	4-gram	Match
on	1	on the	1	on the mat	1	on the mat is	0
the	1	the mat	1	the mat is	0	the mat is a	0
mat	1	mat is	0	mat is a	0	mat is a cat	0
is	1	is a	0	is a cat	0		
a	0	a cat	0				
cat	1						
P1	**0.83**	**P2**	**0.40**	**P3**	**0.25**	**P4**	**0.00**
Weights	0.25		0.25		0.25		0.25

BLEU = 45.4

62,000,000

Adjustable parameters in the from-scratch model

610,000,000

Adjustable parameters in the heat-and-serve model

BLEU SCORES ENGLISH↦PASHTO

12.8

Commercial system

15.0

From-scratch

18.5

Heat-and-serve

HUMAN SCORES ENGLISH↦PASHTO

68.5	67.5	92.3
Commercial system	From-scratch	Heat-and-serve

BLEU SCORES PASHTO↦ENGLISH

35.0

Commercial system

20.0

From-scratch

25.4

Heat-and-serve

HUMAN SCORES PASHTO↦ENGLISH

83.8

Commercial system

63.5

From-scratch

85.1

Heat-and-serve

TAKEAWAYS

- The surprise language challenge implied crawling and downloading Pashto-English data, and training and evaluating two different neural models.

- The heat-and-serve mBART50-based model attains the best automatic and manual results in the news domain at the expense of speed, even when compared with a general-purpose commercial system.

- Next language, please!

Upcoming paper: “Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months”

Thanks!

Any questions?

Find us at:

@AVMiceliBarone amiceli@ed.ac.uk

@japer3z japerez@ua.es

gourmet

CREDITS

Special thanks to all people who made and share these awesome resources for free:

- Presentation template designed by Slidesmash
- Photographs by unsplash.com and pexels.com
- Vector Icons by Matthew Skiles

Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:

http://www.1001fonts.com/oswald-font.html

https://www.fontsquirrel.com/fonts/open-sans

Colors used

🟦 🟦 🟨 🟩 ⬛