A Multilingual Neural Machine Translation Model for COVID-19

Alexandre Bérard
Zae Myung Kim
Vassilina Nikoulina
Eunjeong Lucy Park
Matthias Gallé
Kweon Woo Jung

NAVER LABS EUROPE
NAVER PAPAGO

NAVER LABS Europe

DEVIEW 2020

Motivation

Help translate COVID-19-related text, by creating a **public**, **biomedical-domain**, **multilingual** machine translation model

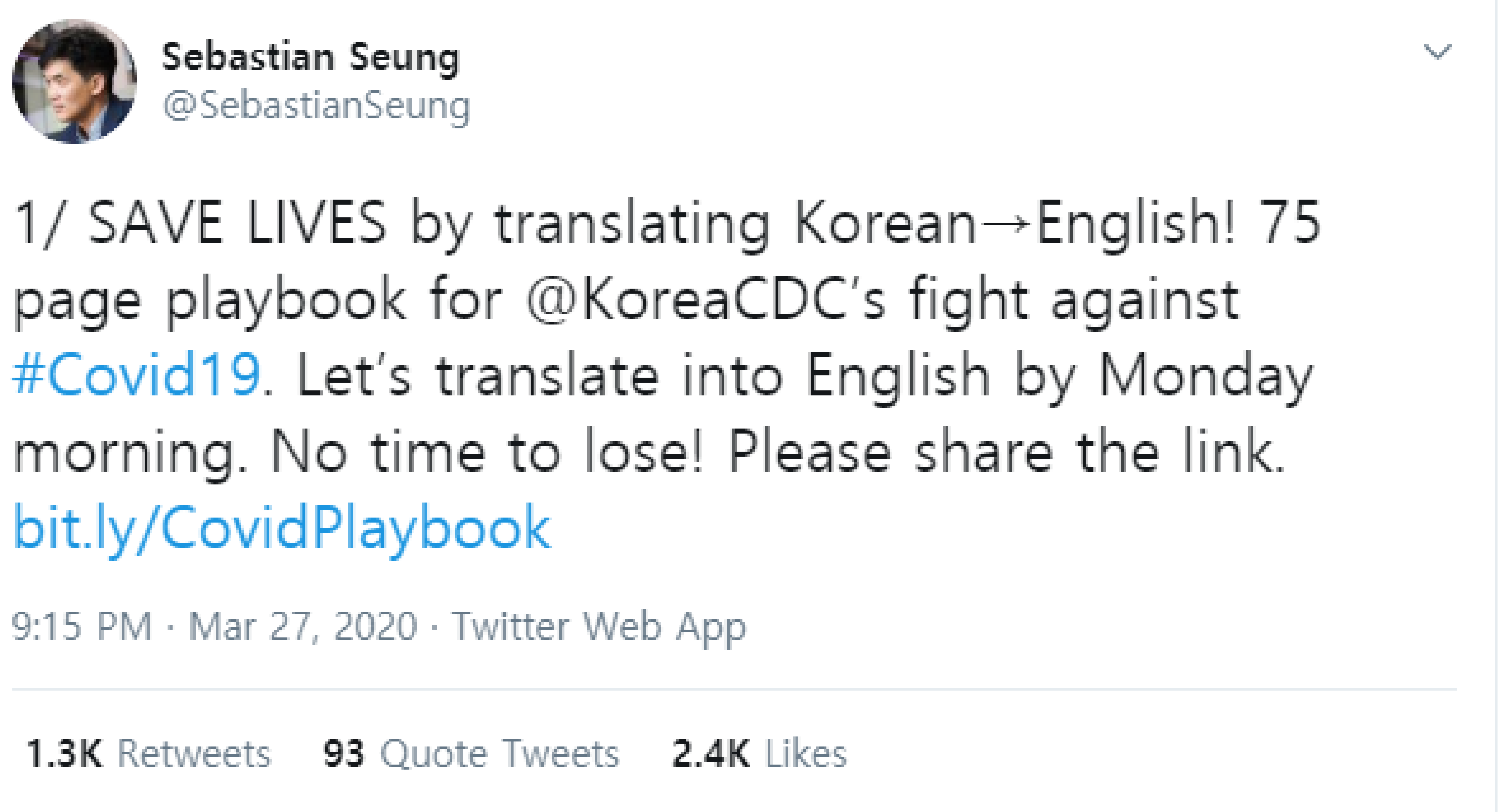
Motivation

1. Help translate COVID-19 guidelines,* news articles and social media reactions

** We recommend expert verification for critical translations*

State-of-the-art public NMT systems are only available for a couple of languages.

None is adapted to the biomedical domain.

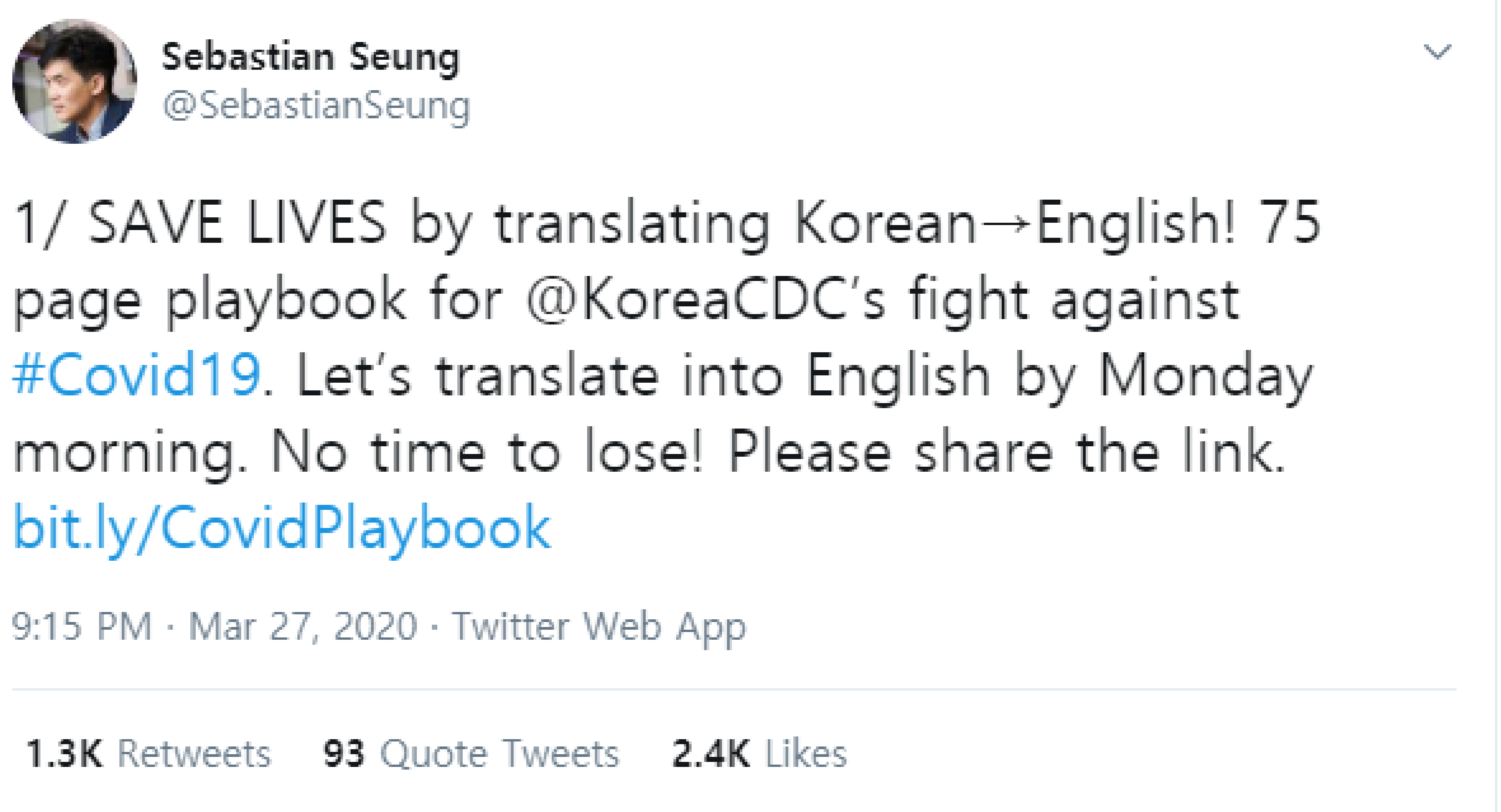


Motivation

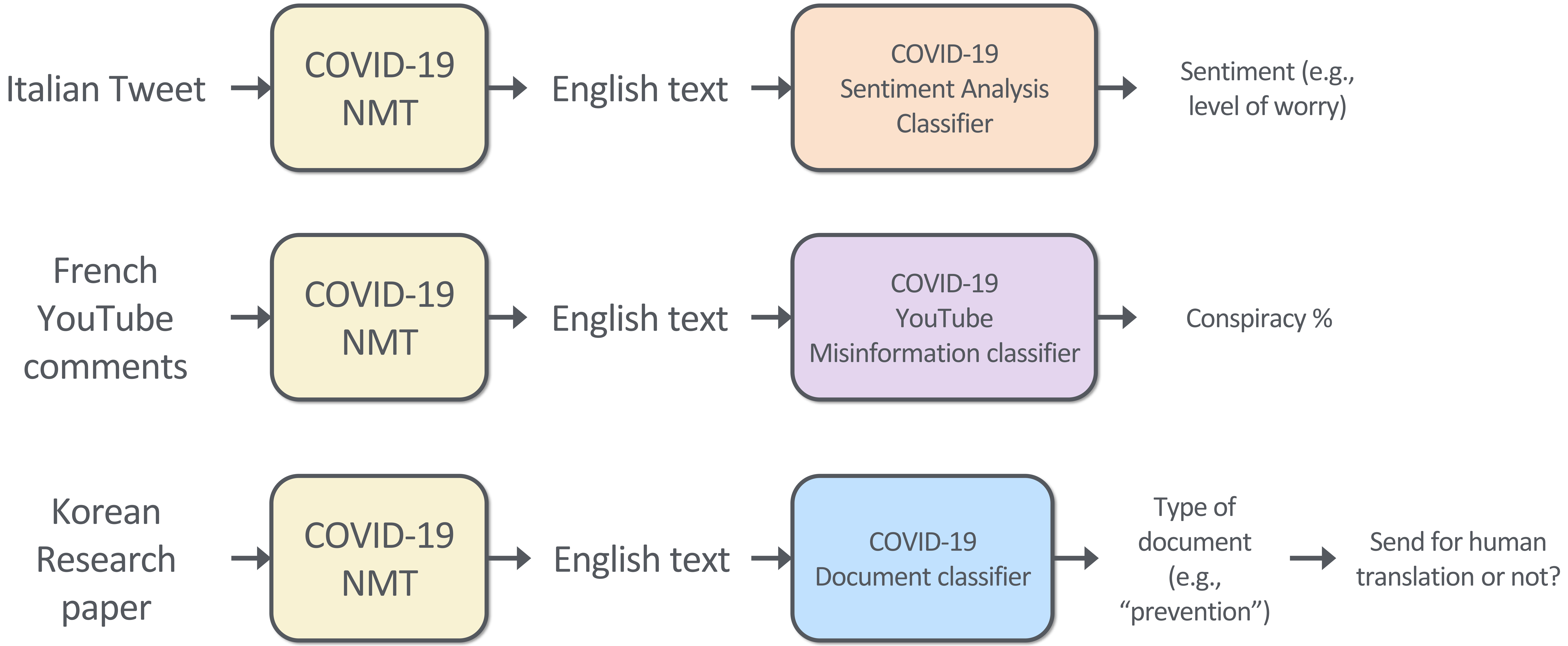
1. **Help translate COVID-19 guidelines,* news articles and social media reactions**
2. **Apply existing NLP tools to other languages than English**

Examples of English-centric NLP tools for COVID-19:

- *Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned*
- *Document Classification for COVID-19 Literature*
- *What Are People Asking About COVID-19? A Question Classification Dataset*
- *Measuring Emotions in the COVID-19 Real World Worry Dataset*
- *NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube*



Motivation


Italian Tweet
COVID-19 NMT
English text
COVID-19 Sentiment Analysis Classifier
Sentiment (e.g., level of worry)
French YouTube comments
COVID-19 NMT
English text
COVID-19 YouTube Misinformation classifier
Conspiracy %
Korean Research paper
COVID-19 NMT
English text
COVID-19 Document classifier
Type of document (e.g., “prevention”)
Send for human translation or not?

Challenges

In-domain data availability

Multilinguality

Compatibility with an existing framework

Evaluation

Related Work

- TAUS Corona Crisis corpora
- SYSTRAN Corona Crisis models
- TICO-19 resources
- NLP COVID-19 Workshop

Publications:

- *Facilitating Access to Multilingual COVID-19 Information via Neural Machine Translation,* Way et al. (2020)
- *A System for Worldwide COVID-19 Information Aggregation,* Aizawa et al. (2020)

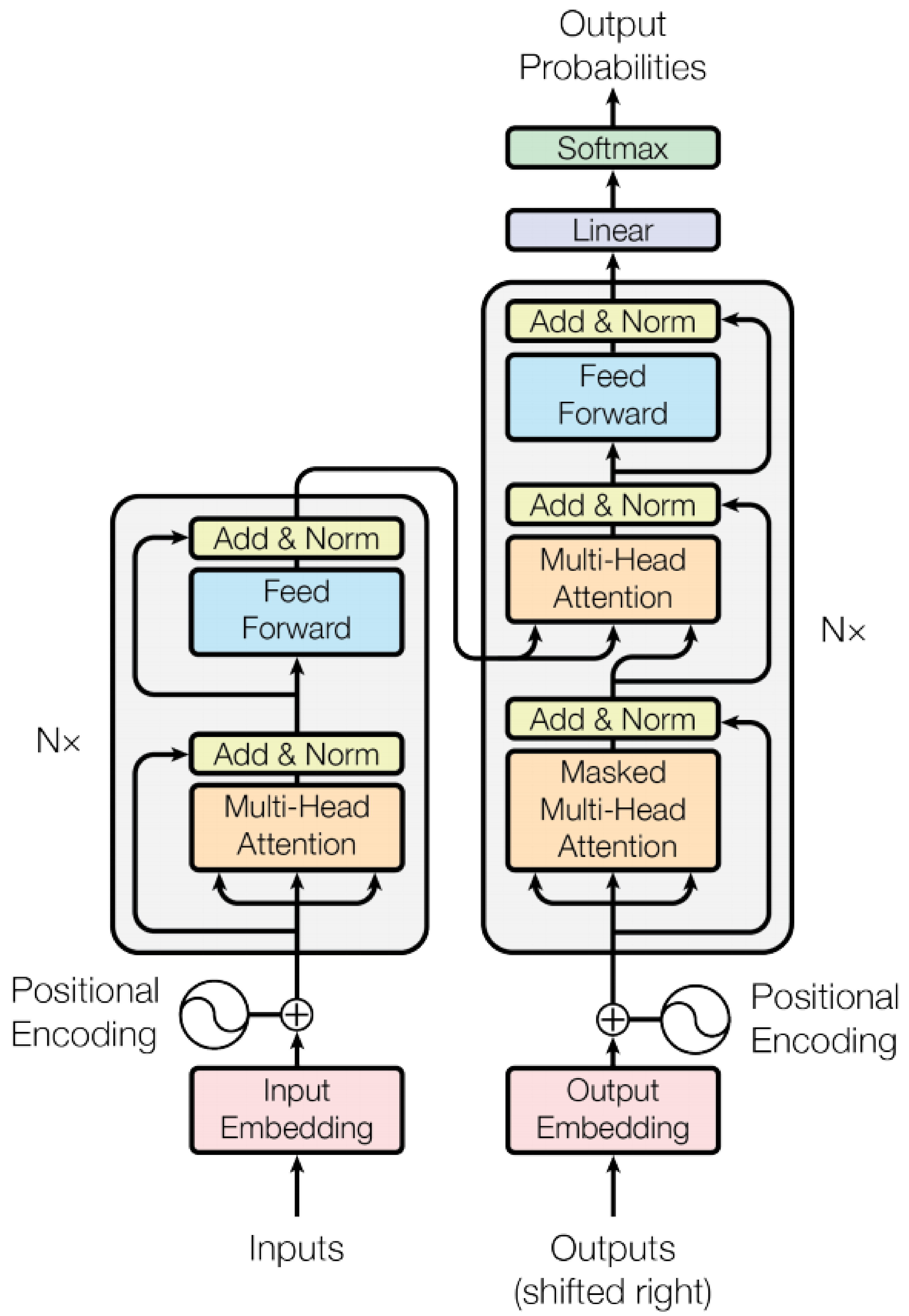
CONTENTS

1. Introduction to Neural Machine Translation

1.1 Transformer

Transformer *(Vaswani et al., 2017)* is now the standard in MT

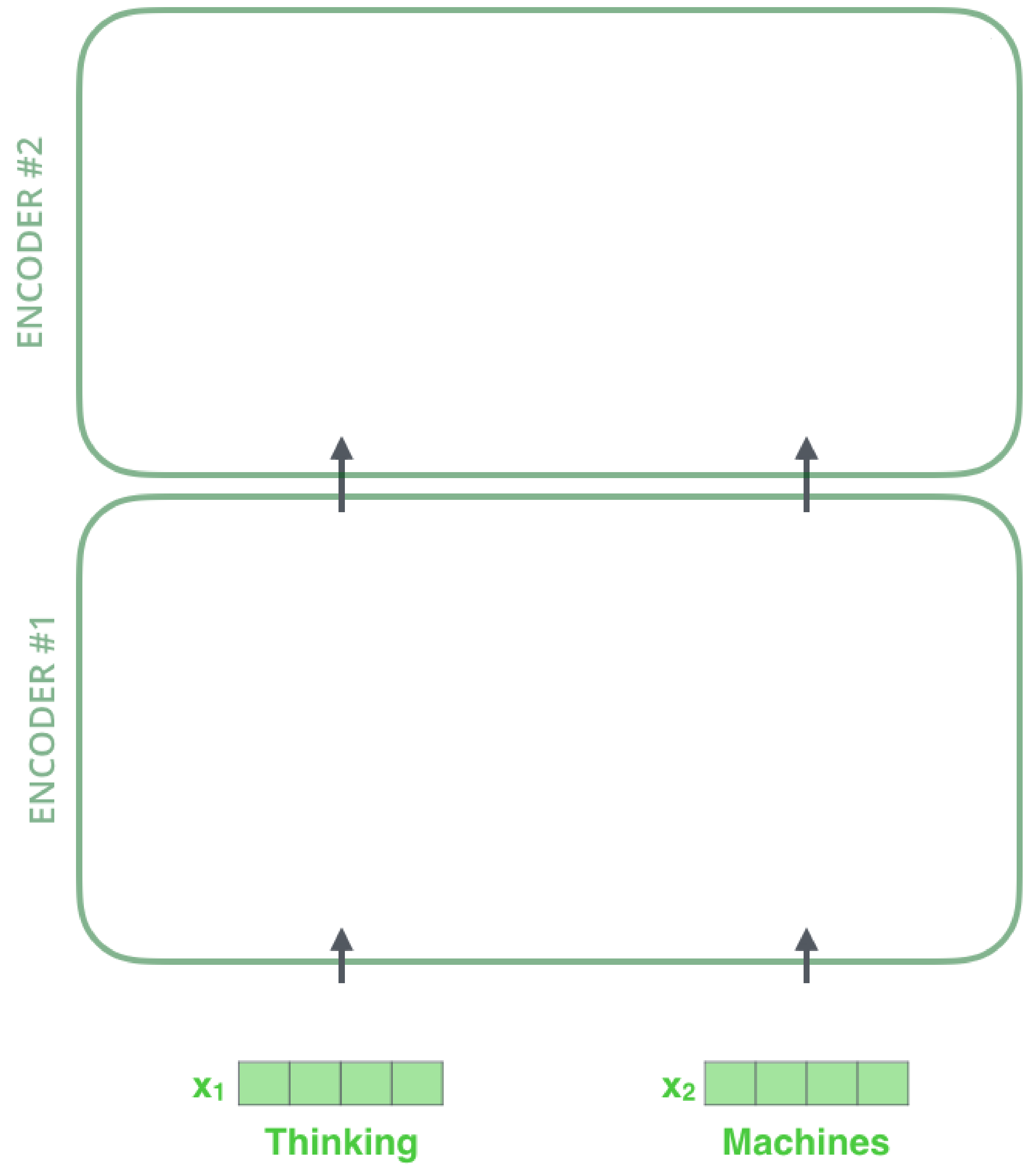
It is also the architecture of BERT *(Devlin et al., 2018)*, the basis of most current NLP work



1.1 Transformer

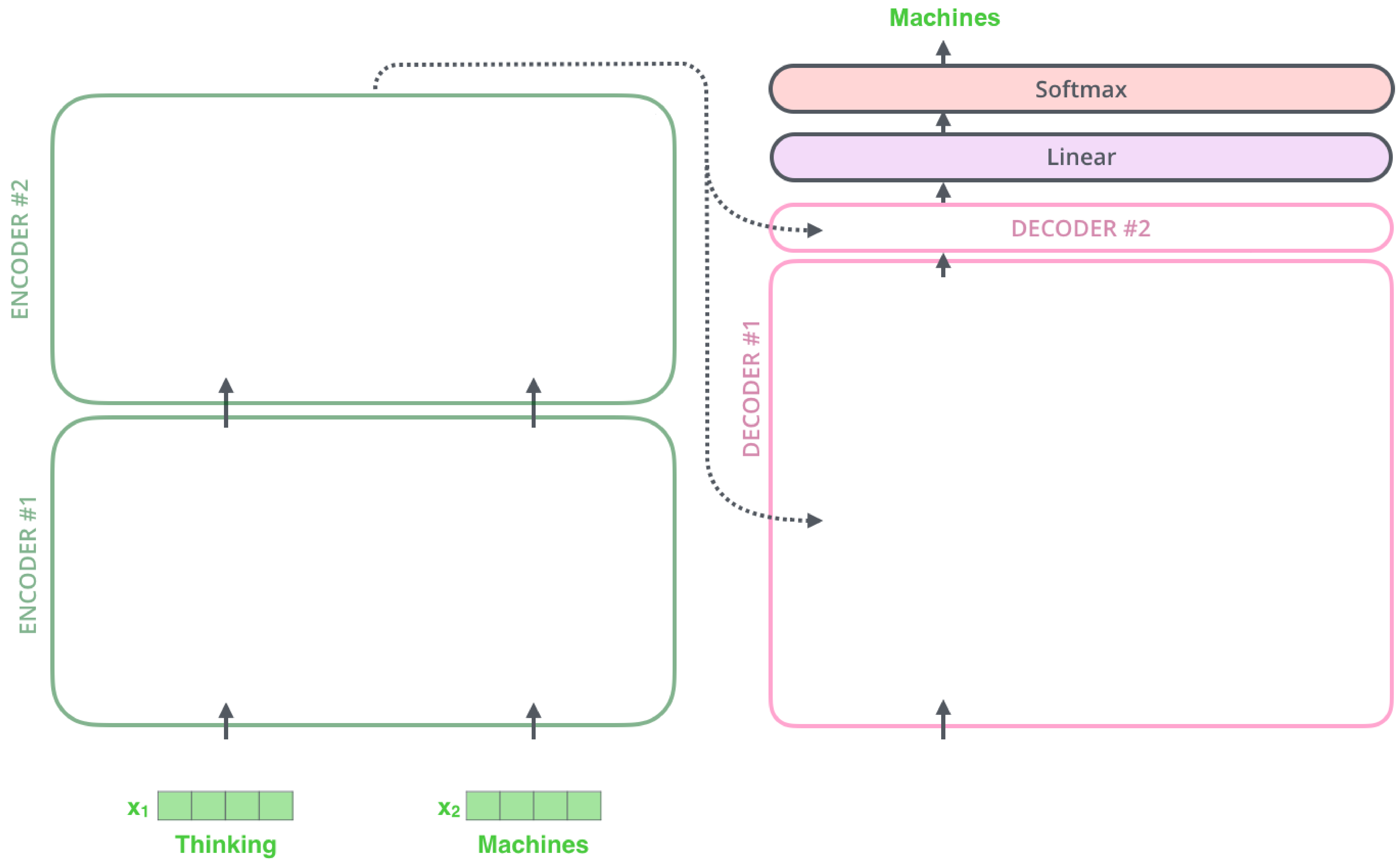
Figures from "The Illustrated Transformer" blog post

1.1 Transformer



Figures from "The Illustrated Transformer" blog post

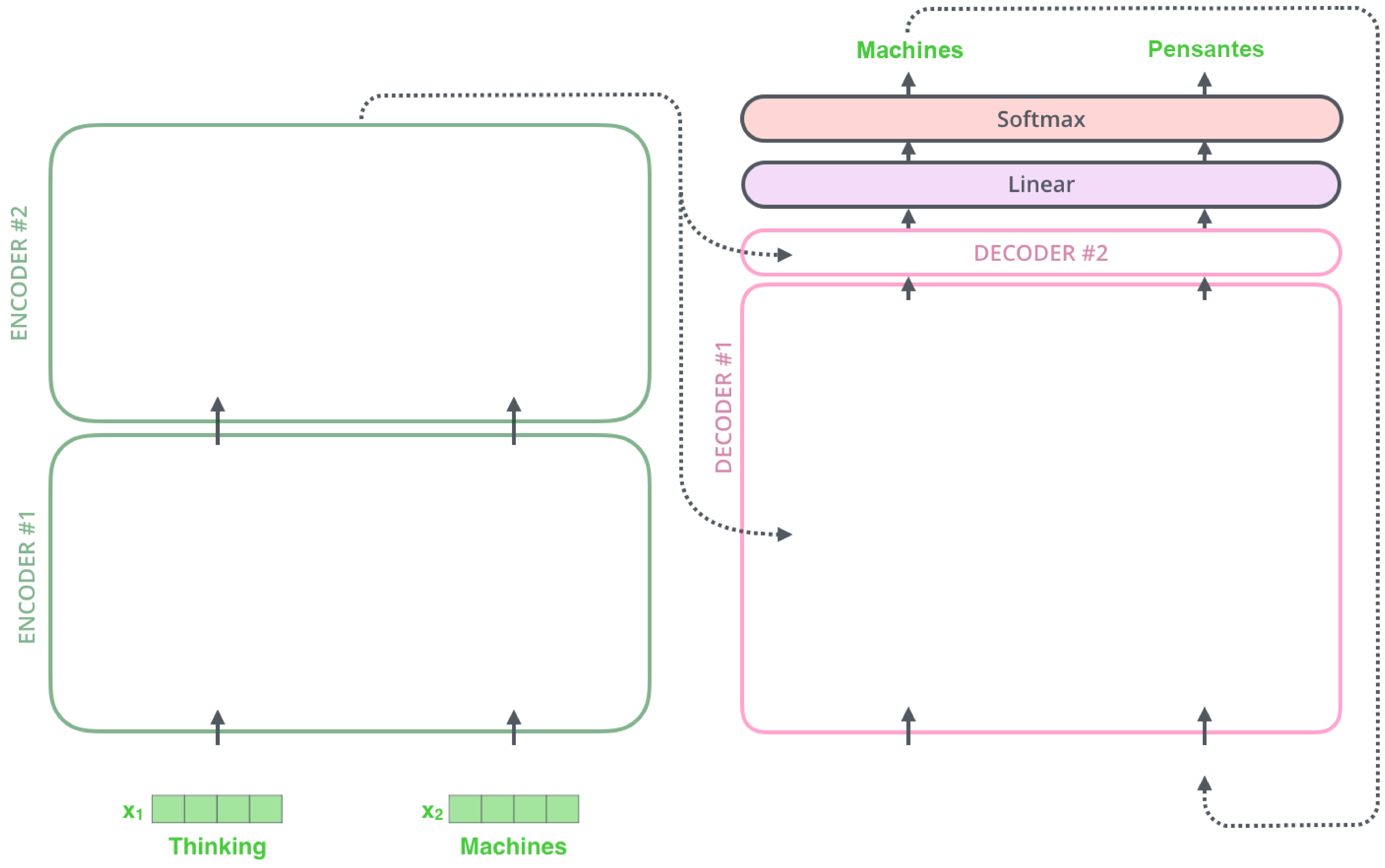
1.1 Transformer



Figures from "The Illustrated Transformer" blog post

1.1 Transformer

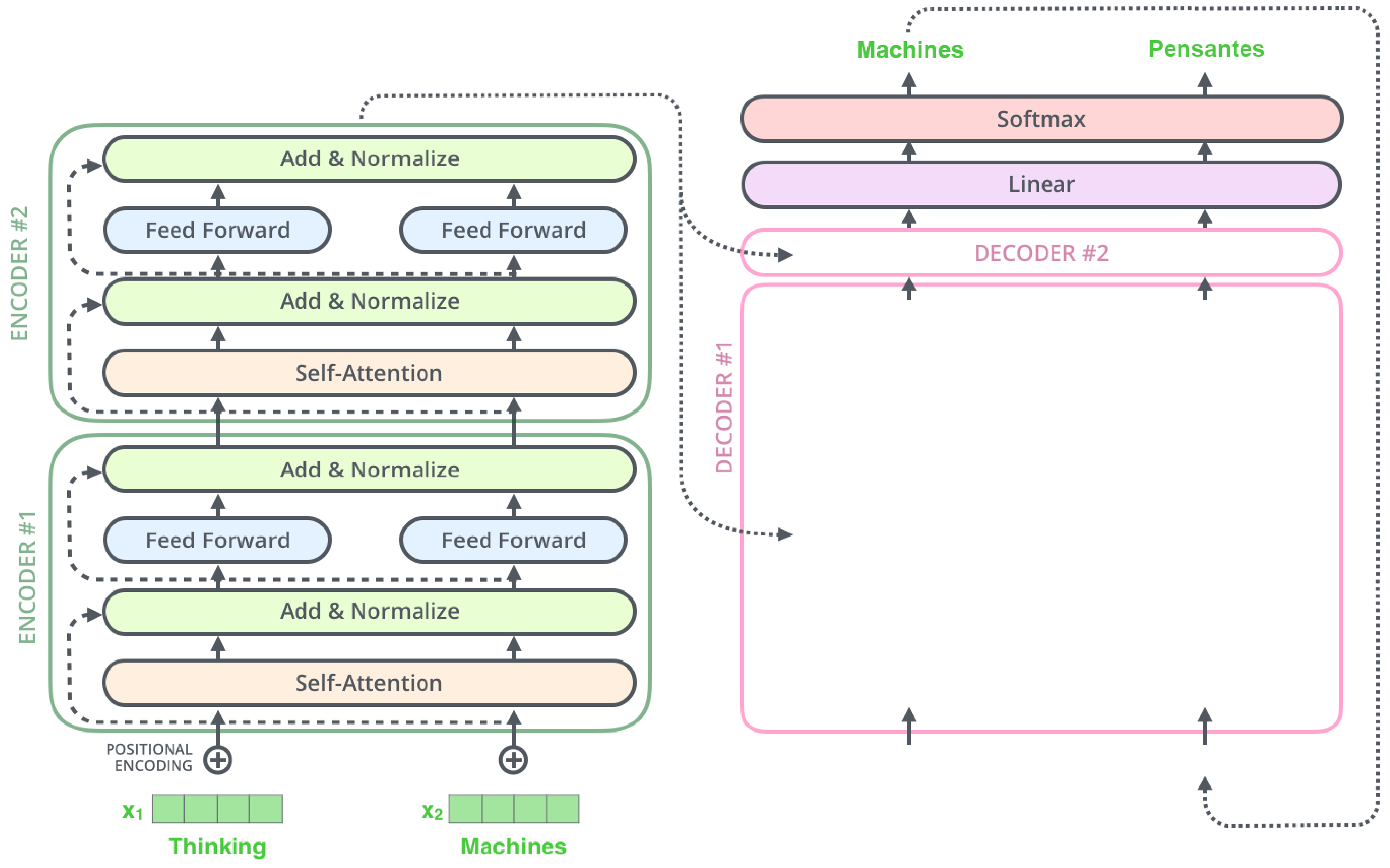
DEVIEW 2020



Figures from "The Illustrated Transformer" blog post

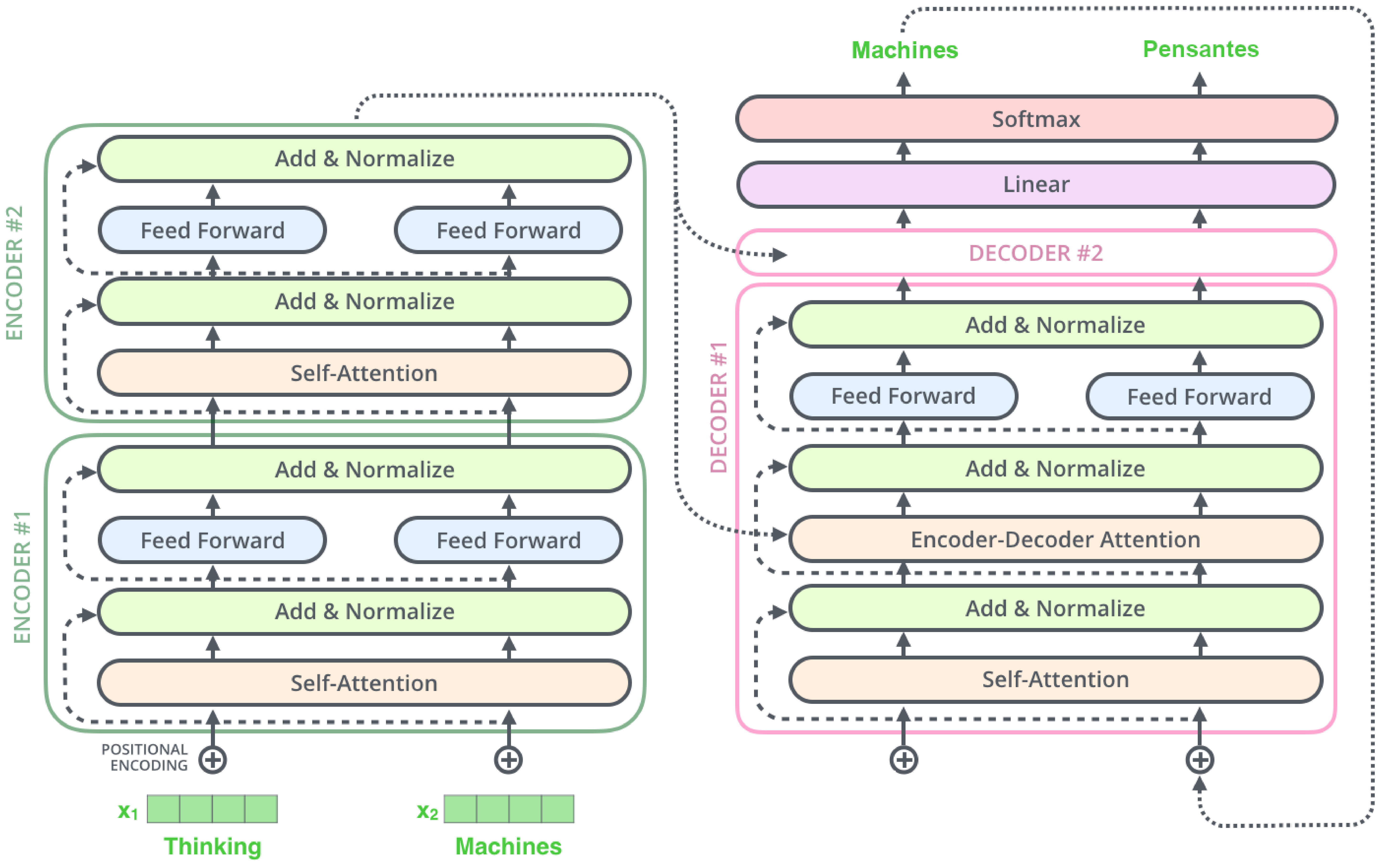
1.1 Transformer

DEVIEW 2020



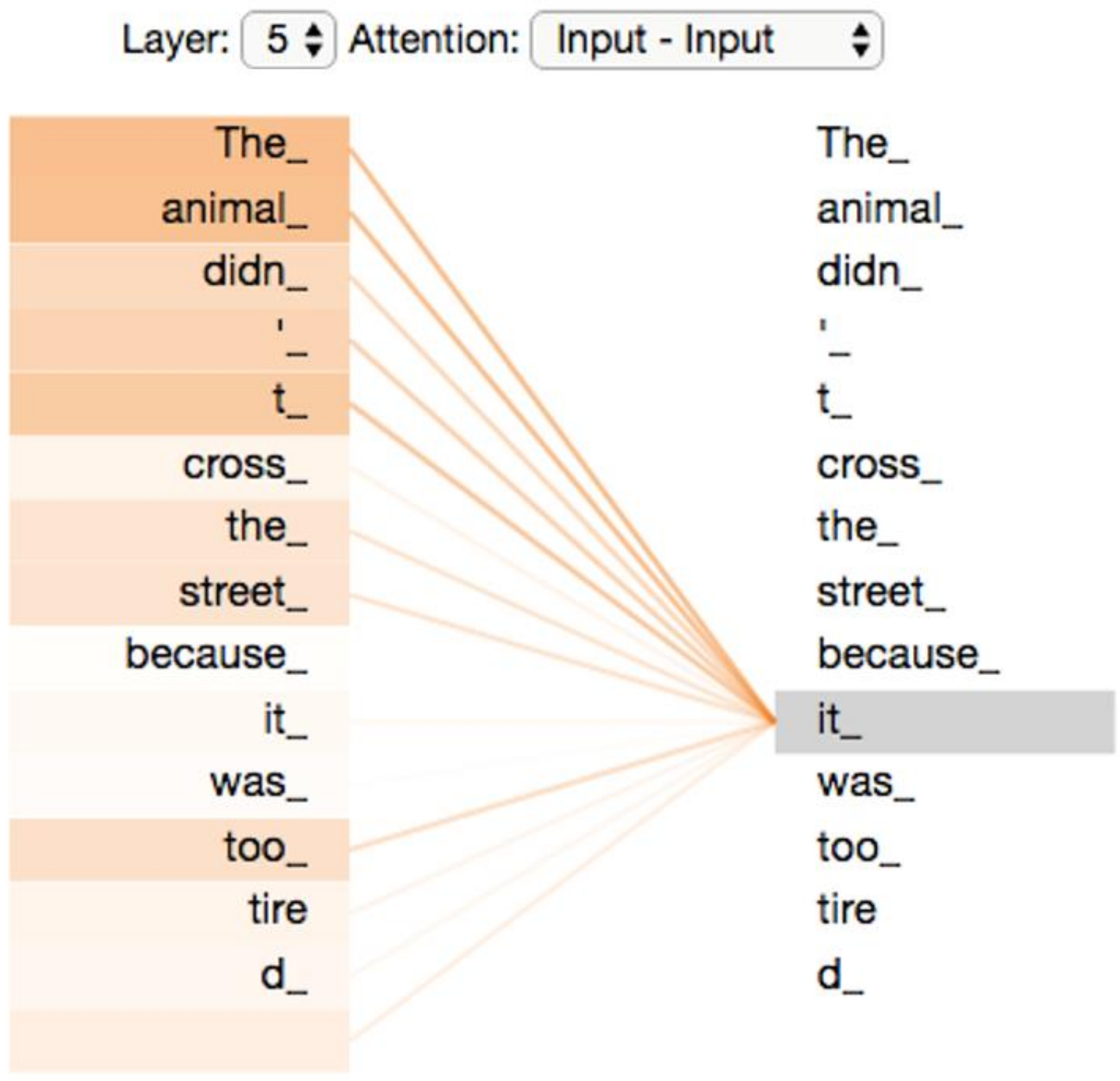
Figures from "The Illustrated Transformer" blog post

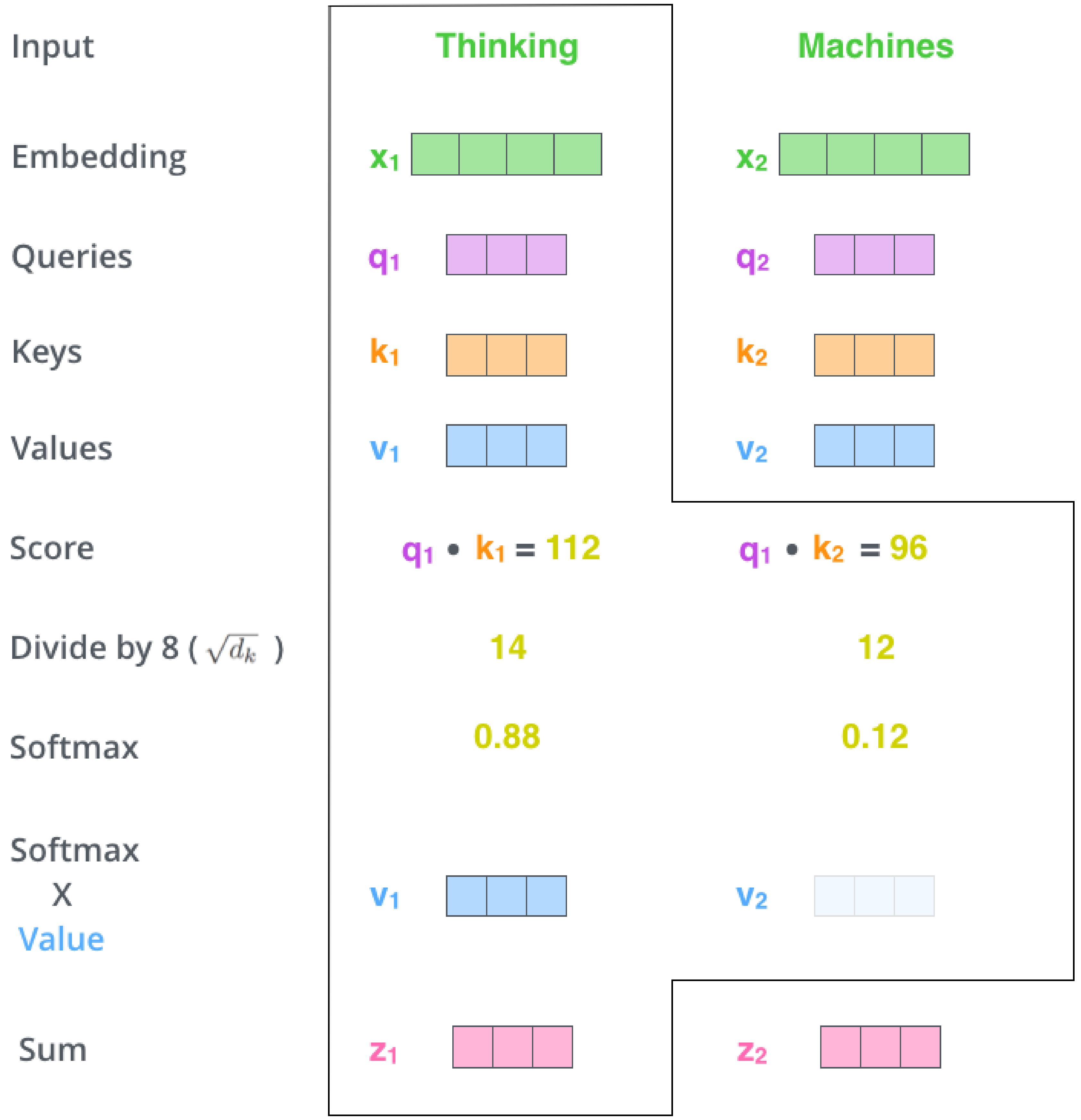
1.1 Transformer



Figures from "The Illustrated Transformer" blog post

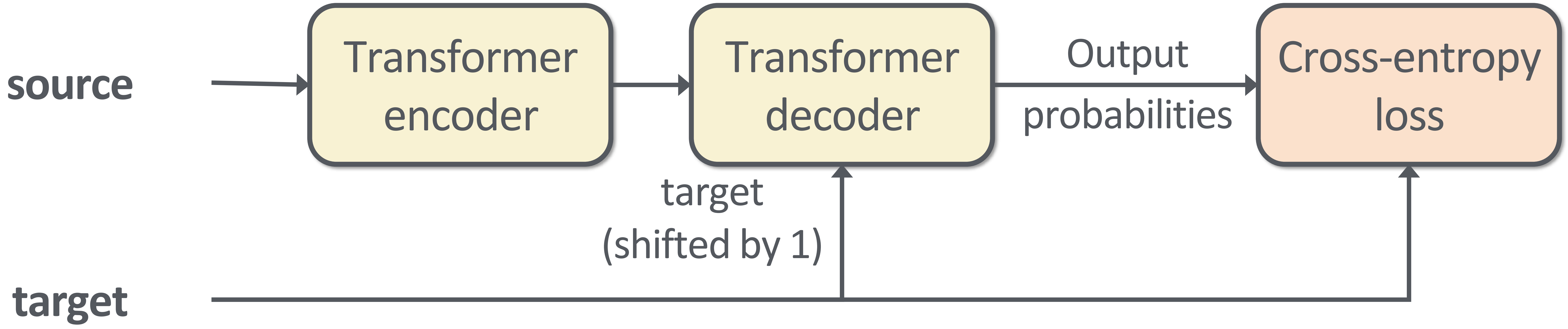
1.1 Transformer





Figures from "The Illustrated Transformer" blog post

1.2 Training



Training data: thousands/millions of (**source**, **target**) pairs of translated sentences

1.2 Training

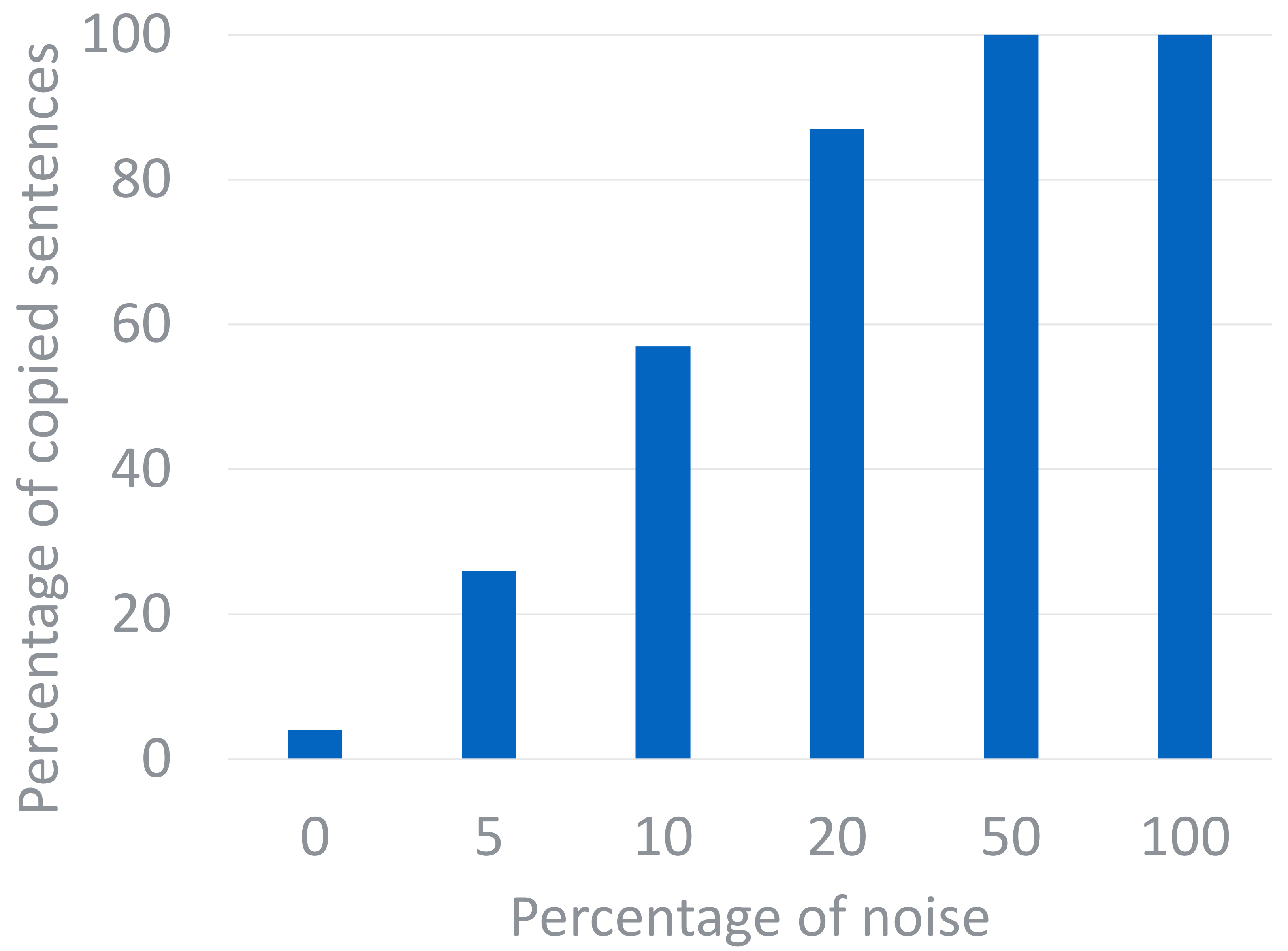
```
for epoch in range(N):
    for source, target in data:
        probs = model(source, target[1:])    # forward pass (model outputs)
        loss = cross_entropy(probs, target)  # forward pass (loss)
        optimizer.backward(loss)             # backward pass: compute gradients
        optimizer.step()                     # update the model weights
```

In practice, **source** and **target** are tensors corresponding to batches of multiple sentences

High parallelism is achieved thanks to matrix multiplications on GPUs

1.3 Pre-processing

1. **Training data filtering:** language identification & length-based filtering



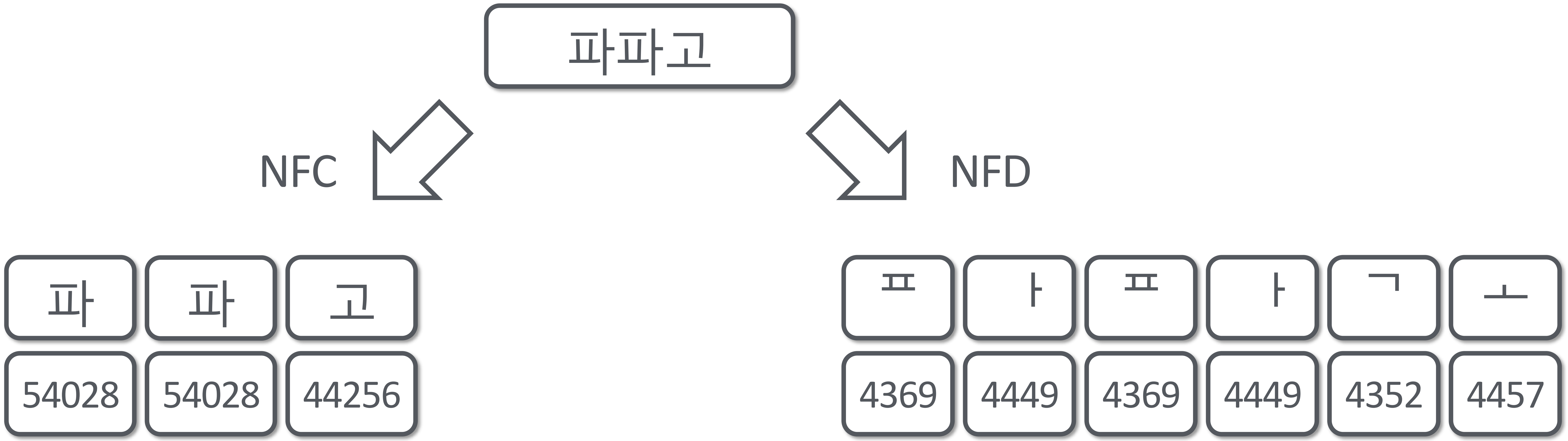
Numbers from *Khayrallah and Koehn (2018)*

Language id	Length filtering	Copies	Hallucinations
No	No	12%	5%
Yes	No	0%	3%
Yes	Yes	0%	1%

Table from *Berard et al. (2019)*

1.3 Pre-processing

1. **Training data filtering:** language identification & length-based filtering
2. **Unicode normalization**



1.3 Pre-processing

1. **Training data filtering:** language identification & length-based filtering
2. **Unicode normalization**
3. **BPE (Byte Pair Encoding)**

newest (6) low (5) widest (3) lower (2)	n e w **e** **s** t (6) l o w (5) w i d **e** **s** t (3) l o w e r (2)	n e w **es** **t** (6) l o w (5) w i d **es** **t** (3) l o w e r (2)	n e w est (6) **l** **o** w (5) w i d est (3) **l** **o** w e r (2)	n e w est (6) lo w (5) w i d est (3) lo w e r (2)
BPE vocabulary:	l, o, w, e, r, n, w, s, t, i, d	l, o, w, e, r, n, w, s, t, i, d, **es**	l, o, w, e, r, n, w, s, t, i, d, **es**, **est**	l, o, w, e, r, n, w, s, t, i, d, **es**, **est**, **lo**

Credits to Chris Manning's "NLP with Deep Learning" lecture

1.3 Pre-processing

1. **Training data filtering:** language identification & length-based filtering
2. **Unicode normalization**
3. **BPE (Byte Pair Encoding)**

BPE vocabulary: l, o, w, e, r, n, w, s, t, i, d, **es**, **est**, **lo**

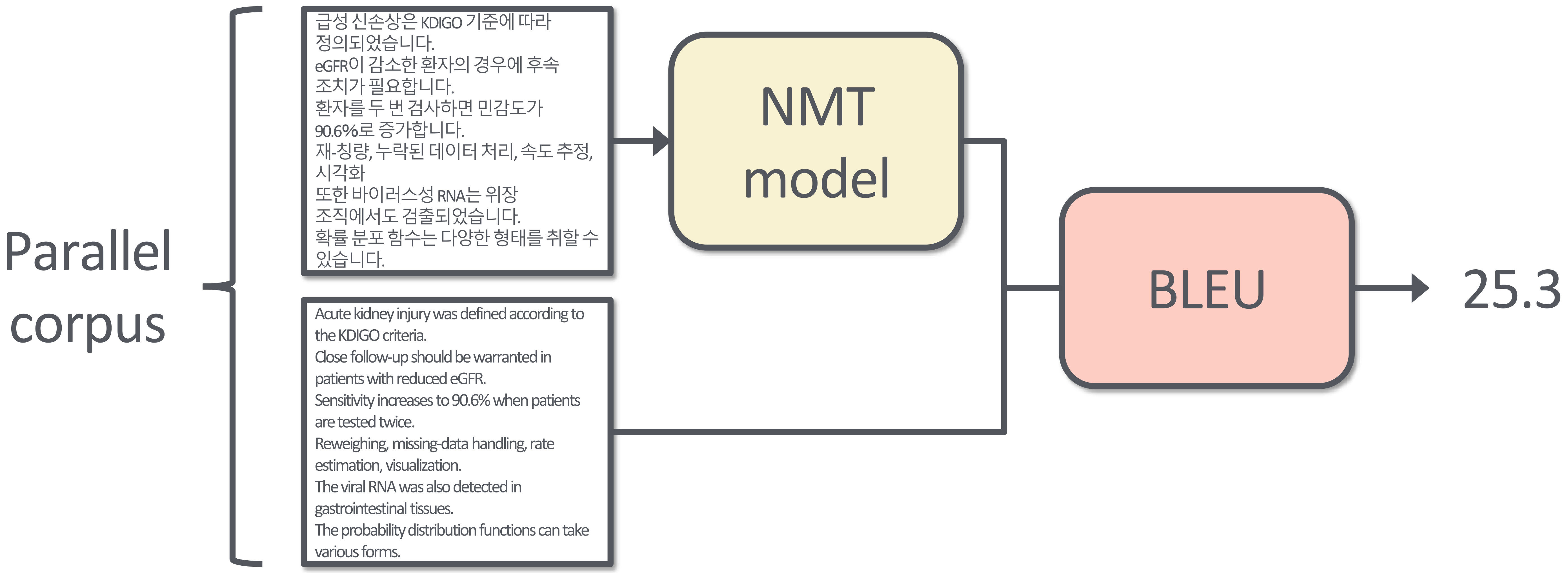
```
lowest losses
      ⇩
l o w e s t _ l o s s e s
      ⇩
l o w es t _ l o s s es
      ⇩
l o w est _ l o s s es
      ⇩
lo w est _ lo s s es
```

1.4 Evaluation

How to measure progress in MT? How to compare two systems?

1. Automatic metrics

→ BLEU, METEOR, chrF, etc.

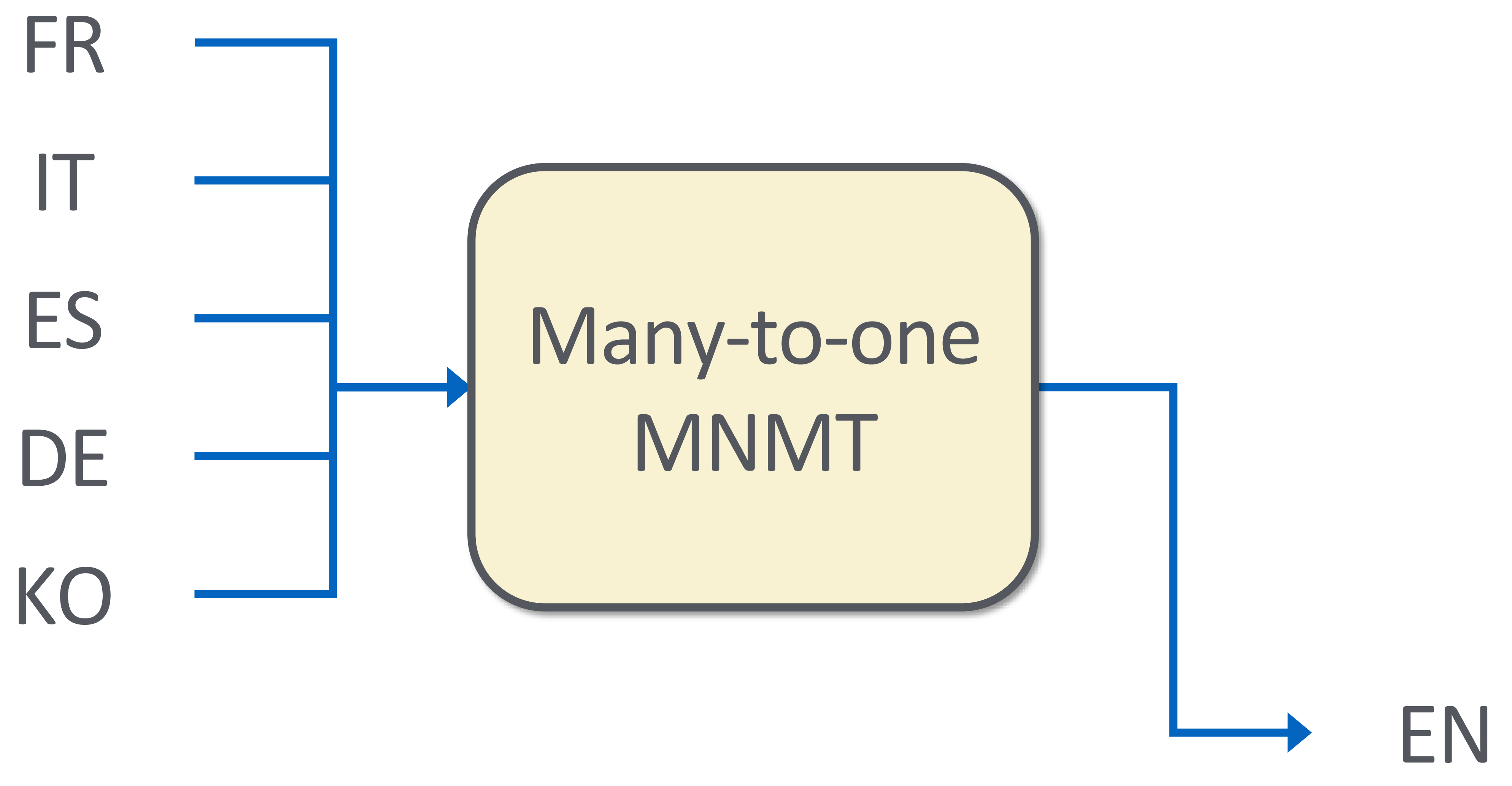


2. Human evaluation

→ Relative ranking

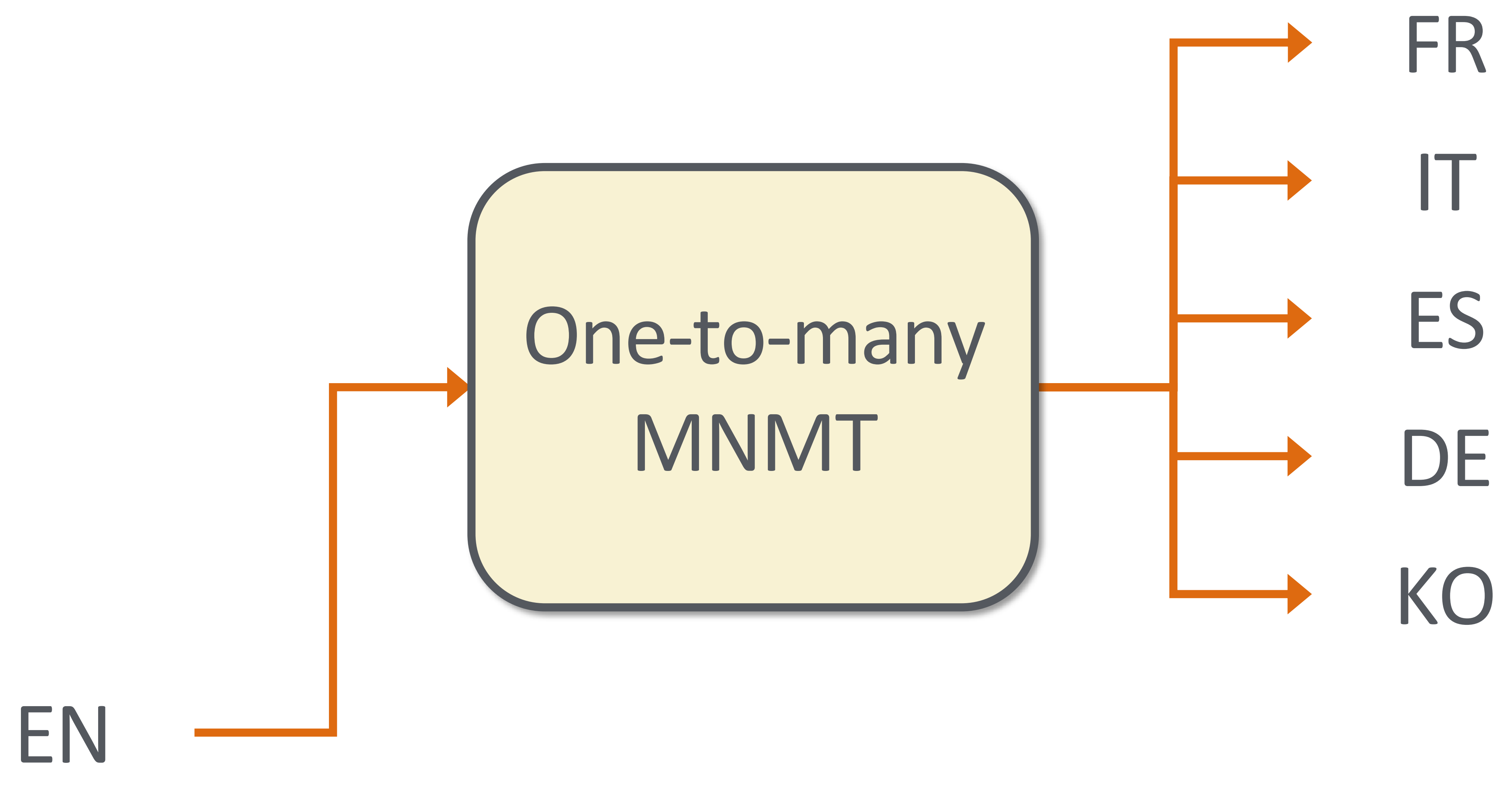
→ Direct assessment

1.5 Multilingual NMT



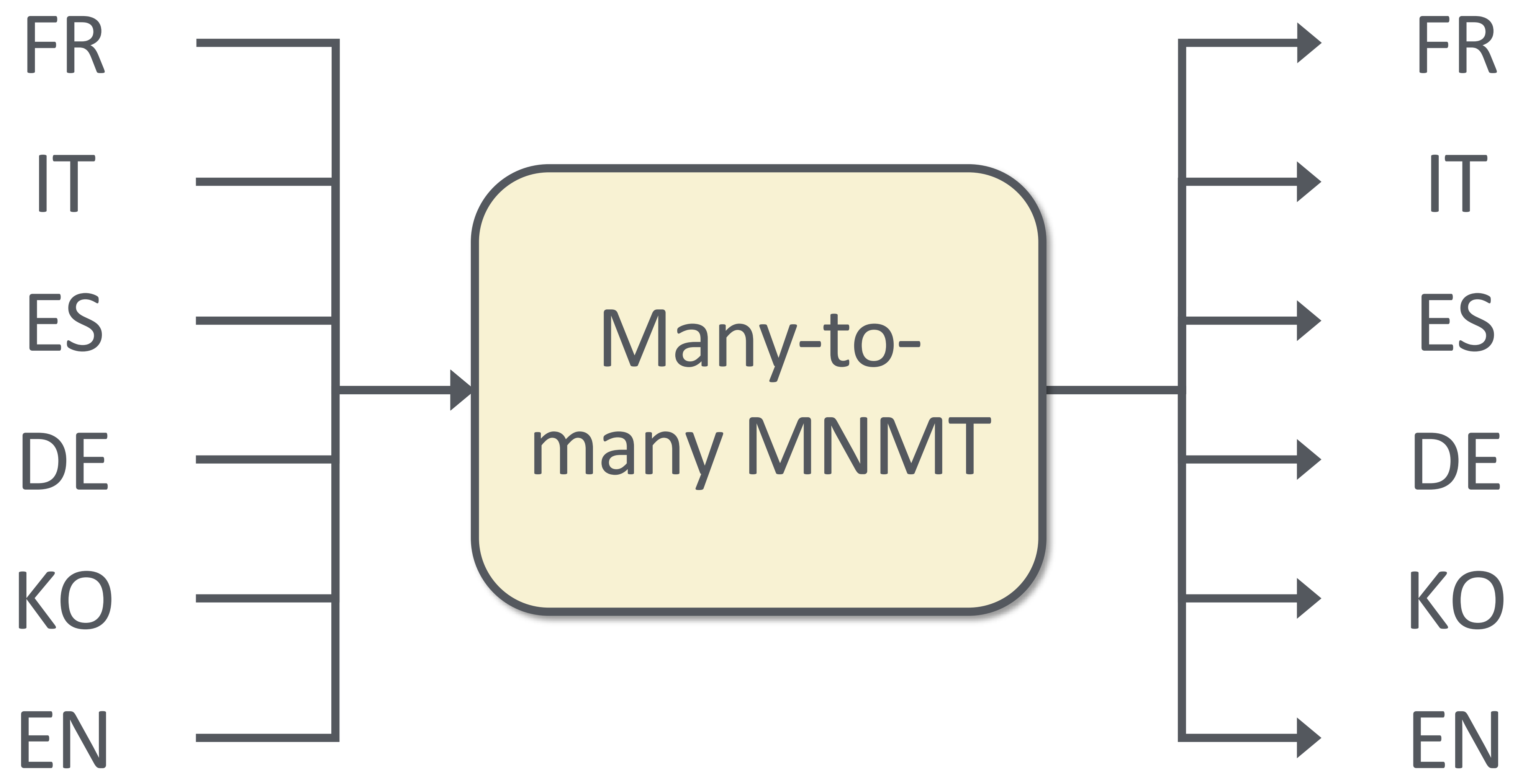
- One single model for multiple source languages
- Joint multilingual BPE model & shared vocabulary and embeddings

1.5 Multilingual NMT



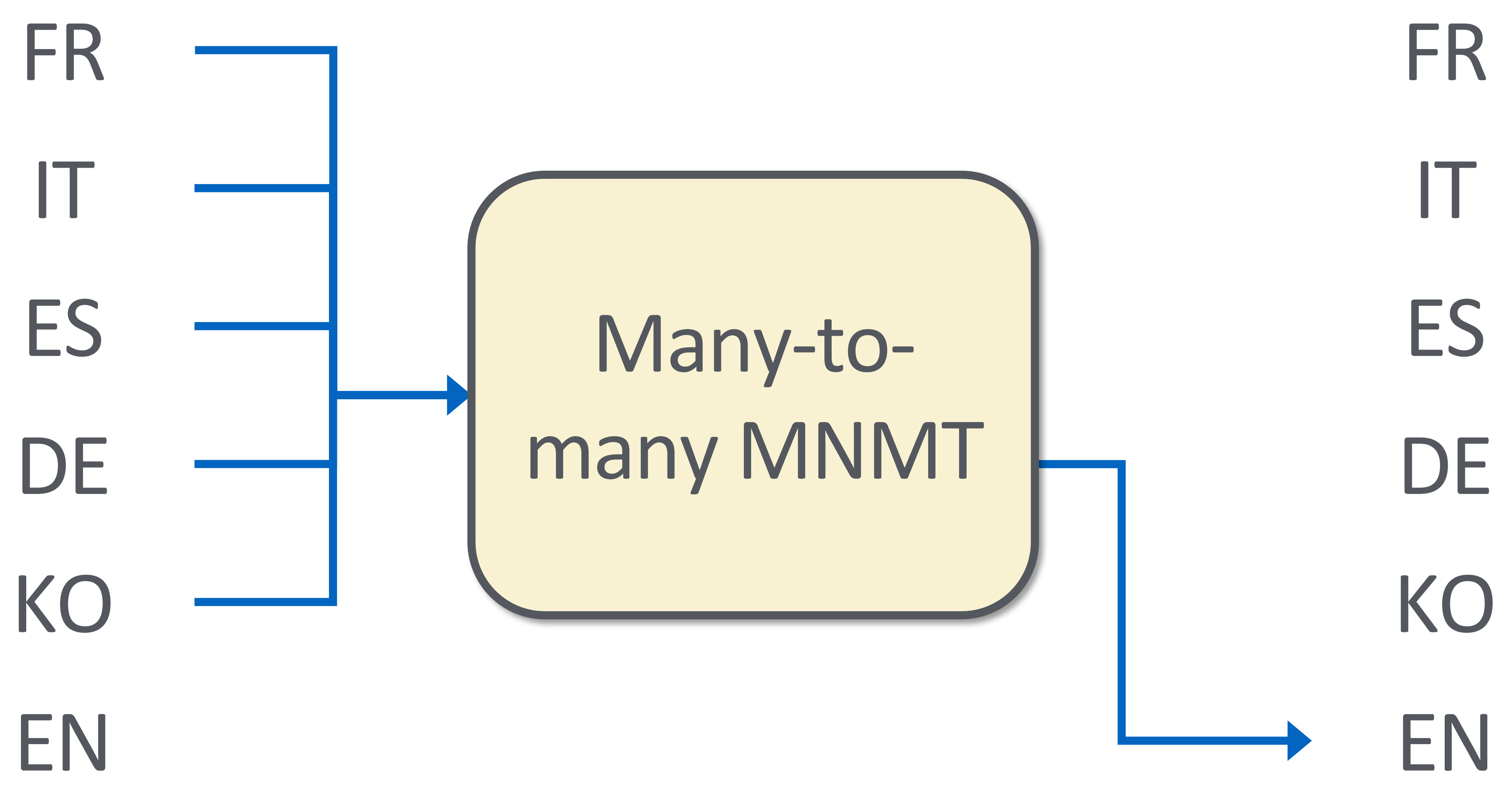
Choose target language with source-side tag: <2FR>, <2IT>, etc.

1.5 Multilingual NMT

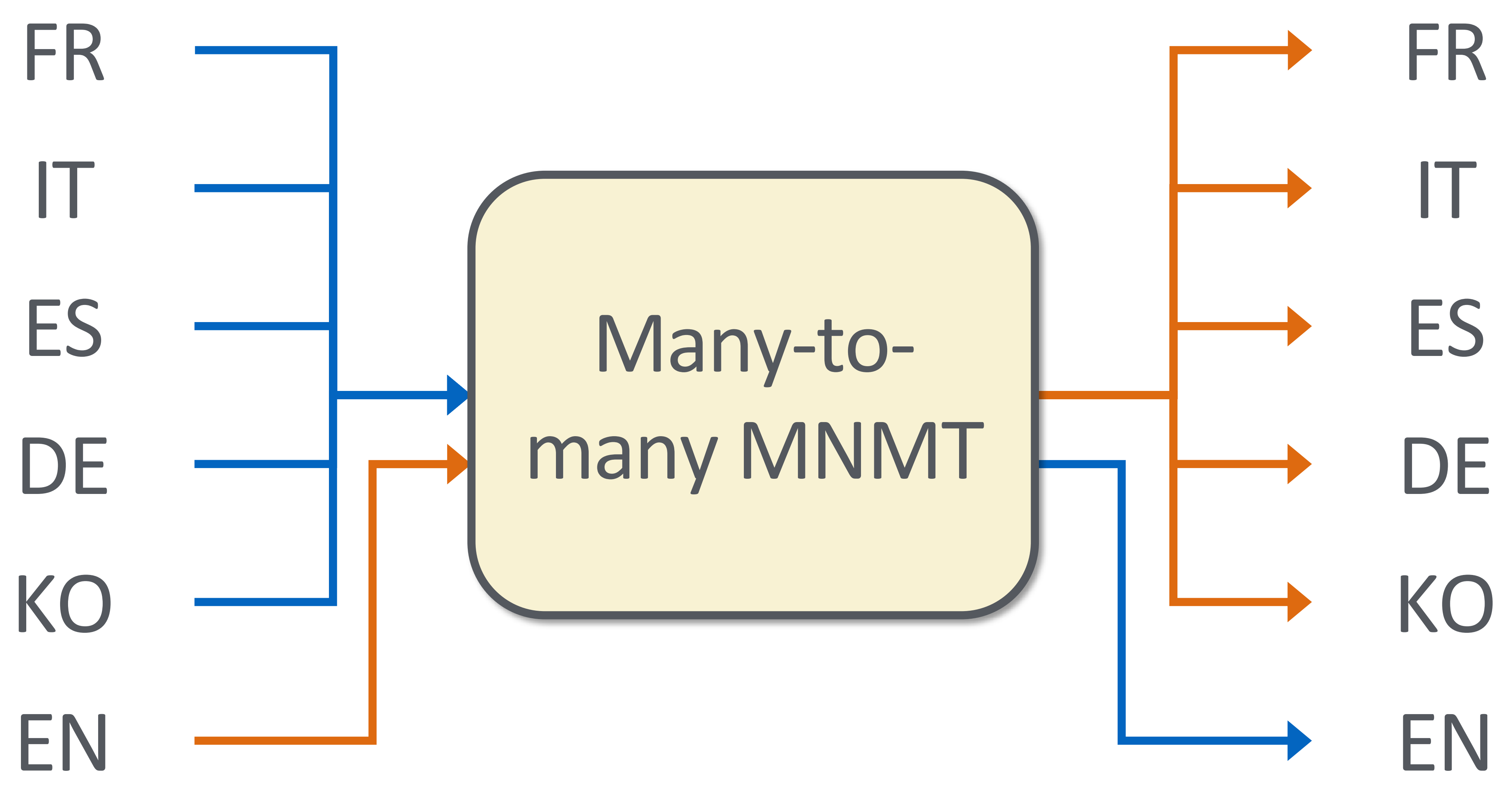


But most training data is paired with English...

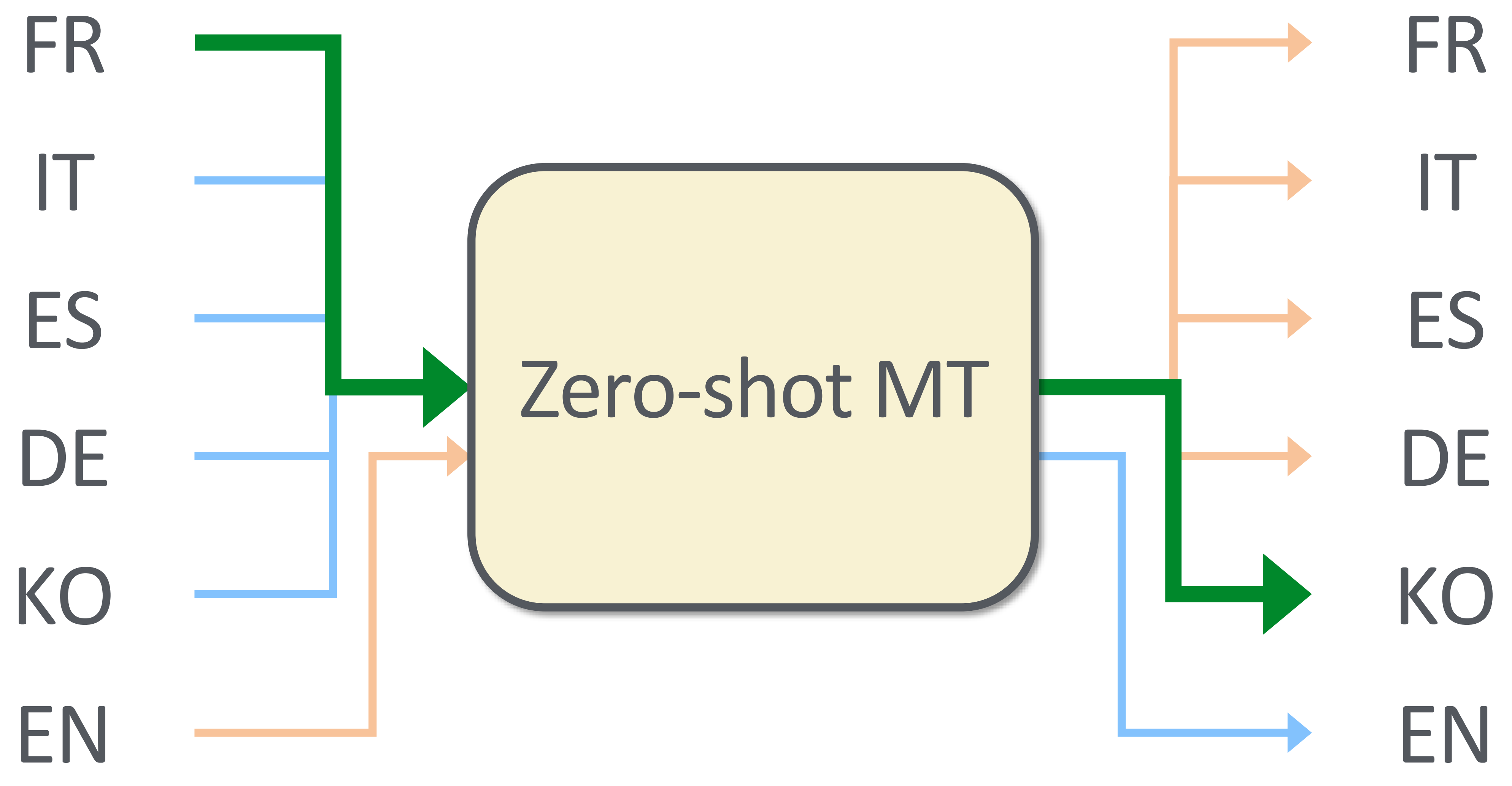
1.5 Multilingual NMT



1.5 Multilingual NMT



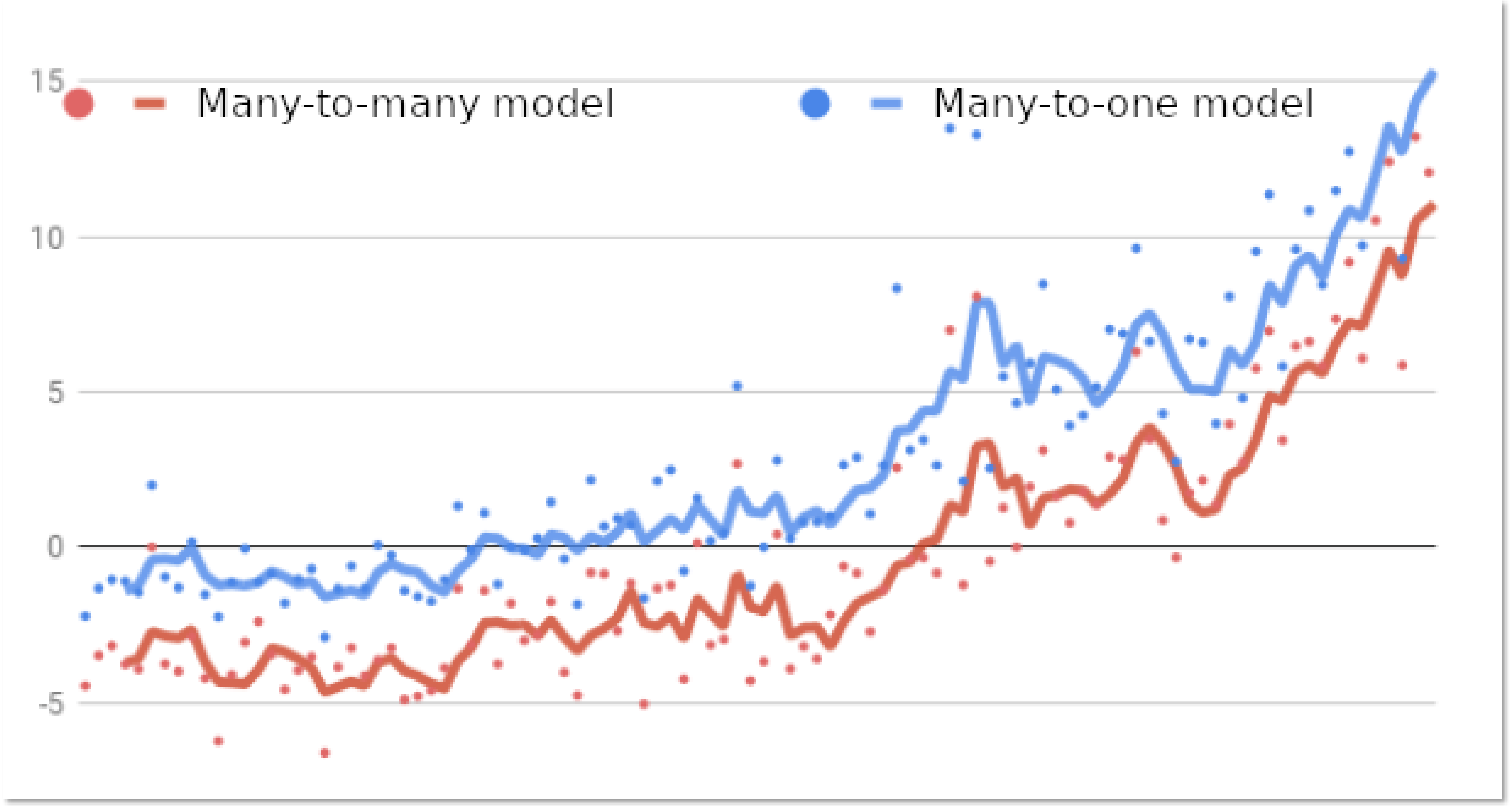
1.5 Multilingual NMT



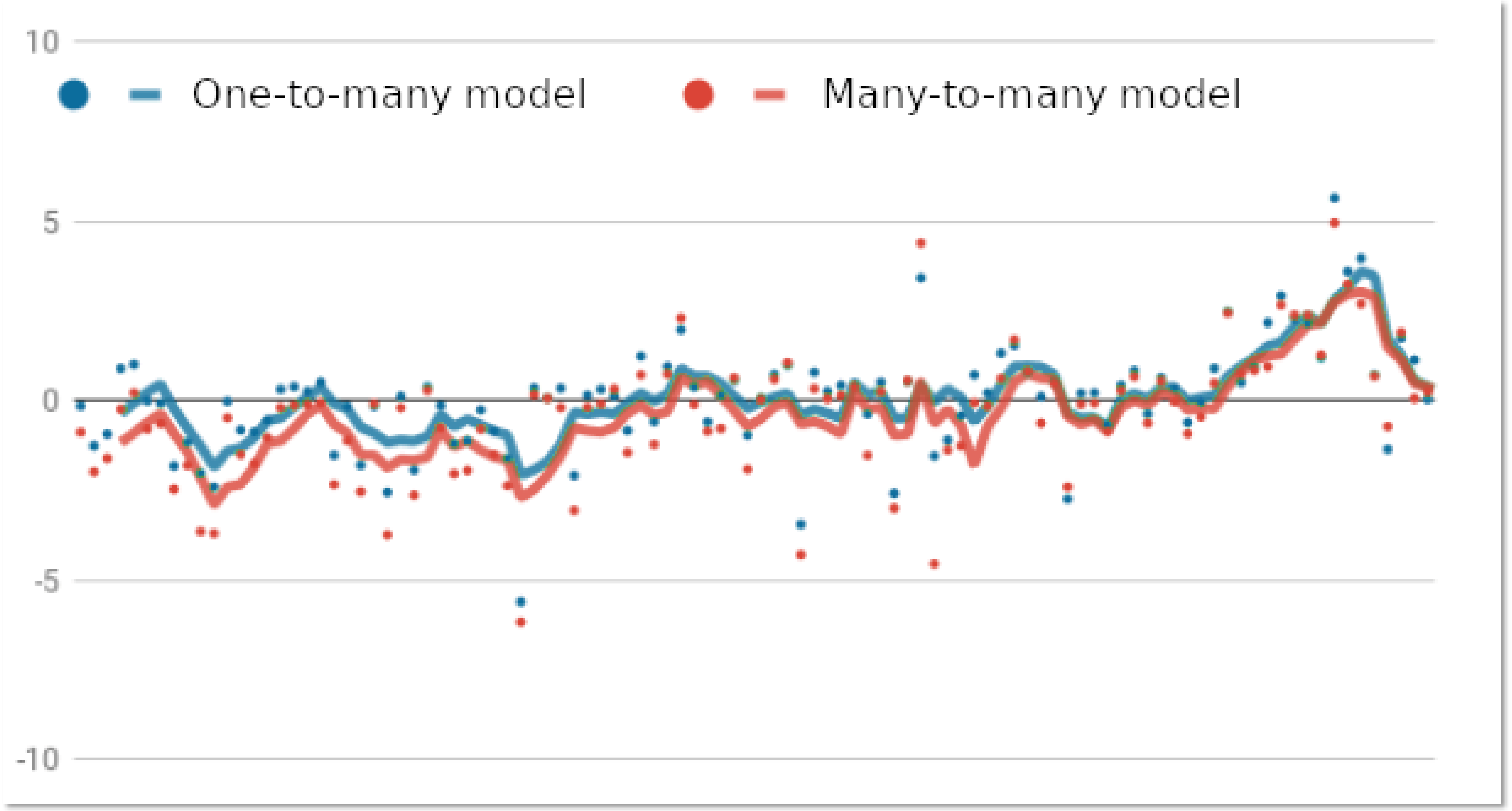
→ "Monolingual Adapters for Zero-Shot Neural Machine translation", Philip et al. (2020)

1.5 Multilingual NMT

BLEU differences with bilingual models, from high to low-resource languages



Any-to-English (BLEU Δ)



English-to-Any (BLEU Δ)

Graphs from Google's "Massively Multilingual NMT in the Wild" paper

1.6 MT Framework: fairseq

pytorch/fairseq

Facebook AI Research Sequence-to-Sequence Toolkit written in Python.

python pytorch artificial-intelligence

9.9k · Python · MIT license · Updated 1 hour ago · 82 issues need help

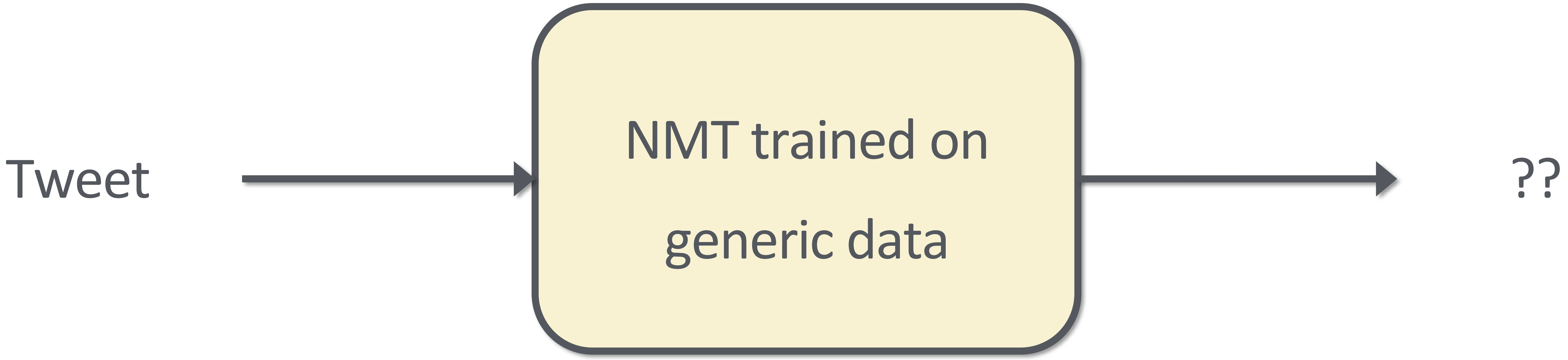
Fast training, thanks to:

1. Multi-GPU and delayed updates
2. Large batches: increased learning rate and faster convergence
3. Mixed-precision training

2. Domain Adaptation

2.1 Domain shift

DEVIEW 2020



2.2 Domain-specific issues

- Lexical ambiguity → *carte* in French (means *card*, *map*, or *menu*)
- Unknown words or expressions → "noix de Saint-Jacques"
- Wrong level of politeness → *tu / vous* in French
- Robustness to noise → typos, missing punctuation, etc.
- Others → dialect, punctuation norms, etc.

2.3 Examples of domains

INPUT: 방탄소년단 보라해

GENERIC: BTS Borahae

DOMAIN-ADAPTED: BTS, I purple you.

INPUT: 여자친구 노래 들어보자

GENERIC: Let's listen to your girlfriend's song.

DOMAIN-ADAPTED: Let's listen to GFRIEND's song.

2.3 Examples of domains

INPUT: 노래 틀고 놀자 그래 좋은 것 같아 노래 틀어줘

GENERIC: Let's play some music. Yeah. Sounds good. Play some music.

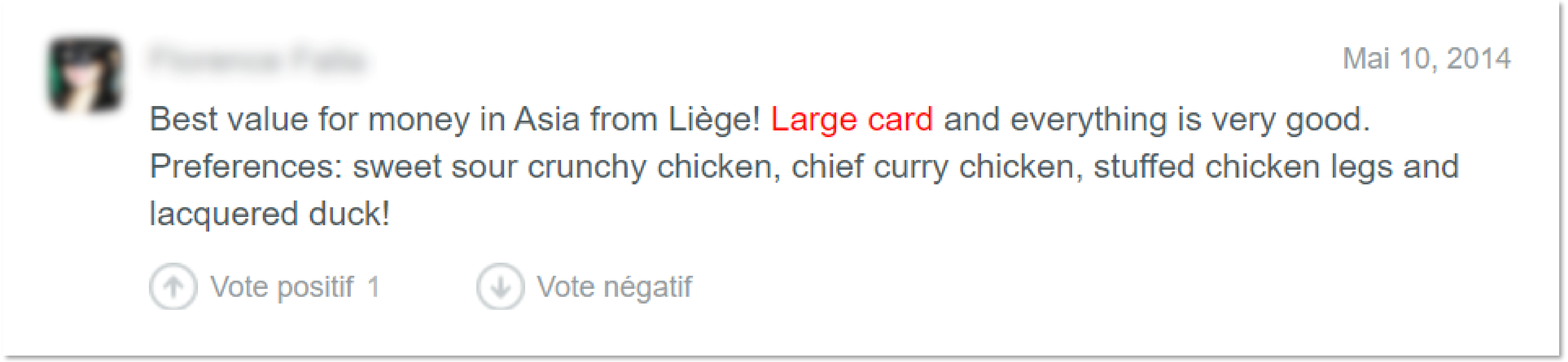
DOMAIN-ADAPTED: Let's play some music. I think it's good. Play some music.

INPUT: 그렇지x2 멋있어x2

GENERIC: That's right. x2 It's cool. x2

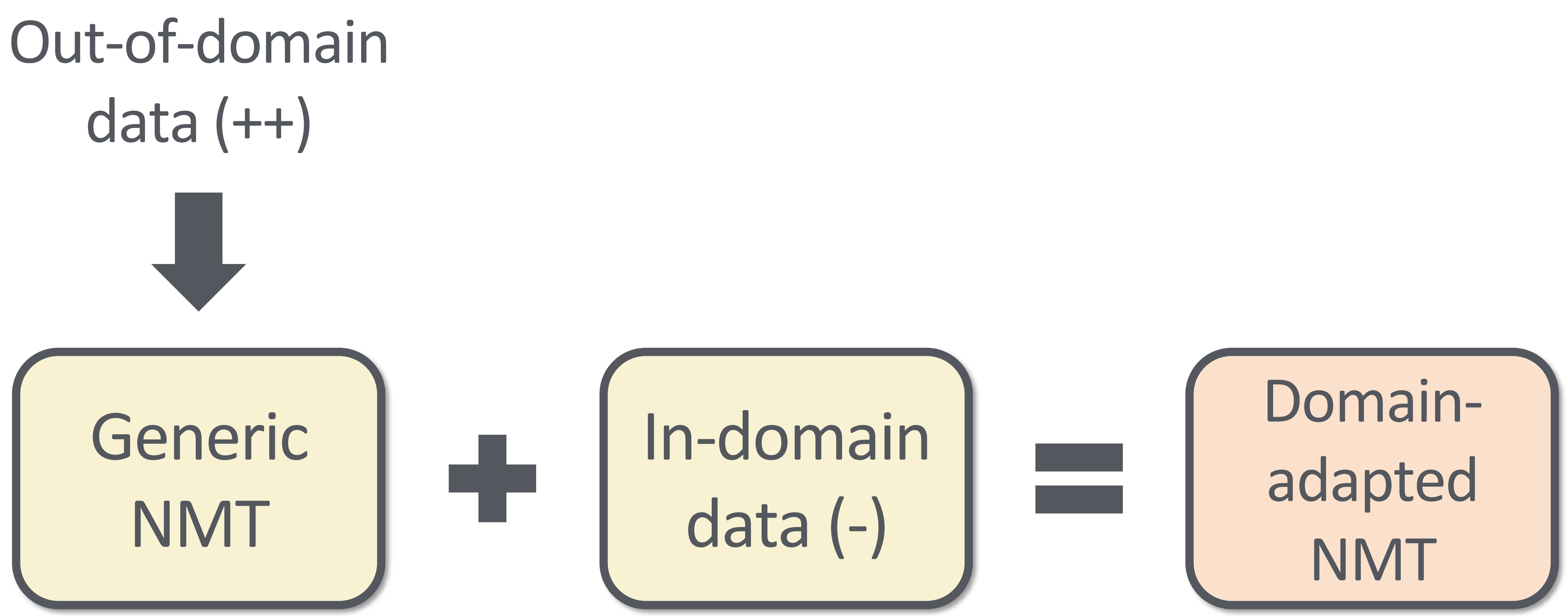
DOMAIN-ADAPTED: That's right. x2 Cool. x2

2.3 Examples of domains



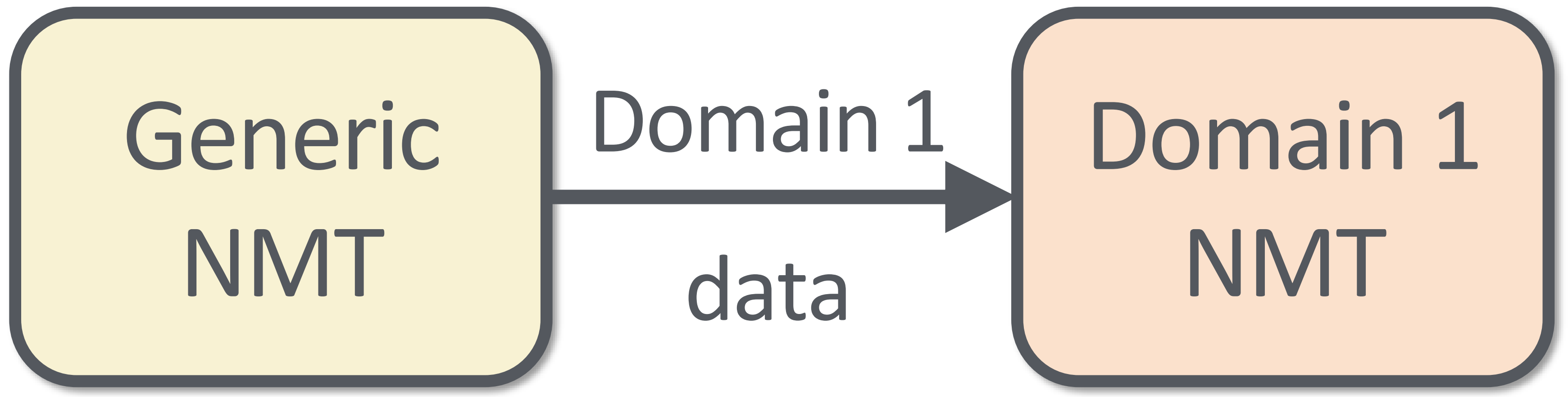
→ "Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness", *Berard et al. (2019)*

2.4 What is domain adaptation?



2.5 Fine-tuning

Continue training the generic model on the in-domain data only



Advantages:

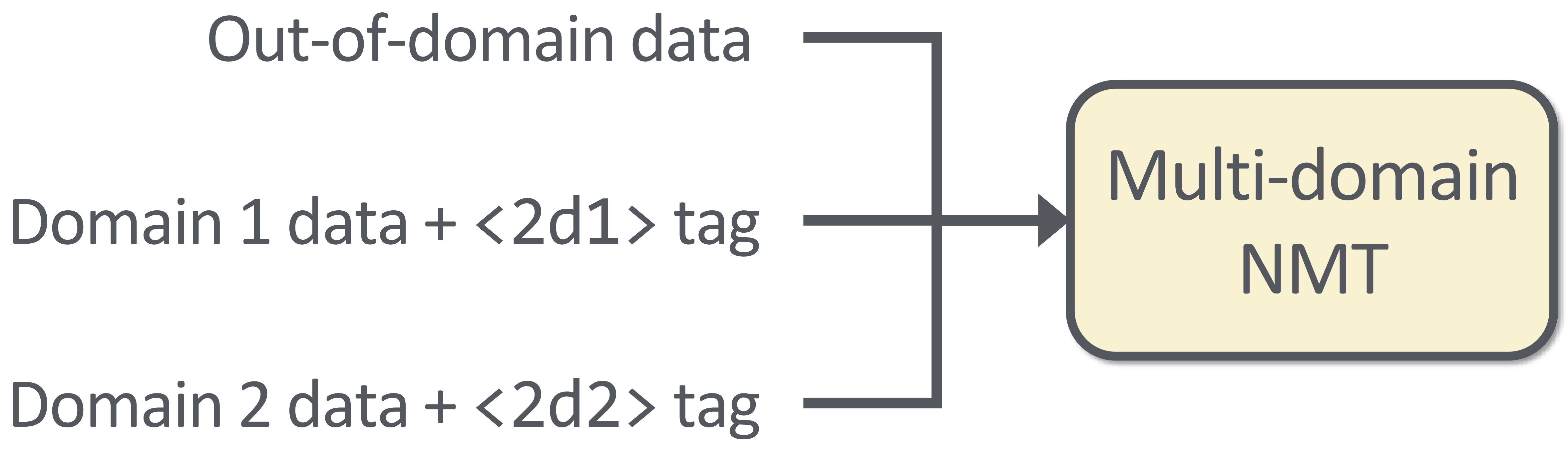
- Quick and easy
- Best in-domain performance

Drawbacks:

- Catastrophic forgetting → quality degradation on the other domains
- One model per domain

2.6 Domain tags

Train a model from scratch with data from multiple domains, with control tags indicating the domain



Advantages:

- Easy to implement
- Multi-domain models with easy control at inference time

Drawbacks:

- Very costly → need to re-train the model from scratch each time a new domain comes in

3. Our Model

3.1 Training data

DEVIEW 2020

Generic: *ParaCrawl, United Nations, Europarl, AI Hub, OpenSubtitles, WikiMatrix, IWSLT (TED Talks), etc.*

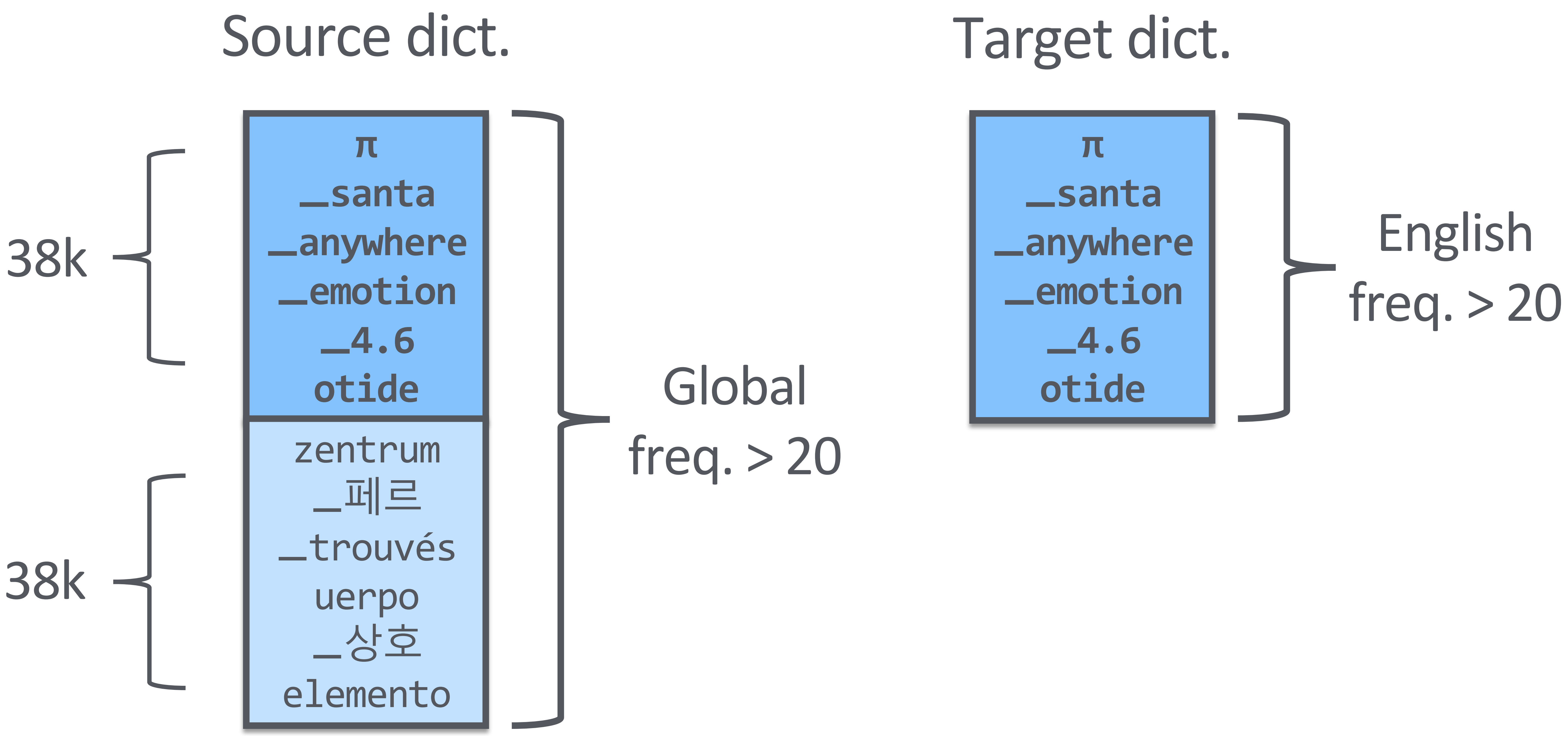
Biomedical: *Medline, TAUS, UMLS, etc.*

Language	Total	Generic	Back-translated	Biomedical
French	124.0	119.7		4.3
Spanish	88.9	86.2		2.7
German	84.5	81.6		2.9
Italian	43.9	43.1		0.8
Korean	13.4	5.5	7.8	0.1
Total	**354.8**	**336.2**	**7.8**	**10.8**

Numbers in millions of sentence pairs (with **English** as target)

3.2 Pre-processing

- Joint BPE with SentencePiece → *partially shared* source/target vocabulary & embeddings




```
target_embed = source_embed[:38000]    # share parameters
```

3.2 Pre-processing

- Joint BPE with SentencePiece → *partially shared* source/target vocabulary & embeddings
- **Inline casing:** robustness to capitalization

NAVER Labs Europe is in Grenoble

⇩

▁na | <U> | ver | <U> | ▁labs | <T> | ▁europe | <T> | ▁is | ▁in | ▁grenoble | <T>

→ "NAVER LABS Europe's Systems for the WMT19 Machine Translation Robustness Task", *Berard et al. (2019)*

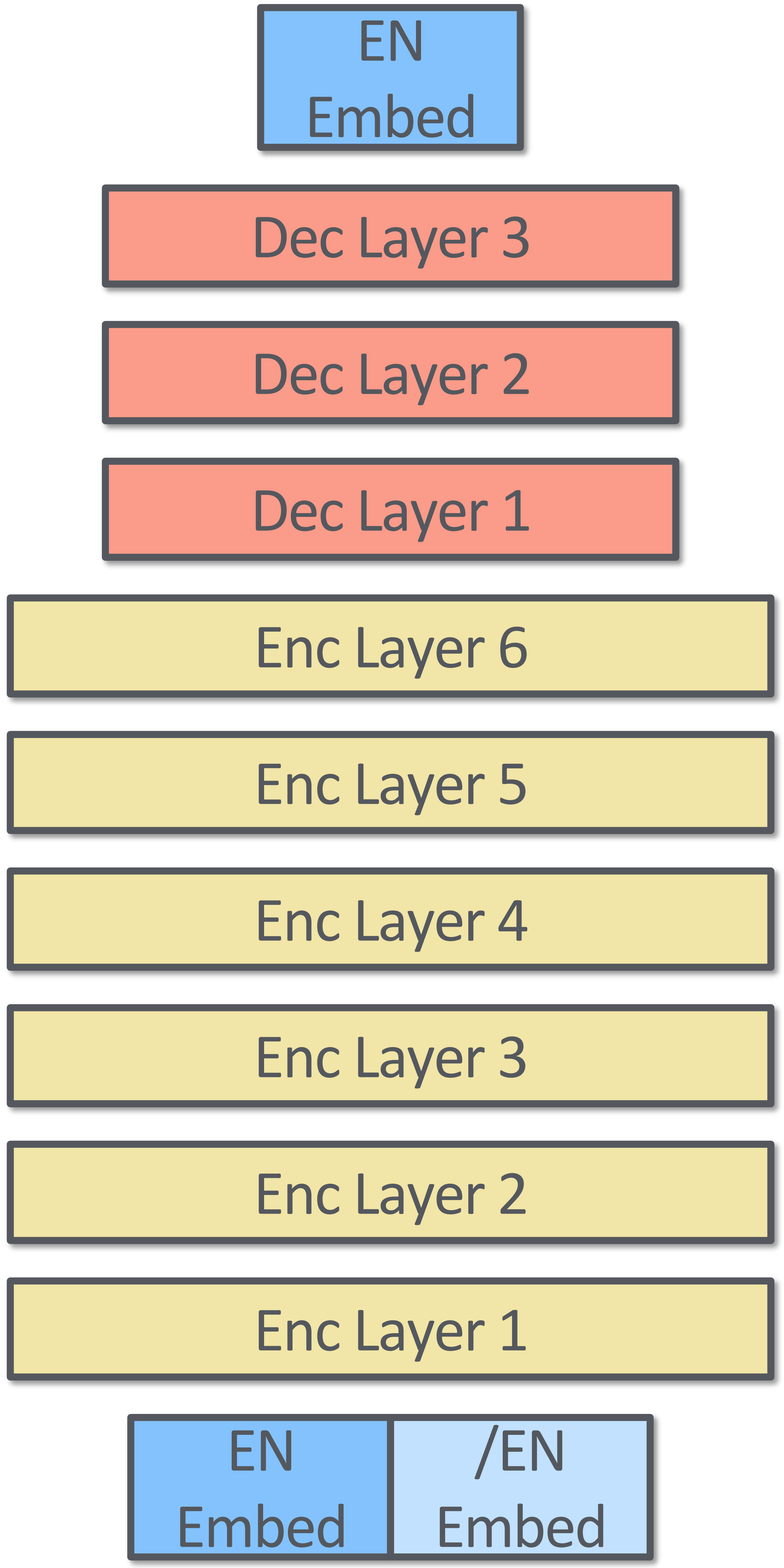
3.3 Hyper-parameters

Transformer Big with:

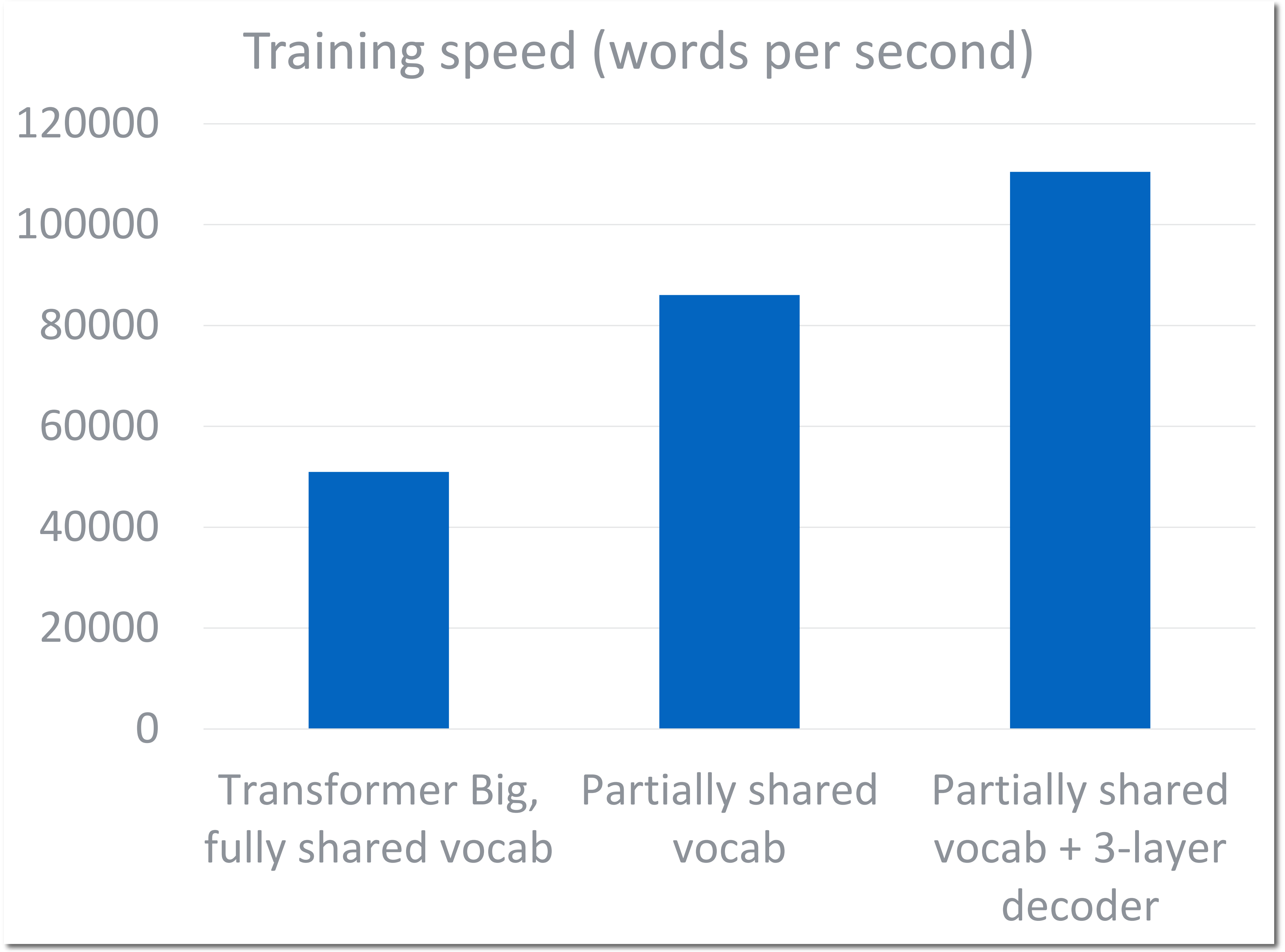
- *Wide* 6-layer encoder of size 8192
- *Shallow* 3-layer decoder

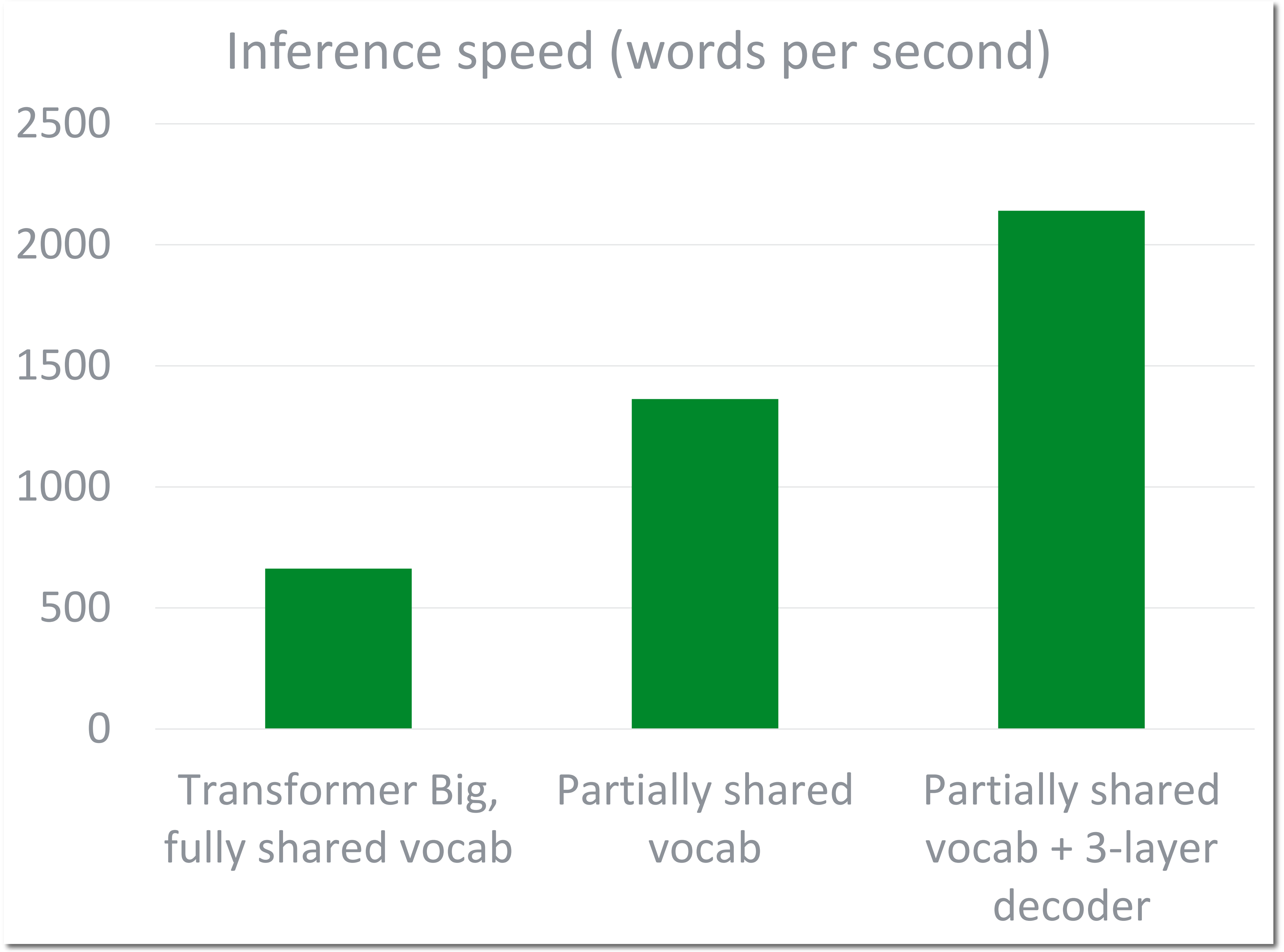
Total parameters: 254M

Checkpoint size: 2.7GiB



3.3 Hyper-parameters





3.3 Hyper-parameters

- Up-sampling biomedical data by × 2
- `<medical>` tag before each source biomedical sentence

Language	Generic	Back-translated	Biomedical	Total
French	119.7		4.3 × 2	128.3
Spanish	86.2		2.7 × 2	91.6
German	81.6		2.9 × 2	87.4
Italian	43.1		0.8 × 2	44.7
Korean	5.5	7.8	0.1 × 2	13.5
Total	**336.2**	**7.8**	**10.8**	**365.5**

3.3 Hyper-parameters

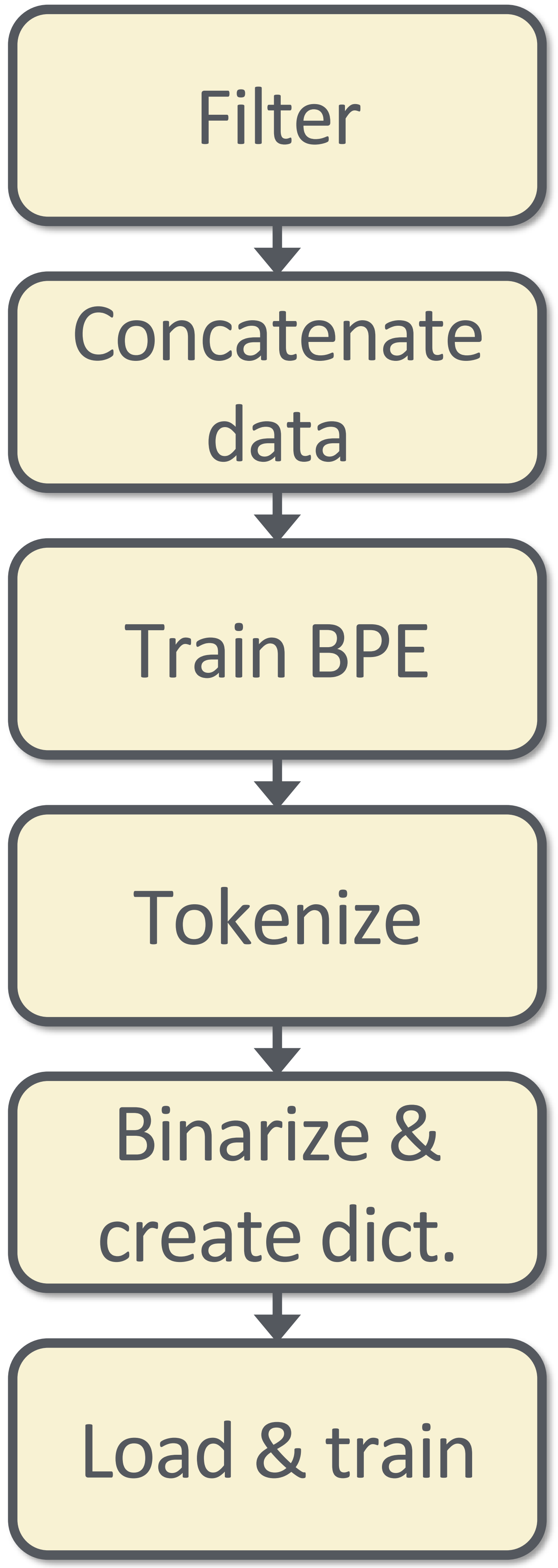
DEVIEW 2020

- Up-sampling biomedical data by × 2
- `<medical>` tag before each source biomedical sentence
- Per-language sampling with temperature 5

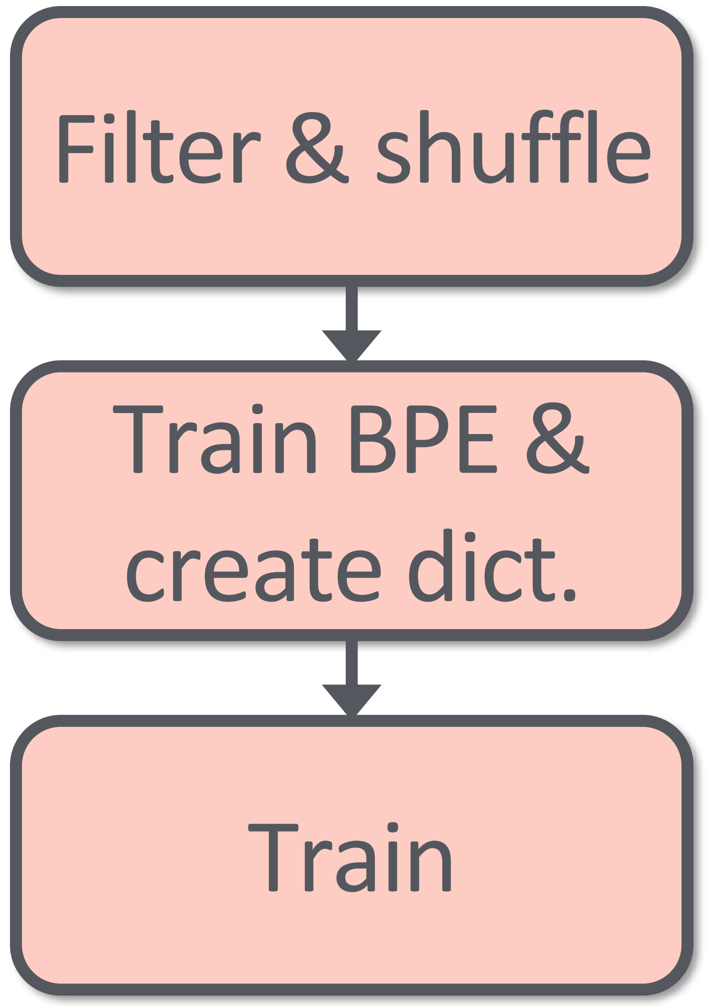
Language	Generic	Back-translated	Biomedical	Total	Frequency	Sampling probability
French	119.7		4.3 × 2	128.3	0.35	**0.23**
Spanish	86.2		2.7 × 2	91.6	0.25	**0.22**
German	81.6		2.9 × 2	87.4	0.24	**0.21**
Italian	43.1		0.8 × 2	44.7	0.12	**0.19**
Korean	5.5	7.8	0.1 × 2	13.5	0.04	**0.15**
Total	**336.2**	**7.8**	**10.8**	**365.5**	**1.0**	**1.0**

3.4 On-the-fly preprocessing

fairseq's static pipeline



Our dynamic pipeline



3.4 On-the-fly preprocessing

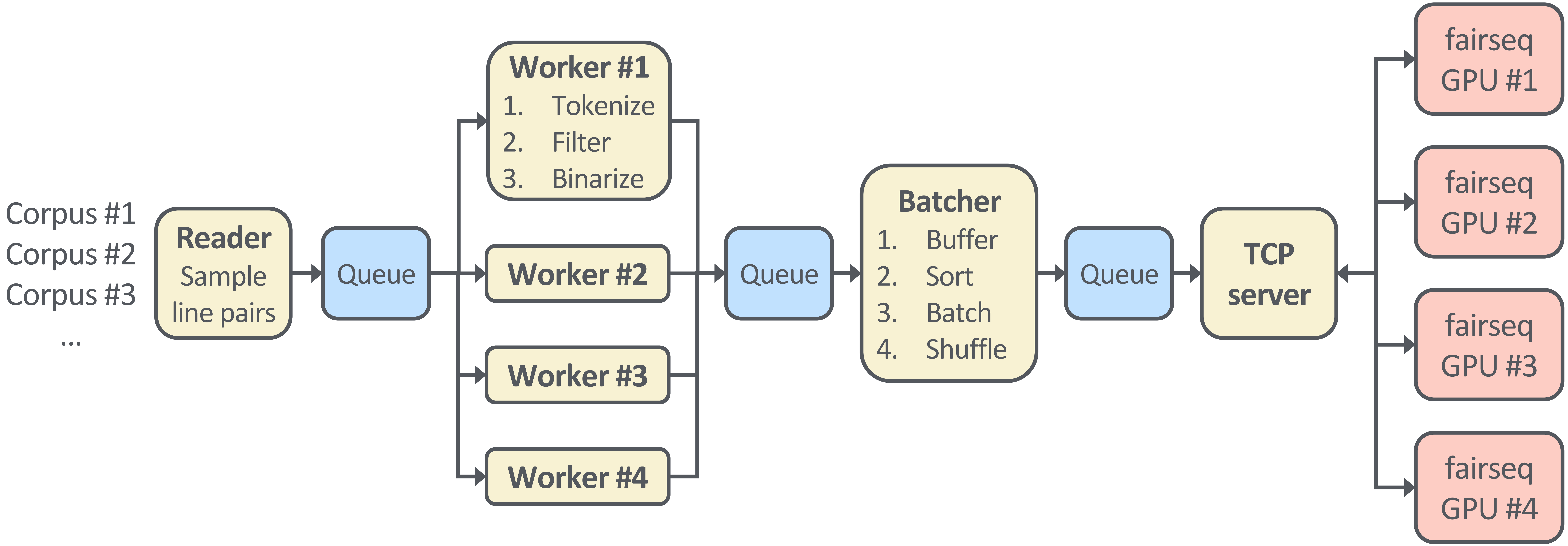
Advantages:

- Less user-side pre-processing hassle
- Easy to randomly sample from multiple corpora
- Easy to apply stochastic pre-processing (e.g., BPE dropout)

Implementation issues:

- How to efficiently pre-process and batch to avoid creating a bottleneck
- How to shuffle

3.4 On-the-fly preprocessing



3.5 How to use the model

→ https://github.com/naver/covid19-nmt

1. **Install *fairseq***

```
git clone https://github.com/pytorch/fairseq
cd fairseq python3 -m venv env
source env/bin/activate python3 -m pip install --editable .
python3 -m pip install sentencepiece
```

3.5 How to use the model

→ https://github.com/naver/covid19-nmt

2. Download the model

```
git clone https://github.com/naver/covid19-nmt.git
cat covid19-nmt/Covid19/checkpoint_best.pt.part* \
> covid19-nmt/Covid19/checkpoint_best.pt
```

3.5 How to use the model

DEVIEW 2020

→ https://github.com/naver/covid19-nmt

3. Good to go!

```
fairseq-interactive covid19-nmt/Covid19 --user-dir covid19-nmt/Covid19 \
--path covid19-nmt/Covid19/checkpoint_best.pt --medical \
-s src -t en --bpe covid19 --buffer-size 1000 --max-tokens 8000 --fp16 \
< INPUT | grep "^D" | cut -f3 > OUTPUT
```

→ "A Multilingual Neural Machine Translation Model for Biomedical Data", *Berard et al. (2020)*

4. Evaluation and examples

4.1 New Korean-English test set

Human translation of handpicked English sentences into Korean

Covers: *safety guidelines, gov. briefings, clinical tests, biomedical exp., etc.*

Source	Sentence pairs
arXiv	500
KCDC	258
All	758

arXiv → Abstracts of biomedical papers on SARS-CoV-2 and COVID-19

KCDC → Official guidelines and reports from Korea Centers for Disease Control

4.2 Generic evaluation

Language	Model	News
French	Ours	**41.0**
	NLE@WMT19	40.2
	OPUS-MT	38.0
German	Ours	**41.3**
	FAIR@WMT19	41.0
	OPUS-MT	39.5
Spanish	Ours	**36.6**
	OPUS-MT	35.2

Helsinki-NLP/**Opus-MT**

Open neural machine translation models and web services

translation-interface machine-learning natural-language-processing machine-translation neural-machine-translation language-technology language-translation-service translation-service

72 Python MIT license Updated on 12 Sep

Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task

Alexandre Bérard **Ioan Calapodescu** **Claude Roux**
Naver Labs Europe
first.last@naverlabs.com

Facebook FAIR's WMT19 News Translation Task Submission

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov
Facebook AI Research,
Menlo Park, CA & New York, NY.

4.2 Generic evaluation

Language	Model	News	TED Talks
French	Ours	**41.0**	**41.1**
	NLE@WMT19	40.2	-
	OPUS-MT	38.0	38.9
German	Ours	**41.3**	31.6
	FAIR@WMT19	41.0	**32.0**
	OPUS-MT	39.5	30.4
Spanish	Ours	**36.6**	**48.8**
	OPUS-MT	35.2	47.3
Italian	Ours	-	**42.2**
	OPUS-MT	-	40.4
Korean	Ours	-	**21.3**
	OPUS-MT	-	17.8

Helsinki-NLP/**Opus-MT**

Open neural machine translation models and web services

translation-interface machine-learning natural-language-processing machine-translation neural-machine-translation language-technology language-translation-service translation-service

☆ 72 Python MIT license Updated on 12 Sep

Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task

Alexandre Bérard **Ioan Calapodescu** **Claude Roux**
Naver Labs Europe
first.last@naverlabs.com

Facebook FAIR's WMT19 News Translation Task Submission

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov
Facebook AI Research,
Menlo Park, CA & New York, NY.

4.3 Biomedical-domain evaluation

Participation in the **WMT20 Biomedical task**

Official BLEU results (Medline 2020):

Language	Our model	(Next) best model
French	43.1	**44.1**
German	34.1	**34.8**
Spanish	**50.6**	46.4
Italian	**42.5**	42.0

BLEU on our new test set:

Language	Our model	OPUS-MT
Korean	**37.4**	20.0

Remarks:

- Single model, not an ensemble
- Multilingual
- Not tuned on Medline → also good on other domains

→ "NAVER LABS Europe's Participation in the Robustness, Chat, and Biomedical Tasks at WMT 2020", *Berard et al. (2019)*

4.4 Examples

DEVIEW 2020

SOURCE: 결과는 ACE2 발현이 코점막 상피세포에서 나타난다는 것을 보여줍니다.

REFERENCE: The result indicates that the **ACE2 expression appears** in **nasal epithelial** cells.

OUR MODEL: The results show that **the expression of ACE2 occurs** in the **nasal epithelial** cells.

OPUS-MT: The result is that **ACE2 spins out of** the **epidural upper blood** cell.

4.4 Examples

SOURCE: 근육통 또는 관절통(OR=0.28; 95%CI=0.10~0.66), 미각장애(OR=0.28; 95%CI=0.05~0.92) 및 후각소실증(OR=0.23; 95%CI=0.04~0.75)은 보호 인자였습니다.

REFERENCE: **Myalgia or artralgia** (OR=0.28; 95%CI=0.10 to 0.66), **dysgeusia** (OR=0.28; 95%CI=0.05 to 0.92) and **anosmia** (OR=0.23; 95%CI=0.04 to 0.75) were protective factors.

OUR MODEL: **Muscle or joint pain** (OR=0.28; 95%CI=0.10~0.66), **taste disturbance** (OR=0.28; 95%CI=0.05~0.92) and **olfactory loss** (OR=0.23; 95%CI=0.04~0.75) were protective factors.

OPUS-MT: A **muscle or joint pain (OR) 0.89% CI** (0.10 to 0.66), a **taste disorder (. 28; 95% CI = 0.05.92)** and a **olfactory osmosis** (OR = 0.23; 95%CI = **0.04.0.0.75**) were protective factors.

4.4 Examples

SOURCE: 벨기에의 경우, 코로나바이러스감염증-19에 대한 보고된 사망률(확인된 사례 및 잠재적 사례)과 초과 사망률은 거의 차이가 나지 않습니다.

REFERENCE: Belgium has **virtually no discrepancy** between **COVID-19** reported mortality **(confirmed and possible cases)** and excess mortality.

OUR MODEL: In Belgium, there is **little difference** between the reported mortality rate **(confirmed and potential cases)** and the excess mortality rate for **coronavirus-19.**

OPUS-MT: In Belgium, the reported mortality rate for **Coronar virus infection -- 19 --** and the excess mortality rate **is very little different**.

4.4 Examples

SOURCE: 반대로, 저발현 유전자의 경우 경로는 림프구 분화 억제 및 T 세포 활성화를 나타냈습니다.

REFERENCE: Conversely, for **under-expressed** genes, **pathways** indicated **repression of lymphocyte differentiation** and T cell activation.

OUR MODEL: Conversely, in the case of the **hypoplastic** gene, the **pathway** showed **lymphocyte differentiation inhibition** and T cell activation.

OPUS-MT: On the other hand, in the case of a **low-end** gene, the **path** showed **lymph nodes** and T cells **activated**.

4.4 Examples

SOURCE: CSF는 68%의 경우에 hyperproteinorrachia 및 / 또는 다구증을 보인 반면, RT-PCR에 의한 SARS-CoV-2 RNA는 음성으로 나타났다.

REFERENCE: CSF showed **hyperproteinorrachia** and/or **pleocytosis** in 68% of cases whereas SARS-CoV-2 RNA by RT-PCR resulted negative.

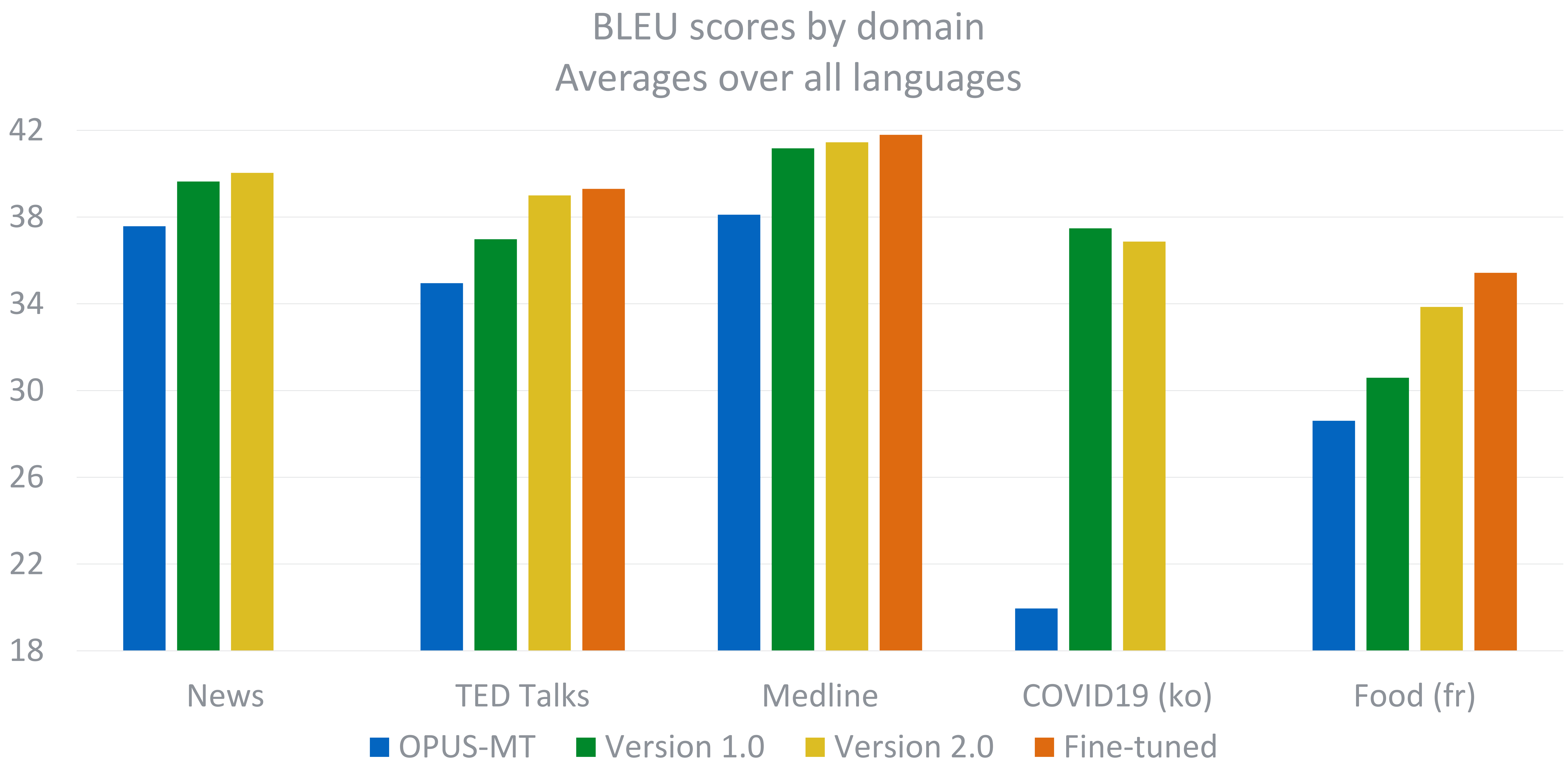
OUR MODEL: CSF showed **hyperproteinorrhea** and/or **multiple sclerosis** in 68% of cases, while SARS-CoV-2 RNA by RT-PCR was negative.

OPUS-MT: While CSF shows **hyperproteinocia** and / or **multiplexia** in 68% of cases, SARS-CoV-2 RNA in RT-PCR is negative.

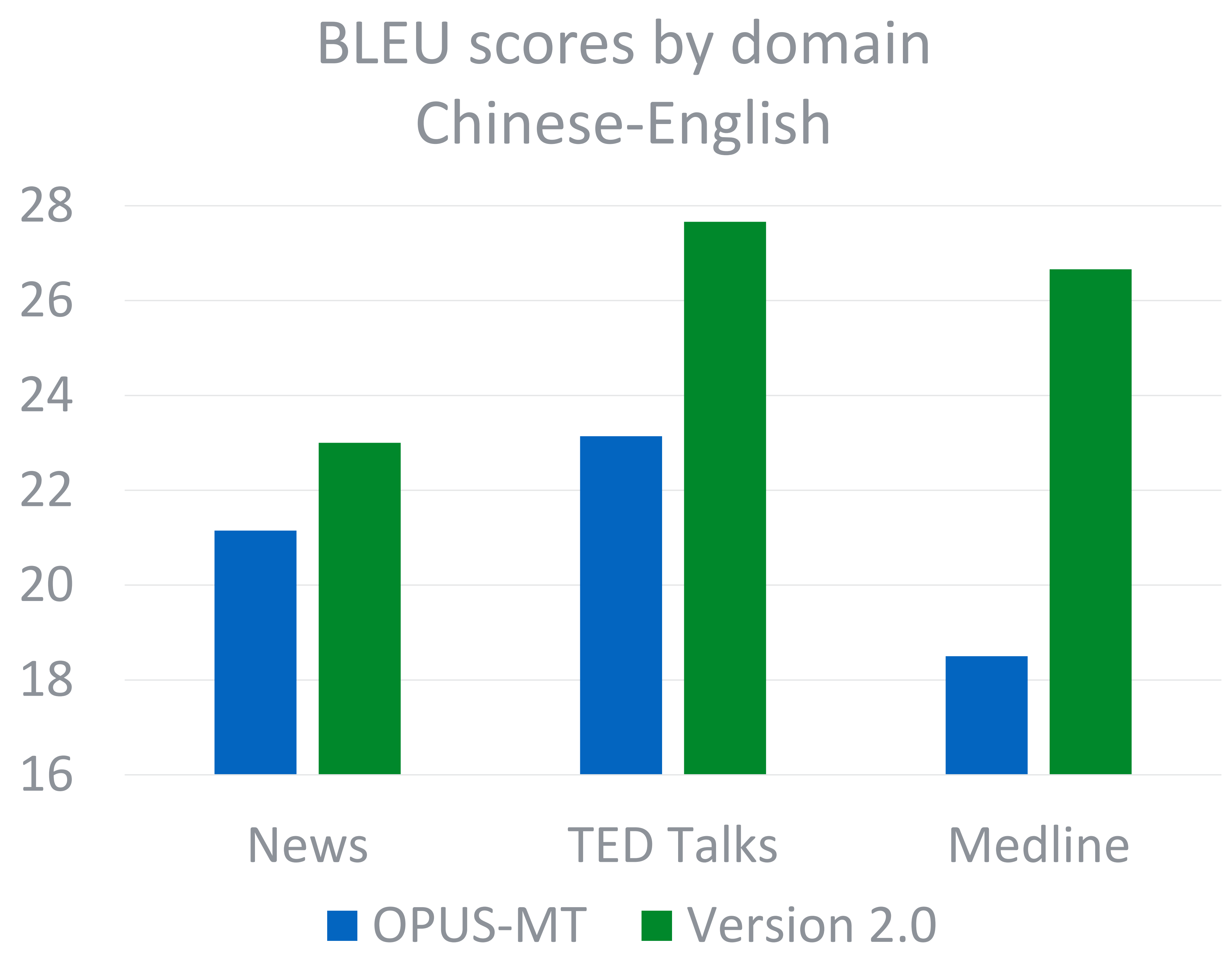
4.5 Version 2.0

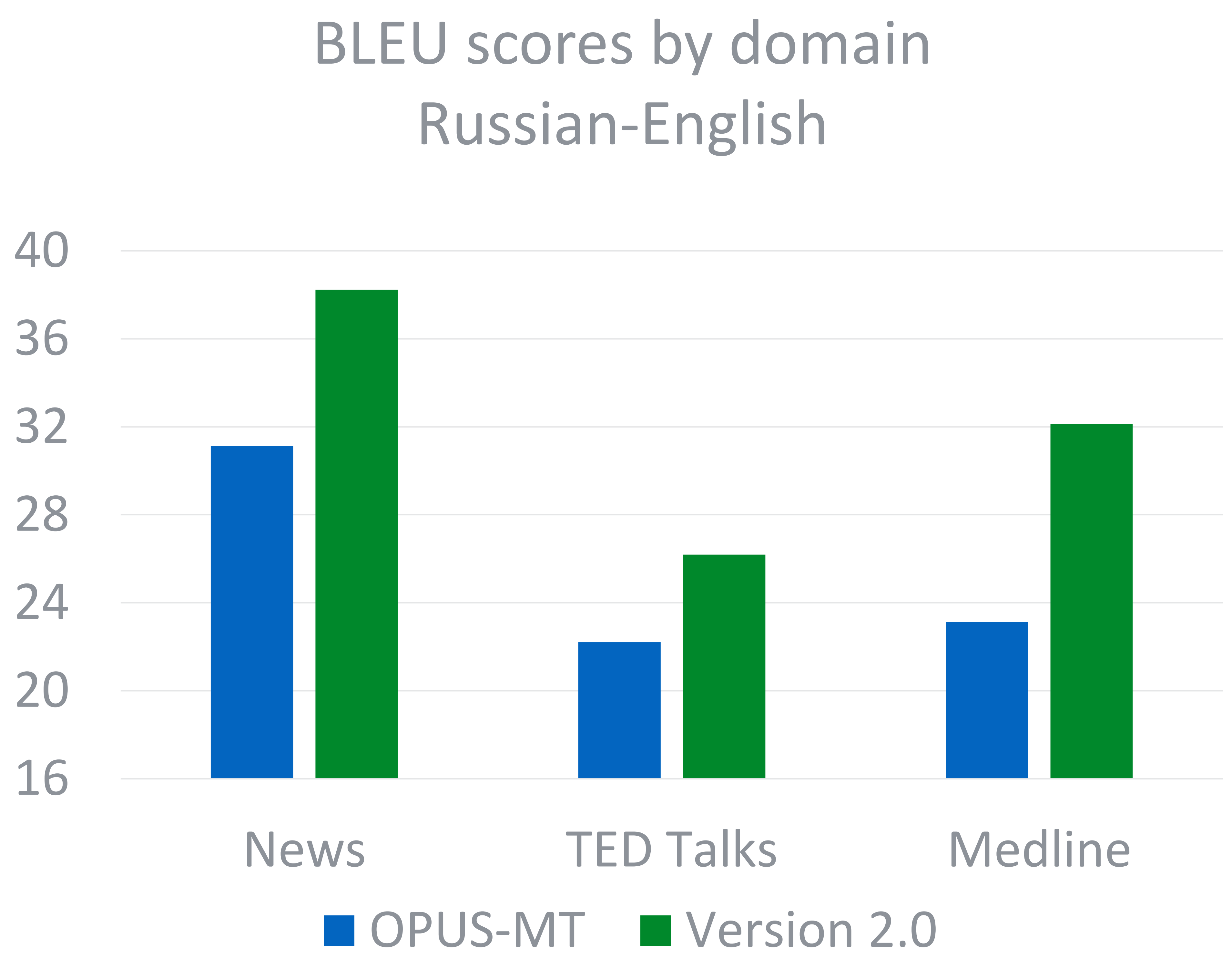
- New languages: **Chinese** and **Russian** (available in TAUS Corona Crisis corpora)
- More training data (689M)
- New domain tags: `<medical>`, `<medline>`, `<ted>`, `<food>`, `<subtitles>`, `<religious>`, `<wiki>`

4.5 Version 2.0

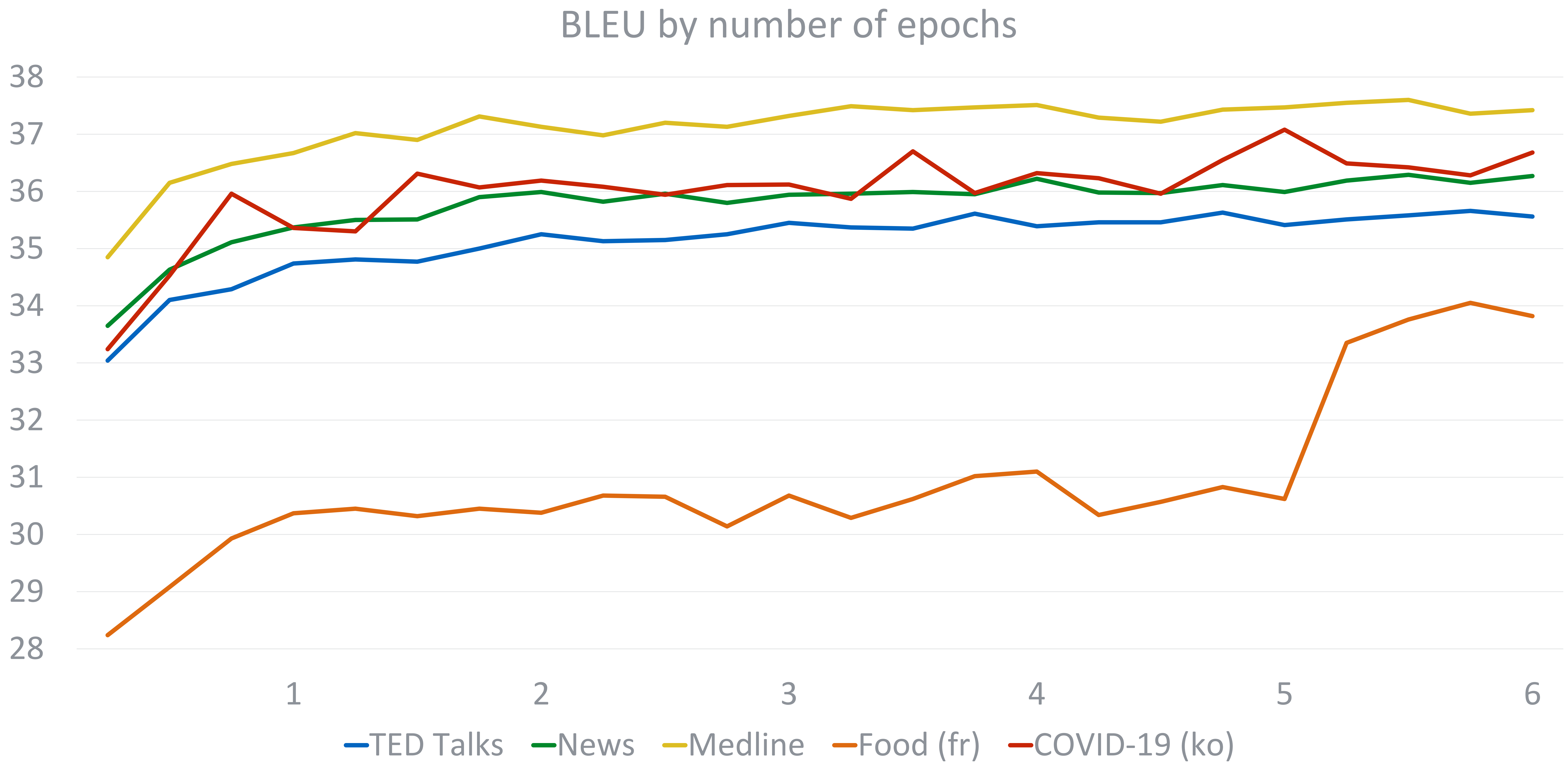


4.5 Version 2.0



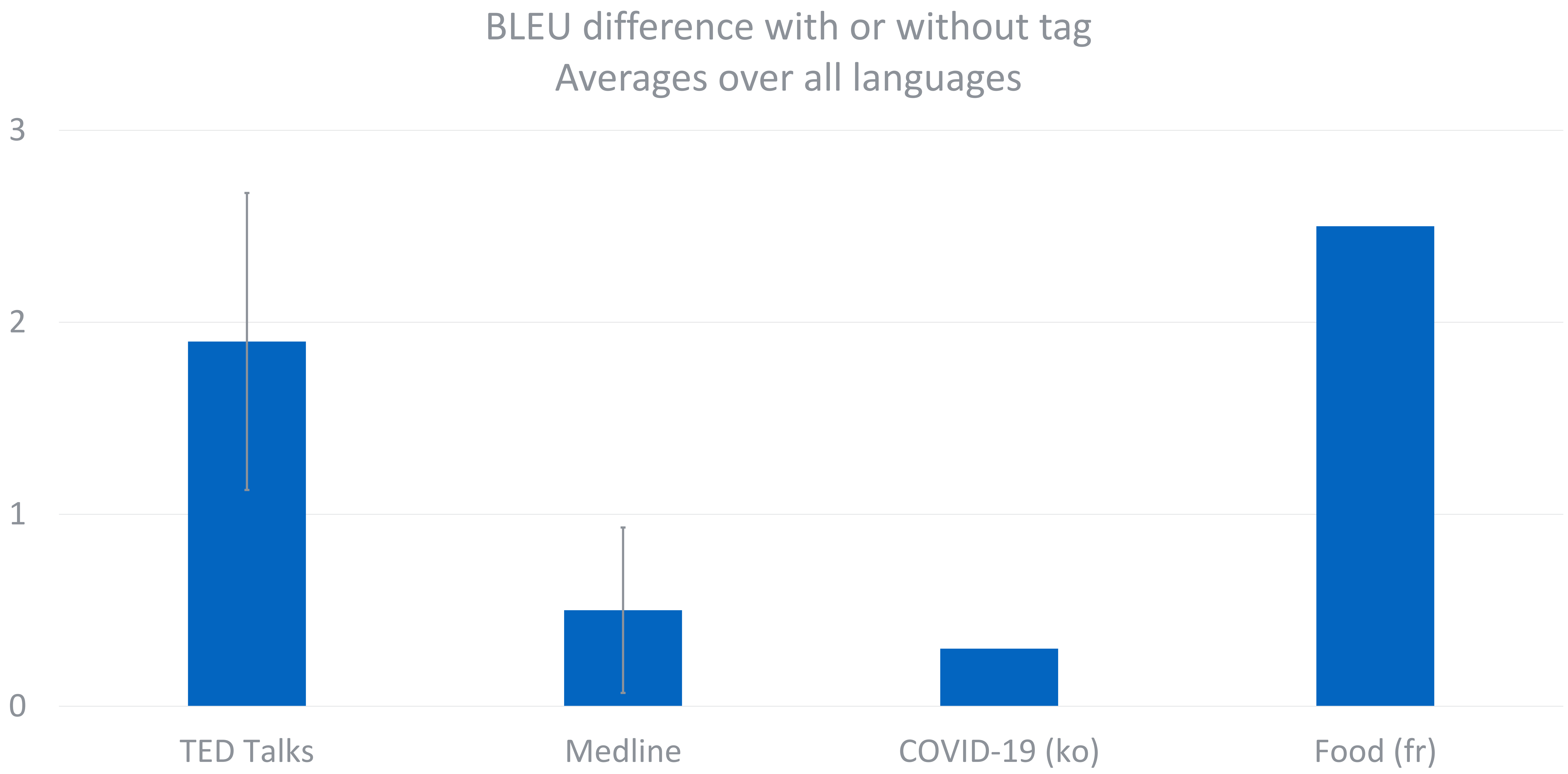


4.5 Version 2.0



- 1 epoch = 51 hours (on 4 V100s)
- 6 epochs = 13 days

4.5 Version 2.0



Conclusion

Conclusion

Contributions:

- Multilingual multi-domain model, compatible with *fairseq*
- New Korean-English COVID-19 test set
- Participation in WMT20 Biomedical Task → top performance in 2/4 languages
- Version 2.0 of this model with more languages and more domains

Future work:

- Add more languages
- Share a one-to-many model
- Share domain-specific adapter layers

Thank You