papago

BERT and NMT

Stéphane Clinchant, Kweon Woo Jung, Vassilina Nikoulina

NAVER LABS EUROPE and Papago
NAVER

Motivation

How

can **BERT** improve Machine Translation Models ?

Why

CONTENTS

1. Neural Machine Translation
2. BERT
3. Combining BERT and NMT
4. Experiments

A brief introduction to Neural MT

Neural Machine Translation

SOURCE : 존은 메리를 사랑합니다

TARGET :

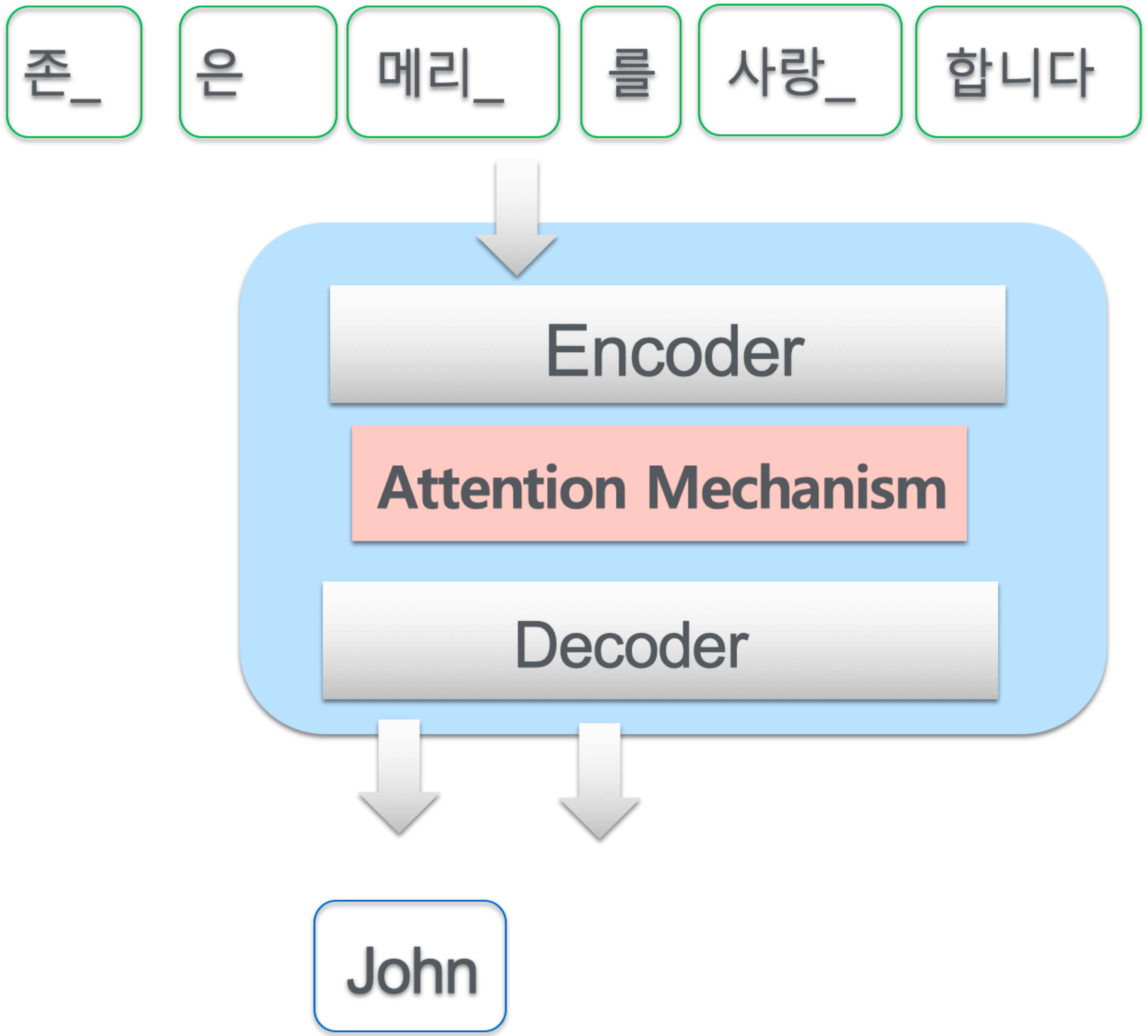
Neural Machine Translation

존_ 은 메리_ 를 사랑_ 합니다

Byte Pair encoding (BPE)

TARGET :

Neural Machine Translation



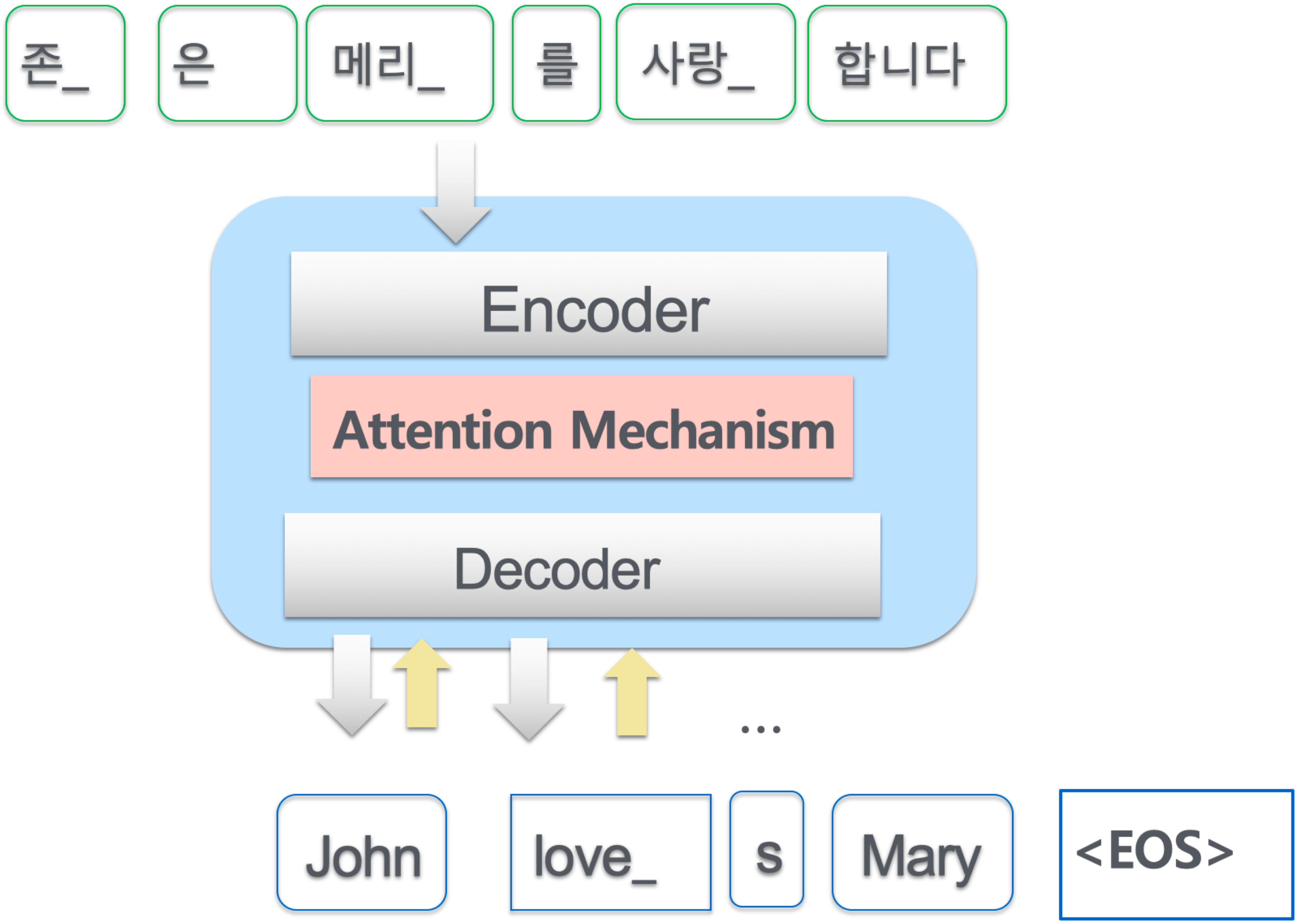
Byte Pair encoding (BPE)

TARGET :

Neural Machine Translation

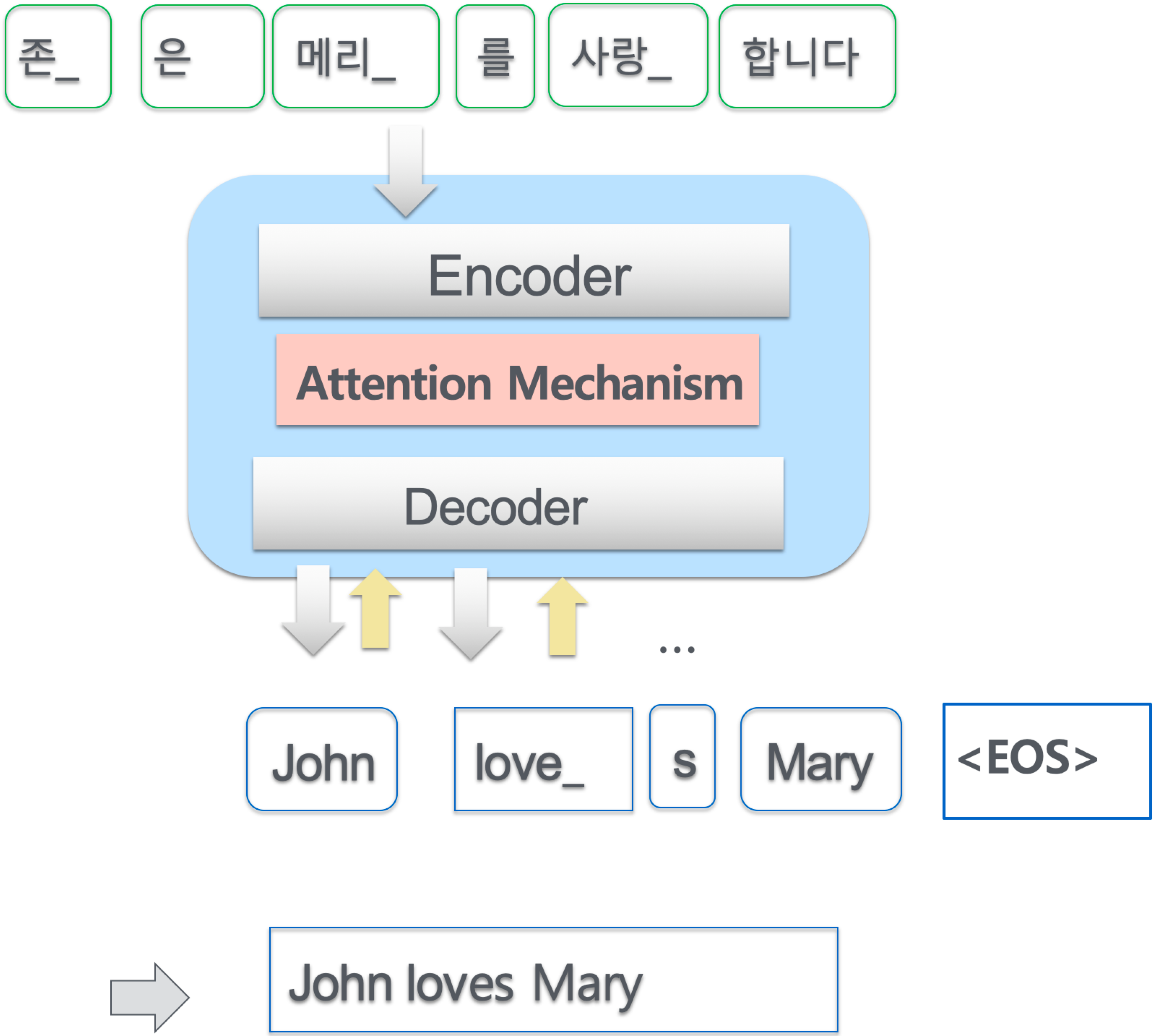
DEVIEW 2019

Byte Pair encoding (BPE)



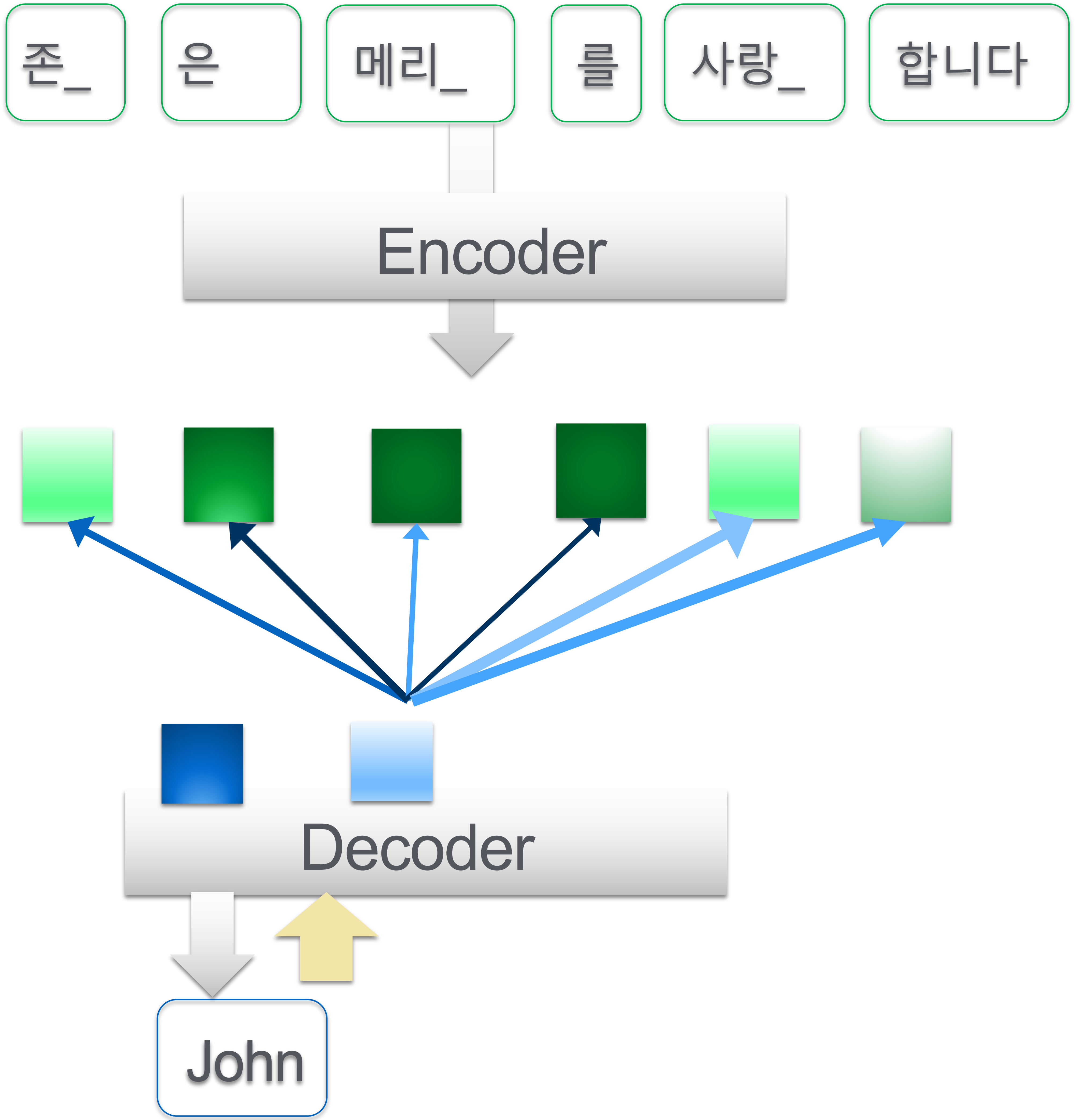
TARGET :

Neural Machine Translation



Byte Pair encoding (BPE)

Attention Mechanism



- Intermediate layer
- Learn linear combination given a query (“word” = vector)
- Flexible
- Model Contexts

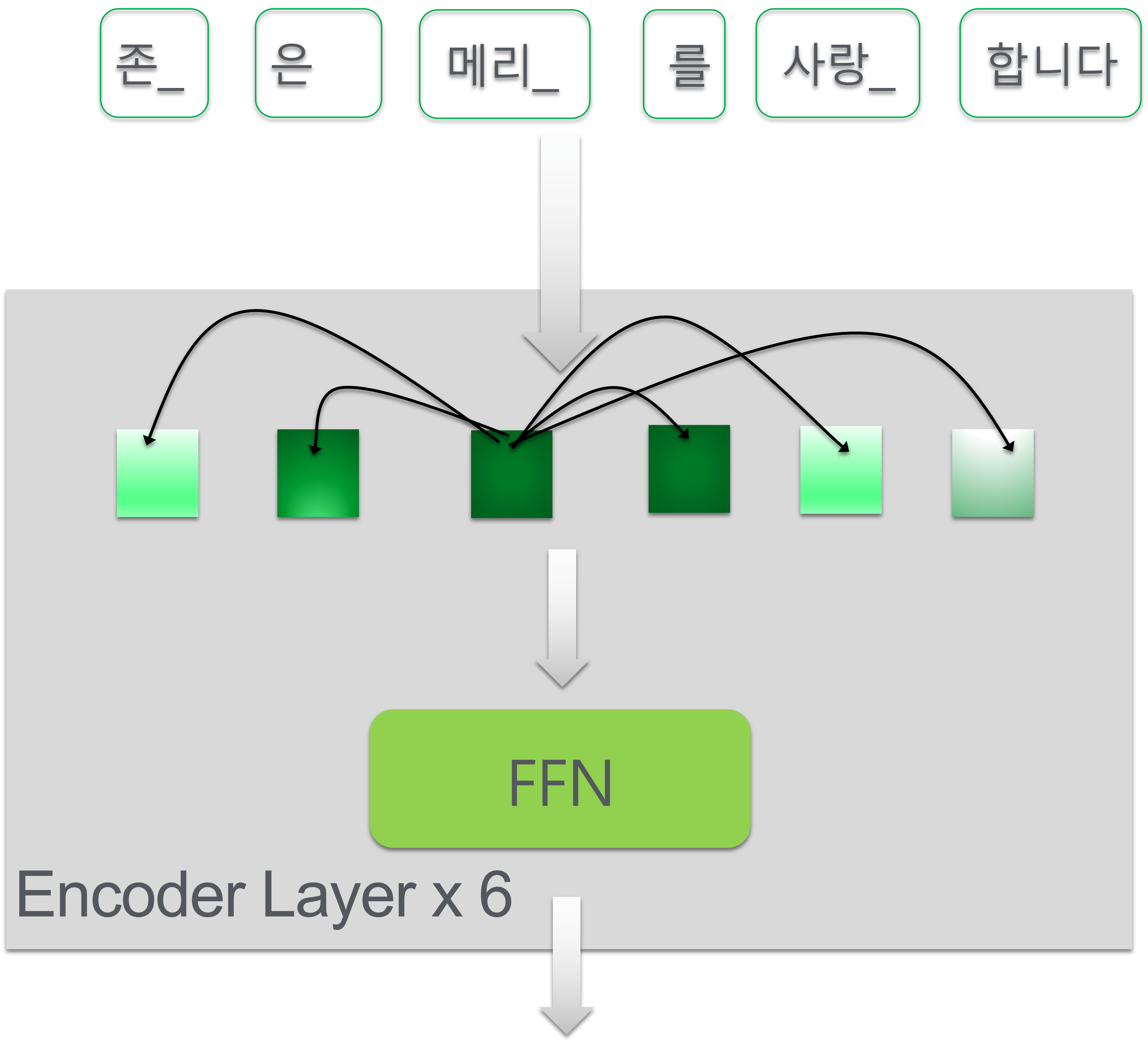
Transformer Models, Vaswani et al. 2017

- ~~RNN~~ → ~~Convolution~~ → "Attention is all you Need"

Encoder
  Self attention: Each word "pays attention" to all other words

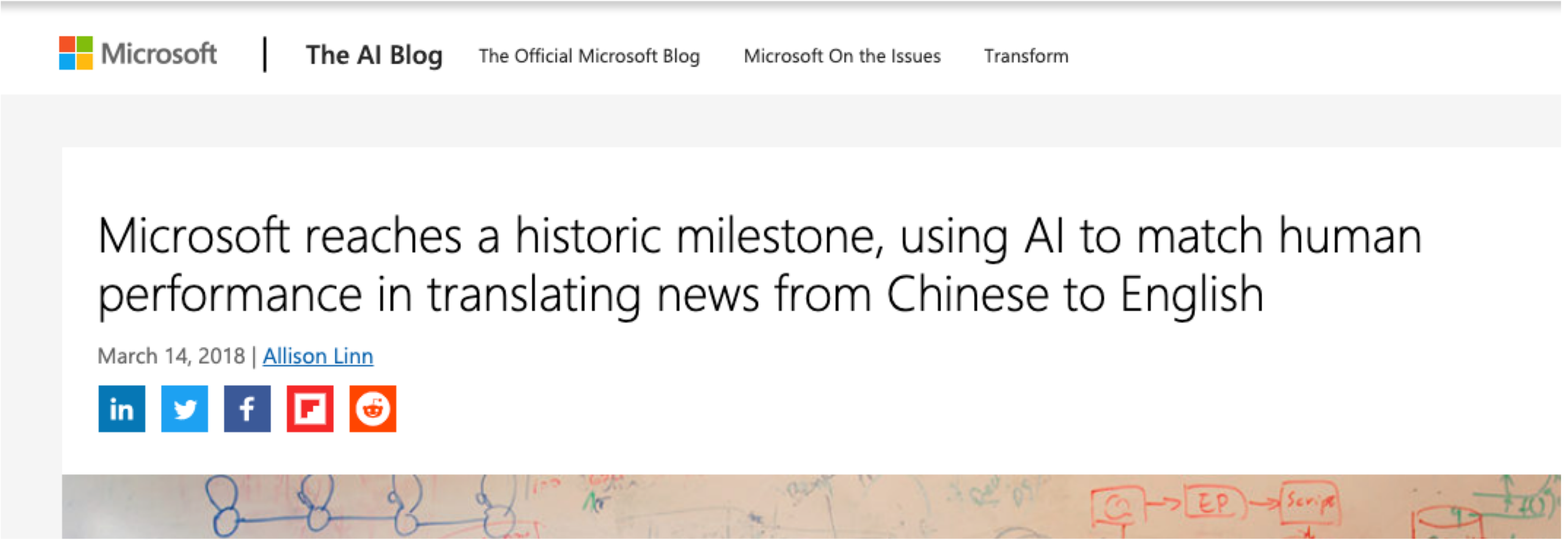
Decoder
  Each layer has 'self attention' and attention to encoder

Significant improvement !



MT reaches human parity ?

DEVIEW
2019


Microsoft | The AI Blog The Official Microsoft Blog Microsoft On the Issues Transform

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn


Microsoft MT reaches parity with (bad) human translation

Published on March 18, 2018

Tommi Nieminen
Translation Technology Developer and Translator at Own Company

1 article + Follow

Robustness of MT Models

Source sentence	"In home cooking, there is much to be discovered - with a few minor tweaks you can achieve good, if not sometimes better results," said Proctor.
translation (src)	"Beim Kochen zu Hause gibt es viel zu entdecken - mit ein paar kleinen nderungen kann man gute, wenn nicht sogar manchmal bessere Ergebnisse erzielen", sagte Proktor.
translation(UNK + src)	• *"In home cooking; there is much to be discovered- with few minor tweaks you can achieve good, if not sometimes better results", **sagte Proktor***

Robustness of MT Models

Source sentence	Freundschaft schließen durch Backen
translation (src)	Make friends through baking.
translation(ich + src) Fluent	Should you want to join us?

Example Taken from : Hallucinations in NMT https://pdfs.semanticscholar.org/9768/5859d4bcbc3b893425e6cb8fda8e9c15cfcb.pdf

Some problems with NMT

DEVIEW 2019

Very good Fluency- Adequacy ?

Too Good Language Model

Model never saw its own errors

Exposure Bias

Machine Translation Challenges

- Context Based Translations
- Model Robustness
- Evaluation is difficult
- No click logs
- Difficult Problem with a rich literature
- Experiments were/are/will be time consuming
- …

Introducing BERT

BERT

Bidirectional Encoder Representations from Transformers (Devlin et al. 2018)

- 🥇 Machine Reading
- Key ingredient of many NLP models/papers
- **excels** at transferring sentences representations

- Word Embedding → Bert Embedding
- "ResNet for Text"

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table from the original BERT paper Devlin et al.

The Masked Language Model Task (MLM)
(Cloze Task)

Predict randomly masked tokens from sentences

Sentence : 나는_ 비빔_ [MASK] 좋_ 아한다_.

Transformer Encoder

Predict : 밥을

Contextualized representation thanks to self-attention

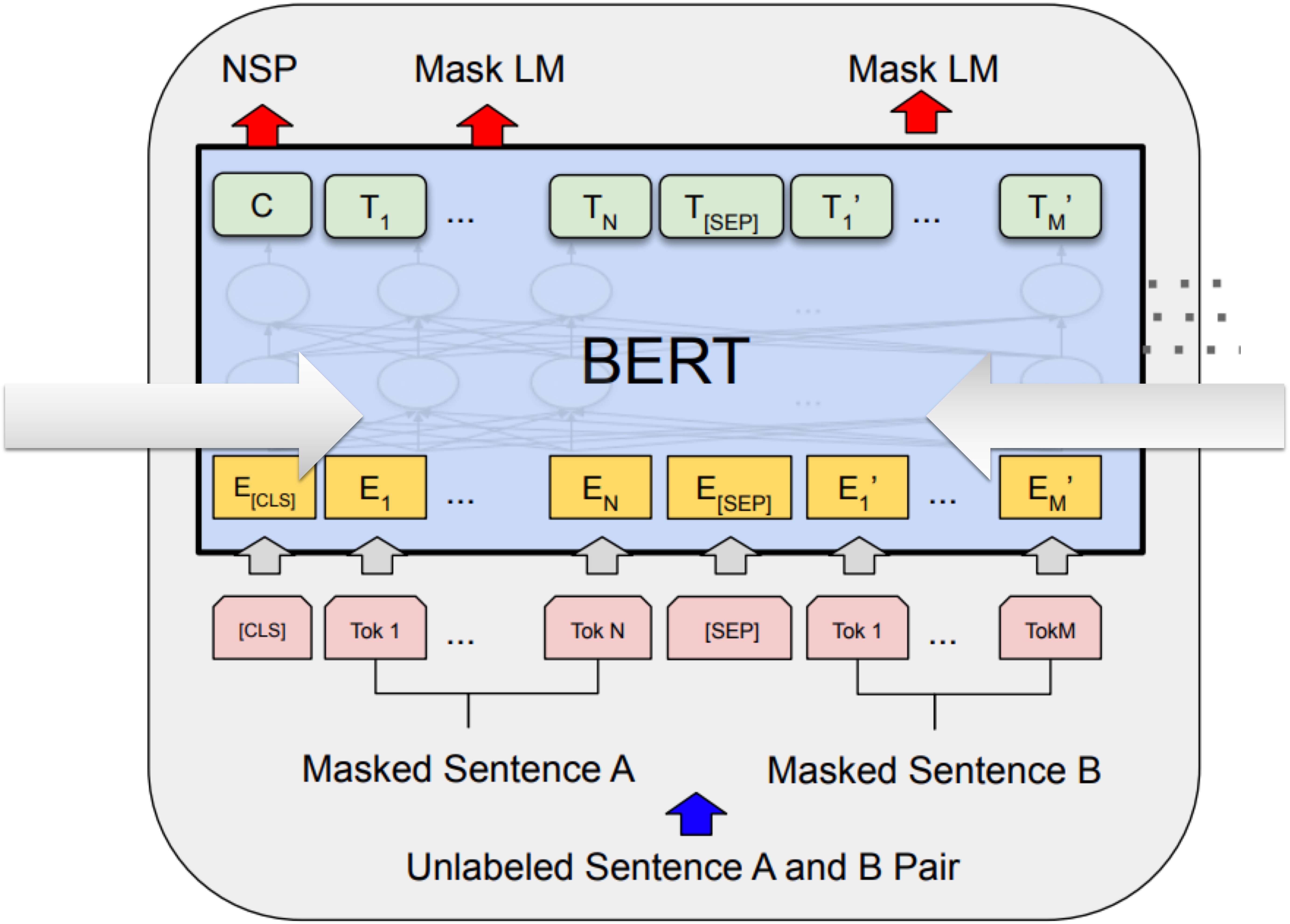
The Next Sentence Prediction Task

Are those two sentences consecutive ? Yes/No

저는 비빔밥을 좋아합니다. 하지만 저는 KPOP이 싫어요.

BERT

(Shared) Transformer Encoder

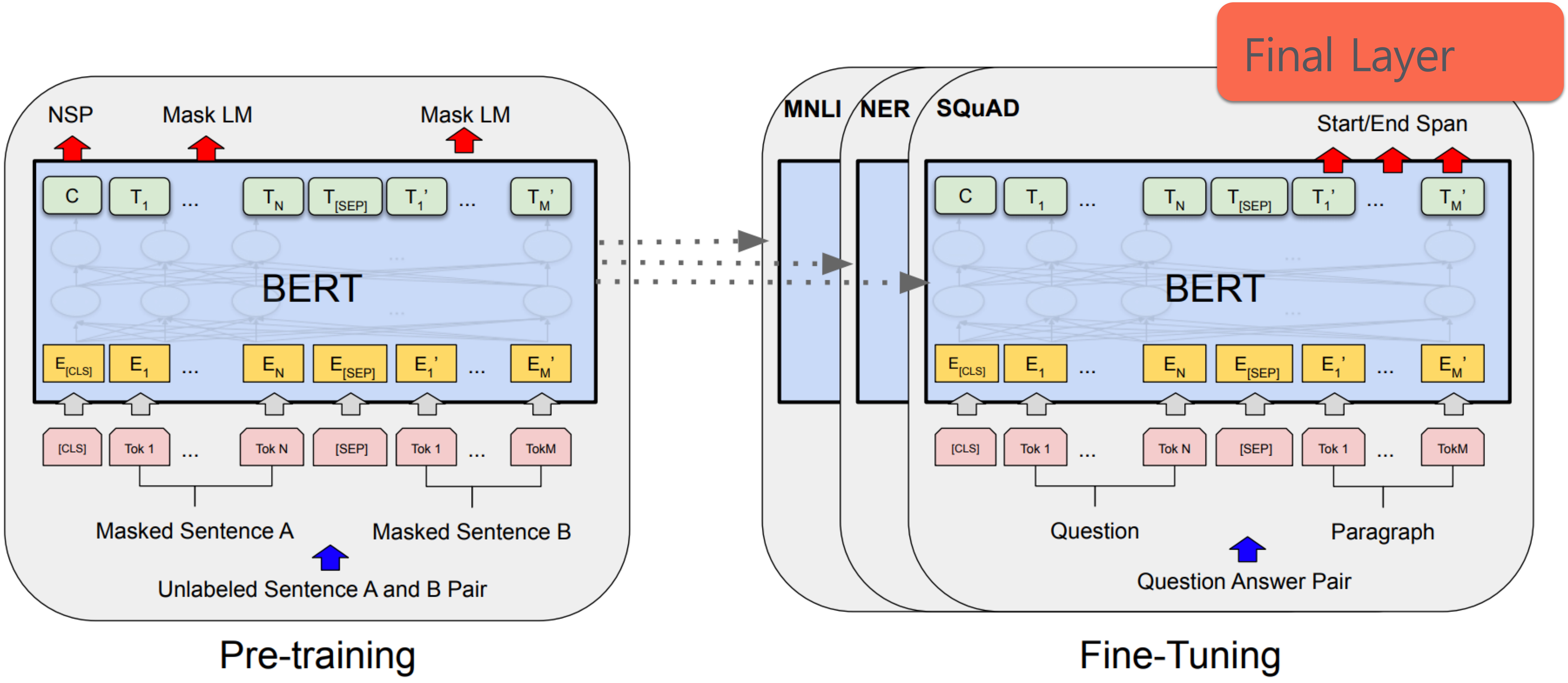


Pre-training

(Shared) Transformer Encoder

Schema from Devlin et al.

Finetuning with BERT



Schema from the original BERT paper (Devlin et al.)

Practical Details

Probability of Masking Tokens	15%
Number of Layers	12-24
Vocabulary	~30k
Parameters	110M-330M
Training Corpora	3,300 Million words
Training Time	BERT Large: 64 TPU 4 days

Two sides of the same … Encoder

BERT Transformer Encoder

« Auto Encoder »

12+ Layers

GELU, Position Embedding, Segment Embedding

NMT Transformer Encoder

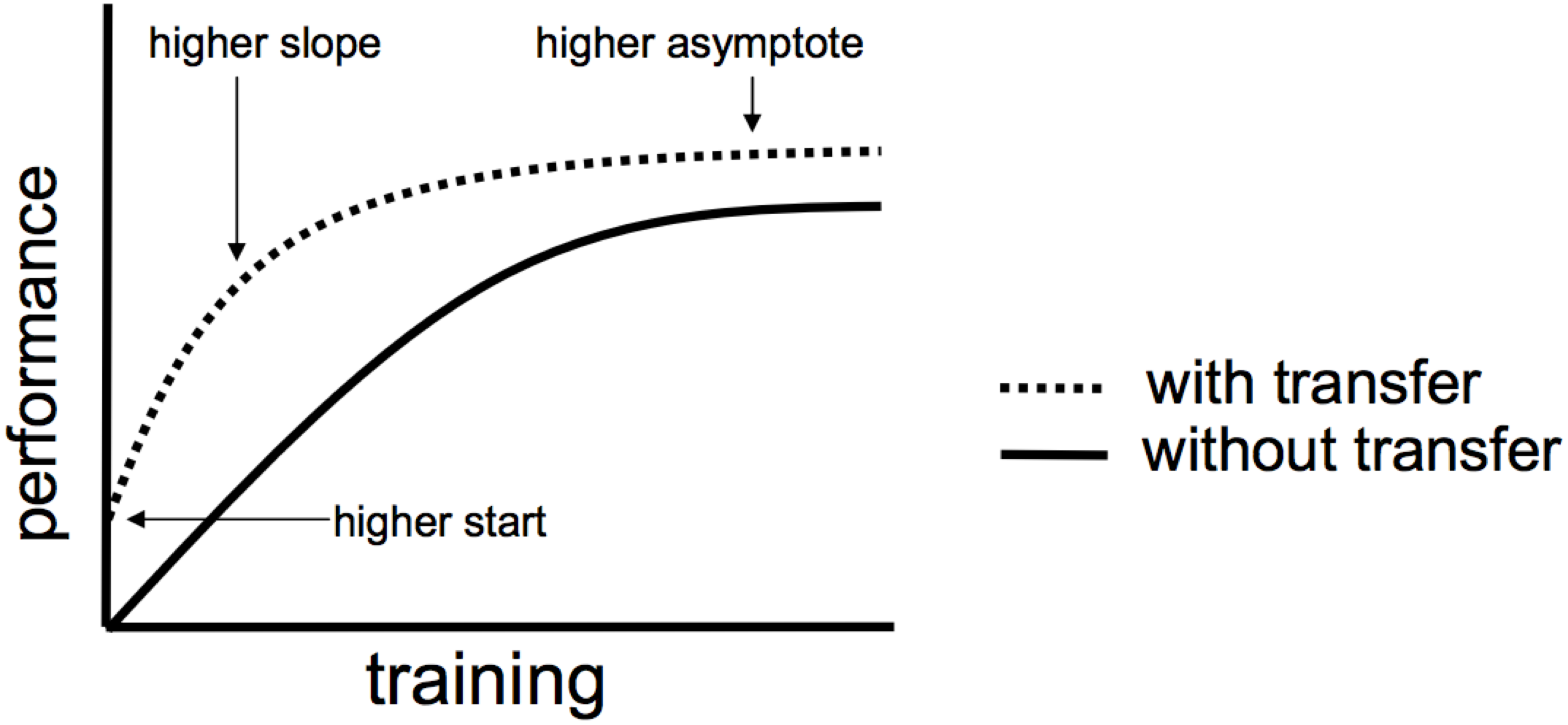
« Encoder for Translation»

6 Layers

RELU, Sinusoidal Position Embedding

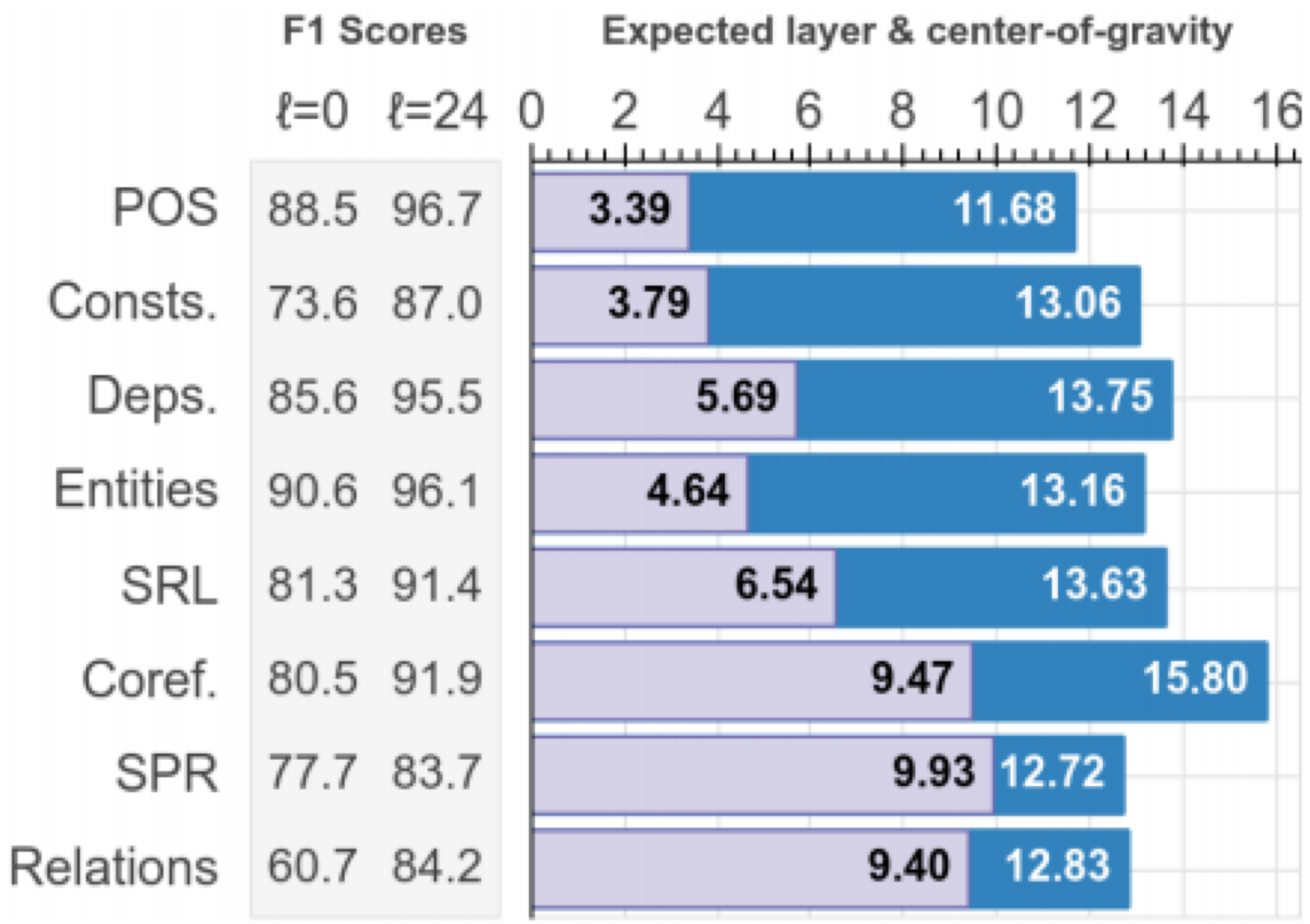
Transfer Learning with BERT ?

- **Similar Encoders: can we transfer sentences representations for NMT ?**



BERT Rediscovers the Classical NLP Pipeline, Tenney et al.

BERT is also very good at **capturing, syntactic and semantic information.**



Combining BERT and NMT

Hypothesis and Questions

- Human translation = text understanding + text generation

- BERT model learns 'text understanding' task

- **Question**: Is NMT encoder restricted to *understanding* only?
- **Hypothesis**: « *Encoder is already translating* »
  NMT encoder has an self encoding effect and translation effect:

Hypothesis and Questions

Why would pretraining with BERT work better for NMT ?

- More data → better 'understanding'
- BERT and source Domain Adaptation (Transfer Learning)?
- Can we make the encoder more robust ?

Is BERT encoder more robust?

나는_ 비빔_ [MASK] 좋_ 아한다_

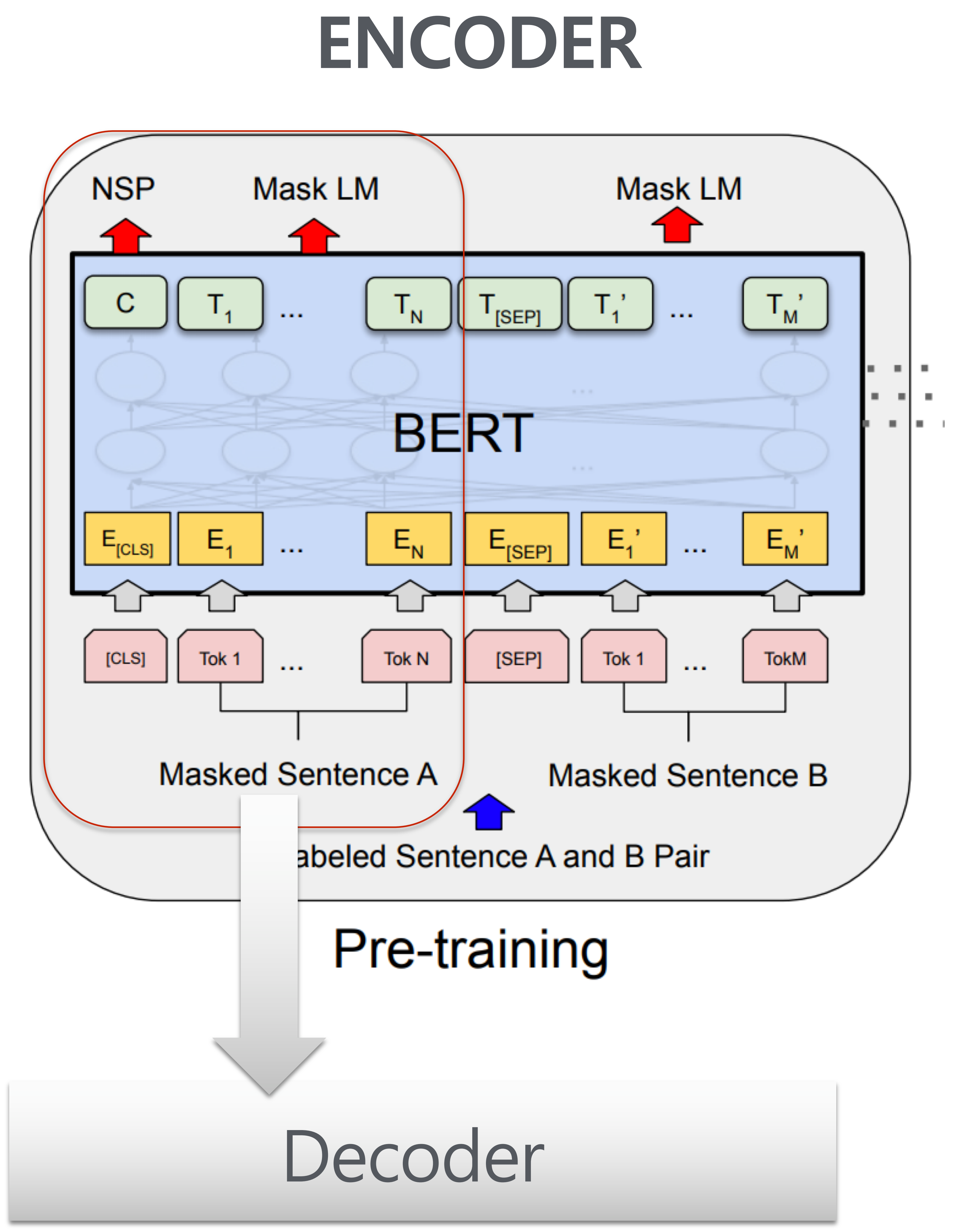
ENCODER

- BERT is trained *to deal with* missing token and find possible replacements
- Does pretraining impact rare/unknown word translation, noisy input ?

1 Finetuning approach

DEVIEW 2019

- Initialize and update BERT model
- Simplest
- Tricky for decoder

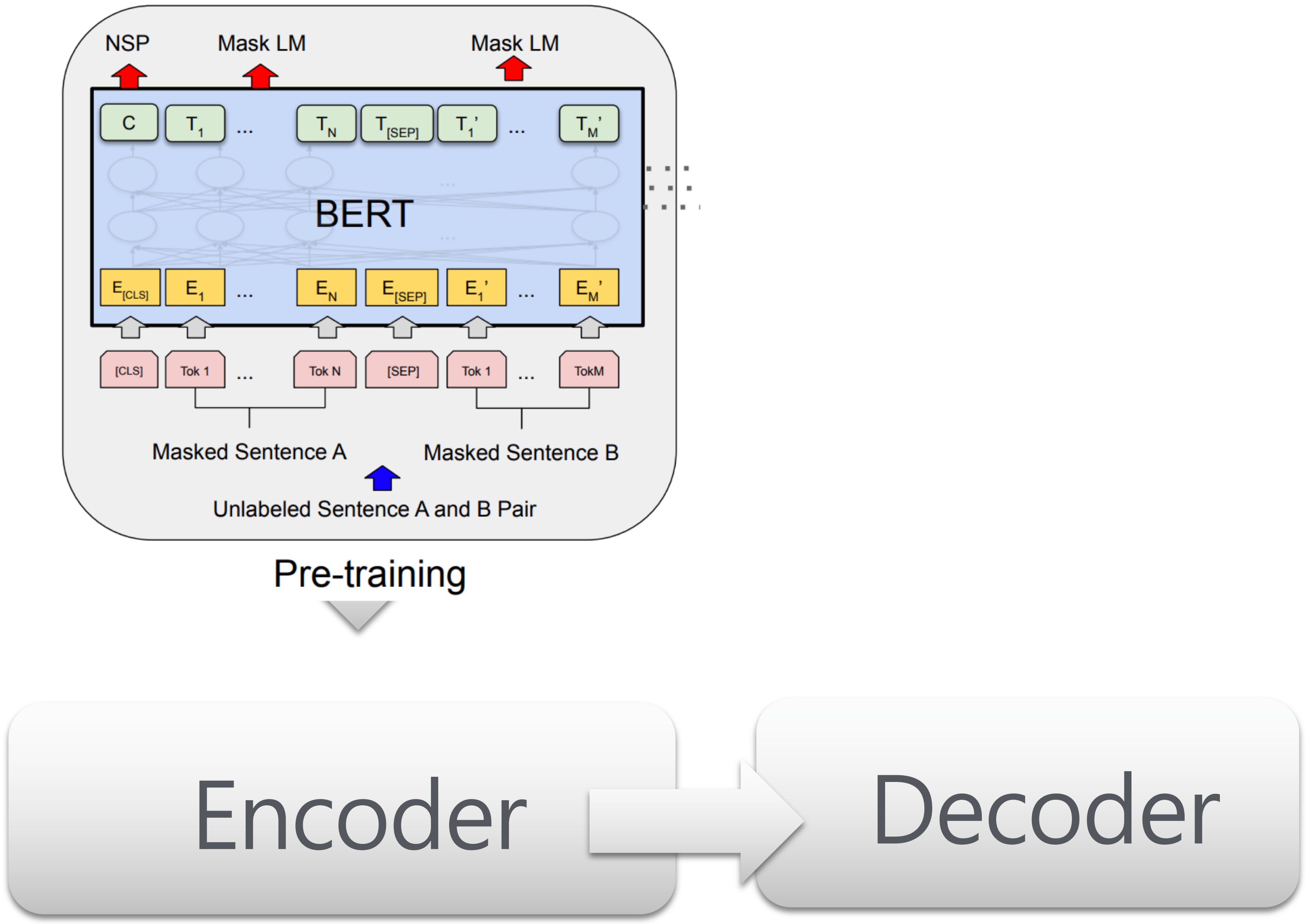


- *Cross-lingual Language Model Pretraining, Lample at al, 2019*
- *MASS: Masked Sequence to Sequence Pre-training for Language Generation, Song et al.*

2.Embedding approach

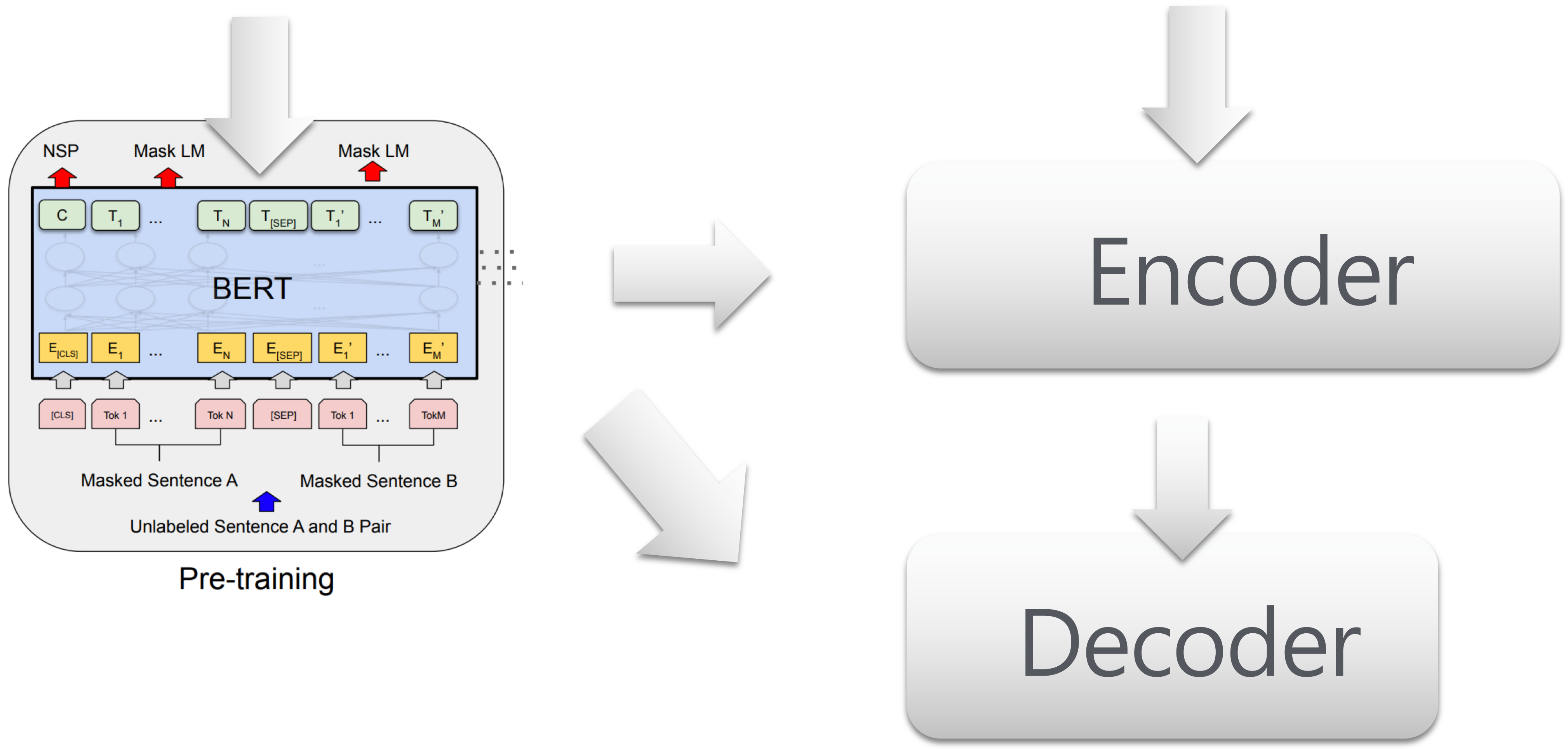
- Use BERT as the first layers of NMT- encoder
- Can easily work for encoder and decoder
- Can Reuse BERT /ELMO etc
- Deep Encoders

Pre-trained Language Model Representations for Language Generation , Edunov et al, 2019



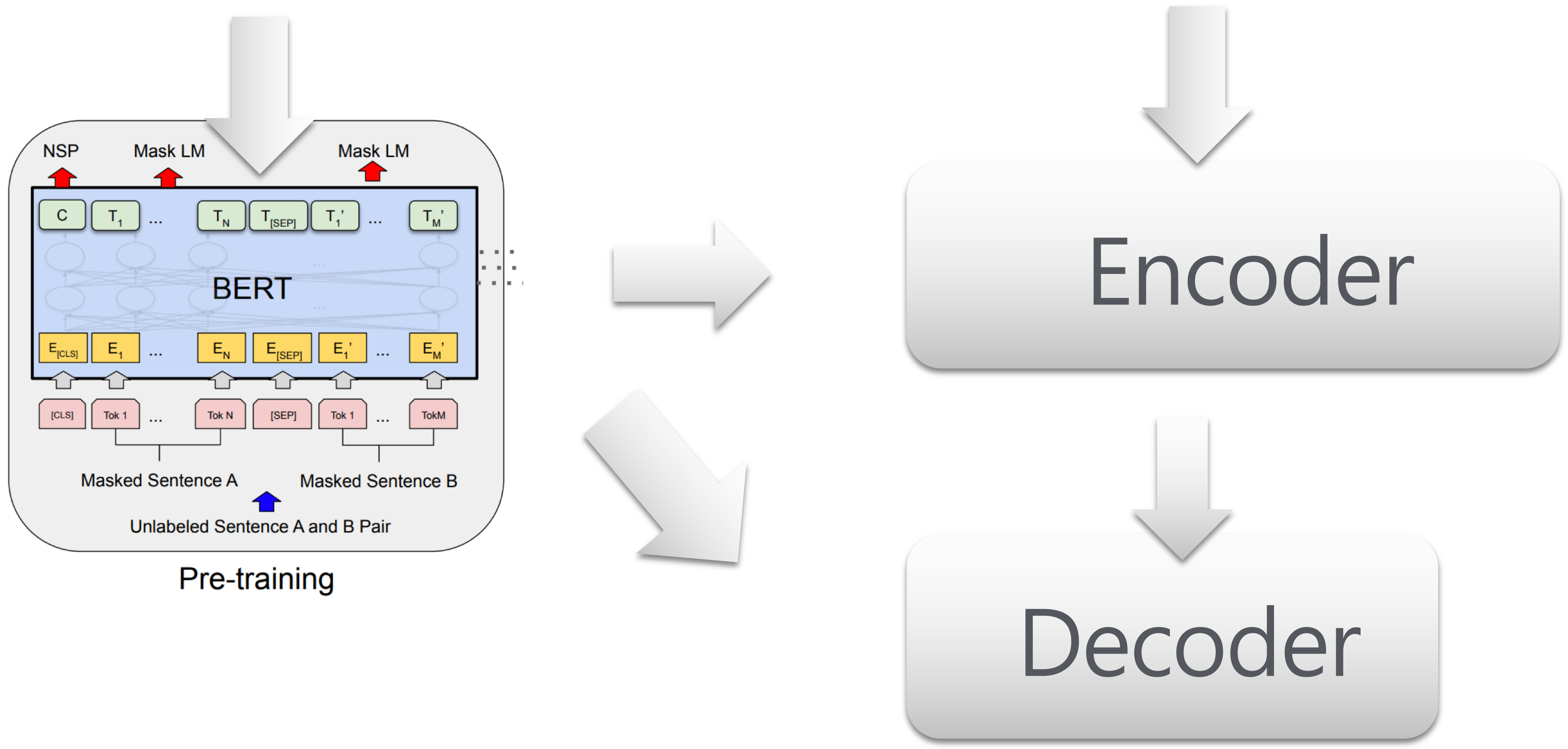
3.Multi Encoder Approach

DEVIEW 2019



3.Multi Encoder Approach

DEVIEW 2019



- Different ways to Transfer BERT encoder to NMT Encoder …
- *Towards Making the Most of BERT in Neural Machine Translation, Yang et al*
- *Incorporating BERT into NMT* https://openreview.net/pdf?id=Hyl7ygStwB
- Promising Results … but not included here

Experiments

Related Works and Motivation

- Reusing encoder ✔ but decoder ✖
- Tasks, Datasets Models are not always comparable 🍎 ≠ 🍐
- Experimental study aiming for systematic comparisons
- Beyond BLEU benefit ? (Domain Adaptation and Robustness)

BERT+NMT architectures

- **BERT Setup**

How ?	Why ?
6 Layers BERT Encoder	to be fair with NMT encoder
Relu and Sinusoidal embedding	like original transformer
MASK=UNK token	to test robustness
MLM Task only	NSP had no impact
Frequency Sampling	As Lample et al. XLM
Iterations	300k

BERT training datasets

What is the impact of pretraining data ?

Dataset	~ Number of Sentences	Description
IWSLT	800K	all IWSLT data available in English
WMT14-En-De.Src	4M	source side of parallel corpus
Wiki	70M	English wikipedia dump
News	210M	70M from News Crawl, News Commentary and Common Crawl[1]

[1] provided by WMT 2019 http://www.statmt.org/wmt18/translation-task.html

BERT+NMT architectures

- **BERT+NMT architectures**

How ?	Why ?
Freeze: initialize NMT encoder (with BERT) and **freeze**	Is BERT encoder enough ?
FT: initialize NMT encoder and **fine-tune**	Simplest approach
Embedding: use BERT encoder output as an input to NMT encoder and finetune	Benefit from Deeper Model ?

Experiments

DEVIEW 2019

- *medium-high* resource: WMT 2014 English-German: 4M sentences
- *low resources* : IWSLT 2014 English-German: 200k sentences

WMT 14 English-German

Experimental settings

Preprocessing	no tokenization, no normalization
BPE	**no joint** BPE: en: BPE with 32K vocabulary trained on concatenation of Wiki+News (~280M sent) de: 32K BPE for German learnt on target part of WMT 14 parallel data
Transformer	- transformer-big model for BERT and NMT - shared in-out embeddings - dropout 0.3,...

→ baseline is slightly different from official baseline

WMT 14 English-German

Evaluation

MT Evaluation: hard task

BLEU : modified precision of **n-gram co-occurences**

- between *reference translation and hypothesis translation*;
- averaged over 1,2,3,4-grams

Test Sets

Different domains test sets

News: WMT-14, WMT-18

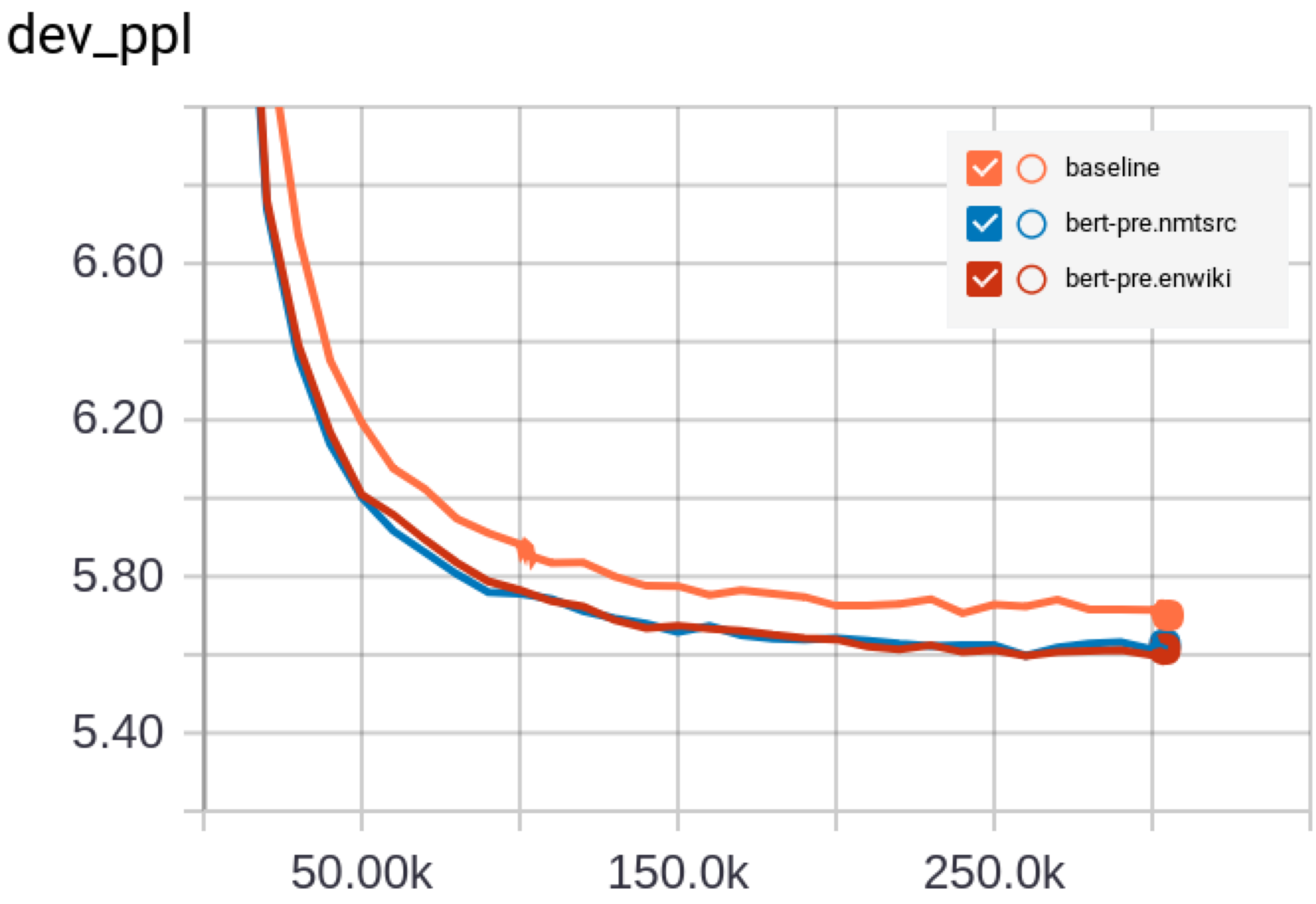
Speech: IWSLT-15, OpenSub

Technical: KDE

Wikipedia: wiki

Training Curves

Perplexity: lower is better



WMT English-German: Results

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1

WMT English-German: Results

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1
News.Freeze	23.6	35.5	15.0	13.8	26.5	15.1

WMT English-German: Results

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1
News.FT	**27.9**	**40.2**	**18.8**	**15.7**	**29.1**	17.9
News.Emb	**27.7**	**39.9**	**18.9**	**16.0**	**29.3**	**18.2**

- FT ≈ Emb
- Best Improvement on News and Wiki test set

WMT English-German: Results

DEVIEW 2019

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1
News.FT	**27.9**	**40.2**	$\boxed{\textbf{18.8}}$	**15.7**	**29.1**	17.9
Wiki.FT	**27.7**	**40.6**	$\boxed{\textbf{18.4}}$	**15.4**	28.7	**19.0**

- no domain adaptation effect observed

WMT English-German: Results

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1
News.FT	**27.9**	**40.2**	**18.8**	**15.7**	**29.1**	17.9
Wiki.FT	**27.7**	**40.6**	**18.4**	**15.4**	28.7	**19.0**

- no domain adaptation effect observed
- wiki.FT seems to be weaker on Speech domain
- News.FT slightly better : is it due to bigger data?

WMT English-German: Results

	wmt14	wmt18	wiki	OpenSub	iwslt15	kde
Baseline	27.3	39.5	17.6	15.3	28.9	18.1
News.FT	**27.9**	**40.2**	**18.8**	**15.7**	**29.1**	17.9
Wiki.FT	**27.7**	**40.6**	**18.4**	**15.4**	28.7	**19.0**
WMT.En-de.Src.FT	**27.7**	**40.1**	**18.3**	15.3	28.7	**18.4**

- WMT.EN-De.Src.FT: same data, better performance !
- Better initialization helps training (better source encoding)

Lessons learnt up to now

DEVIEW 2019

- BERT provides good initialization point: even with same data we achieve better performance
- More data >= In-domain data
- What about robustness to noise ?

How to measure robustness ?

Evaluate all the models on synthetic noise test sets:

raw sentence	John loves Mary
UNK.S	<UNK> John loves Mary
UNK.E	John loves Mary <UNK>
chswap	John lvoes Mary
chrand	Johnw loves Mary Jhn loves Mary
uppercase	John LOVES Mary

WMT English-German: robustness

	wmt14	+unk.s	+unk.e	+chswap	+chrand	+up
Baseline	27.3	24.8	24.4	24.2	24.7	23.5
WMT.En-de.Src.FT	27.7	24.9	22.9	24.4	25.2	24.5
Wiki.FT	27.7	25.8	24.9	24.4	24.9	24.4
News.FT	27.9	24.9	24.9	24.5	25.3	24.5
News.Emb	27.7	24.7	24.8	24.6	25.3	24.2

NMT+BERT models mostly have higher BLEU scores

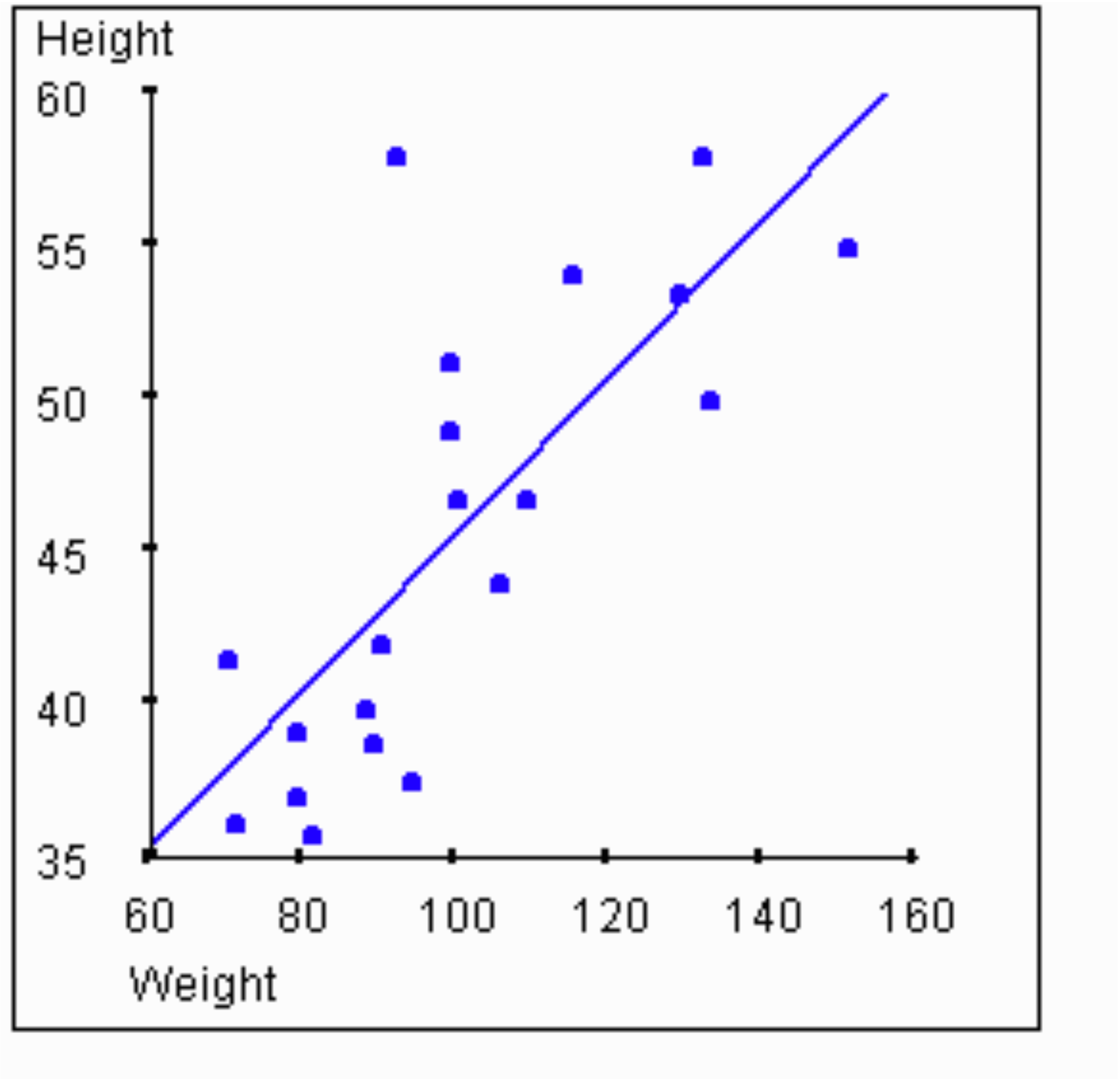
But that was already the case for clean test

Assessing Robustness

Problem: Test Set and Noisy Test are correlated. So are their BLEU scores …

Why a model is better on noisy test sets ?

- Is the model better in general ?
- Is the model more robust ?
- Decrease in number of 'hallucinations' ?



WMT English-German: ΔBLEU

	wmt14	+unk.s	+unk.e	+chswap	+chrand	+up
Baseline	27.3	24.8 / -2.5	24.4 / -2.9	24.2 / -3.1	24.7 / - 2.5	23.5 / -3.8
WMT.En-de.Src.FT	27.7	24.9 / -2.6	22.9 / -4.8	24.4 / -3.3	25.2 / -2.5	24.5 / -3.2
Wiki.FT	27.7	25.8 / -1.9	24.9 / -2.8	24.4 / -3.3	24.9 / -2.8	24.4 / -3.3
News.FT	27.9	24.9 / - 3.0	24.9 / -3.0	24.5 / -3.4	25.3 / -2.6	24.5 / -3.4
News.Emb	27.7	24.7 / -3.0	24.8 / - 2.9	24.6 / - 3.1	25.3 / -2.6	24.2 / -3.5

- BLEU is higher, but BLEU delta is lower (or the same) for most of the models !

- Problem with ΔBLEU: is it really interpretable?

Δchrf : robustness

chrF : characters n-gram based F-measure
(Showed good correlation at sentence-level)[1]

$$\Delta(chrF) = chrF(M(src_{noisy})) - chrF(M(src_{clean}))$$

Distribution of $\Delta(chrF)$ for each model:

- *more sentence* with negative $\Delta(chrF)$ indicate *less robust model*

[1] Qingsong Ma, Ondej Bojar, and Yvette Graham. Results of the wmt18 metrics shared task: Both characters and embeddings achieve good performance. WMT 2018

What we would expect …
Δchrf Distribution

Less sentences have big decrease in chrf.

(Note some sentences could be improved when adding noise, possibly correcting undertranslations)

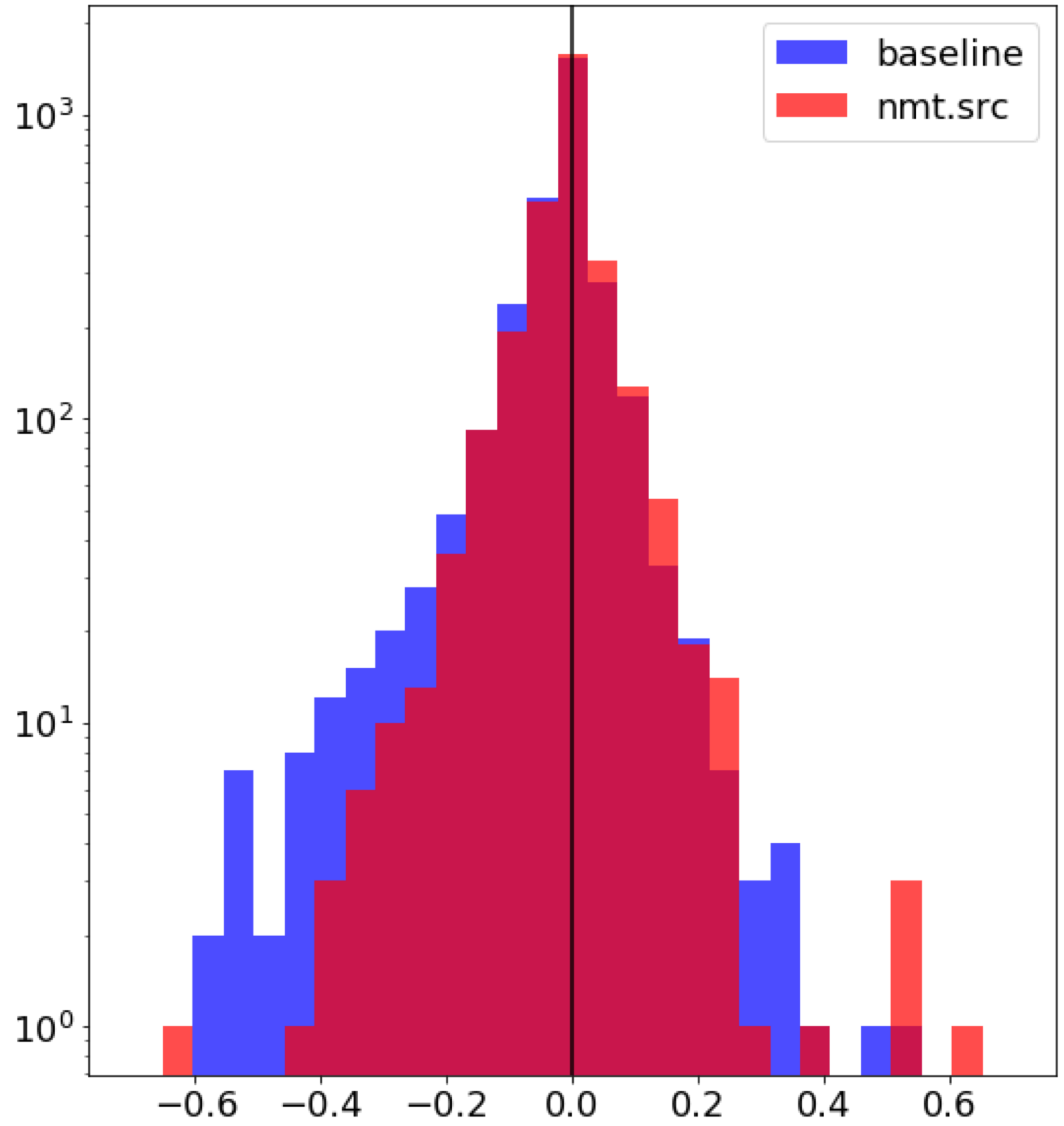
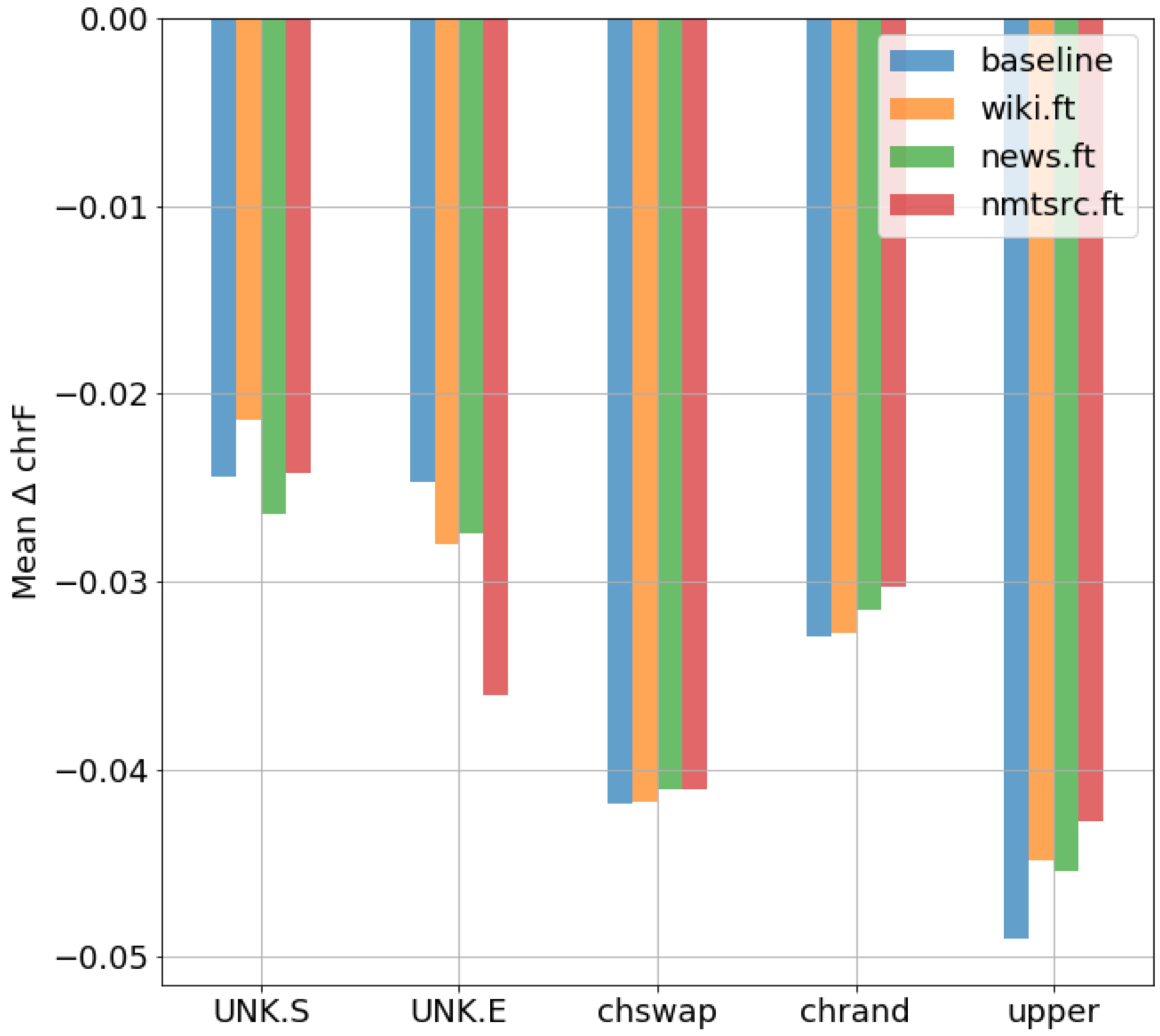


figure obtained with models trained until 100k iterations

Mean Δchrf



higher mean Δchrf → better

- UNK.S, UNK.E : BERT+NMT model are not really more robust
- Upper: NMT+BERT is more robust
- Chswap, chrand: BERT + NMT slightly more robust

Experiments

- *medium-high* resource settings: WMT 2014 English-German
- *low resources* settings: IWSLT 2014 English-German, English-Russian

IWSLT 15 English-German, English-Russian

Motivation:

- Check that BERT pretrained model can be reused in different domain, different language pairs
- low resource settings

IWSLT 15 English-German, English-Russian

Experimental settings

- Monolingual: ~800K English sentences;
- Bilingual : ~200K sentences

- Baseline: BPE 10K vocabulary, transformer base model
- IWSLT BERT: transformer based with 10K vocabulary

- News, Wiki BERT: same as previously (32k BPE, transformer big)
- BERT+NMT: Initilized encoder with BERT and finetune it with NMT

IWSLT 15 English-German, English-Russian

	en-de	en-ru
	Baseline	
tbase.bpe10k	25.9	9.6
tbase.dec3.bpe10k	26.4	16.3

	BERT+NMT	
IWSLT.FT.*tbase.bpe10k*	27.4	17.6
IWSLT.FT.*tbase.dec3.bpe10k*	27.2	**18.1**
Wiki.FT.*tbig.bpe32k*	26.9	17.6
Wiki.FT.*tbig.dec3.bpe32k*	**27.7**	17.8
News.FT.*tbig.bpe32k*	27.1	17.9
News.FT.*tbig.dec3.bpe32k*	27.6	17.9

- Baseline: smaller model → better performance (due to data size)
- ***All*** BERT + NMT better
- ***Domain*** of pretrained BERT does not matter
- ***Convergence***: we can train big model if we have good initialization point (≠ Divergence)

Similar (or better) improvements for other language pairs

Conclusion &
Lessons Learned

Motivation

How

can **BERT** improve Machine Translation Models ?

Why

Lessons Learned

*How can **BERT** improve Machine Translation Models ?*

- Finetuning simple and convenient
- Train Deeper NMT Models (cf PreNorm and PostNorm in Transformers, ACL'19)
- Multi-Encoder Approach will be the best

- MLM on NMT source bring improvement for various language pairs
- Not enough GPU? MLM on your task and dataset may already bring improvement

Lessons Learned

*Why can **BERT** improve Machine Translation Models ?*

- BERT provides a better initialization point for NMT encoder :
  More data, better text ‘understanding’

- Role of NMT Encoder (→ multi encoder)

- But BERT pretraining is not enough to correct robustness issue and exposure bias.

Interested by NMT ?

bit.ly/papago-mt-recruit-201908
europe.naverlabs.com/careers/

Q & A

Thank You