NATURAL LANGUAGE PROCESSING

LECTURE 13: Neural Machine Translation

goorm

KAIST AI
Graduate School of AI

DAVIAN

Index

Machine Translation

- Machine translation is the task of translating a sentence x from source language to a sentence y in target language
- Most biggest Research Area
  - ACL 2021 Accepted Papers
  - EMNLP 2021 Accepted Papers

Source Sentence (x)
나는 학교에 간다

Target Sentence (y)
I go to school

Machine Translation

- Statistical Machine Translation
  - Learn a probabilistic model from data
  - If we want to find best English sentence y, given Korean sentence x

$$argmax_y P(y|x)$$

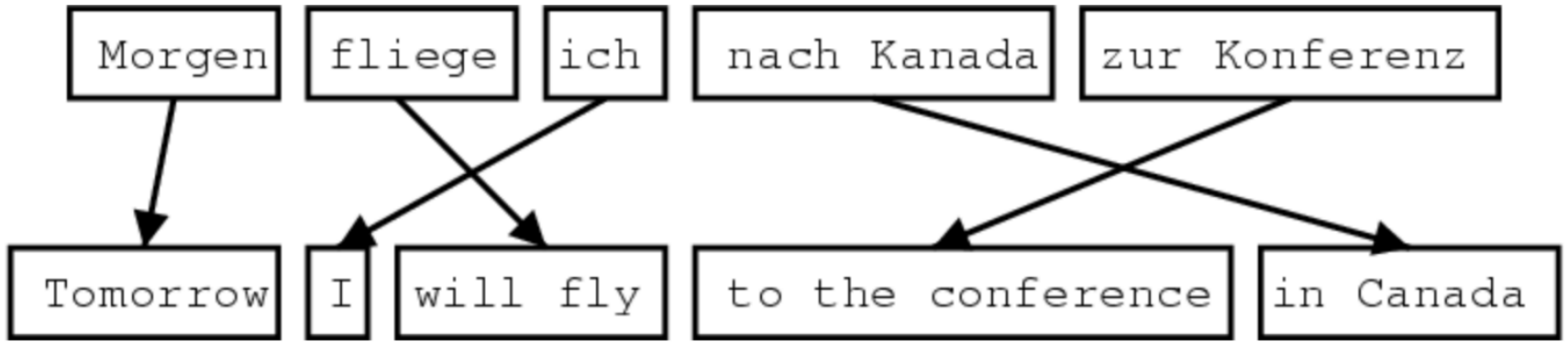
  - Let's use Bayes Rule to break the expression down.

$$argmax_y P(x|y)P(y)$$

  - $P(x|y)$: Translation Model
  - $P(y)$: Language Model

Machine Translation

- Statistical Machine Translation
  - How to learn a translation model $P(x|y)$ from the parallel corpus?
  - Latent a variable intro the model: $P(x, a|y)$, where a is the alignment
  - Alignment: word-level correspondence between source sentence x and target sentence y
  - But, Some words have no counterpart



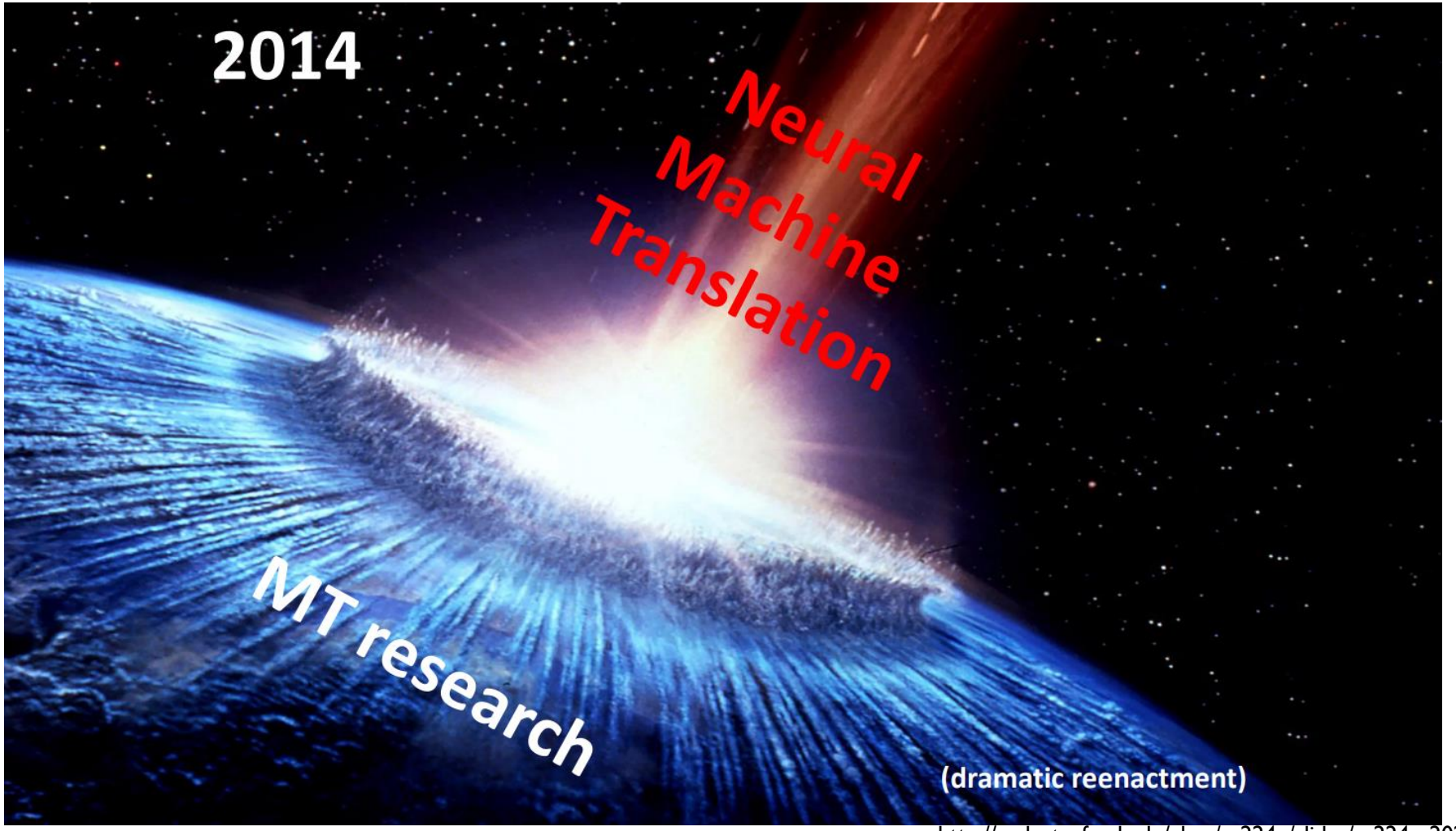
source: http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture07-nmt.pdf

Machine Translation

- Statistical Machine Translation
  - Extremely complex
  - Systems had many subcomponents
  - Feature engineering
  - Human effort to maintain

Machine Translation

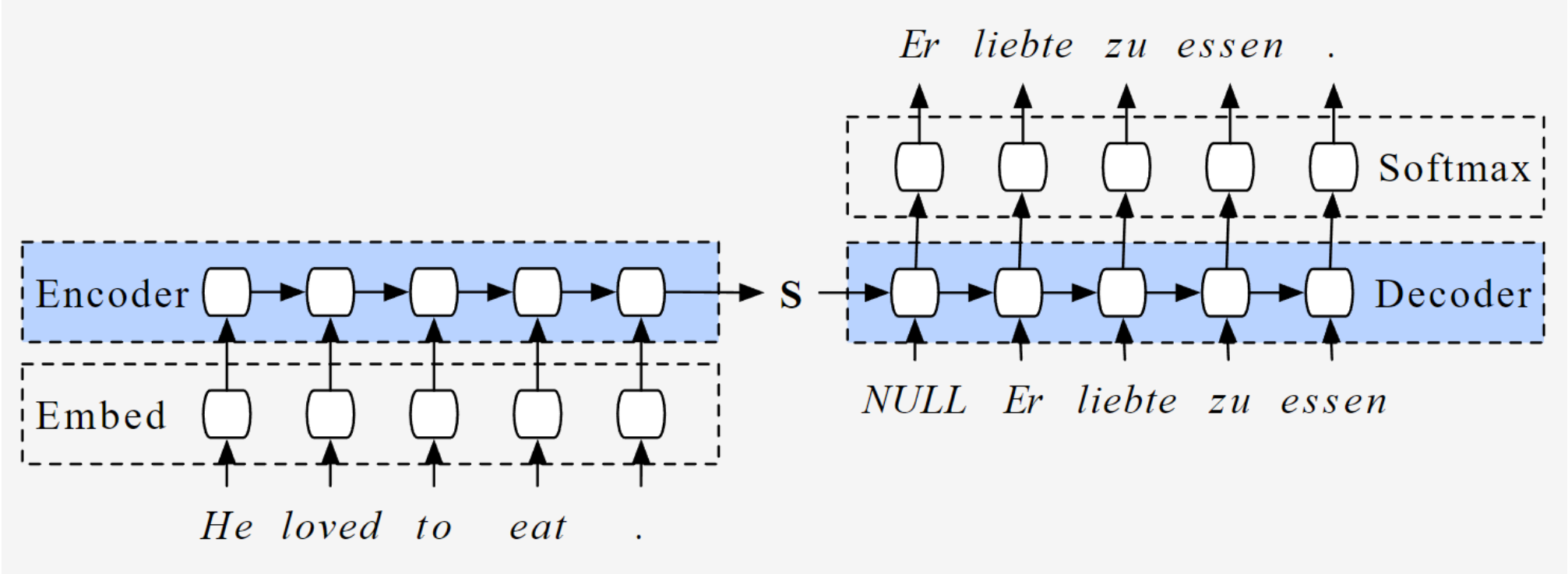
- Neural Machine Translation



source: http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture07-nmt.pdf

Machine Translation

- Neural Machine Translation
  - 2014 Neurips: Sequence-to-sequence Learning with Neural Networks
  - 2015 ICLR: Neural Machine Translation by Jointly Learning to Align and Translate



source: https://smerity.com/articles/2016/google_nmt_arch.html

Machine Translation

- How to evaluate Machine Translation?
  - BLEU (Bilingual Evaluation Understudy)
    - Reference: Half of my heart is in Hanana ooh na na
    - Predicted: Half as my heart is in Obama ooh na

$$precision = \frac{number\ of\ correct}{length\ of\ prediction} = \frac{7}{9} = 78\%$$

$$recall = \frac{number\ of\ correct}{length\ of\ reference} = \frac{7}{10} = 70\%$$

$$F - measure = \frac{precision \times recall}{\frac{1}{2}(precision + recall)} = \frac{0.78 \times 0.7}{0.5 \times (0.78 + 0.7)} = 73.78\%$$

Machine Translation

- How to evaluate Machine Translation?
  - BLEU (Bilingual Evaluation Understudy)
    - N-gram overlap between machine translation output and reference sentence
    - Compute precision for n-grams for one to four
    - Add brevity penalty for too short translations

$$BLEU = \min(1, \frac{length\ of\ prediction}{length\ of\ reference})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$$

source: https://www.statmt.org/book/

Machine Translation

Predicted (from model 1): Half as my heart is in Obama ooh na

Reference: Half of my heart is in Havana ooh na na

Predicted (from model 2): Havana na in heart my is Half ooh of na

Metric	Model 1	Model 2
Precision (1-gram)	$^{7}/_{9}$	$^{10}/_{10}$
Precision (2-gram)	$^{4}/_{8}$	$^{0}/_{9}$
Precision (3-gram)	$^{2}/_{7}$	$^{0}/_{8}$
Precision (4-gram)	$^{4}/_{6}$	$^{0}/_{7}$
Brevity penalty	$^{9}/_{10}$	$^{10}/_{10}$
BLEU	$0.9 \times \sqrt[4]{0.07}$	0

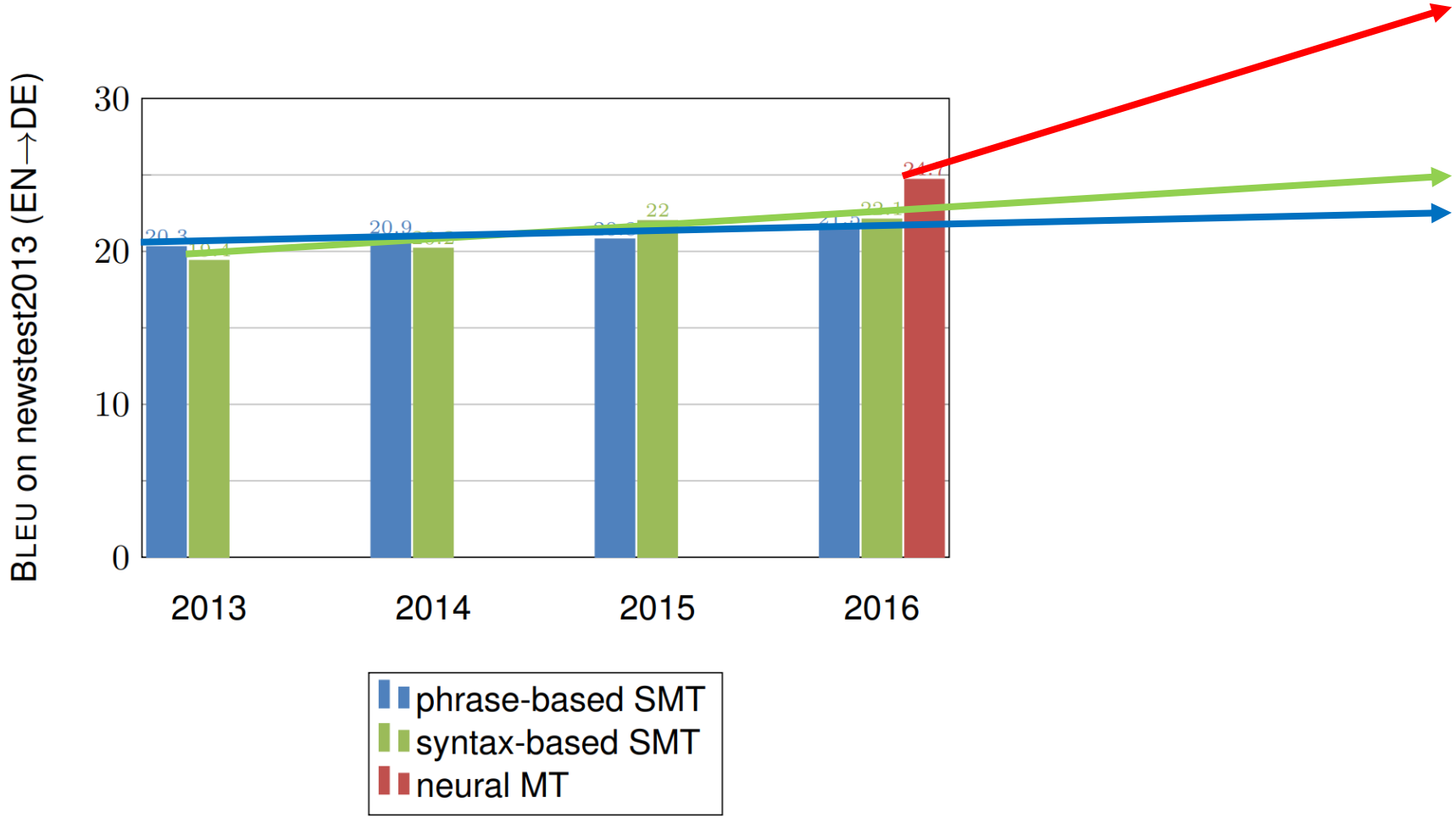
Machine Translation

- Advanced/Other Evaluation Criteria
  - SacreBLEU
  - YiSi
  - ESIM

source: https://www.statmt.org/book/

Machine Translation

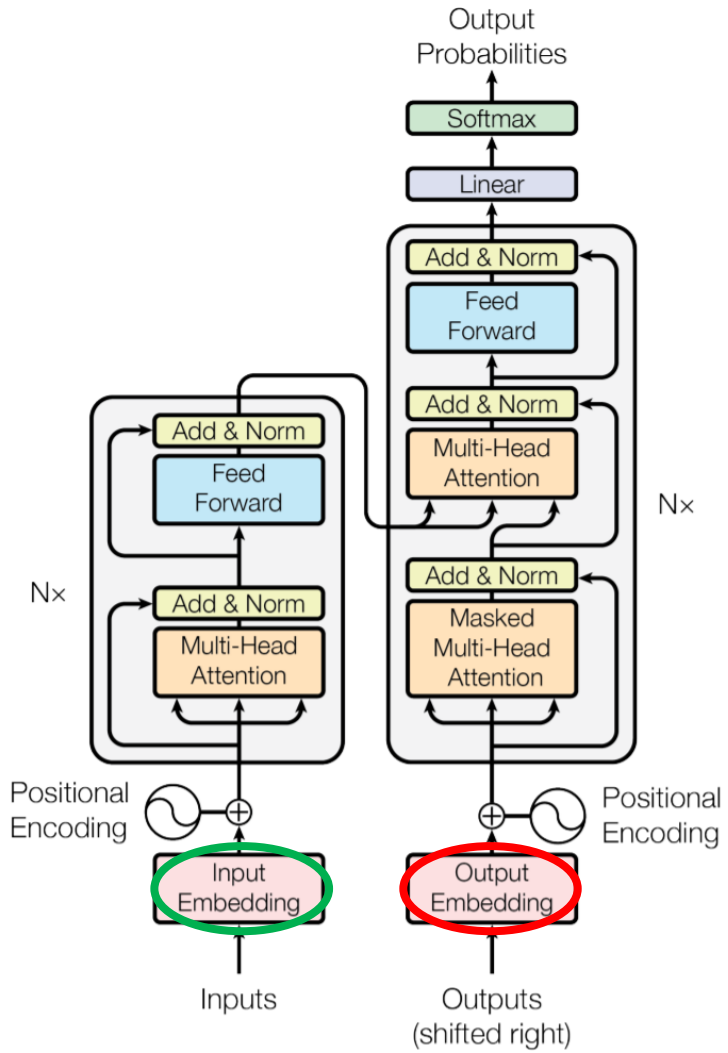
- Neural Machine Translation



source: https://smerity.com/articles/2016/google_nmt_arch.html

Supervised Neural Machine Translation

- Neural Machine Translation
  - How to train with parallel corpus?
  - Open source parallel corpus: OPUS



source: https://arxiv.org/abs/1706.03762

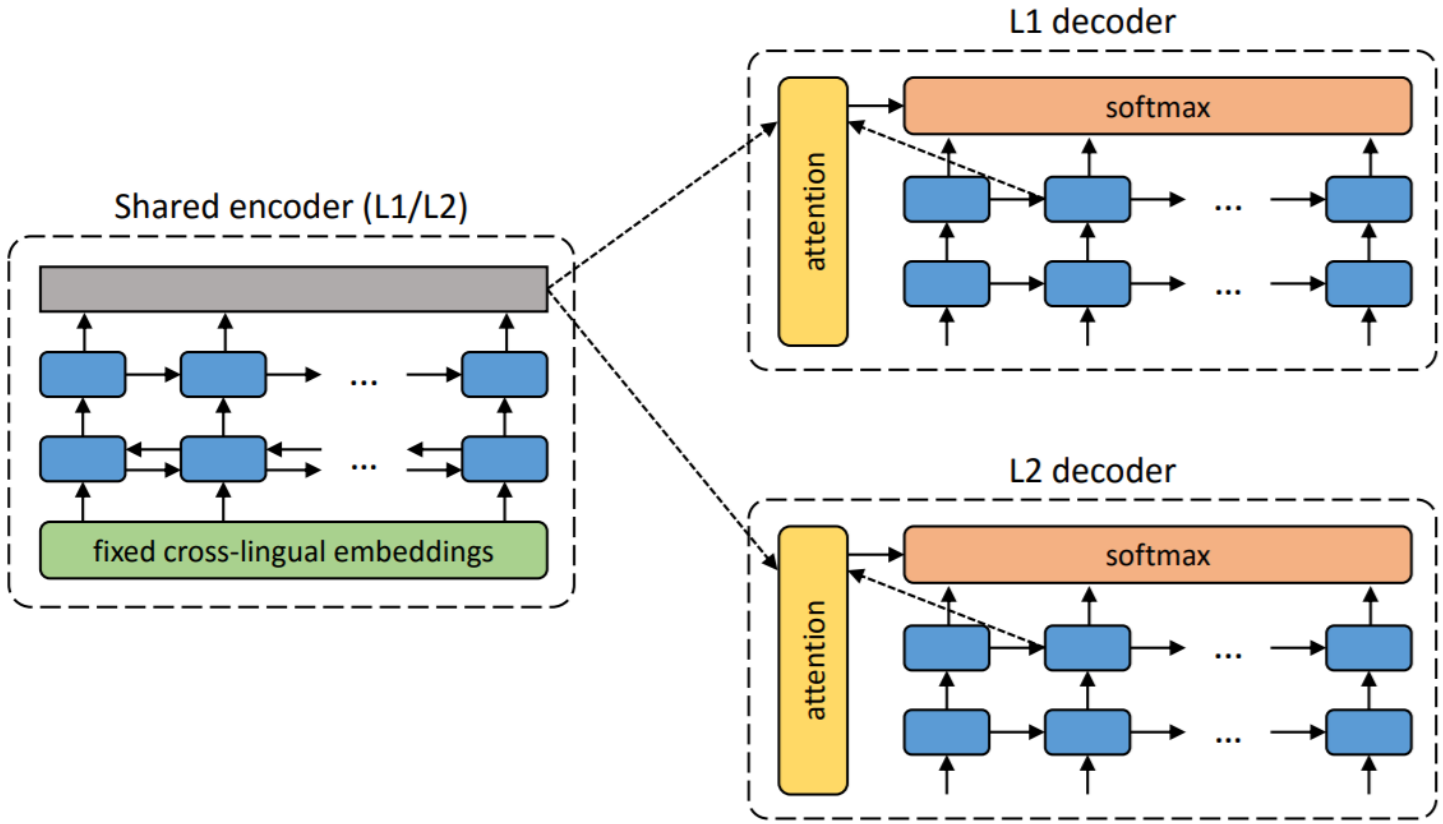
Supervised Neural Machine Translation

- Drawback
  - There are many monoliongual corpus, but few parallel corpus

source: https://arxiv.org/abs/1706.03762

Unsupervised Neural Machine Translation

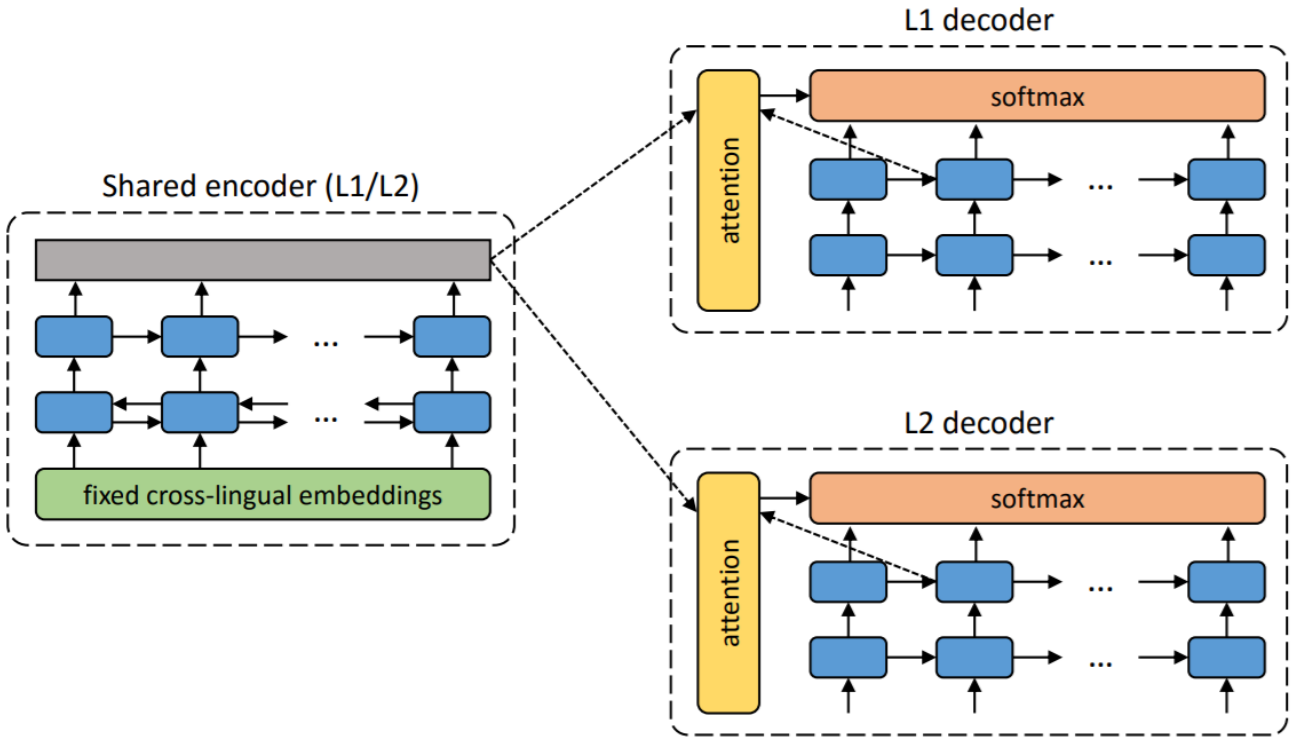
- Unsupervised neural machine translation with monolingual corpus
  - denoising
  - backtranslation



source: https://arxiv.org/abs/1706.03762

Unsupervised Neural Machine Translation

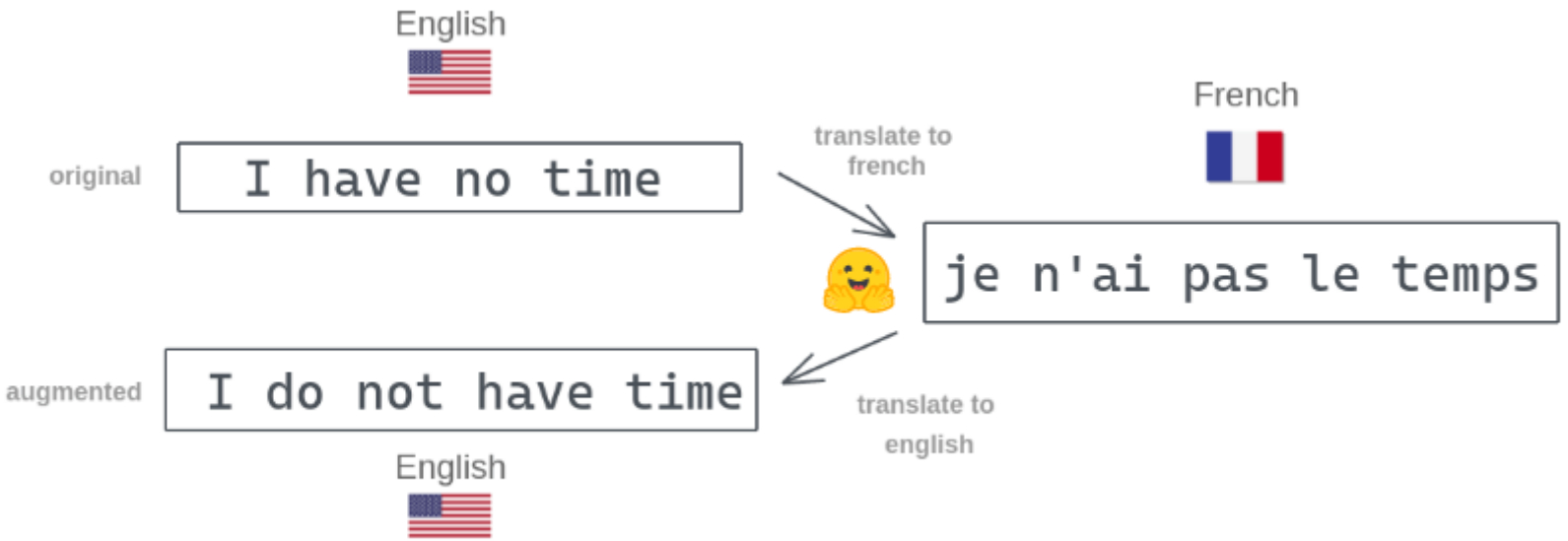
- Denoising
  - Optimizes the probability of encoding a noised version of the sentence with the shared encoder and reconstructing it with the decoder



source: https://arxiv.org/abs/1706.03762

Unsupervised Neural Machine Translation

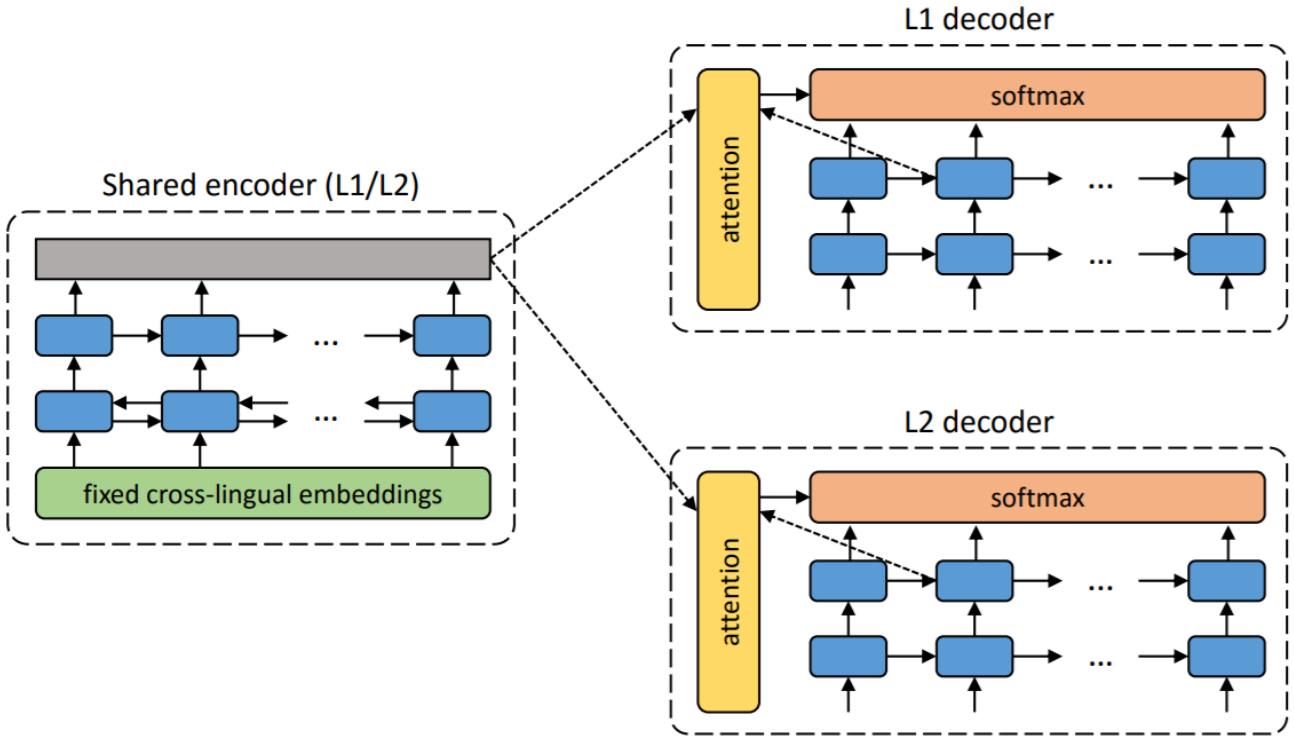
- Backtranslation
  - Retranslating sentence from the target language back to its source language



source: https://amitness.com/back-translation/

Unsupervised Neural Machine Translation

- Backtranslation
  - Translate a sentence in inference mode and then optimizes the probability of encoding this translated sentence with the shared encoder and recovering the original sentence with the decoder



source: https://arxiv.org/abs/1706.03762

Difficulties

- Six Challenges for Neural Machine Translation
  - Domain Mismatch
  - Amount of Traning Data
  - Rare Words
  - Long Sentences
  - Word Alignment
  - Decoding(Beam Search)

source: https://arxiv.org/abs/1706.03872

Difficulties

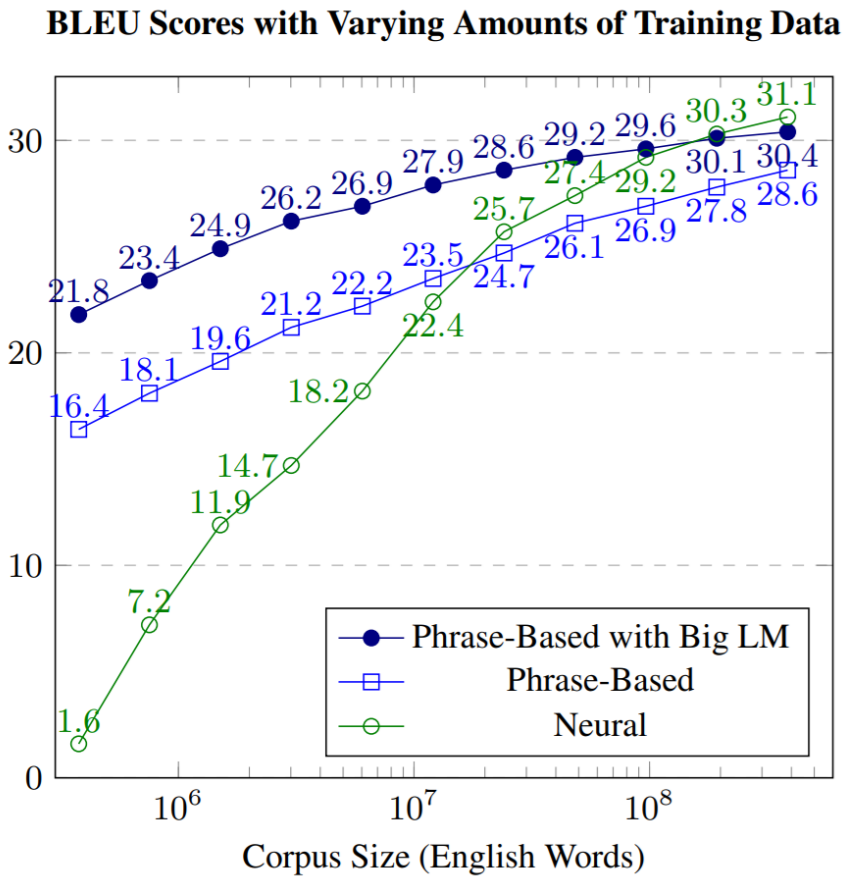
- Domain Mismatch
  - In-domain
  - Out-domain

System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 / 32.8	45.1 / 42.2	35.3 / 44.7	17.9 / 17.9	26.4 / 20.8
Law	31.1 / 34.4	12.1 / 18.2	3.5 / 6.9	1.3 / 2.2	2.8 / 6.0
Medical	3.9 / 10.2	39.4 / 43.5	2.0 / 8.5	0.6 / 2.0	1.4 / 5.8
IT	1.9 / 3.7	6.5 / 5.3	42.1 / 39.8	1.8 / 1.6	3.9 / 4.7
Koran	0.4 / 1.8	0.0 / 2.1	0.0 / 2.3	15.9 / 18.8	1.0 / 5.5
Subtitles	7.0 / 9.9	9.3 / 17.8	9.2 / 13.6	9.0 / 8.4	25.9 / 22.1

source: https://arxiv.org/abs/1706.03872

Difficulties

- Amount of Training Data
  - Outperforms SMT under high-resource conditions



source: https://arxiv.org/abs/1706.03872

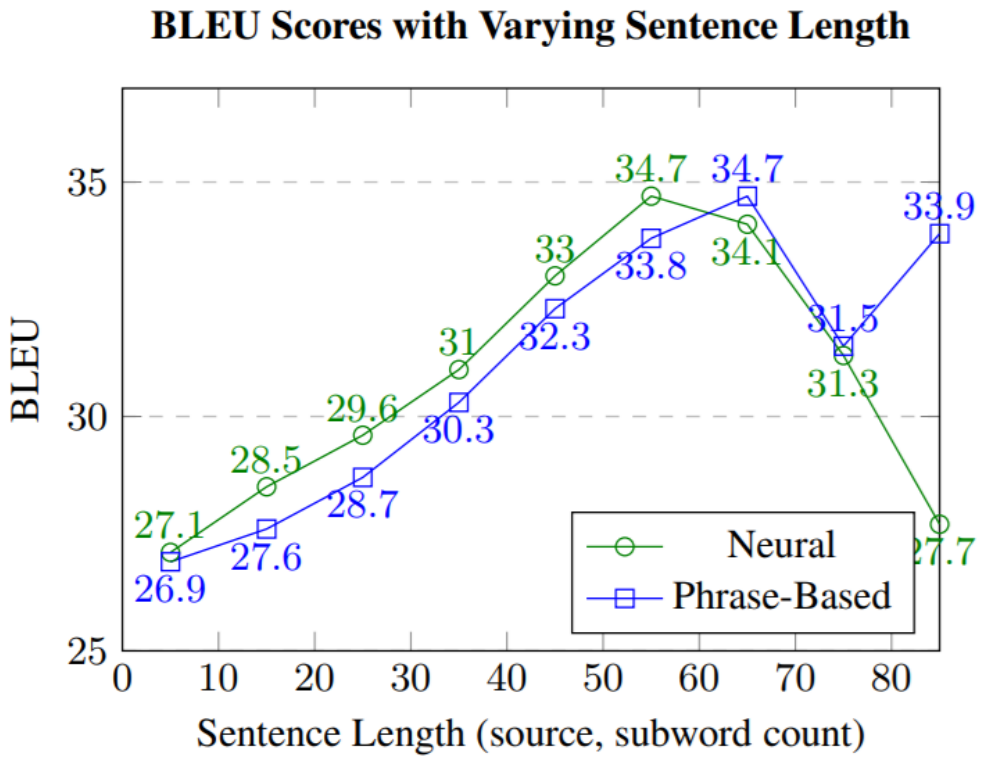
Difficulties

- Rare Words
  - NMT models perform poorly on rare words
  - Solution: Sub-word Embedding(?)

source: https://arxiv.org/abs/1706.03872

Difficulties

- Long Sentences
  - Early encoder-decoder NMT models poorly translate long sentences
  - Attention is all you need
  - Transformer-XL
  - Train short, test long



source: https://arxiv.org/abs/1706.03872

Difficulties

- Word Alignment
  - Attention scores do not correspond with out intuition
  - Attention is not Explanation
  - Attention is not not Explanation

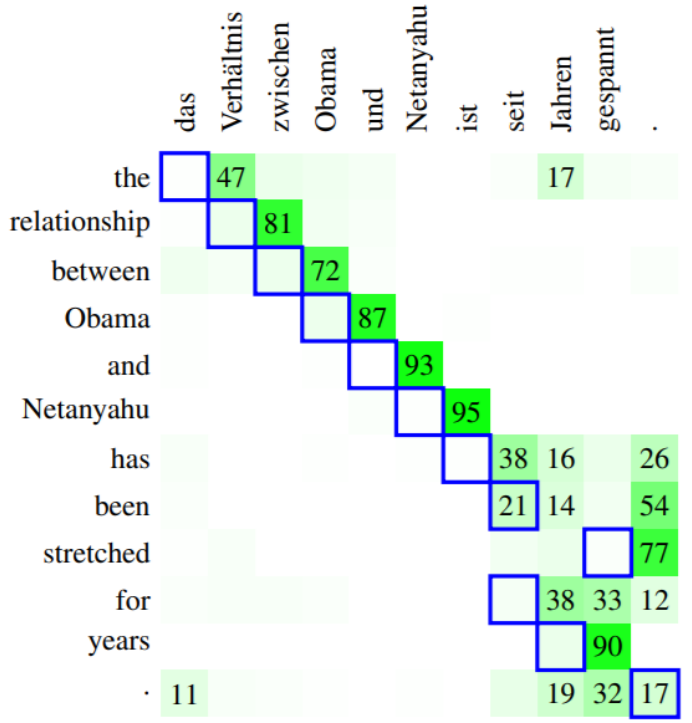
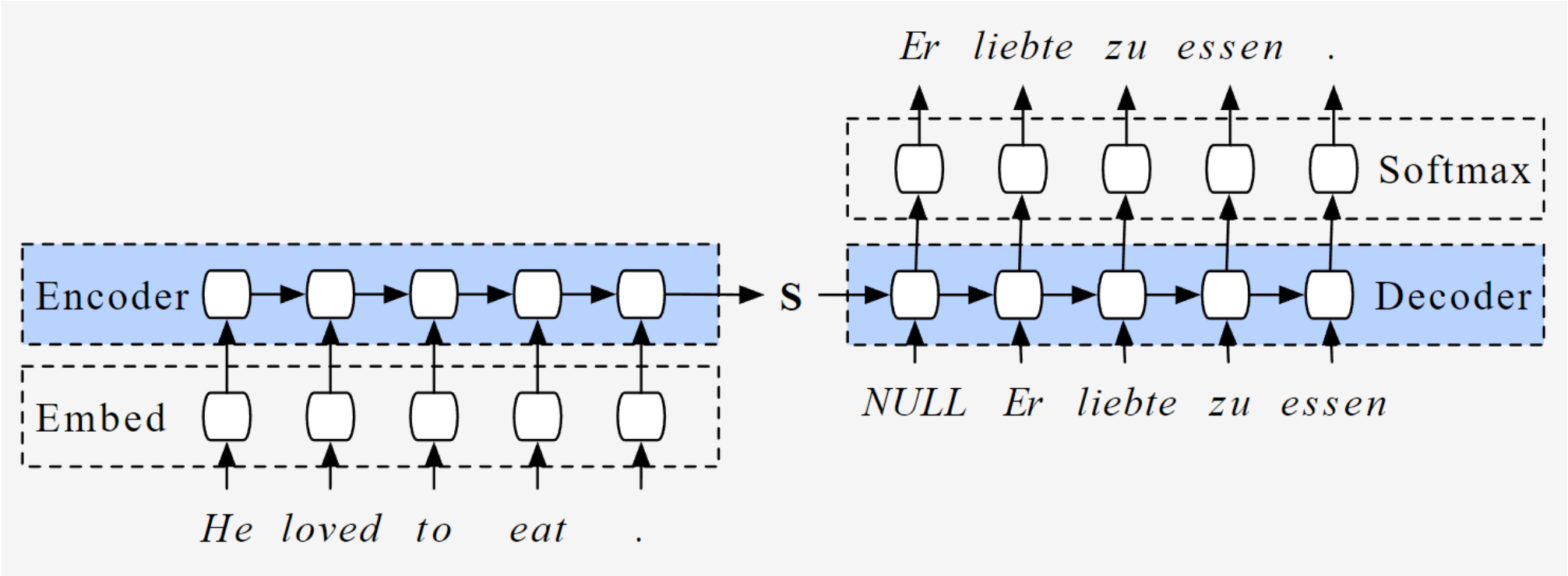


Figure 9: Mismatch between attention states and desired word alignments (German–English).

source: https://arxiv.org/abs/1706.03872

Difficulties

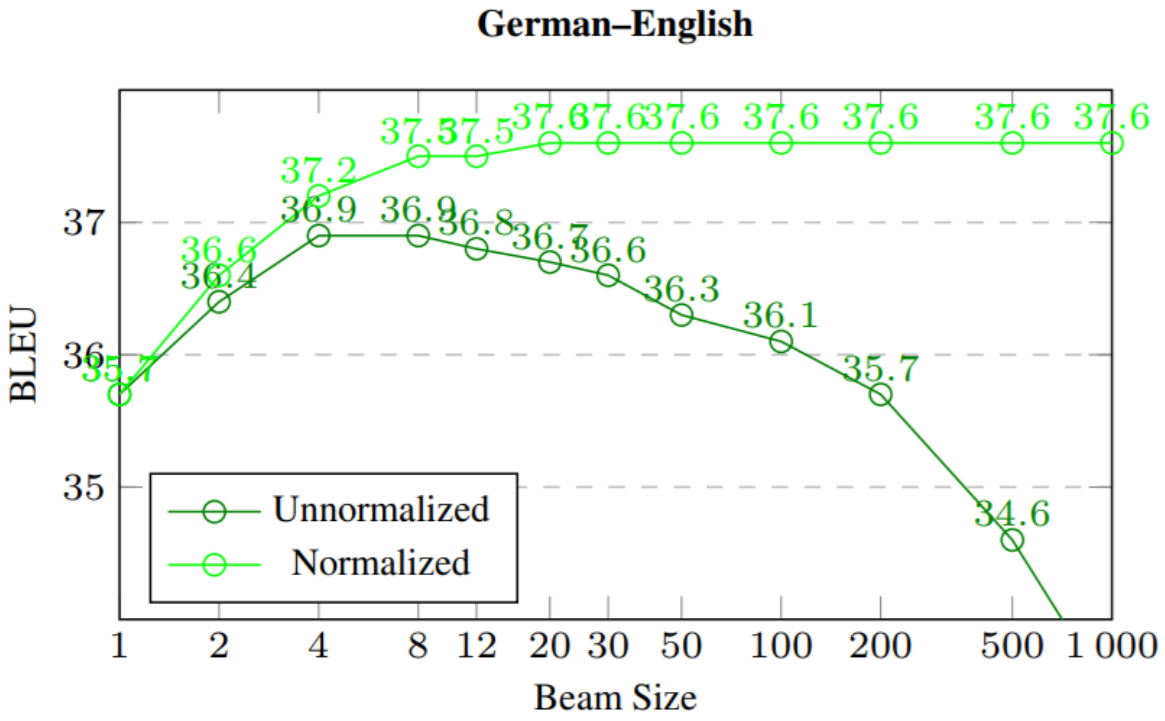
- Beam search
  - Too slow

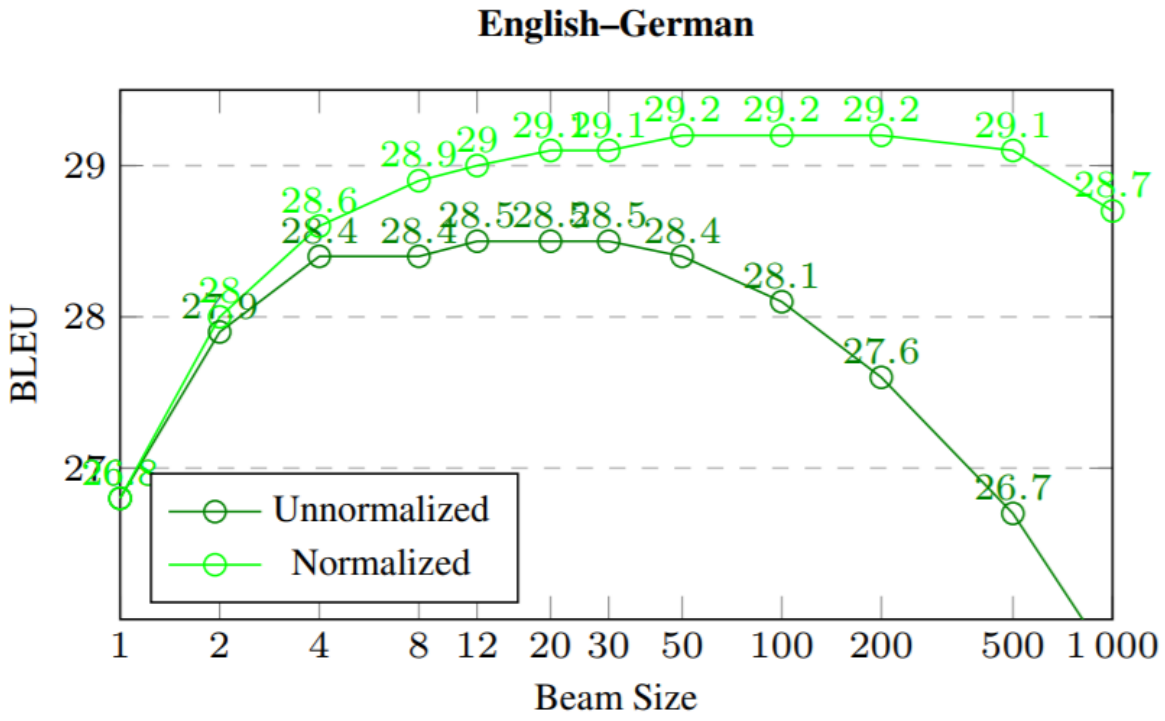


source: https://arxiv.org/abs/1706.03872

Difficulties

- Beam search
  - Increasing the beam size does not consistently improve translation quality

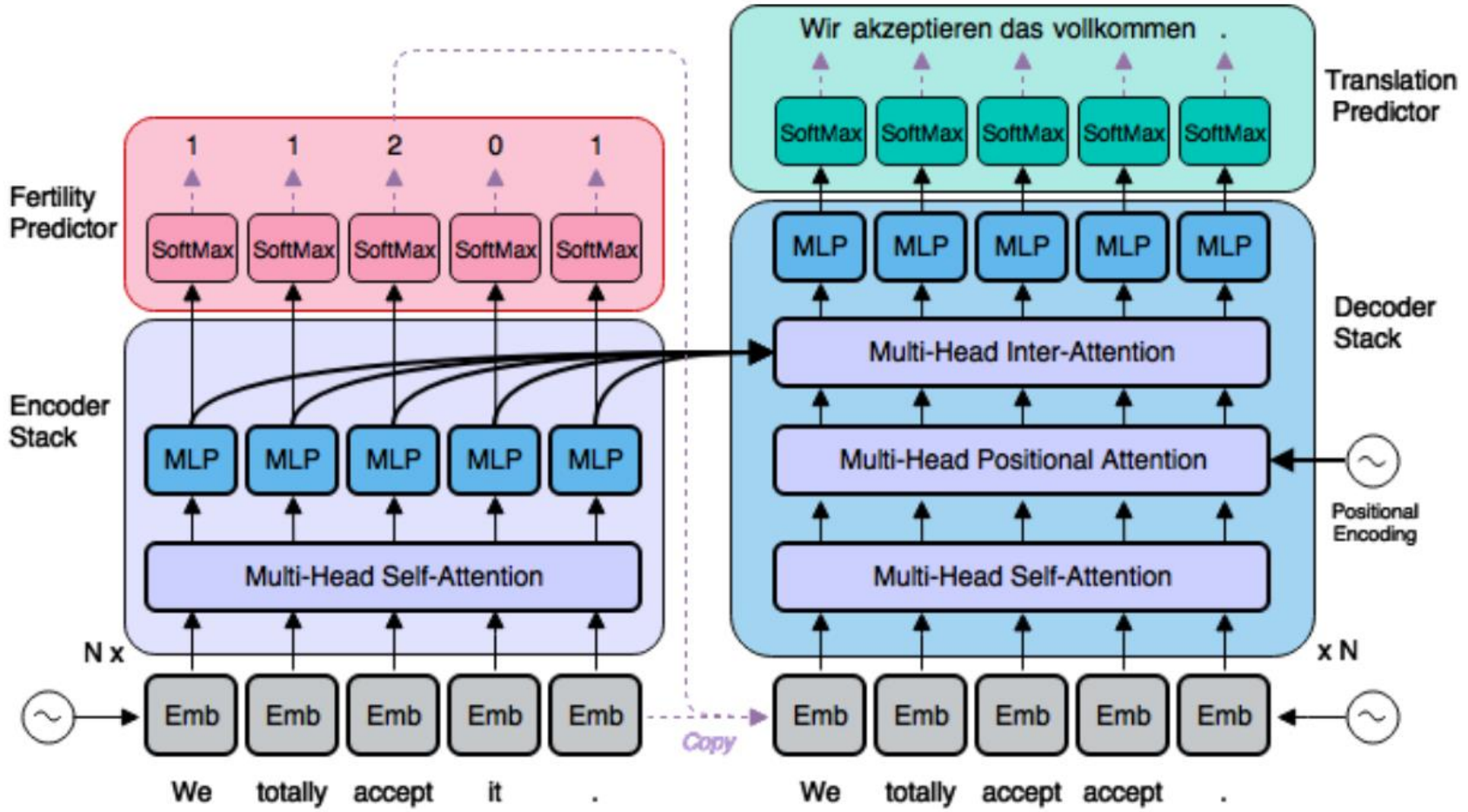




source: https://arxiv.org/abs/1706.03872

Difficulties

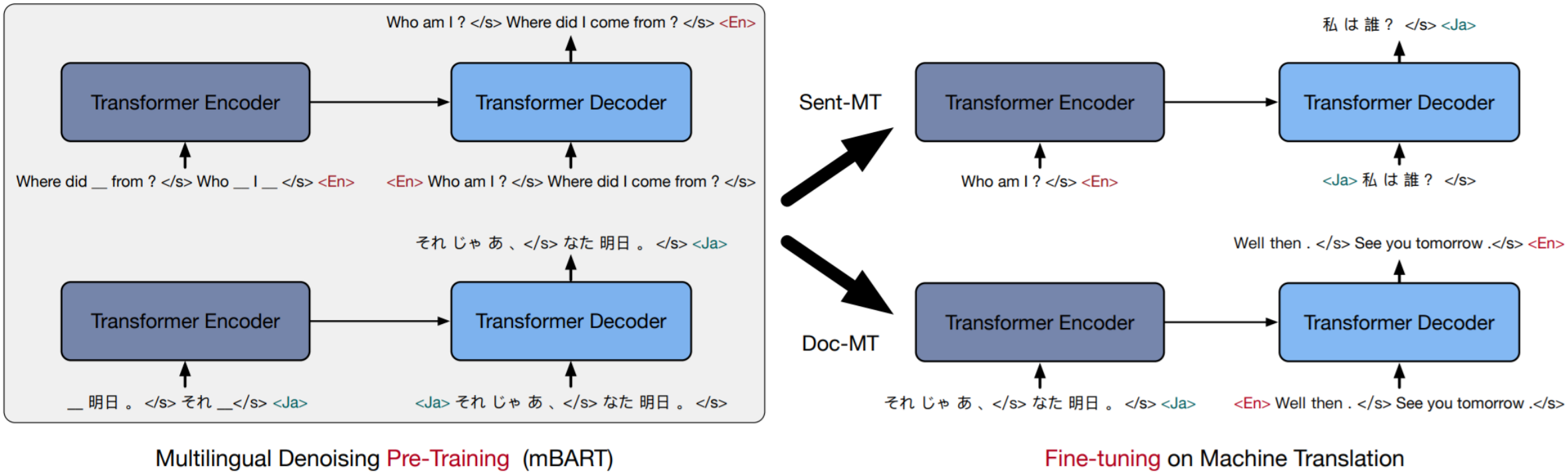
- Non-autoregressive text generation



source: https://blog.einstein.ai/fully-parallel-text-generation-for-neural-machine-translation/

Multilingual Neural Machine Translation

- [Multilingual Denoising Pre-training for Neural Machine Translation]



source: https://arxiv.org/pdf/2001.08210.pdf