# Non-autoregressive machine translation

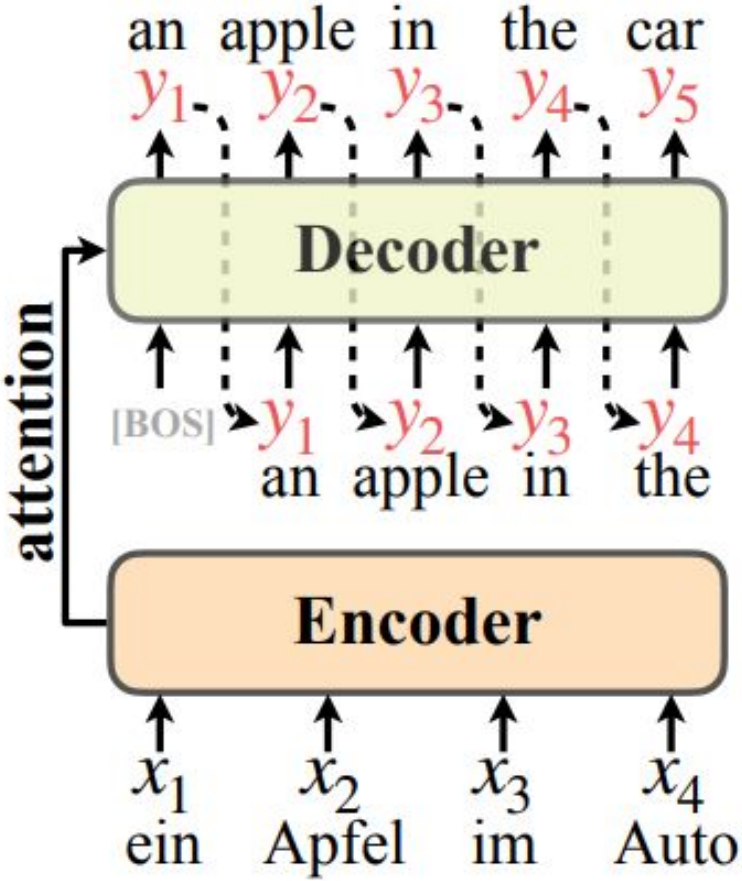
Turchyna Olga

# Seq-to-seq translation task

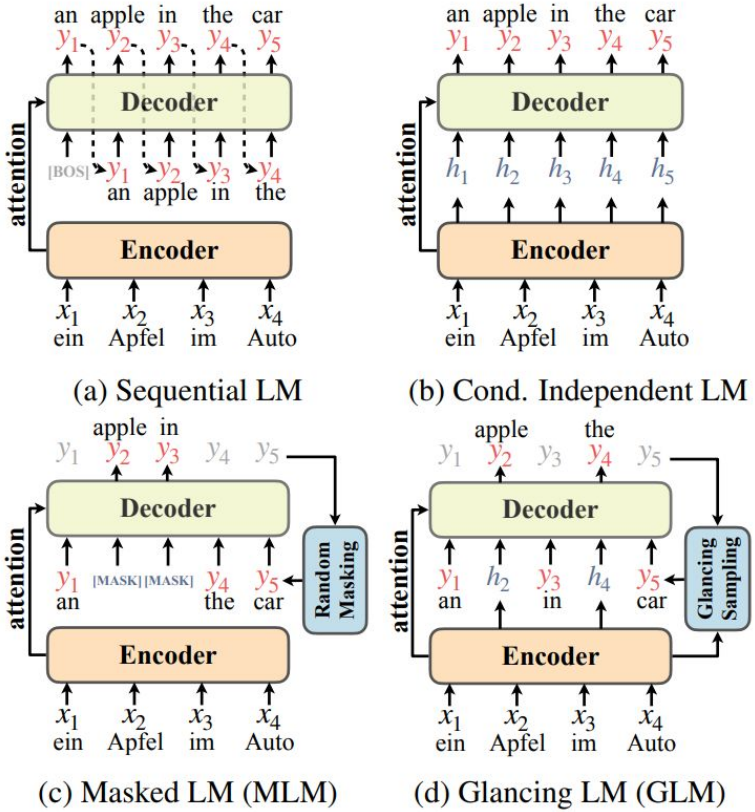
Auto Regressive translation:

- high translation quality
- capture distribution of real translations
- word-by-word nature of human language
- context-aware translations
- effective for long sentences

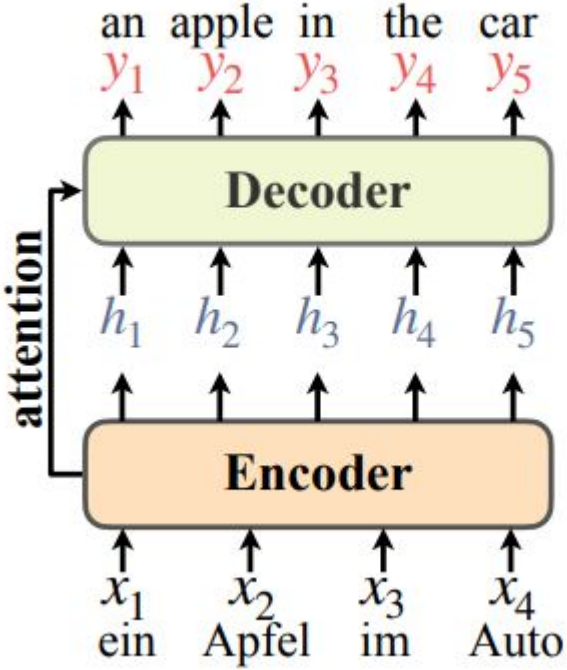
- lack of parallelism
- error propagation
- slow inference speed



# Non-autoregressive decoding

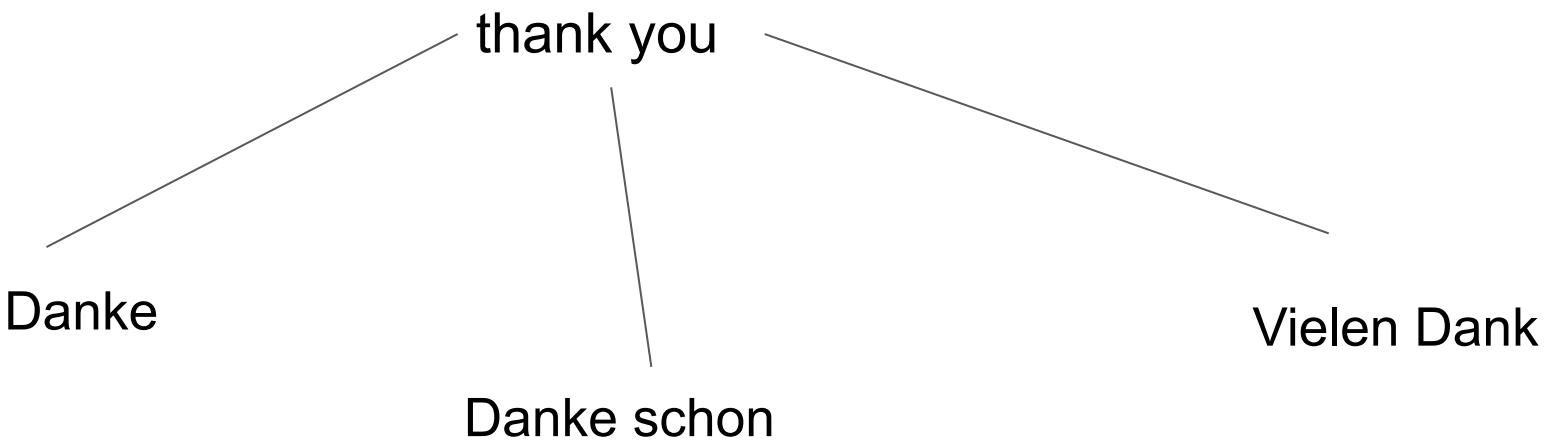


# NAT architecture (2018)




https://arxiv.org/pdf/1711.02281.pdf

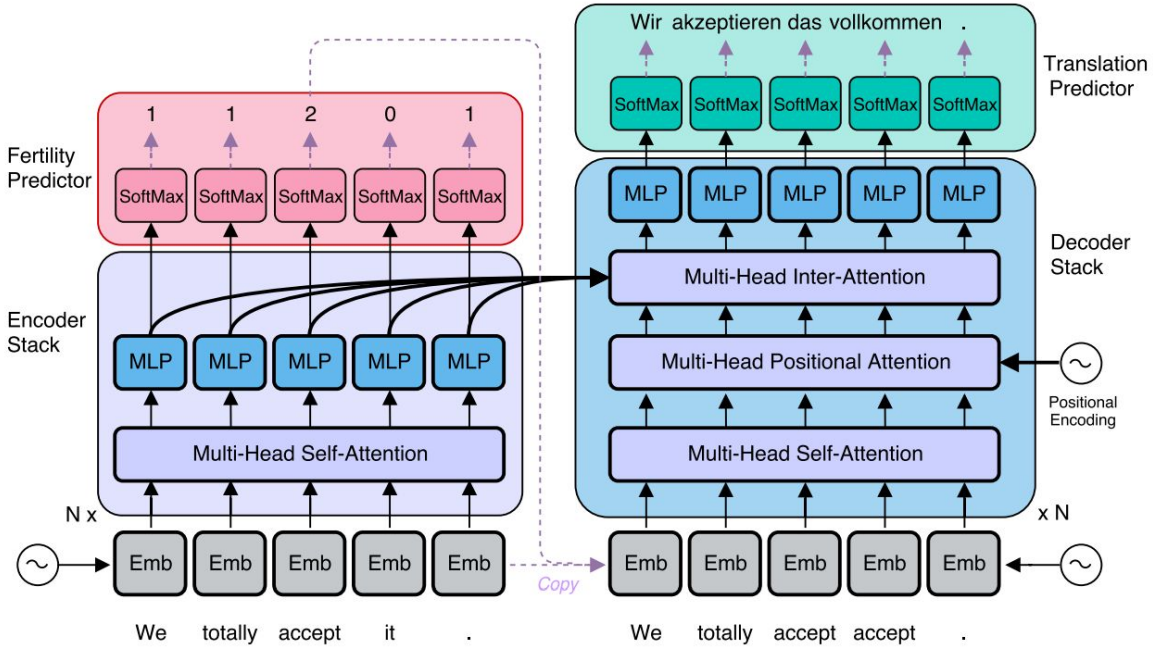
# THE MULTIMODALITY PROBLEM



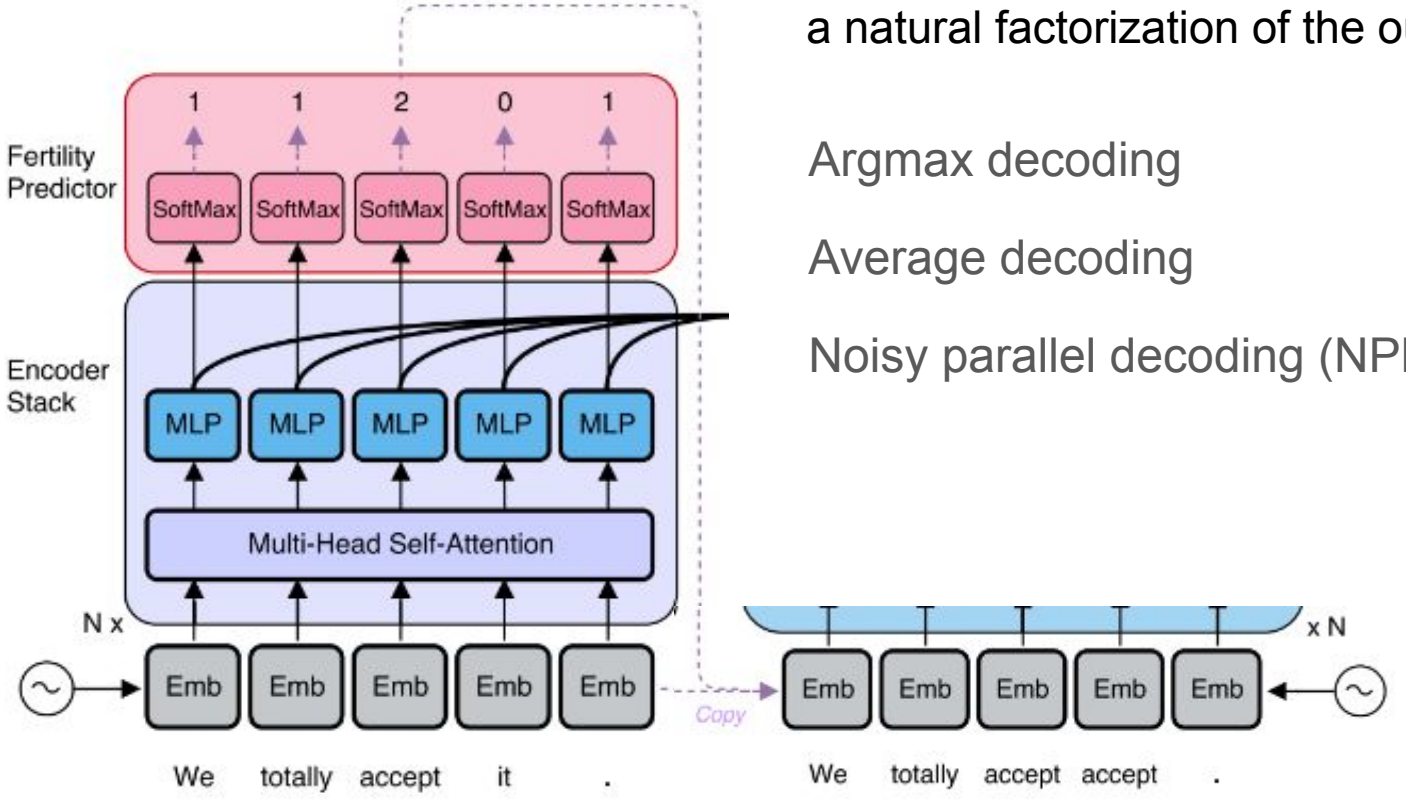

“Thank you.” can be accurately translated into German as any one of “Danke.”, “Danke schon.”, or “Vielen Dank.”. If last two options are possible and distributions are independent options “Danke Dank.” and “Vielen schon.” will also be acceptable

# Fertility prediction

Before decoding starts, the NAT needs to know how long the target sentence will be in order to generate all words in parallel



# Fertility prediction



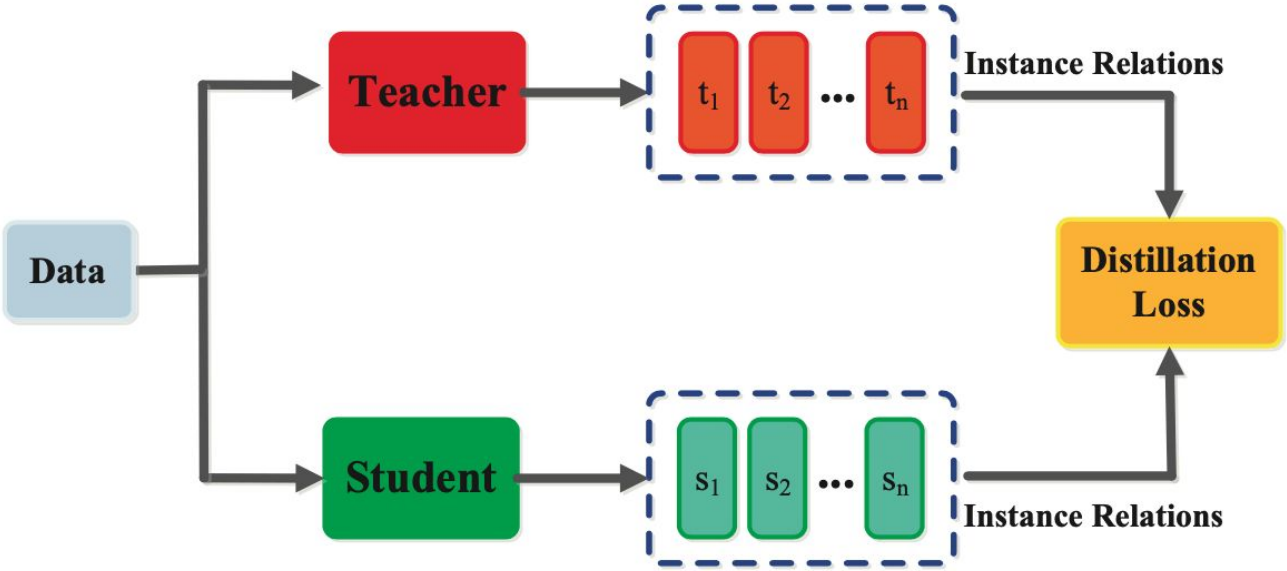

a natural factorization of the output space

Argmax decoding

Average decoding

Noisy parallel decoding (NPD)

# knowledge distillation




Sequence-level knowledge distillation is applied to alleviate multimodality in the training dataset, using autoregressive models as the teachers. The same teacher model used for distillation is also used as a scoring function for fine-tuning and noisy parallel decoding

# Noisy parallel decoding (NPD)

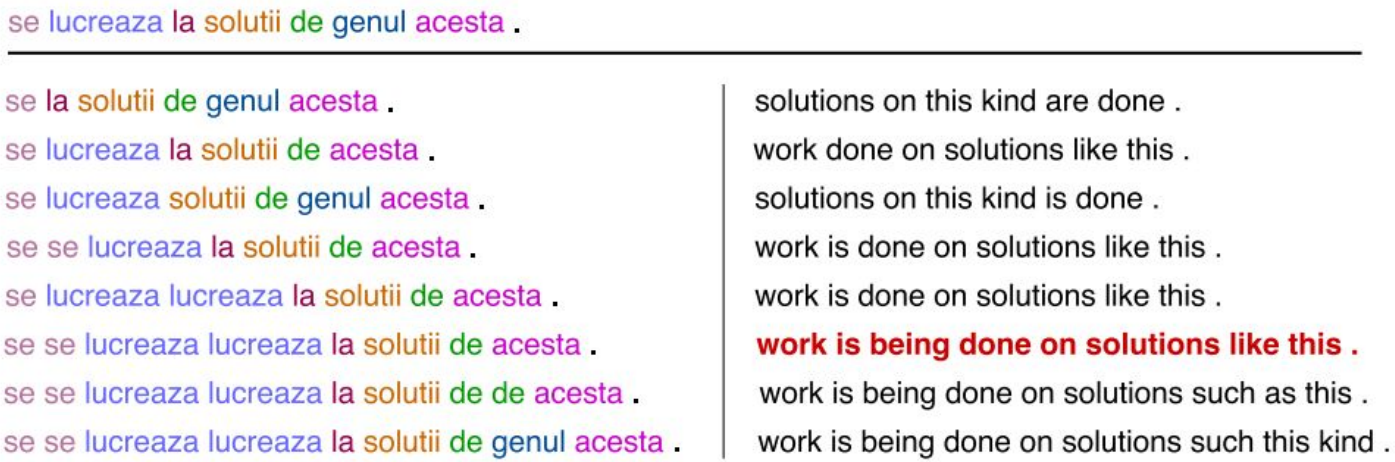
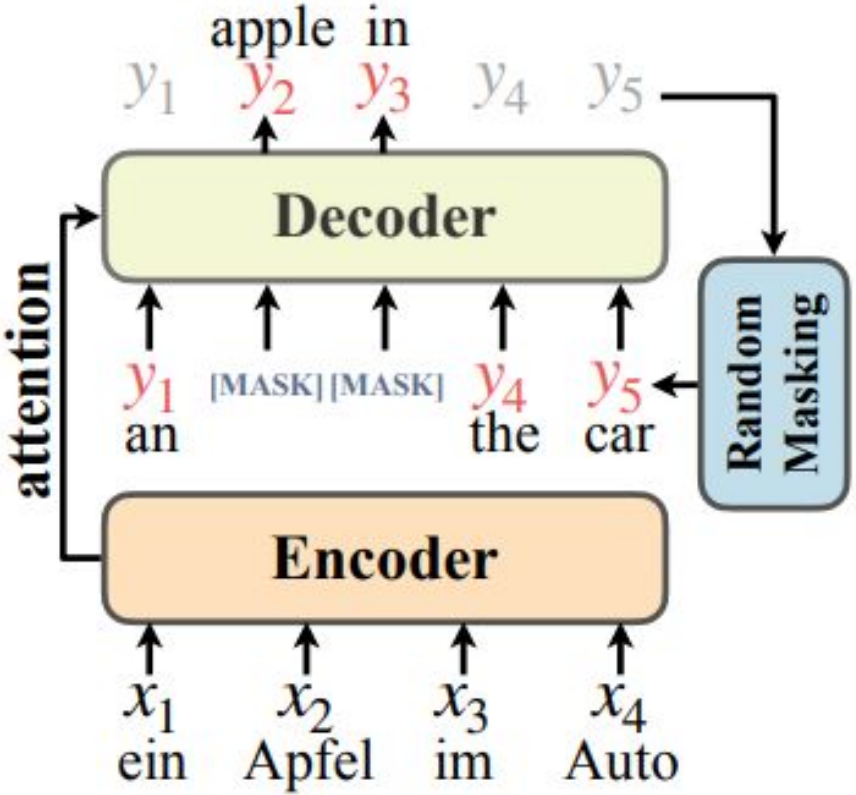



Figure 5: A Romanian–English example translated with noisy parallel decoding. At left are eight sampled fertility sequences from the encoder, represented with their corresponding decoder input sequences. Each of these values for the latent variable leads to a different possible output translation, shown at right. The autoregressive Transformer then picks the best translation, shown in red, a process which is much faster than directly using it to generate output.

# Results

| Models | WMT14 | | WMT16 | | IWSLT16 | | |
|---|---|---|---|---|---|---|---|
| | En→De | De→En | En→Ro | Ro→En | En→De | Latency | Speedup |
| NAT | 17.35 | 20.62 | 26.22 | 27.83 | 25.20 | 39 ms | 15.6× |
| NAT (+FT) | 17.69 | 21.47 | 27.29 | 29.06 | 26.52 | 39 ms | 15.6× |
| NAT (+FT + NPD $s = 10$) | 18.66 | 22.41 | 29.02 | 30.76 | 27.44 | 79 ms | 7.68× |
| NAT (+FT + NPD $s = 100$) | 19.17 | 23.20 | 29.79 | **31.44** | 28.16 | 257 ms | 2.36× |
| Autoregressive ($b = 1$) | 22.71 | 26.39 | 31.35 | 31.03 | 28.89 | 408 ms | 1.49× |
| Autoregressive ($b = 4$) | 23.45 | 27.02 | 31.91 | 31.76 | 29.70 | 607 ms | 1.00× |

# Conditional Masked Language Models




https://arxiv.org/pdf/1904.09324.pdf

# Formal Description

**Mask** For the first iteration ($t = 0$), we mask all the tokens. For later iterations, we mask the $n$ tokens with the lowest probability scores:

$$Y^{(t)}_{mask} = \arg\min_{i}(p_i, n)$$

$$Y^{(t)}_{obs} = Y \setminus Y^{(t)}_{mask}$$

| | |
|---|---|
| $src$ | Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen . |
| $t = 0$ | The departure of the French combat completed completed on 20 November . |
| $t = 1$ | The departure of French combat troops was completed on 20 November . |
| $t = 2$ | The withdrawal of French combat troops was completed on November 20th . |

# Predicting Target Sequence Length

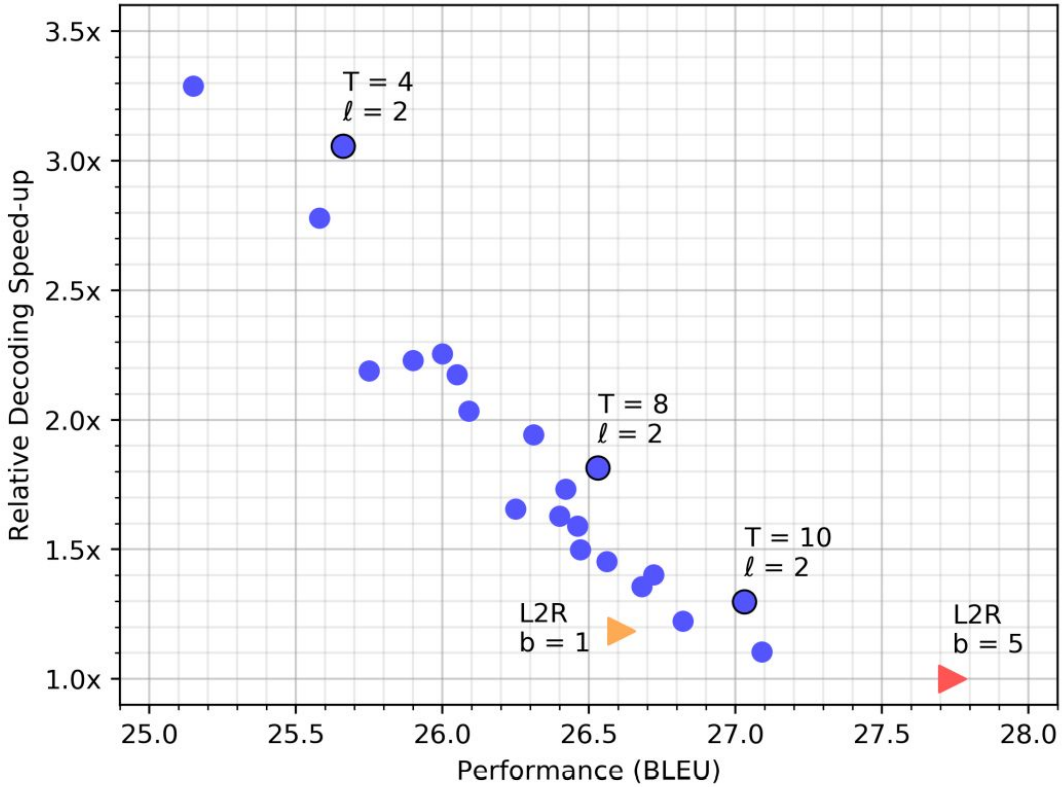
We add a special LENGTH token to the encoder, akin to the CLS token in BERT. The model is trained to predict the length of the target sequence N as the LENGTH token's output, similar to predicting another token from a different vocabulary, and its loss is added to the cross-entropy loss from the target sequence.

We select the top  length candidates with the highest probabilities, and decode the same example with different lengths in parallel. We then select the sequence with the highest average log-probability as our result

# Results

| Model | Dimensions (Model/Hidden) | Iterations | WMT'14 EN-DE | WMT'14 DE-EN | WMT'16 EN-RO | WMT'16 RO-EN |
|---|---|---|---|---|---|---|
| NAT w/ Fertility (Gu et al., 2018) | 512/512 | 1 | 19.17 | 23.20 | 29.79 | 31.44 |
| CTC Loss (Libovický and Helcl, 2018) | 512/4096 | 1 | 17.68 | 19.80 | 19.93 | 24.71 |
| Iterative Refinement (Lee et al., 2018) | 512/512 | 1 | 13.91 | 16.77 | 24.45 | 25.73 |
| | 512/512 | 10 | 21.61 | 25.48 | 29.32 | 30.19 |
| (Dynamic #Iterations) | 512/512 | ? | 21.54 | 25.43 | 29.66 | 30.30 |
| *Small CMLM with Mask-Predict* | 512/512 | 1 | 15.06 | 19.26 | 20.12 | 20.36 |
| | 512/512 | 4 | **24.17** | **28.55** | **30.00** | 30.43 |
| | 512/512 | 10 | **25.51** | **29.47** | **31.65** | **32.27** |
| *Base CMLM with Mask-Predict* | 512/2048 | 1 | 18.05 | 21.83 | 27.32 | 28.20 |
| | 512/2048 | 4 | **25.94** | **29.90** | **32.53** | **33.23** |
| | 512/2048 | 10 | **27.03** | **30.53** | **33.08** | **33.31** |
| Base Transformer (Vaswani et al., 2017) | 512/2048 | $N$ | 27.30 | — — | — — | — — |
| Base Transformer (Our Implementation) | 512/2048 | $N$ | 27.74 | 31.09 | 34.28 | 33.99 |
| Base Transformer (+Distillation) | 512/2048 | $N$ | 27.86 | 31.07 | — — | — — |
| Large Transformer (Vaswani et al., 2017) | 1024/4096 | $N$ | 28.40 | — — | — — | — — |
| Large Transformer (Our Implementation) | 1024/4096 | $N$ | 28.60 | 31.71 | — — | — — |

# The trade-off between speed-up and translation quality



# Number of iterations and knowledge distillation study

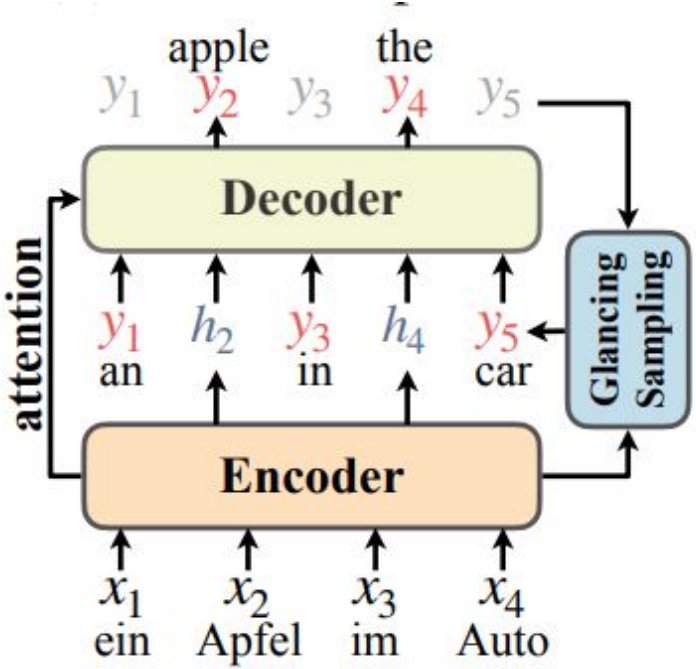
| Iterations | WMT'14 EN-DE | | WMT'16 EN-RO | |
|---|---|---|---|---|
| | BLEU | Reps | BLEU | Reps |
| $T = 1$ | 18.05 | 16.72% | 27.32 | 9.34% |
| $T = 2$ | 22.91 | 5.40% | 31.08 | 2.82% |
| $T = 3$ | 24.99 | 2.03% | 32.19 | 1.26% |
| $T = 4$ | 25.94 | 1.07% | 32.53 | 0.87% |
| $T = 5$ | 26.30 | 0.72% | 32.62 | 0.61% |

Table 3: The performance (BLEU) and percentage of repeating tokens when decoding with a different number of mask-predict iterations ($T$).

| Iterations | WMT'14 EN-DE | | WMT'16 EN-RO | |
|---|---|---|---|---|
| | Raw | Dist | Raw | Dist |
| $T = 1$ | 10.64 | **18.05** | 21.22 | **27.32** |
| $T = 4$ | 22.25 | **25.94** | 31.40 | **32.53** |
| $T = 10$ | 24.61 | **27.03** | 32.86 | **33.08** |

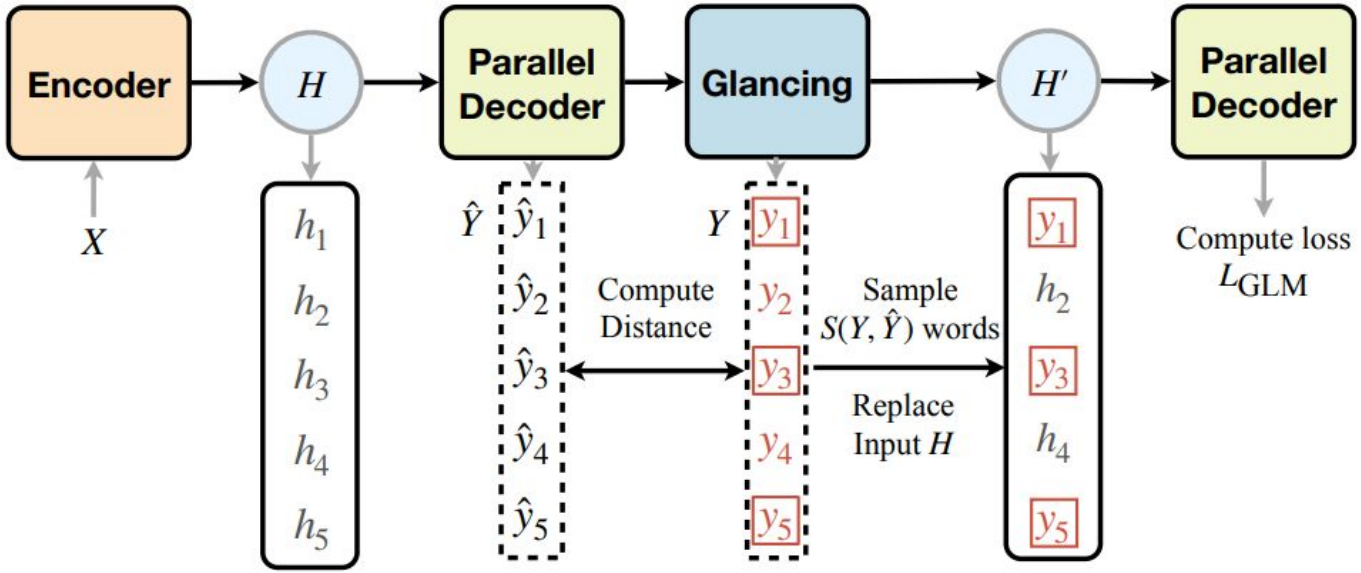
Table 6: The performance (BLEU) of base CMLM, trained with either raw data (Raw) or knowledge distillation from an autoregressive model (Dist).

# Glancing Transformer for Non-Autoregressive Neural Machine Translation



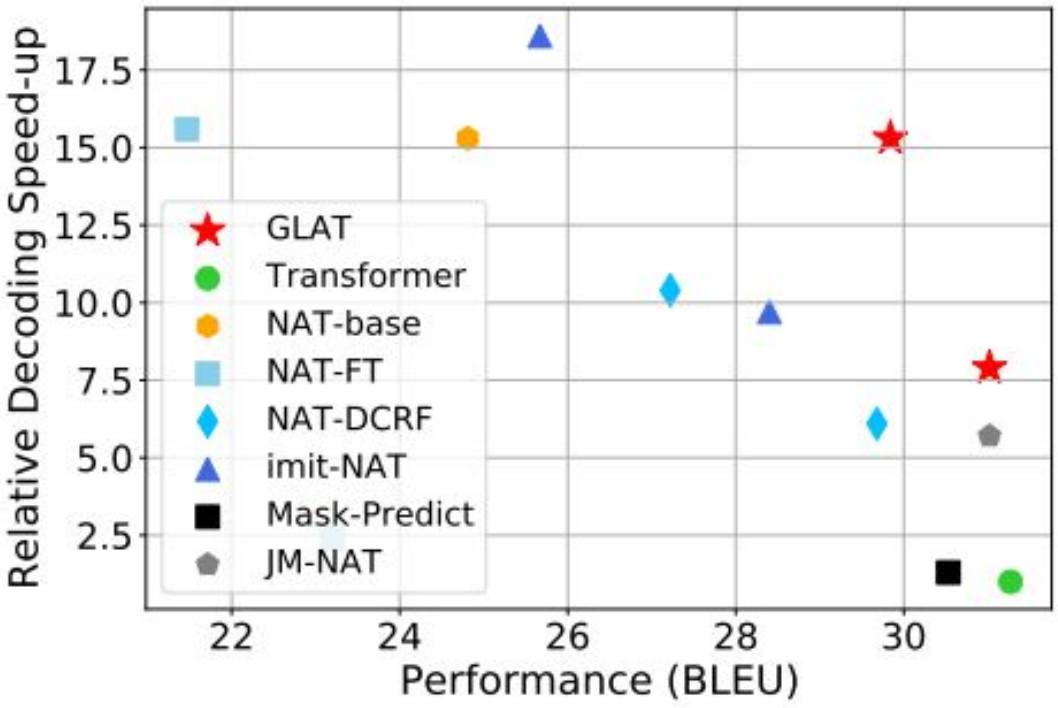

https://arxiv.org/pdf/2008.07905.pdf

# Glancing Transformer for Non-Autoregressive Neural Machine Translation

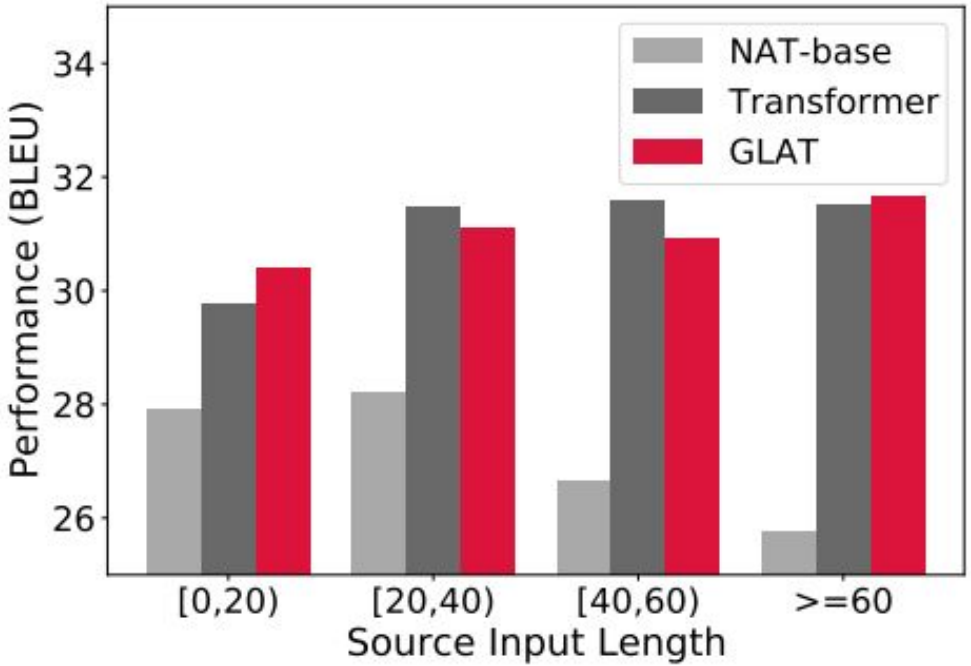


| | | Models | $I_{\text{dec}}$ | WMT14 EN-DE | WMT14 DE-EN | WMT16 EN-RO | WMT16 RO-EN | Speed Up |
|---|---|---|---|---|---|---|---|---|
| AT Models | | Transformer (Vaswani et al., 2017) | T | 27.30 | / | / | / | / |
| | | Transformer (ours) | T | 27.48 | 31.27 | 33.70 | 34.05 | 1.0× |
| Iterative NAT | | NAT-IR (Lee et al., 2018) | 10 | 21.61 | 25.48 | 29.32 | 30.19 | 1.5× |
| | | LaNMT (Shu et al., 2020) | 4 | 26.30 | / | / | 29.10 | 5.7× |
| | | LevT (Gu et al., 2019) | 6+ | 27.27 | / | / | 33.26 | 4.0× |
| | | Mask-Predict (Ghazvininejad et al., 2019) | 10 | 27.03 | 30.53 | **33.08** | **33.31** | 1.7× |
| | | JM-NAT (Guo et al., 2020) | 10 | **27.31** | **31.02** | / | / | 5.7× |
| Fully NAT | | NAT-FT (Gu et al., 2018) | 1 | 17.69 | 21.47 | 27.29 | 29.06 | 15.6× |
| | | Mask-Predict (Ghazvininejad et al., 2019) | 1 | 18.05 | 21.83 | 27.32 | 28.20 | / |
| | | imit-NAT (Wei et al., 2019) | 1 | 22.44 | 25.67 | 28.61 | 28.90 | 18.6× |
| | | NAT-HINT (Li et al., 2019) | 1 | 21.11 | 25.24 | / | / | / |
| | | Flowseq (Ma et al., 2019) | 1 | 23.72 | 28.39 | 29.73 | 30.72 | 1.1× |
| | | NAT-DCRF (Sun et al., 2019) | 1 | 23.44 | 27.22 | / | / | 10.4 × |
| | w/ CTC | NAT-CTC (Libovický and Helcl, 2018) | 1 | 16.56 | 18.64 | 19.54 | 24.67 | / |
| | | Imputer (Saharia et al., 2020) | 1 | 25.80 | 28.40 | 32.30 | 31.70 | 18.6× |
| | w/ NPD | NAT-FT + NPD (m=100) | 1 | 19.17 | 23.20 | 29.79 | 31.44 | 2.4× |
| | | imit-NAT + NPD (m=7) | 1 | 24.15 | 27.28 | 31.45 | 31.81 | 9.7× |
| | | NAT-HINT + NPD (m=9) | 1 | 25.20 | 29.52 | / | / | / |
| | | Flowseq + NPD (m=30) | 1 | 25.31 | 30.68 | 32.20 | 32.84 | / |
| | | NAT-DCRF + NPD (m=9) | 1 | 26.07 | 29.68 | / | / | 6.1× |
| | **Ours** | NAT-base* | 1 | 20.36 | 24.81 | 28.47 | 29.43 | 15.3× |
| | | CTC* | 1 | 25.52 | 28.73 | 32.60 | 33.46 | 14.6 × |
| | | GLAT | 1 | 25.21 | 29.84 | 31.19 | 32.04 | 15.3× |
| | | GLAT + CTC | 1 | 26.39 | 29.54 | 32.79 | **33.84** | 14.6 × |
| | | GLAT + NPD (m=7) | 1 | **26.55** | **31.02** | **32.87** | 33.51 | 7.9× |

# The trade-off between speed-up and BLEU



Performance under different source input length



# References


- NON-AUTOREGRESSIVE NEURAL MACHINE TRANSLATION
  https://arxiv.org/pdf/1711.02281.pdf

- Mask-Predict: Parallel Decoding of Conditional Masked Language Models
  https://arxiv.org/pdf/1904.09324.pdf

- Glancing Transformer for Non-Autoregressive Neural Machine Translation
  https://arxiv.org/pdf/2008.07905.pdf