Deep Encoder, Shallow Decoder: Reevaluating Non-autoregressive MT

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, Noah A. Smith
Paul G. Allen School of CSE, University of Washington
Facebook AI

Summary

- Non-autoregressive MT (**NAR**) is a recent fast alternative to **AR** MT.

Summary

- Non-autoregressive MT (**NAR**) is a recent fast alternative to **AR** MT.
- Parallel generation often **underperforms** yet **outpaces** left-to-right generation on a GPU.

Summary

- Non-autoregressive MT (**NAR**) is a recent fast alternative to **AR** MT.
- Parallel generation often **underperforms** yet **outpaces** left-to-right generation on a GPU.
- Reexamines the speed-accuracy tradeoff.

Summary

- Non-autoregressive MT (**NAR**) is a recent fast alternative to **AR** MT.
- Parallel generation often **underperforms** yet **outpaces** left-to-right generation on a GPU.
- Reexamines the speed-accuracy tradeoff.
    - Suboptimal Layer Allocation
    - Insufficient speed Measurement
    - Lack of Knowledge Distillation for AR Baselines

Reevaluating NAR

Layer Allocation

- Equal depths in the encoder and decoder are typically assumed.
- They have different accuracy and speed implications.

Layer Allocation

- Equal depths in the encoder and decoder are typically assumed.
- They have different accuracy and speed implications.
- Experiments with varying depths.
- **Deep-Shallow** speeds up AR MT with accuracy retained.
  - AR’s speed disadvantage is overestimated.

Speed Measure

- **S1 (Most NAR Works)**
  - 1 sentence (utterance) at a time
  - Instantaneous Translation, Simultaneous Translation,...

Speed Measure

- **S1 (Most NAR Works)**
  - 1 sentence (utterance) at a time
  - Instantaneous Translation, Simultaneous Translation,...
- **Smax**
  - Maximum Batch Size
  - Translate Wikipedia, EU Documents, ...

Knowledge Distillation

- Mitigates Multimodality (Gu et al. 2018).
  - Almost all NAR models need KD.
  - AR MT output is less diverse than human (Shen et al. 2019).

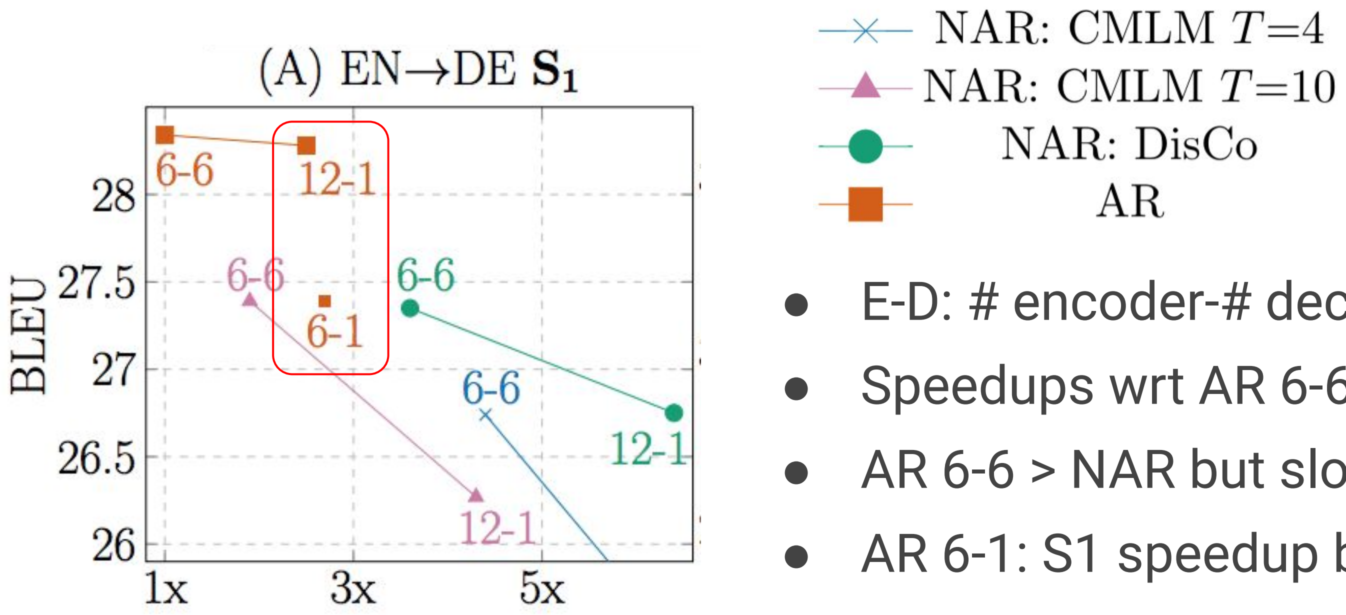
Experiments

Setups: Benchmarks

- Follow prior NAR works (Ghazvininejad et al., 2019; Kasai et al., 2020)
- BPE subwords

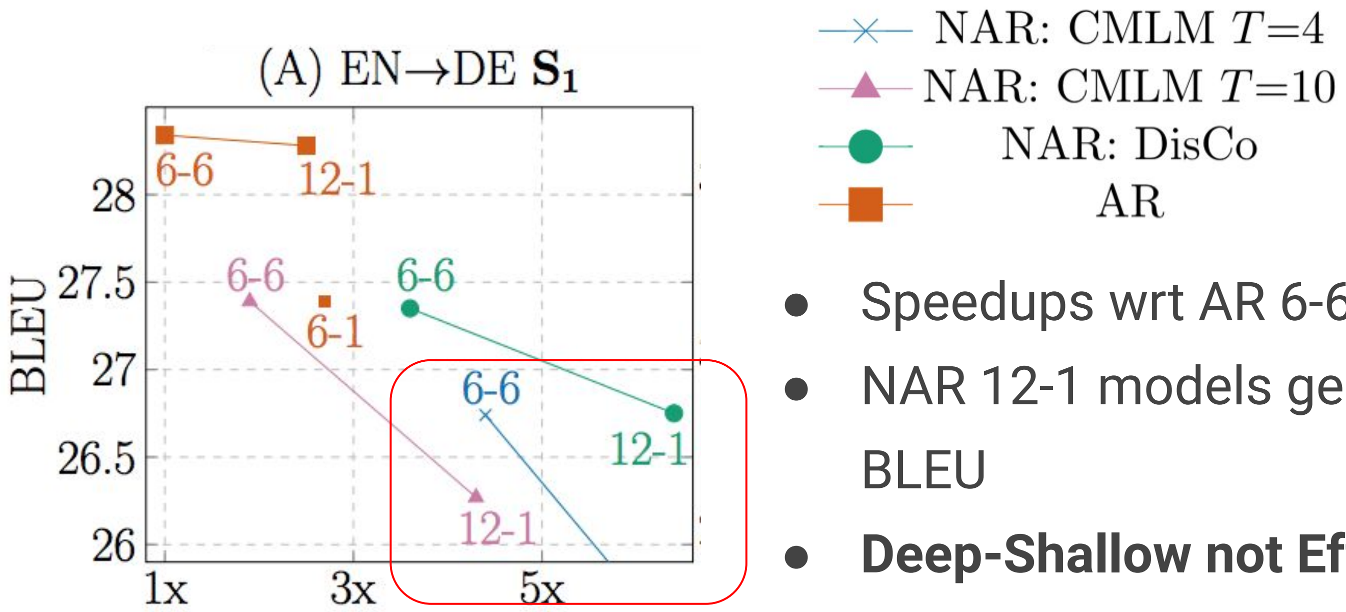
	Train Pairs	Teacher Transformer	Model
WMT 2016 EN-DE	4.5M	Large	Base
WMT 2016 EN-RO	610K	Base	Base
WMT 2017 EN-ZH	20M	Large	Base
WMT 2014 EN-FR	36M	Large	Base

Speed-Accuracy Tradeoff S1



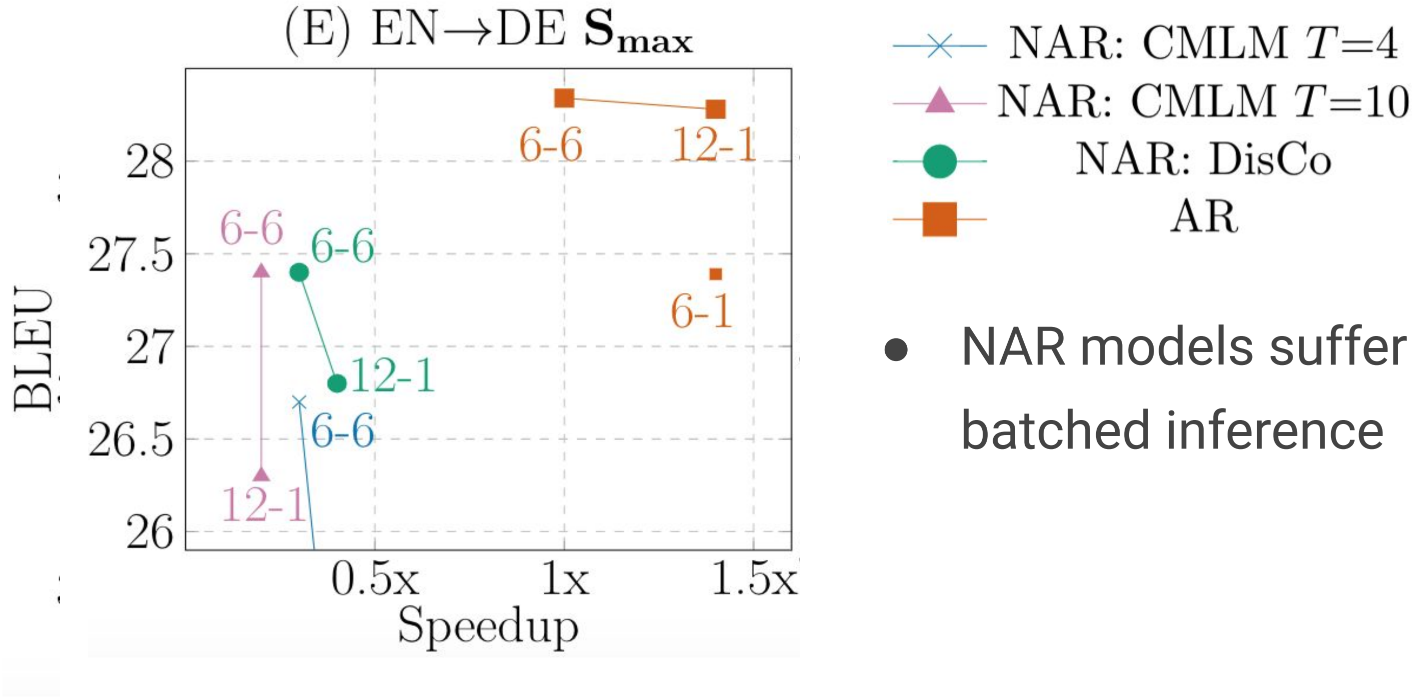
- E-D: # encoder-# decoder
- Speedups wrt AR 6-6 Baseline
- AR 6-6 > NAR but slow in S1.
- AR 6-1: S1 speedup but loss in BLEU.
- **AR 12-1: a balanced middle ground.**

Speed-Accuracy Tradeoff S1



- Speedups wrt AR 6-6 Baseline
- NAR 12-1 models generally suffer in BLEU
- **Deep-Shallow not Effective for NAR**

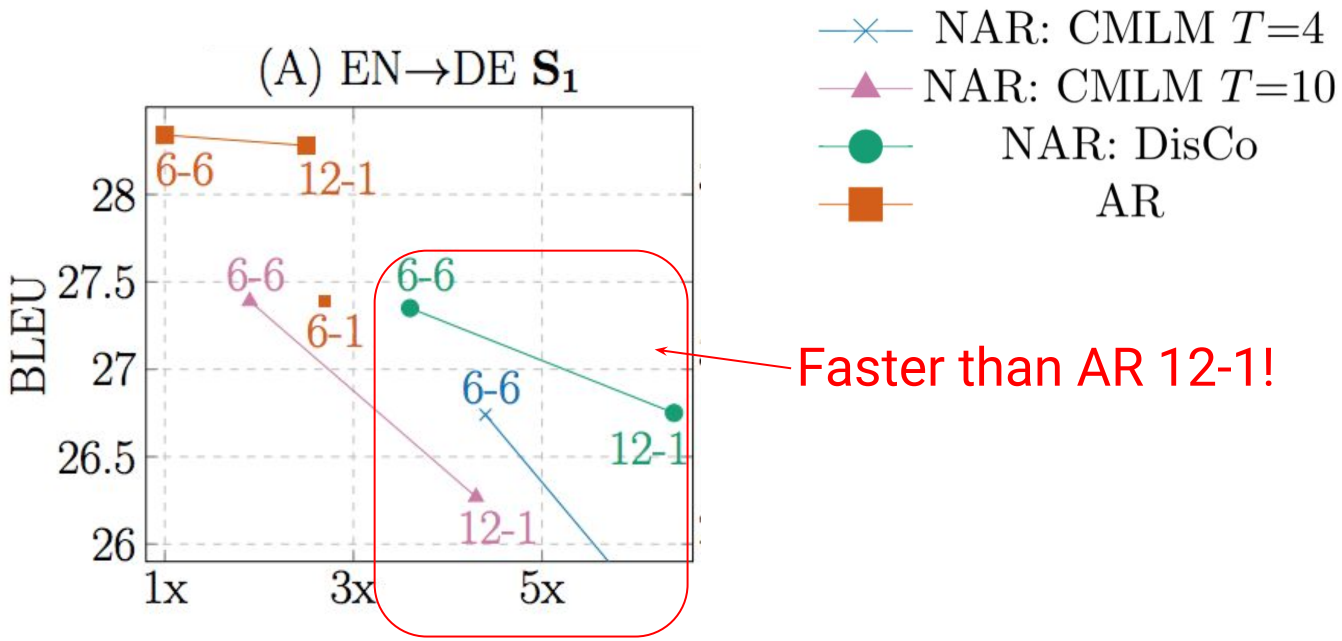
Speed-Accuracy Tradeoff Smax



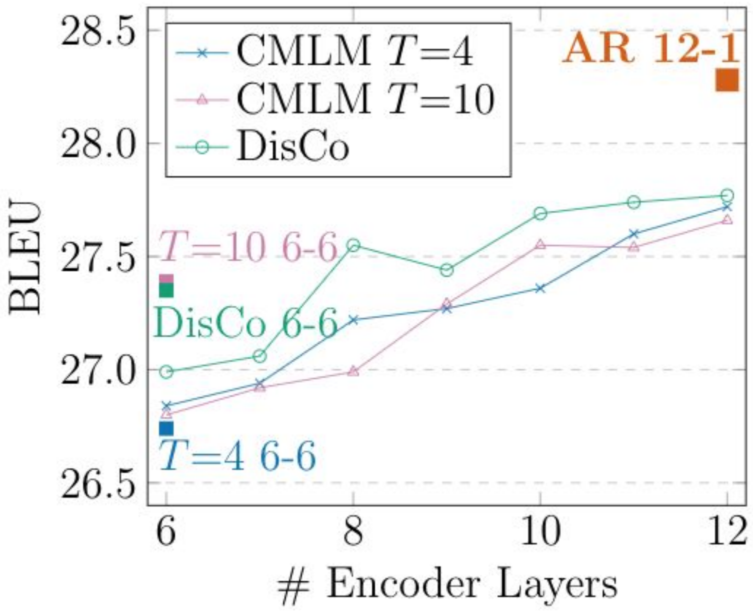
- NAR models suffer in large batched inference

Compare AR and NAR

S1 Speed Constraint

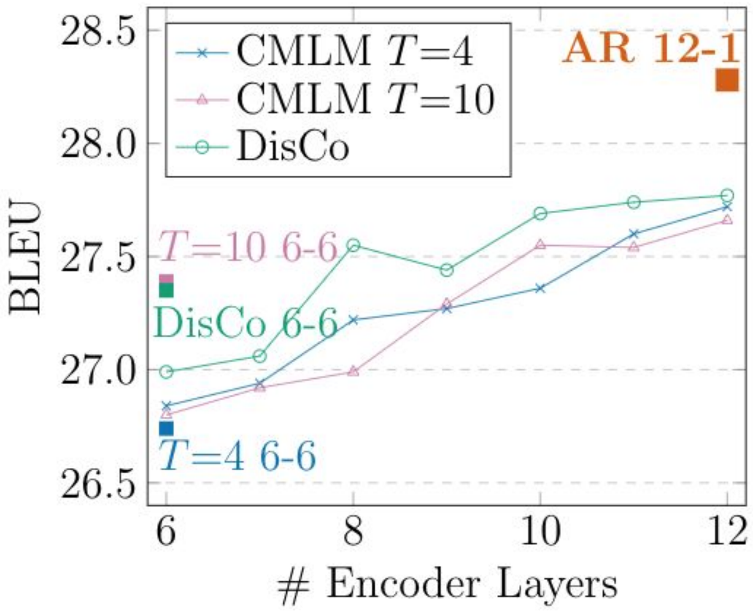


S1 Speed Constraint



- WMT EN-DE Test
- Maximize Decoder Depth in the budget
  - E.g., DisCo 12-9

S1 Speed Constraint



- WMT EN-DE Test
- Maximize Decoder Depth in the budget
  - E.g., DisCo 12-9
- Accuracy still far from AR 12-1 under the same S1 Budget

Conclusion and Future Prospects

Conclusion

- AR’s speed-accuracy balance improves with deep-shallow configurations.

Conclusion

- AR’s speed-accuracy balance improves with deep-shallow configurations.
- Future work in NAR should consider layer allocation, knowledge distillation, and speed measurement.

Conclusion

- AR’s speed-accuracy balance improves with deep-shallow configurations.
- Future work in NAR should consider layer allocation, knowledge distillation, and speed measurement.
- Deep-shallow configurations for other seq2seq tasks? Seq2seq pretraining like T5 or BART?

Thank you!

https://github.com/jungokasai/deep-shallow