ENGINE: Energy-Based Inference Networks for Non-Autoregressive Neural Machine Translation

Lifu Tu[1] Richard Yuanzhe Pang[2] Sam Wiseman[1] Kevin Gimpel[1]

[1]Toyota Technological Institute at Chicago

[2]New York University

ACL 2020

Motivation

Autoregressive Neural Machine Translation

- state-of-art models for neural machine translation
- left-to-right decoding

Non-Autoregressive Neural Machine Translation [Gu et al., 2018]

- parallel decoding, much faster
- large performance gap with autoregressive models

Motivation

Autoregressive Neural Machine Translation

- state-of-art models for neural machine translation
- left-to-right decoding

Non-Autoregressive Neural Machine Translation [Gu et al., 2018]

- parallel decoding, much faster
- large performance gap with autoregressive models

In this work, train a non-autoregressive model to minimize autoregressive energy!

Inference for Structured Models

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} E_\Theta(\boldsymbol{x}, \boldsymbol{y})$$

where energy function $E(\boldsymbol{x}, \boldsymbol{y})$ is a scalar that measures the compatibility of each configuration $\boldsymbol{x}$ and $\boldsymbol{y}$ [LeCun et al., 2006; Belanger and McCallum, 2016]

Inference for Structured Models

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} E_\Theta(\boldsymbol{x}, \boldsymbol{y})$$

where energy function $E(\boldsymbol{x}, \boldsymbol{y})$ is a scalar that measures the compatibility of each configuration $\boldsymbol{x}$ and $\boldsymbol{y}$ [LeCun et al., 2006; Belanger and McCallum, 2016]

Inference Networks [Tu and Gimpel (2018,2019)]

An inference network $\mathbf{A}_\Psi : \mathcal{X} \rightarrow \mathcal{Y}_R$ is parameterized by Ψ and trained with the goal that

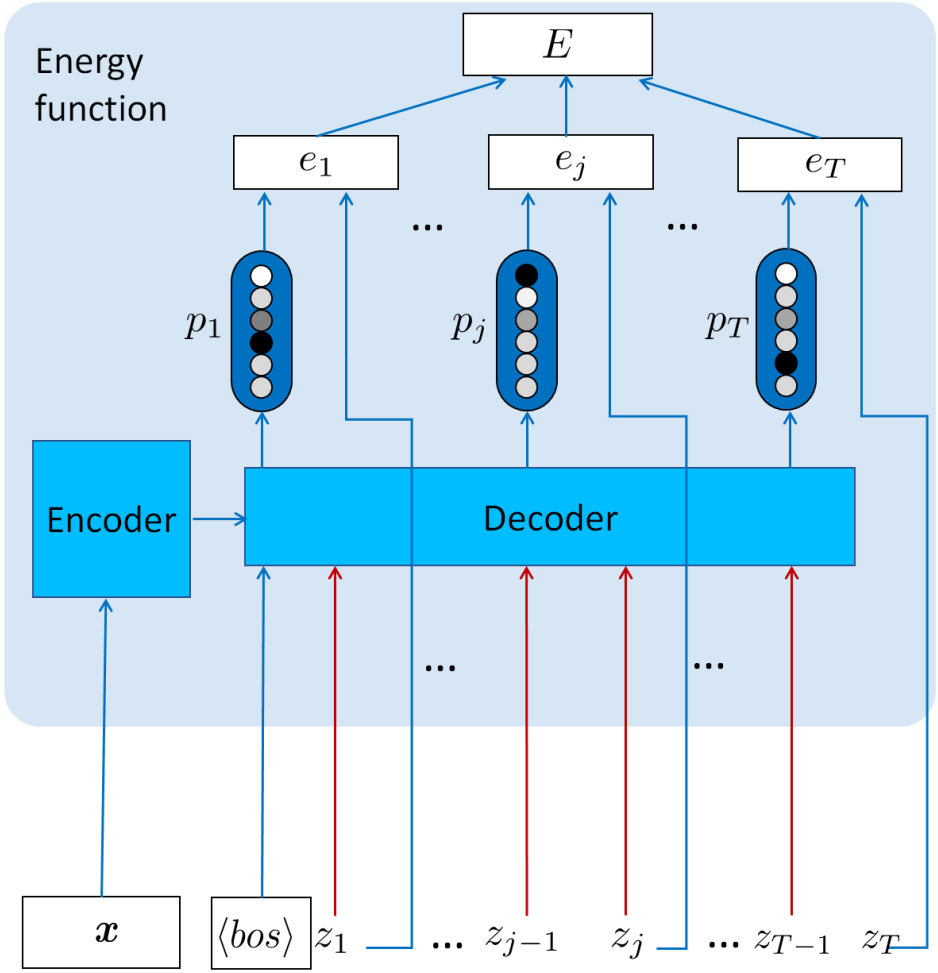
$$\mathbf{A}_\Psi(\boldsymbol{x}) \approx \arg\min_{\boldsymbol{y} \in \mathcal{Y}_R(\boldsymbol{x})} E_\Theta(\boldsymbol{x}, \boldsymbol{y})$$

$\mathcal{Y}_R(\boldsymbol{x})$ is relaxed continuous output space.

The objective for inference networks:

$$\widehat{\Psi} = \arg\min_{\Psi} \sum_{\langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \mathcal{D}} E_{\Theta}(\boldsymbol{x}, \mathbf{A}_{\Psi}(\boldsymbol{x}))$$

ENGINE

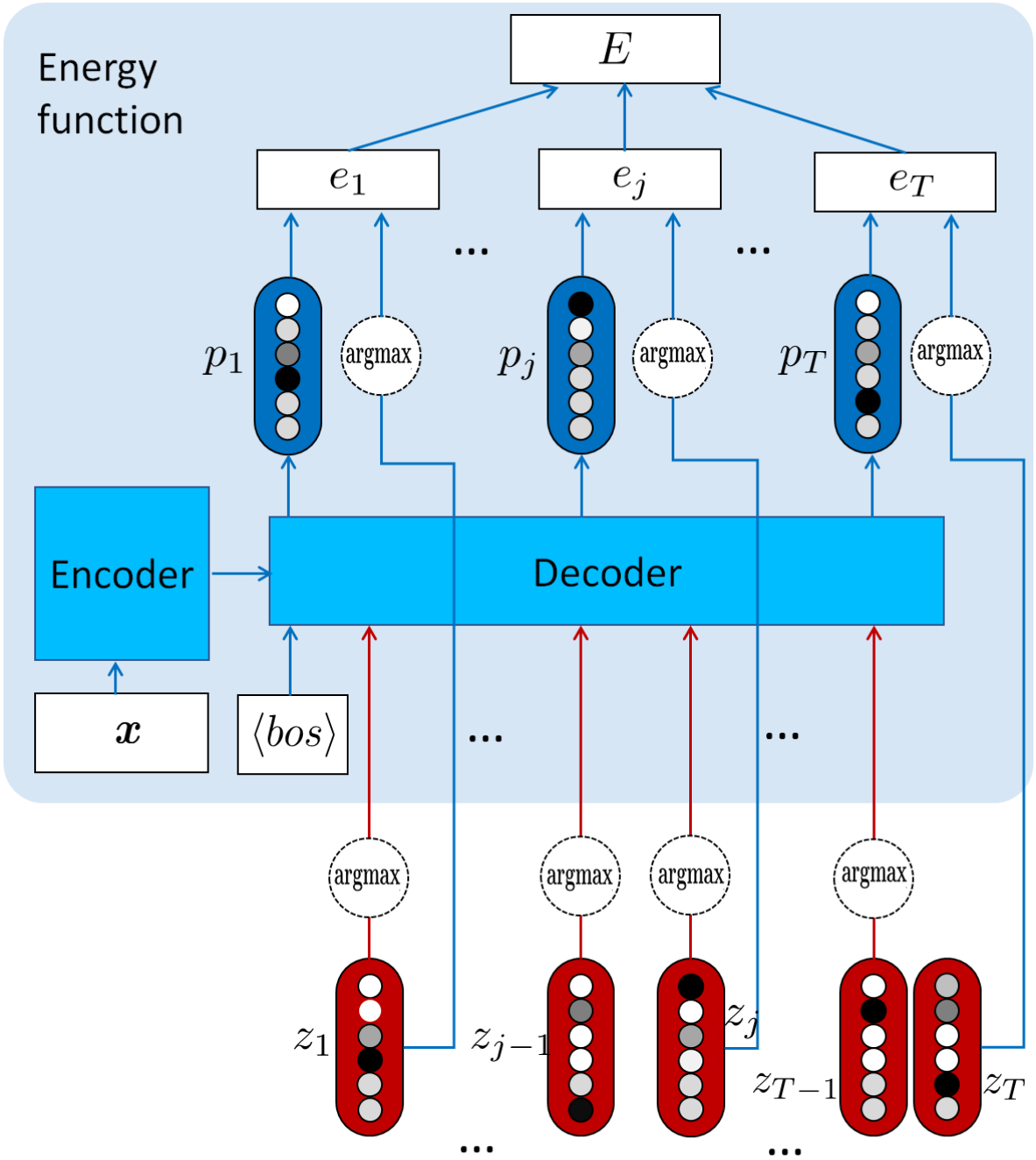


score a sequence of words $\boldsymbol{z}$

$$e_j = -\log p_\Theta(\boldsymbol{z}_t \mid \boldsymbol{z}_{0:j-1}, \boldsymbol{x})$$

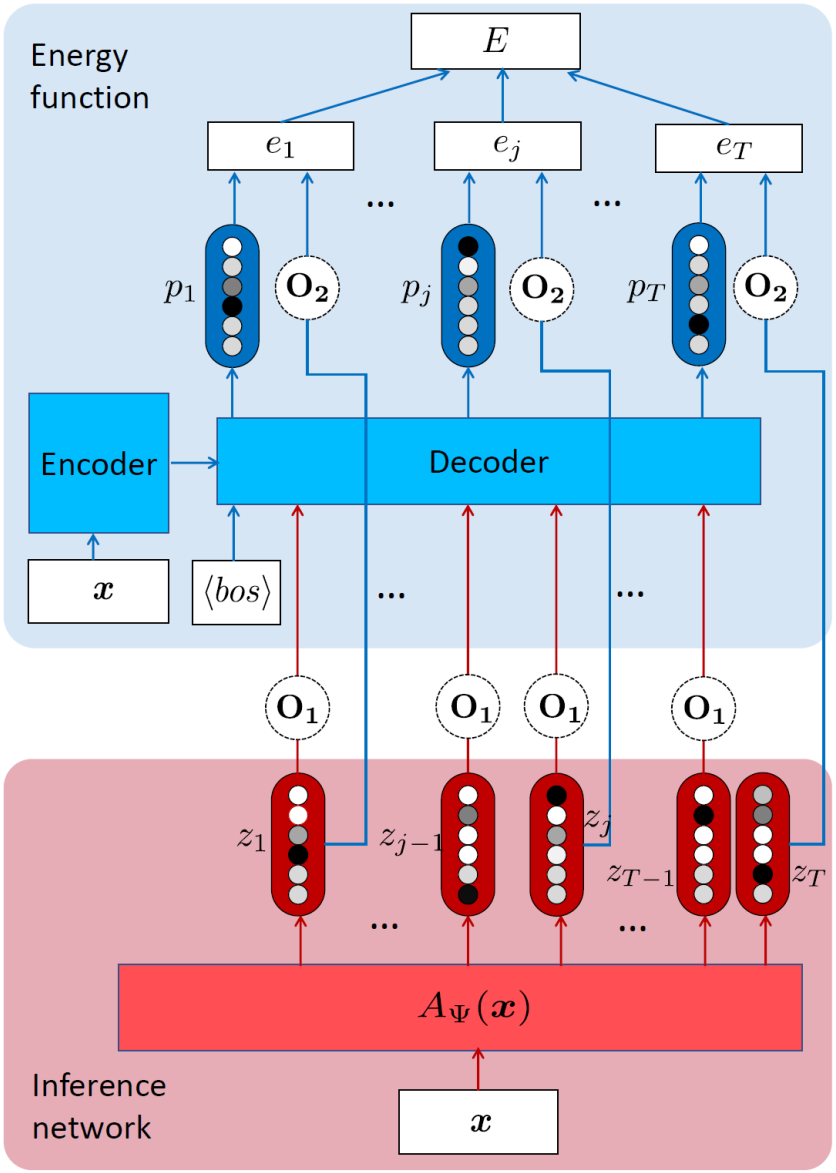
$$E = \sum_{j=1}^{T} e_j$$

The beam search algorithm is to minimize the above energy.



we introduce the generalized energy E

Here $\boldsymbol{z}$ is a sequence of probability vectors!

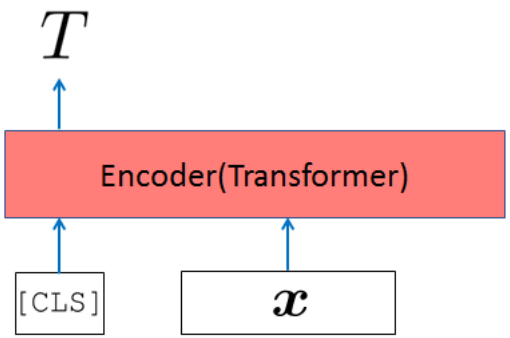


How to allow gradient-based optimization?

argmax operation is replaced by $\mathbf{O_1}$ or $\mathbf{O_2}$

Inference Networks Architecture

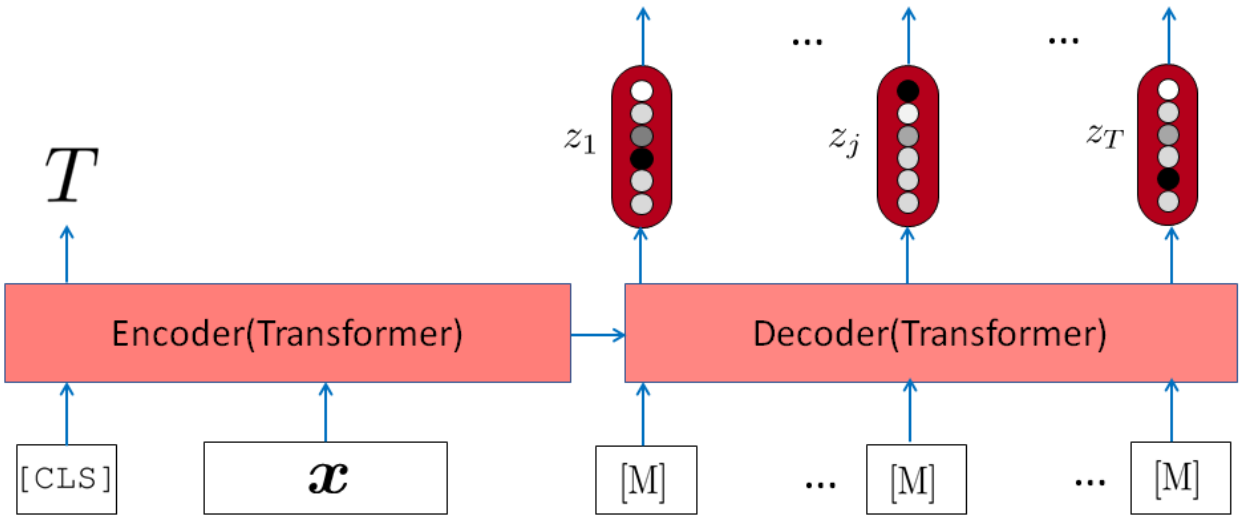
CMLM [Ghazvininejad et al., 2019]



Predicting Length

Inference Networks Architecture

CMLM [Ghazvininejad et al., 2019]

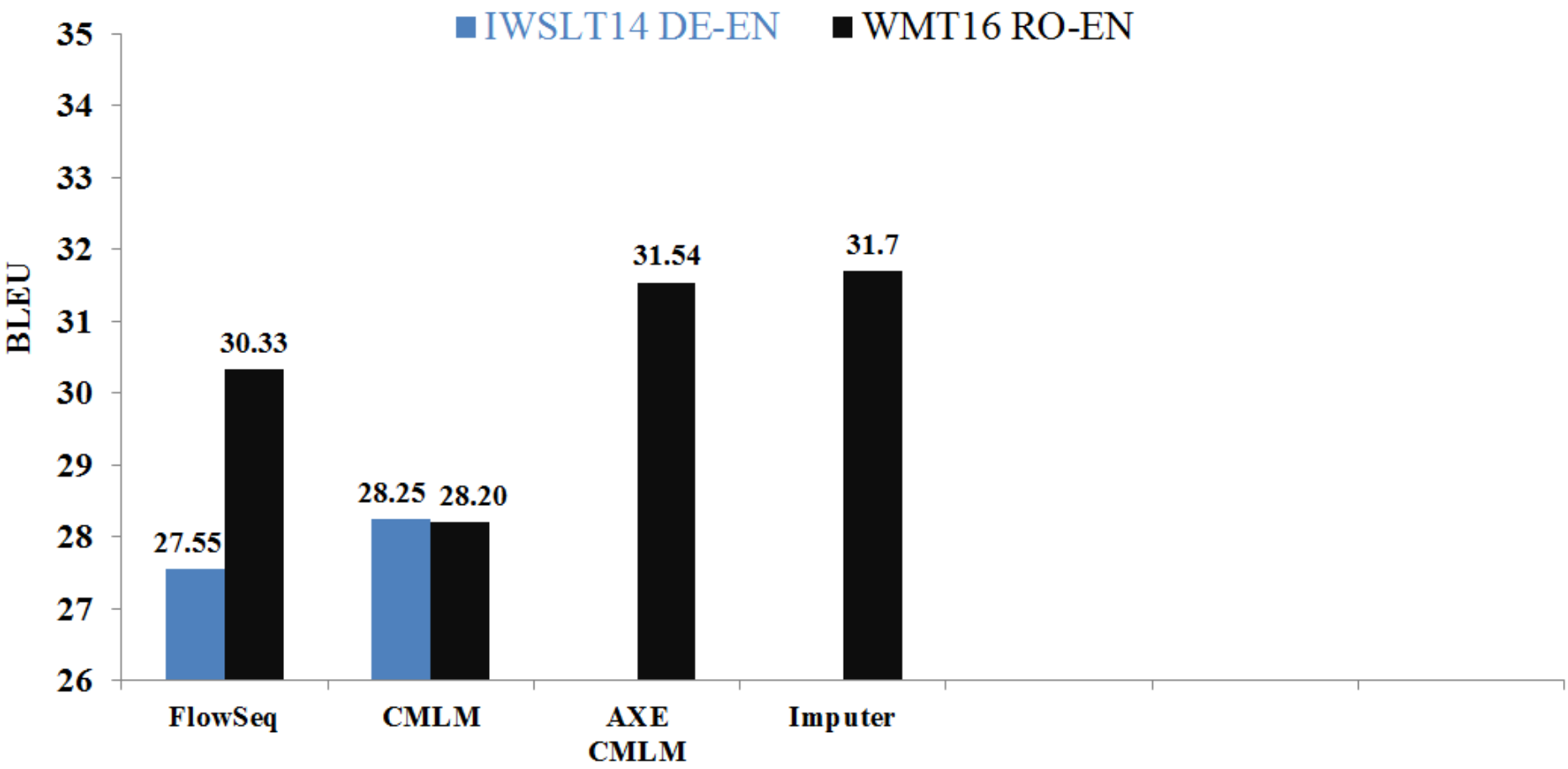


The decoder inputs are the special masked tokens $[M]$.

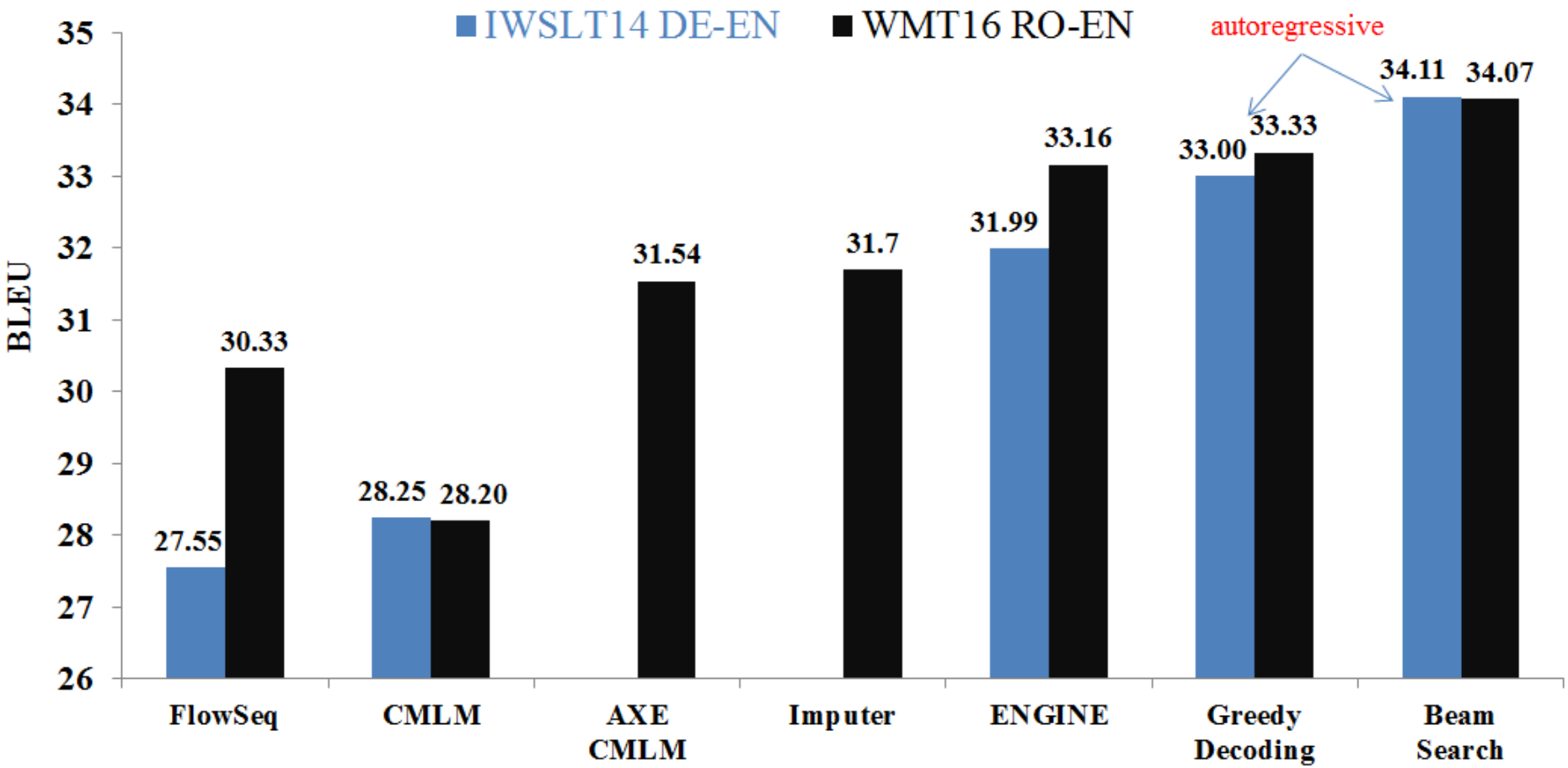
Experimental Setup

- Datasets: IWSLT2014 DE-EN and WMT2016 RO-EN
- Energy function: pretrained autoregressive model
- Inference network architecture: CMLM
- One decoding iteration is used (meaning they are purely non-autoregressive.)

Results


IWSLT14 DE-EN
WMT16 RO-EN
BLEU
35
34
33
32
31
30
29
28
27
26
27.55
30.33
28.25
28.20
31.54
31.7
FlowSeq
CMLM
AXE CMLM
Imputer

Results



- achieves state-of-the-art results for non-autoregressive translation
- approaching the performance of autoregressive models

Example 1

Ground Truth : the u.n. chief again urged all parties , including the divided u.n. security council , to unite and support inclusive negotiations to find a political solution .
CMLM : the un chief again again urged all parties , including the divided un security council to unify and support negotiations in order to find a political solution . **ENGINE (ours)** : the un chief has again urged all parties , including the divided un security council to unify and support negotiations in order to find a political solution .

An example from the WMT'16 RO-EN test set.

Example 2

Ground Truth :
the study 's conductors transmit that " romanians feel the need for a little more adventure in their lives (24 %) , followed by affection (21 %) , money (21 %) , safety (20 %) , new things (19 %) ...

CMLM :

survey survey makers say that ' romanians romanians some something adventadventure ure their lives 24 24 %) followed followed by % % % % % , (21 % %), safety (% % %), new19% %),), 19 % % %), respect 18 % % % % % % % % % , , % % % % % % % , , % , 14 % , 12 % %

ENGINE (ours) :
realisation of the survey say that ' romanians feel a slightly more adventure in their lives (24 %) followed by aff% (21 %) , money (21 %), safety (20 %) , new 19 %) , sex (19 %) , respect 18 % , confidence 17 % , 17 % , connecting 17 % , knowledge % % , 14 % , 14 % , 12 % %

An longer example from the WMT'16 RO-EN test set.

Thanks!