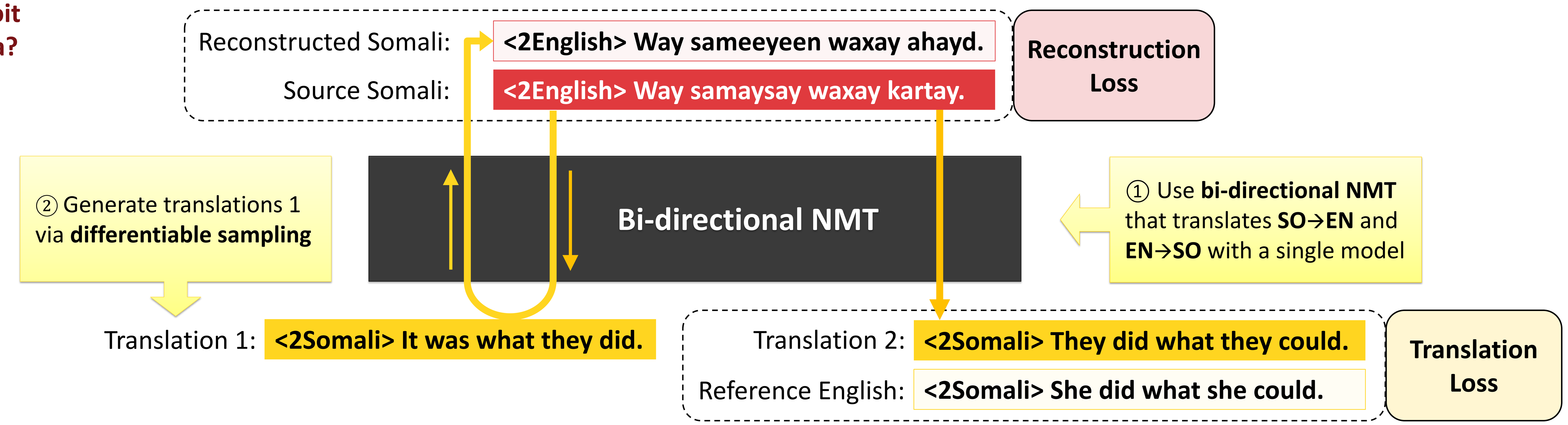
Bi-Directional Differentiable Input Reconstruction for Low-Resource NMT

Xing Niu, Weijia Xu, Marine Carpuat

University of Maryland, College Park

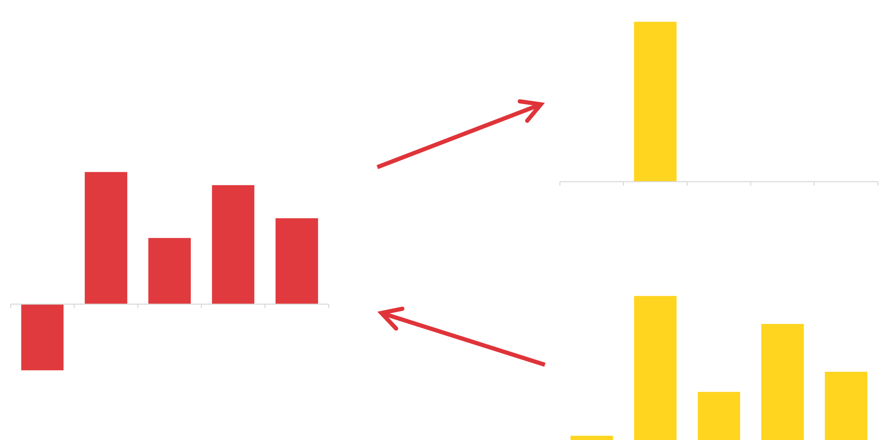
APPROACH

- **How to better exploit limited parallel data?**



- **End-to-end training with differentiable sampling**
  - Approximate the gradient with the Straight-Through Gumbel Softmax

Greedy (β = 0)



$$e_t = \text{one-hot}\left(\arg\max_k \left(a(\boldsymbol{h}_t)_k + G_k\right)\right)$$

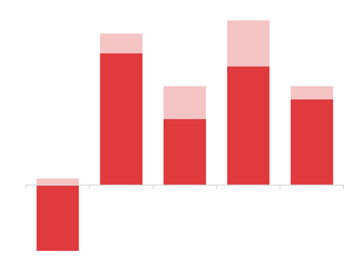
$$G = \text{Gumbel}(0, \beta)$$

Forward prediction

$$\tilde{e}_t = \text{softmax}\left((a(\boldsymbol{h}_t) + G)/\tau\right)$$

Back-propagation

Sampling (β = 0.5)



SETUP

- **Models:**
  - Baseline: optimize the forward translation loss
  - Hidden: fine-tune baselines with translation loss + loss of reconstruction from hidden states
  - Ours: fine-tune baselines with translation loss + loss of reconstruction from sampled translation
- **Tasks:**
  - Swahili, Tagalog, Somali, Turkish <-> English

	Training	Dev.	Test
SW↔EN	60,570	500	3,000
TL↔EN	70,703	704	3,000
SO↔EN	68,550	844	3,000
TR↔EN	207,021	1,001	3,007

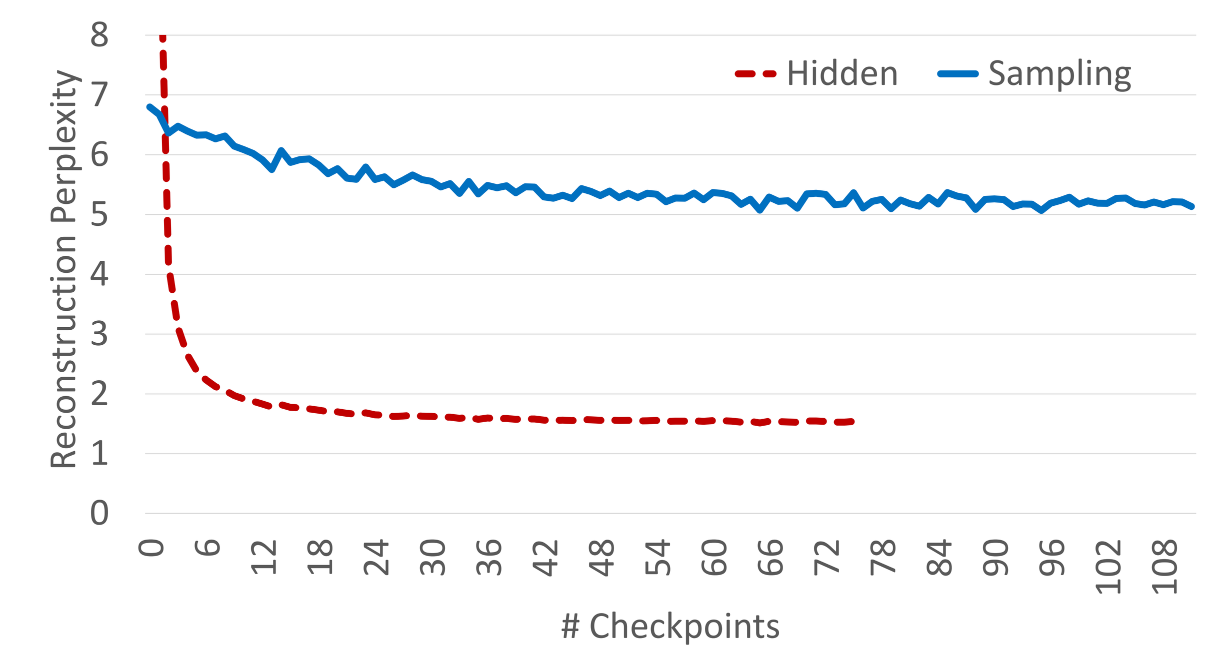
RESULTS

Model	EN→SW		SW→EN		EN→TL		TL→EN		EN→SO		SO→EN		EN→TR		TR→EN	
Baseline	33.60	Δ	30.70	Δ	27.23	Δ	32.15	Δ	12.25	Δ	20.80	Δ	12.90	Δ	15.32	Δ
Hidden	33.41	-0.19	30.91	+0.21	27.43	+0.19	32.20	+0.04	12.30	+0.05	20.72	-0.08	12.77	-0.13	15.34	+0.01
Ours: greedy	33.92	+0.32	31.37	+0.66	27.65	+0.42	32.75	+0.59	12.47	+0.22	21.14	+0.35	13.26	+0.36	15.60	+0.28
Ours: sampling	33.97	+0.37	31.39	+0.69	27.65	+0.42	32.65	+0.50	12.48	+0.23	21.20	+0.41	13.16	+0.25	15.52	+0.19

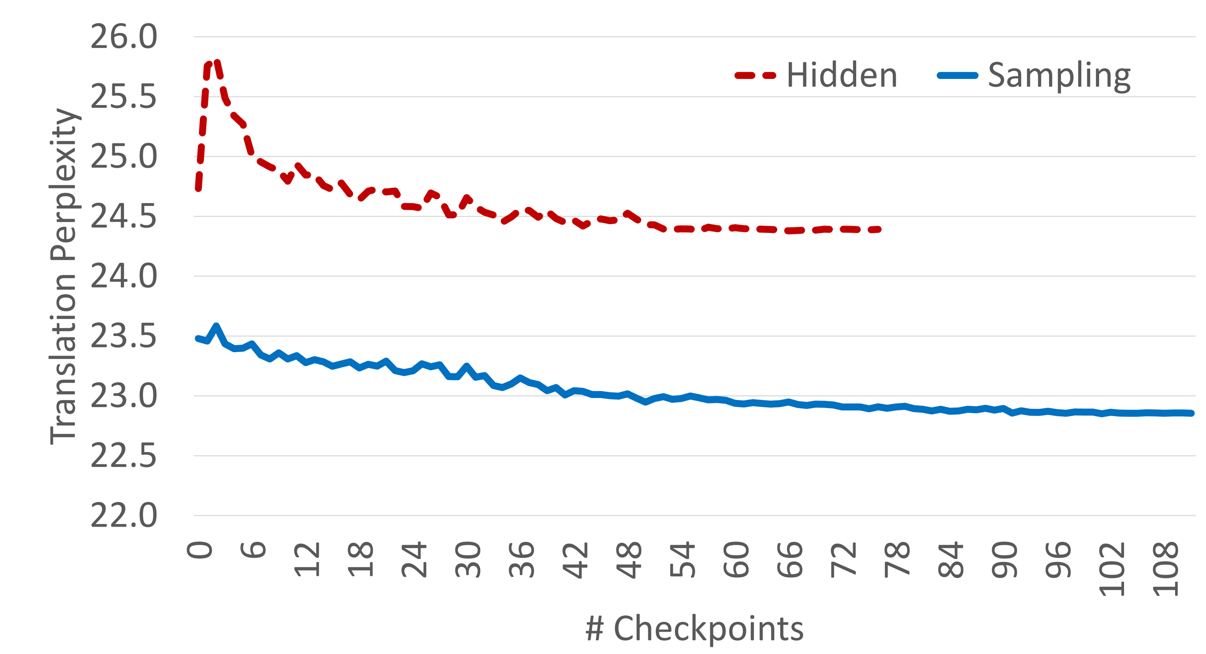
- **Our approach achieves small but consistent BLEU improvements**
- **It is effective even if there is moderate domain mismatch**
  - OOV rate of TR->EN is larger than 20%
  - OOV rate of SW->EN is much smaller and it obtains higher ΔBLEU
- **It outperforms input reconstruction from hidden states**
- **Training becomes more stable (see ANALYSIS)**

ANALYSIS

- **Training curves of perplexity for TR-EN**
  - Training set



  - Development set



- **Hidden: extremely low training perplexity**
  - Hidden representations memorize the input
- **Our approach: more stable training**
  - All parameters are jointly pre-trained

DOWNLOAD

PDF