Bi-Directional Neural Machine Translation with Synthetic Parallel Data

Xing Niu[1], Michael Denkowski[2], Marine Carpuat[1]

[1]University of Maryland [2]Amazon.com, Inc.

INTRODUCTION

- **Problem:**
  - Back-translated monolingual data improves NMT performance [1].
  - But it requires building a reverse NMT system which is expensive.
- **Our solution:**
  - Combine back-translation with bi-directional NMT.
  - Inspired by multilingual NMT which reduces deployment complexity by packing multiple language pairs into a single model [2].

APPROACH

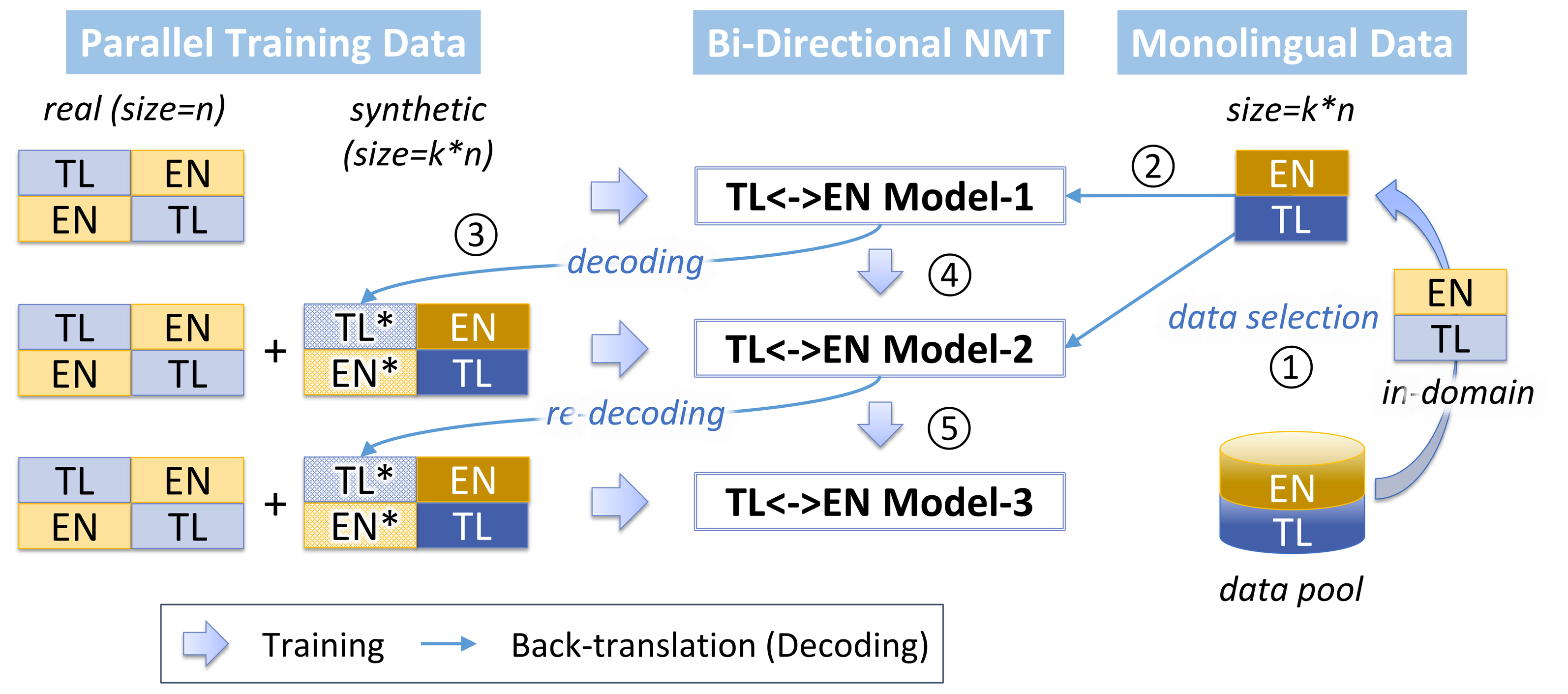
① Select the monolingual data using cross-entropy difference [3].

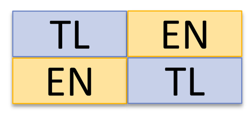
② Back-translate both source and target monolingual data by a single initial bi-directional NMT model (Model-1).

③ Always place the real (monolingual) data on the target side.

④ Fine-tune Model-1 on the augmented training data to get a stronger NMT model (Model-2).

⑤ Re-decode the monolingual data and fine-tune Model-2 to get an even stronger NMT model (Model-3).





Bi-directional parallel training data:

1. Adding a language token (e.g. <2en>) to the source.
2. Swapping the source and target sentences and appending the swapped version to the original.

IN-DOMAIN EVALUATION (BLEU)

ID	Training Data	TL→EN	EN→TL	SW→EN	EN→SW	DE→EN	EN→DE
U-1	L1→L2	31.99	31.28	32.60	39.98	29.51	23.01
U-2	L1→L2 + L1⋆→L2	**24.21**	**29.68**	**25.84**	**38.29**	**33.20**	**25.41**
U-3	L1→L2 + L1→L2⋆	22.13	27.14	24.89	36.53	30.89	23.72
U-4	L1→L2 + L1⋆→L2 + L1→L2⋆	23.38	29.31	25.33	37.46	33.01	25.05
	L1=EN	L2=TL		L2=SW		L2=DE	
B-1	L1↔L2	32.72	31.66	33.59	39.12	28.84	22.45
B-2	L1↔L2 + L1⋆↔L2	32.90	**32.33**	33.70	**39.68**	29.17	**24.45**
B-3	L1↔L2 + L2⋆↔L1	32.71	31.10	33.70	39.17	**31.71**	21.71
B-4	L1↔L2 + L1⋆↔L2 + L2⋆↔L1	**33.25**	**32.46**	**34.23**	38.97	30.43	22.54
B-5	L1↔L2 + L1⋆→L2 + L2⋆→L1	**33.41**	**33.21**	**34.11**	**40.24**	**31.83**	**24.61**
B-5*	L1↔L2 + L1⋆→L2 + L2⋆→L1	33.79	32.97	34.15	40.61	31.94	24.45
B-6*	L1↔L2 + $\underline{\text{L1}\star}$→L2 + $\underline{\text{L2}\star}$→L1	**34.50**	**33.73**	**34.88**	**41.53**	**32.49**	**25.20**

- Synthetic data (i.e. MT output) is annotated by asterisks.
- Largest improvements within each zone are highlighted.

- Uni-directional models (U-x).
  - Models trained on real target language data outperform using synthetic target language data (**U-2** vs. **U-3,4**).
- Bi-directional models (B-x).
  - Combining all synthetic parallel data and always placing the MT output on the source side achieve best overall performance (**B-5**).
  - Bi-directional models outperform the best uni-directional models for low-resource (EN-TL/SW) language pairs (**B-5** vs. **U-1**).
  - Bi-directional models struggle to match performance in the high-resource (EN-DE) scenario (**B-5** vs. **U-2**).
  - Bi-directional models reduce the training time by 15-30% (B-5 vs. U-2).
- Fine-tuning and re-decoding.
  - Instead of training from scratch (B-5), we can continue training baseline models (B-1) on augmented data and achieve comparable translation quality (**B-5***).
  - Fine-tuning significantly reduces cost by up to 20-40% computing time.
  - Re-decoding the same monolingual data using improved models (B-5*) leads to even stronger models (**B-6***).

OUT-OF-DOMAIN EVALUATION (BLEU)

		L2=TL		L2=SW	
ID	Training Data (L1=EN)	TL→EN	EN→TL	SW→EN	EN→SW
A-1	L1↔L2	11.03	10.17	6.56	3.80
A-2	L1↔L2 + L1⋆→L2 + L2⋆→L1	16.49	22.33	8.70	7.47
A-3	L1↔L2 + $\underline{\text{L1}\star}$→L2 + $\underline{\text{L2}\star}$→L1	**18.91**	**23.41**	**11.01**	**8.06**

- A long-distance domain adaptation task: News/Blog to Bible.
  - Domain mismatch is demonstrated by the extremely low BLEU scores of baseline News/Blog systems (A-1).
  - Selecting monolingual data which is closer to Biblical language.
  - After fine-tuning baseline models on augmented parallel data (A-2) and re-decoding (A-3), we see BLEU scores increase by 70-130%.

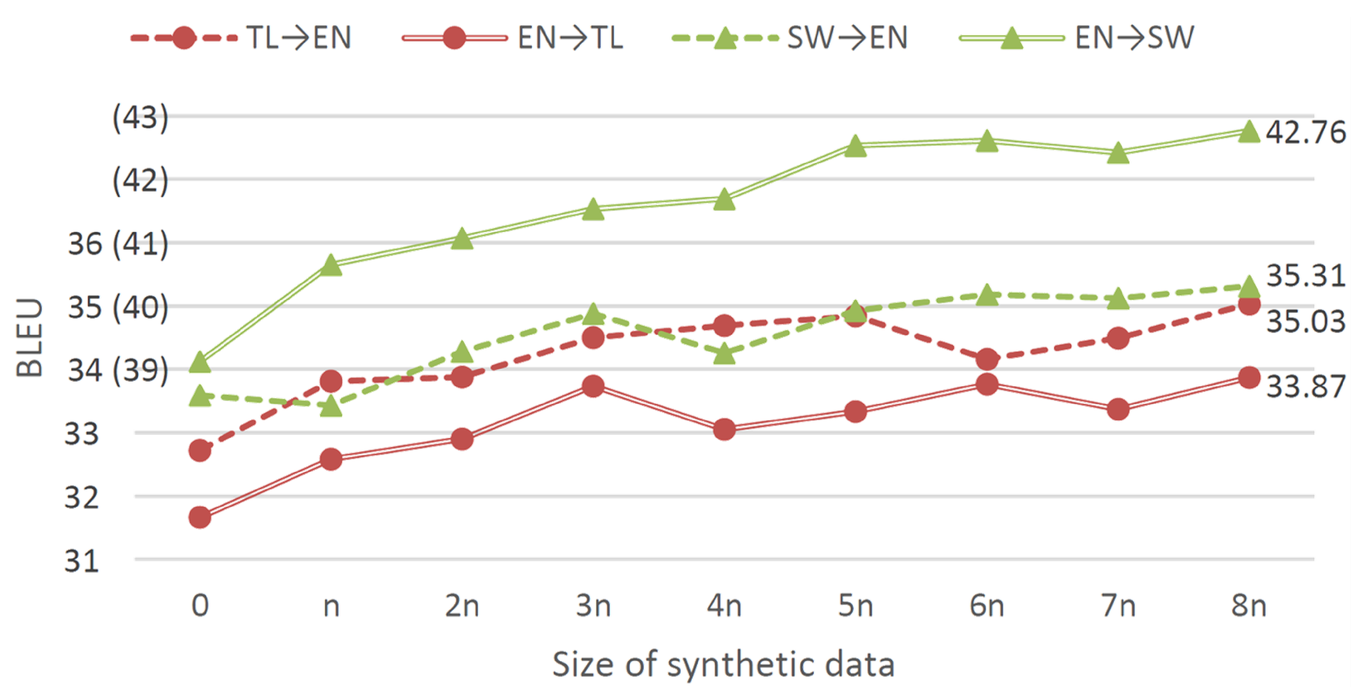
EXPERIMENTAL SETUP

- Training data:

Language Pair		#Sentences	Dataset
English-Tagalog	EN-TL	50,705	News/Blog
English-Swahili	EN-SW	23,900	News/Blog
English-German	EN-DE	4,356,324	WMT News

- In-domain test data:
  - News/Blog for EN-TL and EN-SW
  - News for EN-DE
- Out-of-domain test data:
  - Bible for EN-TL and EN-SW

SIZE OF SYNTHETIC DATA



- Using synthetic parallel data is always helpful, but when the size is larger than 5n, adding more contributes less (i.e. reaching the plateau) for our systems.

CONCLUSION

- We introduce a bi-directional NMT protocol to effectively leverage monolingual data.
- Training and deployment costs are reduced significantly compared to standard uni-directional systems.
- It improves BLEU for low-resource languages, even over uni-directional systems with back-translation.
- It is effective in domain adaptation.

REFERENCES

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In ACL.

[2] Melvin Johnson et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. TACL.

[3] Robert C. Moore and William D. Lewis. 2010. Intelligent selection of language model training data. In ACL.