Exploiting Monolingual Data in Neural Machine Translation

Wang Zhu, Bowen Chen, Zongwei Na, Jiacheng Chen
Simon Fraser University
CMPT 413 Natural Language Processing, Group LGD

SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD

Motivation

- Neural Machine Translation (NMT) models have made great success, but many of them heavily depend on large-scale parallel corpus
- Collecting monolingual data is always easier than collecting parallel corpus, see WMT 18's datasets for example:

File	Size
Europarl v7	62[illegible]MB
Europarl v8	23[illegible]MB
ParaCrawl corpus	2.8GB
Common Crawl corpus	876MB
News Commentary v13	111M

Parallel data (mostly in MB)

Common Crawl	10.5GB	102GB	103 GB	4.0GB	5.3GB	TBC	42GB	18GB	33GB

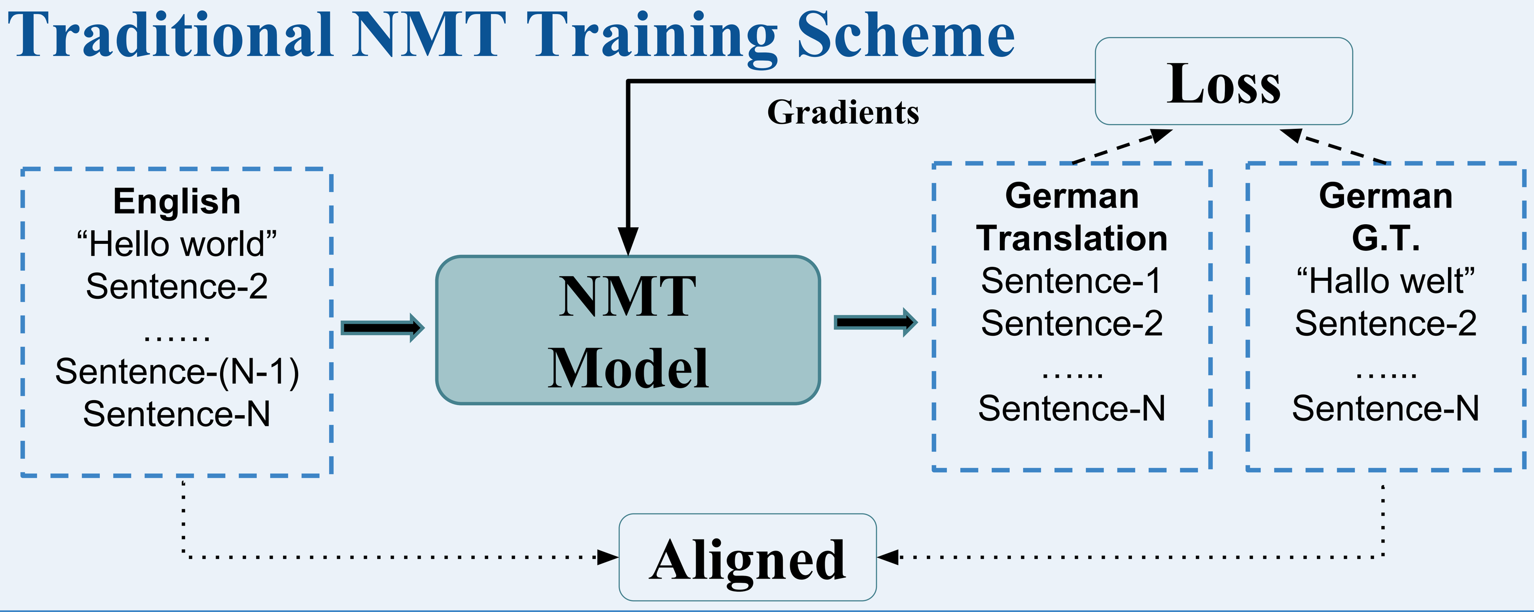
News Crawl: articles from 2015	360MB	2.2GB	1.3GB	43MB	203MB		608MB				4.8G (excludes et)
News Crawl: articles from 2016	252MB	1.6GB	1GB	34MB	163MB		418MB	77MB			3.7G (excludes et)
News Crawl: articles from 2017	323M	1.8GB	1.3GB	36MB	143MB		504MB	135MB			4.2G

Monolingual data (the scale is way larger than parallel datasets)

- Can we utilize rich monolingual data to train NMT models? Or in other word, can we still train NMT models if we don't have supervision (i.e. paired corpus) or only have limited supervision?

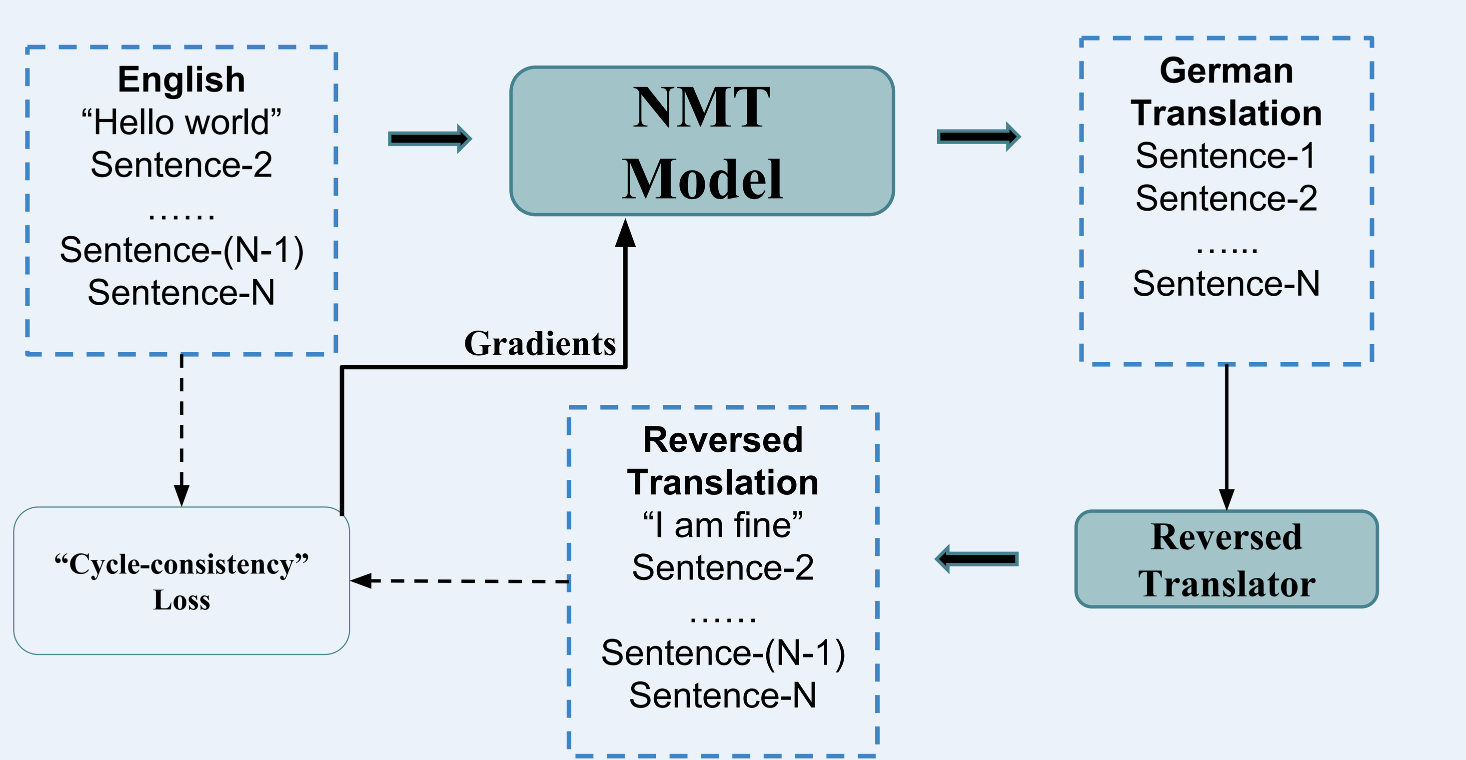
Methodology

Traditional NMT Training Scheme



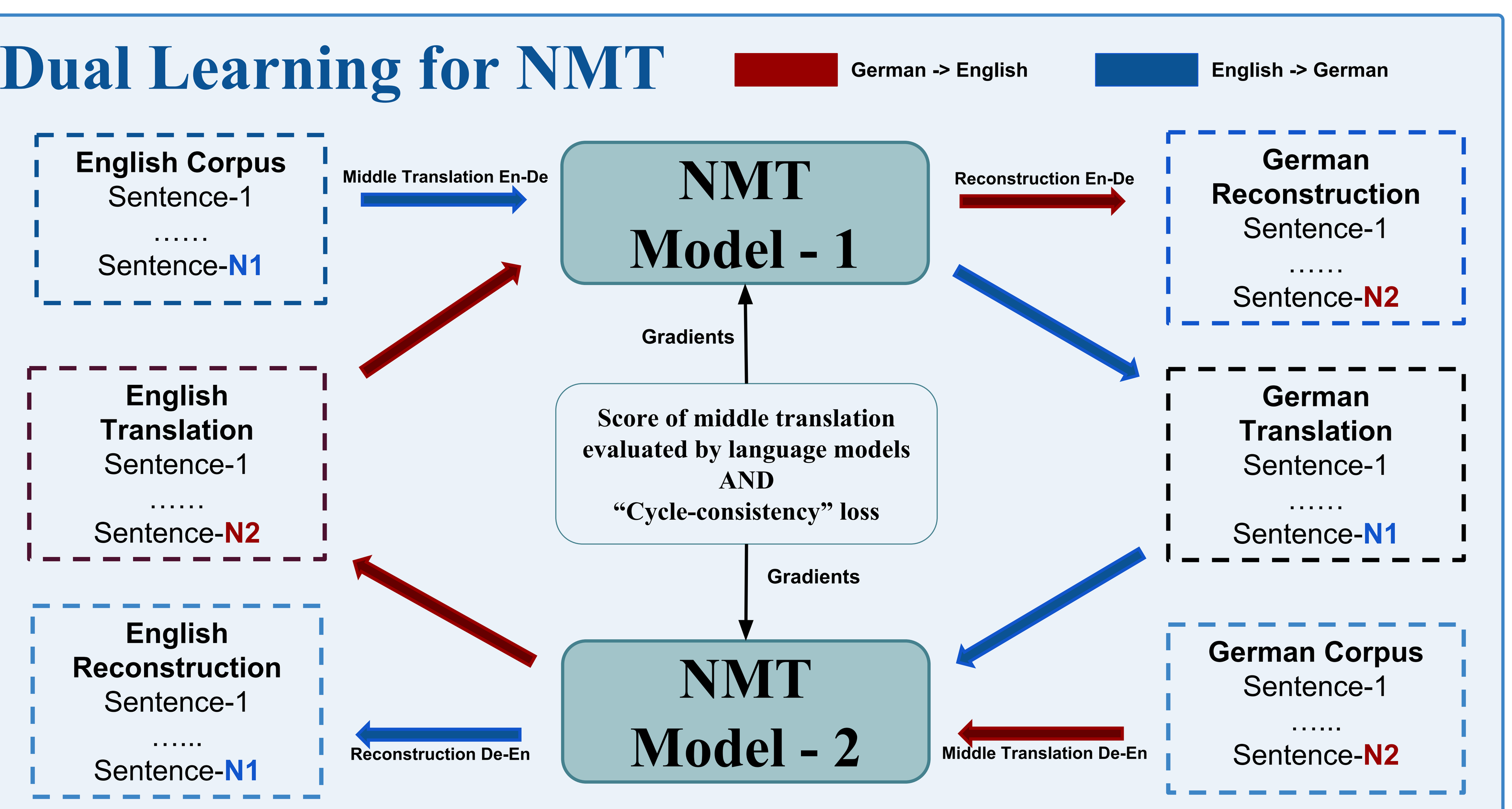
- When we have a **reversed translator** (German-to-English), we don't need paired training data for training English-to-German NMT model:

Training NMT with a Reversed Translator



- But it's **difficult to have a perfect reversed translator** to guide the training (as difficult as training our target NMT model)
- But can we train two NMT models with reversed translating directions together, and let them **guide each other** during the training?

Dual Learning for NMT



Note: Gradients are estimated by **sampling the translation models (policy gradient)**

Algorithm

Input: Monolingual corpora D_A and D_B, initial translation models Θ_{AB} and Θ_{BA}, language models LM_A and LM_B, α , beam search size K, learning rates $\gamma_{1,t}$, $\gamma_{2,t}$.

repeat

$t = t + 1$.

Sample sentence s_A and s_B from D_A and D_B respectively.

Set $s = s_A$ ▹*Model update for the game beginning from A.*

Generate K sentences $s_{mid,1}, \dots, s_{mid,K}$ using beam search according to translation model $P(.|s;\Theta_{AB})$.

for $k = 1, \dots, K$ **do**

Set the language-model reward for the kth sampled sentence as $r_{1,k} = LM_B(s_{mid,k})$.

Set the reconstruction reward for the kth sampled sentence as $r_{2,k} = \log P(s|s_{mid,k};\Theta_{BA})$.

Set the total reward of the kth sample as $r_k = \alpha r_{1,k} + (1-\alpha) r_{2,k}$.

end for

Compute the stochastic gradient of Θ_{AB}: ▹ *Policy Gradient*

$$\nabla_{\Theta_{AB}} \hat{E}[r] = \frac{1}{K} \sum_{k=1}^{K} [r_k \nabla_{\Theta_{AB}} logP(s_{mid,k}|s;\Theta_{AB})]$$

Compute the stochastic gradient of Θ_{BA}:

$$\nabla_{\Theta_{BA}} \hat{E}[r] = \frac{1}{K} \sum_{k=1}^{K} [(1-\alpha) \nabla_{\Theta_{BA}} logP(s|s_{mid,k};\Theta_{BA})]$$

Model updates:

$$\Theta_{AB} \leftarrow \Theta_{AB} + \gamma_{1,t} \nabla_{\Theta_{AB}} \hat{E}[r], \Theta_{BA} \leftarrow \Theta_{BA} + \gamma_{2,t} \nabla_{\Theta_{BA}} \hat{E}[r]$$

Set $s = s_B$ ▹*Model update for the game beginning from B.*

Go through the algorithm symmetrically.

until convergence

Evaluation

Dataset: We use the news and news commentary data from WMT Workshops for training both neural language models and NMT models:

- **Language Models**: ~300,000 monolingual sentences for both English and German
- **NMT Models**: ~200,000 parallel sentences
- **Dual Learning**: ~500,000 monolingual sentences for both languages (only 30,000 are used at the moment of making this poster due to the slow training speed and limited resources)
- **Testing**: 1,000 parallel German-English sentences

Note that there are actually millions of monolingual sentences for German and English in WMT's news dataset, but we don't have enough time and computing resources to fully exploit them.

Baseline: Attention + Bi-LSTM NMT models trained with parallel corpus.
We use the same model architecture in all experiments

German to English

Method	BLEU	Word Accuracy
NMT + normal training	21.46	18.33
NMT + Dual Learning	**24.10**	**21.52**

English to German

Method	BLEU	Word Accuracy
NMT + normal training	17.60	18.27
NMT + Dual Learning	17.78	18.85

Conclusion:

- We got improvements over the baseline on German-to-English translation, but didn't get clear improvements on the reversed task. This might because we don't have enough time to run dual learning on large monolingual data
- Dual Learning can be a potential strategy for overcoming the lack of parallel corpus
- The training of Dual Learning is way slower than normal training schemes due to the sampling step (using beam search) for estimating rewards. Also we cannot use mini-batch in the training of Dual Learning + NMT

References

[1]. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations (ICLR 2015)*

[2]. Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma. Dual Learning for Machine Translation. *Conference on Neural Information Processing Systems (NeurIPS 2016)*