Dual Learning for Machine Translation

Gómez Herrera María Andrea Liliana
Siboyabasore Cédric

University of Liège

November 2021

Outline

Introduction and Problem addressed

- Develop a dual-learning mechanism, which can enable an NMT system to automatically learn from unlabeled data through a dual-learning game, since human labeling is very costly.
- Given that any MT task has a dual task, each task is represented by an agent and they teach each other through a reinforcement learning process.

Objective

Present a dual-learning mechanism that can leverage monolingual data, in the source and target languages, in a more effective way.

- Train translation models from unlabeled data through reinforcement learning.
- Demonstrate the power of deep reinforcement learning for complex real-world applications.

State of the Art

Neural machine translation systems are implemented using RNN based encoder-decoder frameworks that learn probabilistic mapping $P(y|x)$ from source language x to target language y :

- Encoder generates T recurrent states $h_i = f\left(h_{i-1}, x_i\right)$
- Decoder computes probability $P(y_t|y_{<t}, x)$ where y_t is the word at $t-$th position in sentence y and $y_{<t}$ are the words that precede y_t. $P(y|x)$ is computed using the chain rule.
- The learning objective is to find optimal parameters Θ^* such that

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \sum_{(x,y)\in D} \sum_{t=1}^{T_y} \log P\left(y_t \mid y_{<t}, x; \Theta\right)$$

State of the Art

Limitation:

- Parallel data (x, y) are costly to collect. Millions of bilingual sentence pairs are needed to train Neural Machine Translation models.

State of the Art

What has been studied in subsequent researches:

Methods that leverage monolingual data to boost the performance of Machine Translation systems.

State of the Art

- Monolingual corpora in the target language used to train a Language Model, which is then integrated with the Machine Translation models trained from parallel bilingual corpora to improve the translation quality. [1]

Main limitation: Methods that leverage monolingual data to boost the performance of Machine Translation systems.

[1] T. Brants, et al. *Large language models in machine translation*, C.Gulcehre et al. *On using monolingual corpora in neural machine translation.*

State of the Art

- Pseudo bilingual sentence pairs are generated from monolingual data by using the Translation Model trained from aligned parallel corpora, and then these pairs are used to enlarge the training data for subsequent learning. [2]

Main limitation: Although these methods can enlarge the parallel training data, there is no guarantee/control on the quality of the pseudo bilingual sentence pairs.

[2] N. Ueffing, et al. *Semi-supervised model adaptation for statistical machine translation*, R. Sennrich et al. *Improving neural machine translation models with monolingual data.*

State of the Art

What this paper proposes:

- Taking leverage of the immense monolingual data in a more effective way.
- Taking leverage of the duality of translation tasks
  ex: $FR \rightarrow EN$ and $EN \rightarrow FR$

Dual Learning for Neural Machine Translation

- Two monolingual corpora D_A and D_B which contain sentences from language A and B respectively.
- Two neural machine translation models $P(.|s;\Theta_{AB})$ and $P(.|s;\Theta_{BA})$, with Θ_{AB} and Θ_{BA} the parameters to optimize.
- Two language models $LM_A(.)$ and $LM_B(.)$, that take a sentence as input and output a real value to indicate how confident the sentence is a natural sentence in its own language. Trained respectively using D_A and D_B.

Dual Learning for Neural Machine Translation

Reinforcement learning process :

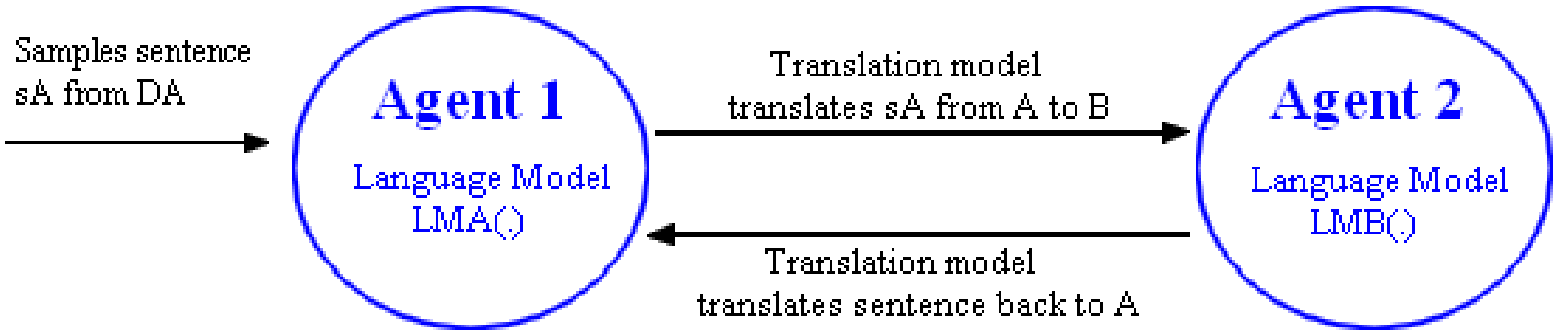


Figure 1: Dual game starting from Agent 1

Dual Learning for Neural Machine Translation

Reinforcement learning process :

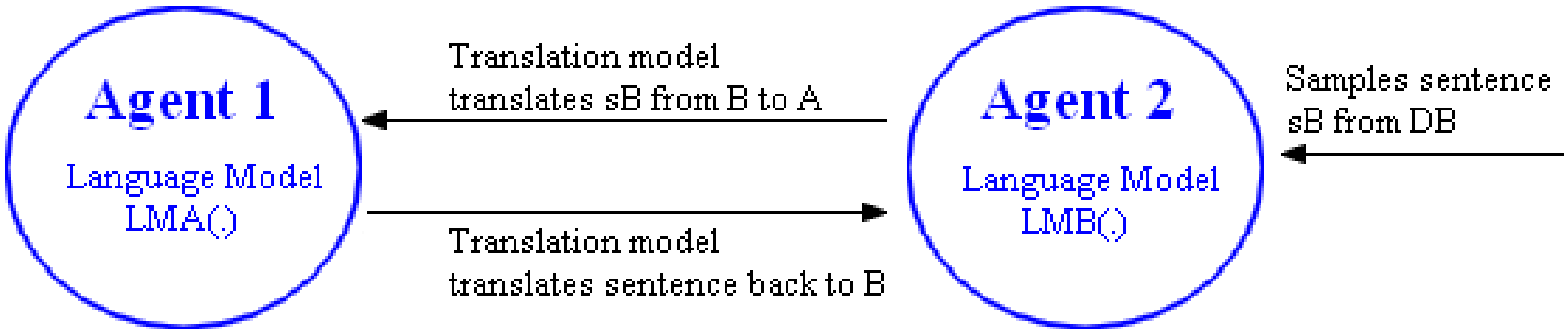


Figure 2: Dual game starting from Agent 2

Dual Learning for Neural Machine Translation

Sample s_A from D_A
Generate K sentences $s_{mid,1}, \ldots, s_{mid,K}$ from s_A using beam search according to translation model $P(.|s; \Theta_{AB})$.
for k=1,...., K **do**:
Compute language model reward $r_{1,k} = LM_B(s_{mid,k})$
Compute reconstruction reward $r2, k = log\ P(s|smid, k; BA)$
Compute total reward $r_k = \alpha r_{1,k} + (1 - \alpha) r_{2,k}$
end for
Compute the stochastic gradient of θ_{AB}:

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

Compute the stochastic gradient of θ_{BA}:

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

Update model:

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

Repeat process symmetrically with s_B sampled from D_B

Repeat entire process until convergence

Models

Language models

- Gated recurrent unit connected to a fully connected layer

Neural machine translation models

- Encoder-decoder architecture with LSTM RNN encoder and decoders, with dot product attention

Reward

- Language model reward $r_{1,k}$: average over square rooted length of string

Beam search size k

- k= 2 during training, 12 during testing such as in previous works

All other hyperparameters during the experiments were set by cross-validation.

- α was finally set to 0.01 such that total reward $r_k = 0.01 r_{1,k} + 0.99 r_{2,k}$

Baselines

- The standard neural machine translation (NMT).
- A recent NMT-based method which generates pseudo bilingual sentence pairs from monolingual corpora to assist training (pseudo-NMT).

Datasets

Neural machine translation models are trained on

Bilingual corpora from WMT'14:

12M sentence pairs extracting from five datasets: Europarl v7, Common Crawl corpus, UN corpus, News Commentary, and 109French-English corpus.

Language models are trained on

Monolingual data from WMT'14:

"News Crawl: articles from 2012", available in both English and French.

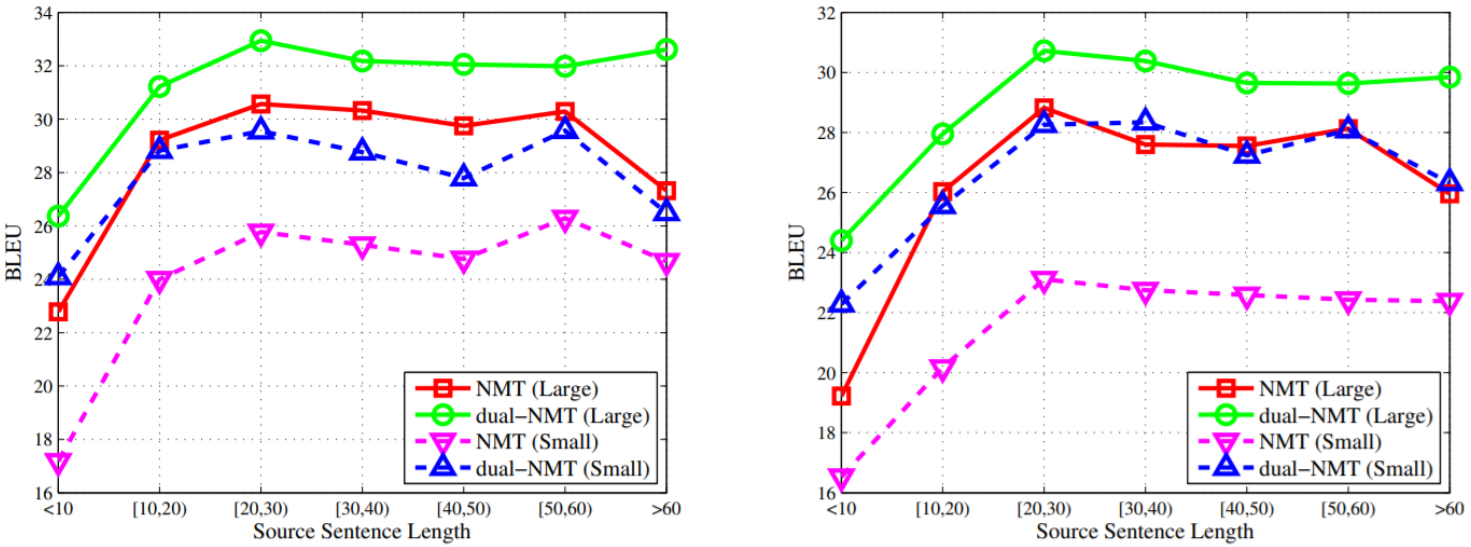
Performance

- The Dual-NMT algorithm surpasses the baseline algorithms.

	En→Fr (Large)	Fr→En (Large)	En→Fr (Small)	Fr→En (Small)
NMT	29.92	27.49	25.32	22.27
pseudo-NMT	30.40	27.66	25.63	23.24
dual-NMT	**32.06**	**29.78**	**28.73**	**27.50**

Figure 3: Translation results of En↔ Fr task. The results of the experiments using all the parallel data for training are provided in the first two columns and the results using 10% parallel data for training are in the last two.

Performance



(a) En→Fr (b) Fr→En

Figure 4: BLEU scores w.r.t lengths of source sentences

Performance

- The translation from French to English is better than the one from English to French.
- With only 10% bilingual data, dual-NMT achieves comparable translation accuracy as vanilla NMT using 100% bilingual data for the $FR \rightarrow EN$ task.

Contributions

This work

- Helped reduce dependency on aligned bilingual data. Results are encouraging.
- Demonstrated the usefulness of deep reinforcement learning for the field. Before that, Neural machine translation was purely based on neural networks.
- Demonstrated the power of deep reinforcement learning for real-life applications, other than games.

Limitations

- More transparency on the effects of the hyperparameters.
- Restricted to Neural Machine Translation systems.
- Research only based on $FR \rightarrow EN$ and $EN \rightarrow FR$ translation tasks

Extensions

Design and test dual-learning algorithms for more dual tasks, such as:

- Speech recognition versus text to speech
- Image caption versus image generation
- Question answering versus question generation,
- Search (matching queries to documents) versus keyword extraction (extracting keywords/queries for documents)

$\Rightarrow$ *Dual Supervised Learning*

Extensions

- If more than two associated tasks can form a closed loop, apply the methodology to improve the model in each task from unlabeled data.

Close-loop learning $\Rightarrow$ *Dual Transfer Learning for Neural Machine Translation with Marginal Distribution Regularization*

$\Rightarrow$ Model-Level Dual Learning

What is next?

- Learn translations directly from monolingual data of two languages (maybe plus lexical dictionary).
- Apply a dual-NMT approach to phrase-based SMT systems.
- Jointly train multiple translation models for a tuple of 3+ languages using monolingual data.