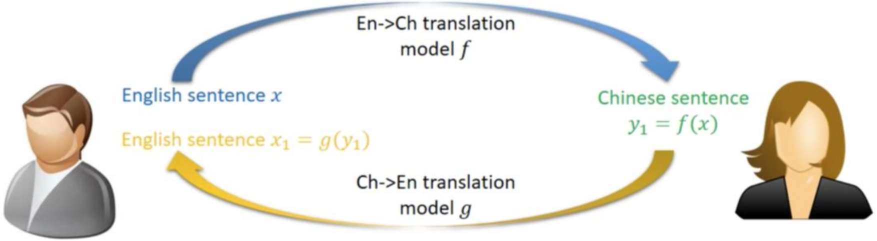
Dual Learning for Machine Translation

Presented by:
Sreeja R Thoom
Jayavardhan Reddy Peddamail

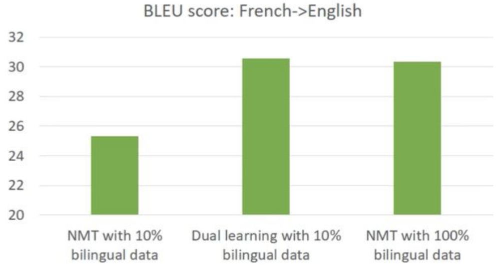
Overview

- What
  - Introduce an autoencoder-like mechanism, "Dual learning", to utilize monolingual datasets



1)

- Results
  - Dual Learning with 10% data ≈ Baseline model with 100% data



1)

1) "Dual Learning: A New Learning Paradigm", https://www.youtube.com/watch?v=HzokNo3g63E&feature=youtu.be

Difficulty in getting large bilingual data

- Solution: utilization of monolingual data
  - Train a language model of the target language, and then integrate it with the MT model[1)2)]

    <- does not fundamentally address the shortage of parallel data.

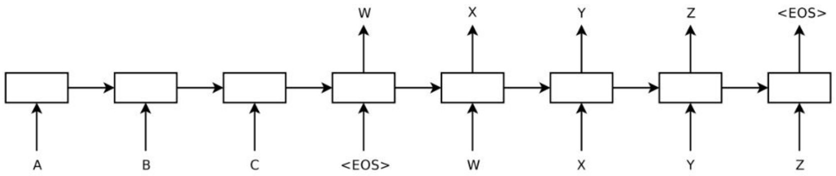
  - Generate pesudo bilingual data from monolingual data[3)4)]

    <- no guarantee on the quality of the pesudo bilingual data

1) T. Brants et al., “Large language models in machine translation”, EMNLP 2007
2) C. Gucehre et al., “On using monolingual corpora in neural machine translation”, arix 2015
3) R. Sennrich et al., “Improving neural machine translation models with monolingual data”, ACL 2016
4) N. Ueffing et al., “Semi-supervised model adaptation for statistical machine translation”, Machine Translation Journal 2008

Neural machine translation

- Learn conditional probability $P(y|x;\Theta)$ from a input $x = \{x_1, x_2, \ldots, x_{T_x}\}$ to an output $y = \{y_1, y_2, \ldots, y_{T_y}\}$



- Maximize the log probability

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

Hidden Vectors(Encoder):

$$h_i = f(h_{i-1}, x_i)$$

Decoder Portion:

$$\begin{aligned} P(y_t|y_{<t}, x) &\propto \exp(y_t; r_t, c_t) \\ r_t &= g(r_{t-1}, y_{t-1}, c_t) \\ c_t &= q(r_{t-1}, h_1, \cdots, h_{T_x}) \end{aligned}$$

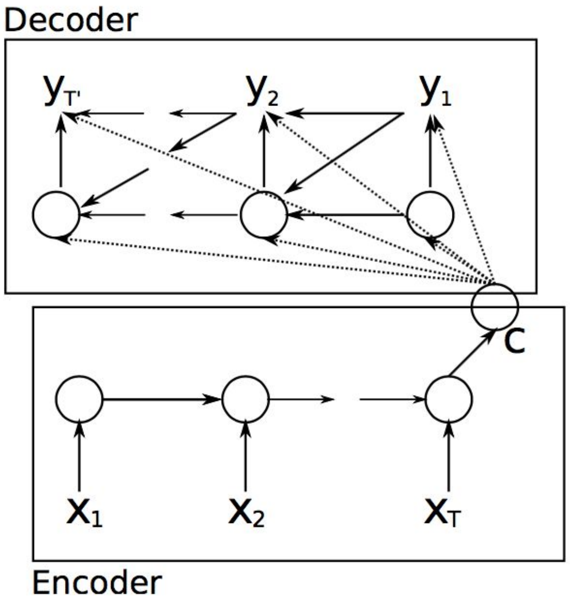
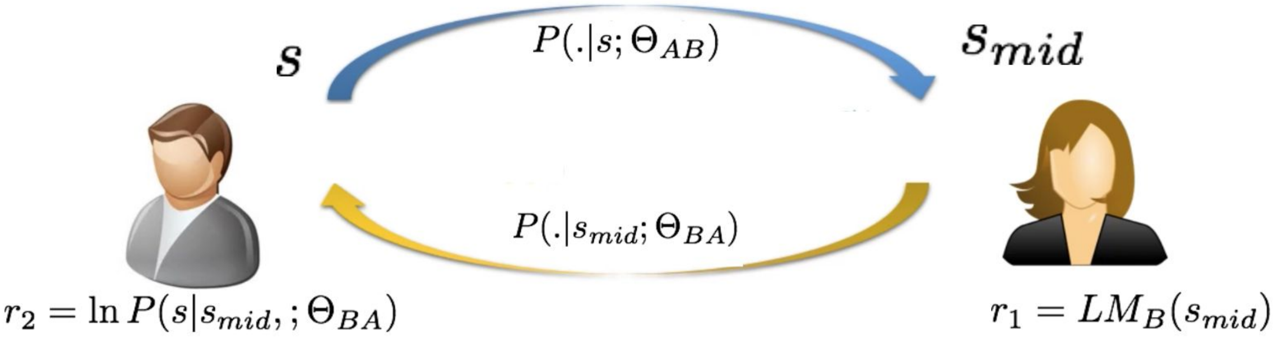


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

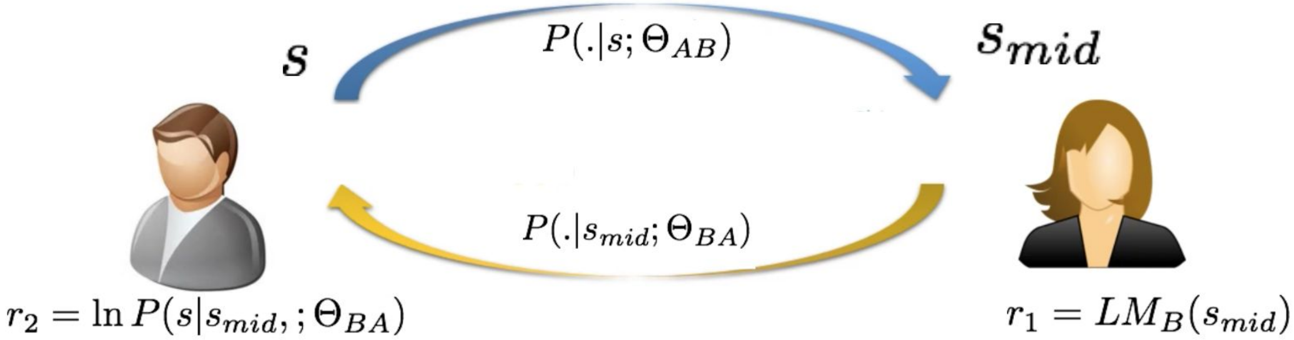
Dual learning algorithm



- Use monolingual datasets to train translation models through dual learning

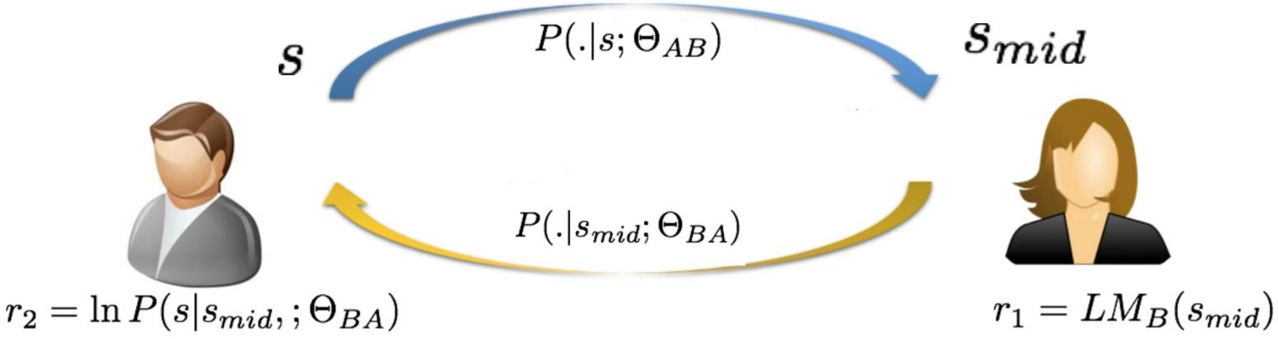
- Things required
  - D_A: corpus of language A
  - D_B: corpus of language B (not necessarily aligned with D_A)
  - $P(.|s;\Theta_{AB})$: translation model from A to B
  - $P(.|s;\Theta_{BA})$: translation model from B to A
  - $LM_A(.)$: learned language model of A
  - $LM_B(.)$: learned language model of B

Dual learning algorithm



1. Generate K translated sentences $s_{mid,1}, s_{mid,2}, \dots, s_{mid,K}$ from $P(.|s;\Theta_{AB})$ based on beam search

Dual learning algorithm



1. Generate K translated sentences

$$s_{mid,1}, s_{mid,2}, \dots, s_{mid,K}$$

from $P(.|s;\Theta_{AB})$ based on beam search

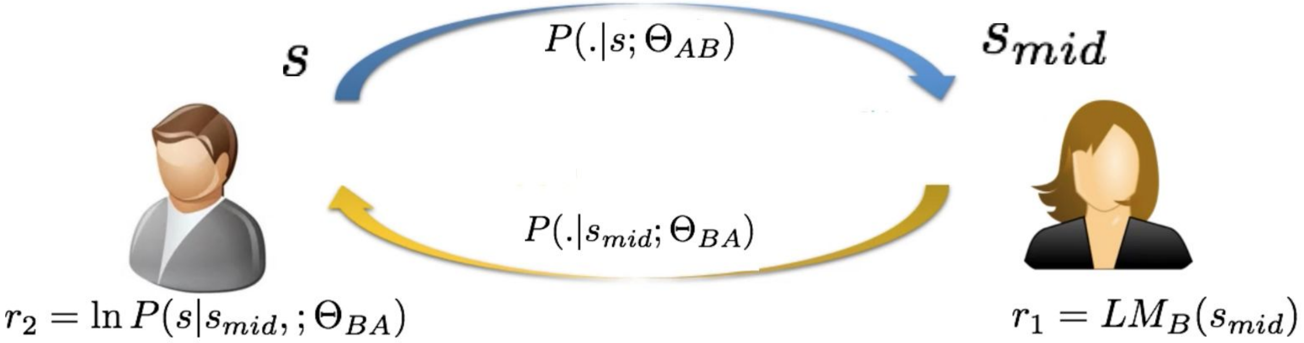
2. Compute intermediate rewards

$$r_{1,1}, r_{1,2}, \dots, r_{1,K}$$

from $LM_B(s_{mid,k})$ for each sentence as

$$r_{1,k} = LM_B(s_{mid,k})$$

Dual learning algorithm

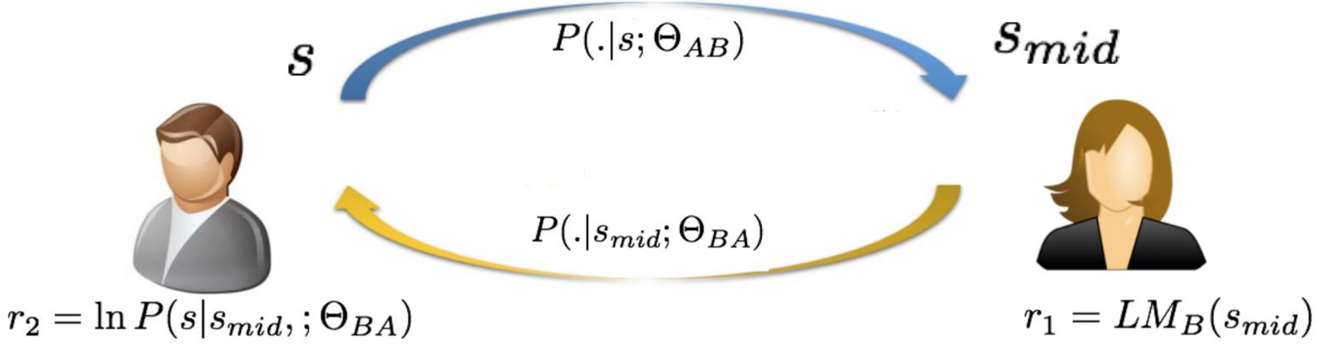


3. Get communication rewards

$r_{2,1}, r_{2,2}, \dots, r_{2,k}$

for each sentence as $r_{2,k} = \ln P(s|s_{mid,k};\Theta_{\mathrm{BA}})$

Dual learning algorithm



3. Get communication rewards

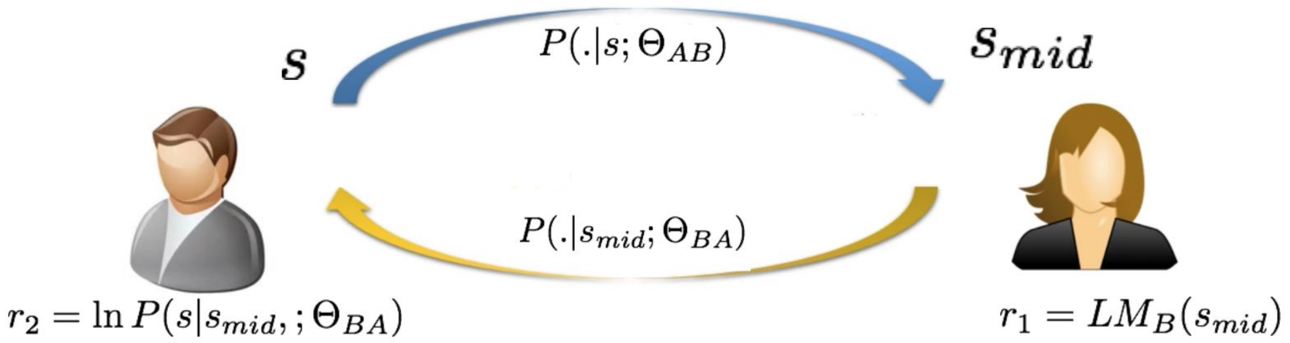
$$r_{2,1}, r_{2,2}, \ldots, r_{2,k}$$

for each sentence as $r_{2,k} = \ln P(s|s_{mid,k};\Theta_{\mathrm{BA}})$

4. Set the total reward of k-th sentence as

$$r_k = \alpha r_{1,k} + (1-\alpha) r_{2,k}$$

Dual learning algorithm

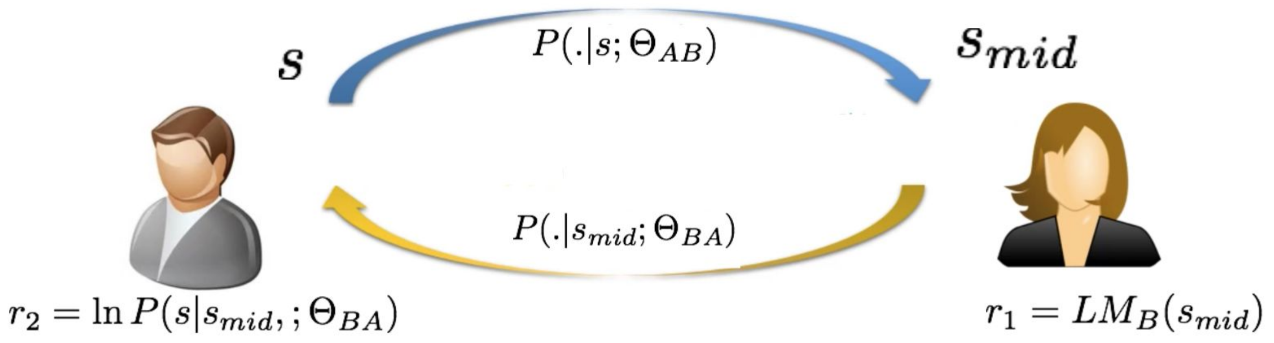


5. Compute the stochastic gradient of Θ_{AB} and Θ_{AB}

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

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

Dual learning algorithm



5. Compute the stochastic gradient of Θ_{AB} and Θ_{AB}

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

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

6. Update model parameters

$$\Theta_{AB} \leftarrow \Theta_{AB} + \gamma_1 \nabla_{\Theta_{AB}} E[r]$$
$$\Theta_{BA} \leftarrow \Theta_{BA} + \gamma_2 \nabla_{\Theta_{BA}} E[r]$$

Stochastic gradient of models

$$\nabla_{\Theta_{AB}} E[r] = \sum_{s_{mid}} [\nabla_{\Theta_{AB}} P(s_{mid}|s;\Theta_{AB}) \cdot r + P(s_{mid}|s;\Theta_{AB}) \cdot \nabla_{\Theta_{AB}} r]$$
$$= \sum_{s_{mid}} P(s_{mid}|s;\Theta_{AB}) \nabla_{\Theta_{AB}} \ln P(s_{mid}|s;\Theta_{AB}) \cdot r$$
$$\approx \frac{1}{K} \sum_{k} \nabla_{\Theta_{AB}} \ln P(s_{mid,k}|s;\Theta_{AB}) \cdot r_k \longrightarrow$$

Beam Search

$$\nabla_{\Theta_{BA}} E[r] = \sum_{s_{mid}} [\nabla_{\Theta_{BA}} P(s_{mid}|s;\Theta_{AB}) \cdot r + P(s_{mid}|s;\Theta_{AB}) \cdot \nabla_{\Theta_{BA}} r]$$
$$= \sum_{s_{mid}} P(s_{mid}|s;\Theta_{AB}) \cdot \nabla_{\Theta_{BA}} (1-\alpha) \ln P(s|s_{mid};\Theta_{BA})$$
$$\approx \frac{1}{K} \sum_{k} \nabla_{\Theta_{BA}} (1-\alpha) \ln P(s|s_{mid,k};\Theta_{BA})$$

$$\nabla_\theta \mathbb{E}[f(x)] = \nabla_\theta \int p_\theta(x) f(x) dx$$
$$= \int \frac{p_\theta(x)}{p_\theta(x)} \nabla_\theta p_\theta(x) f(x) dx$$
$$= \int p_\theta(x) \nabla_\theta \log p_\theta(x) f(x) dx$$
$$= \mathbb{E}\Big[f(x) \nabla_\theta \log p_\theta(x)\Big]$$

Algorithm 1 The dual-learning algorithm

1: **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 $\gamma_{1,t}, \gamma_{2,t}$.
2: **repeat**
3: $t = t + 1$.
4: Sample sentence s_A and s_B from D_A and D_B respectively.
5: Set $s = s_A$. $\triangleright$ *Model update for the game beginning from A.*
6: Generate K sentences $s_{mid,1}, \ldots, s_{mid,K}$ using beam search according to translation model $P(.|s; \Theta_{AB})$.
7: **for** $k = 1, \ldots, K$ **do**
8: Set the language-model reward for the kth sampled sentence as $r_{1,k} = LM_B(s_{mid,k})$.
9: Set the communication reward for the kth sampled sentence as $r_{2,k} = \log P(s|s_{mid,k}; \Theta_{BA})$.
10: Set the total reward of the kth sample as $r_k = \alpha r_{1,k} + (1 - \alpha) r_{2,k}$.
11: **end for**
12: Compute the stochastic gradient of Θ_{AB}:

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

13: Compute the stochastic gradient of Θ_{BA}:

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

14: Model updates:

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

15: Set $s = s_B$. $\triangleright$ *Model update for the game beginning from B.*
16: Go through line 6 to line 14 symmetrically.
17: **until** convergence

Dataset:

- WMT'14
- 12M sentence pairs
- English -> French, French -> English

Data usage (for dual learning):

- Small
  - Train translation models with 10% bilingual data.
  - Train translation models with 10% bilingual data and monolingual data through dual learning algorithm.
- Large:
  - Train translation models with 100% bilingual data.
  - Train translation models with 100% bilingual data and with monolingual data through dual learning algorithm.

Evaluation

- BLEU: geometric mean of n-gram precision

$$\text{BLEU} = \text{BP} \times \left(\prod_{n}^{4} \text{n-gram precision} \right)^{\frac{1}{4}}$$

$$\text{BP} = \begin{cases} 1 & \text{if } |\text{pred}| > |\text{true}| \\ e^{\frac{1-|\text{true}|}{|\text{pred}|}} & \text{if } |\text{pred}| \leq |\text{true}| \end{cases}$$

$$Precision = \exp(\sum_{n=1}^{N} w_n \log p_n), \quad \text{where} \quad w_n = 1/n$$

Experiment settings

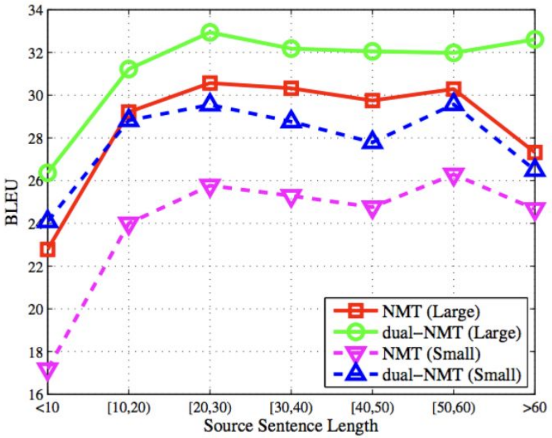
- Baseline models
  - Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate”
  - Sennrich et al., “Improving Neural Machine Translation Models with Monolingual Data”

Results

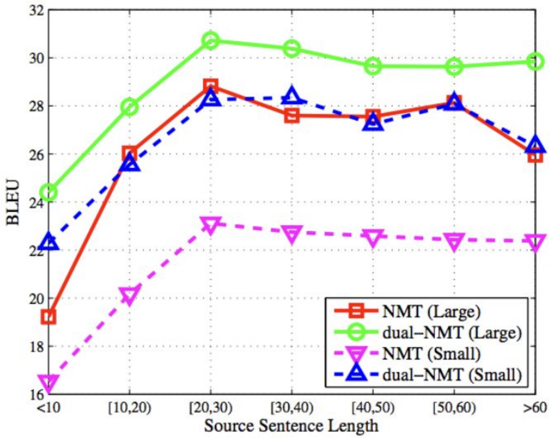
	En→Fr (Large)	Fr→En (Large)	En→Fr (Small)	Fr→En (Small)
NMT	29.92	27.49	25.32	22.27
pseudo-NMT	30.40	27.66	25.63	23.24
dual-NMT	**32.06**	**29.78**	**28.73**	**27.50**

- Outperform the base line models
- In Fr->En, dual learning with 10% data ≈ baseline models with 100% data.
- Dual learning is effective especially in a small dataset.

Results



(a) En→Fr



(b) Fr→En

- For different source sentence length
  - Improvement is significant for long sentences.

Results

	En→Fr→En (L)	Fr→En→Fr (L)	En→Fr→En (S)	Fr→En→Fr (S)
NMT	39.92	45.05	28.28	32.63
pseudo-NMT	38.15	45.41	30.07	34.54
dual-NMT	**51.84**	**54.65**	**48.94**	**50.38**

- Reconstruction performance (BLEU)
  - Huge improvement from baseline models, especially in En->Fr-En(S)

How is Reconstruction BLEU score higher than Translation BLEU Score?

Results

- Reconstruction examples

	Translation-back-translation results before dual-NMT training	Translation-back-translation results after dual-NMT training
Source (En)	The majority of the growth in the years to come will come from its liquefied natural gas schemes in Australia.	
En→Fr	La plus grande partie de la crois- -sance des années à venir viendra de ses systèmes de gaz naturel liquéfié en Australie .	La majorité de la croissance dans les années à venir viendra de ses régimes de gaz naturel liquéfié en Australie .
En→Fr→En	Most of the growth of future years will come from its liquefied natural gas systems in Australia .	The majority of growth in the coming years will come from its liquefied natural gas systems in Australia .
Source (Fr)	Il précise que " les deux cas identifiés en mai 2013 restent donc les deux seuls cas confirmés en France à ce jour " .	
Fr→En	He noted that " the two cases identified in May 2013 therefore remain the only two two confirmed cases in France to date " .	He states that " the two cases identified in May 2013 remain the only two confirmed cases in France to date "
Fr→En→Fr	Il a noté que " les deux cas identifiésen mai 2013 demeurent donc les deux seuls deux deux cas confirmés en France à ce jour "	Il précise que " les deux cas identifiés en mai 2013 restent les seuls deux cas confirmés en France à ce jour ".

Future extensions & words

- Application in other domains

Application	Primal task	Dual task
Speech processing	Speech recognition	Text to speech
Image understanding	Image captioning	Image generation
Conversation engine	Question	Response
Search engine	Search	Query/Keyword suggestion

- Generalization of dual learning
  - Dual -> Triple -> ⋯ -> n-loop
- Learn from scratch
  - only with monolingual data

Summary

- What
  - Introduce "Dual learning algorithm" to utilize monolingual data
- Results
  - With 100% data, the model outperforms the baseline models
  - With 10% data, the model shows the comparable result with the baseline models
- Future
  - Dual learning mechanism can be applied to other domains
  - Learn from scratch

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English- https://blogs.microsoft.com/ai/machine-translation-news-test-set-human-parity/

Questions:

- Removing all Monolingual sentences with out of vocabulary terms.
- Reconstruction BLEU score> Translation BLEU Score